Semiconductor Detector Systems

Helmuth Spieler
Physics Division, Lawrence Berkeley National Laboratory

CLARENDON PRESS • OXFORD
2006

PREFACE

When semiconductor detectors were introduced in the 1960s, their primary use was in energy spectroscopy. By virtue of their dramatically improved energy resolution they revolutionized gamma ray and charged particle spectroscopy. Detector systems typically consisted of a single sensor, often cooled to liquid nitrogen temperature, and a bin filled with electronics, highly optimized for energy resolution and count-rate capability. In the early 1980s a new development began, which would bring about a second revolution. Instead of emphasizing energy resolution, these system exploited the micron-scale patterning capabilities of semiconductor technology to form detectors with many electrodes as position-sensing devices. Efficient readout of these detectors required high-density front-end electronics. Detector systems now consisted of a highly segmented sensor, typically with strip electrodes on a 50 – 100 μm pitch, combined as a unit with an array of custom integrated circuits, often with 128 channels of readout electronics per chip. These systems required a radically different design approach from previous semiconductor detector systems. Rather than emphasizing primarily one or two performance parameters, energy resolution and count-rate performance, these systems needed to fulfill many conflicting requirements, *i.e.* low noise, low power and minimum material. Later – in systems for high-luminosity colliders – additional demands on fast timing and radiation resistance had to be met.

So different were these requirements from the established paradigm, that conventional wisdom among the experts held these systems to be impractical, if not impossible. Today, highly integrated semiconductor detector systems with tens of thousands of channels are routine and new detectors are under construction covering hundreds of square meters with millions of channels. Although advances in technology – especially in integrated circuit density – have played a major role in bringing this about, this development required more than the magic of modern technology. These systems are bounded by rather fundamental constraints, which were already well-understood in the 1970s. The challenge was to take a fresh look at these constraints and develop new architectures that balanced experimental demands with practical technology.

The goal of this book is to show how this balance comes about, so it emphasizes both sensors and electronics as a system. It is written primarily for physicists who devise new detectors and bring them into operation, so I include basic discussions of amplifiers, circuits, and electronic noise that are familiar to engineers, but not covered in a typical physics curriculum. The choice of topics and the organization of the book resulted from courses I have taught in the Physics Department of the University of California at Berkeley and from numerous short courses on detectors and signal processing on six continents, ranging from the undergraduate to the faculty level. Much of the material was developed

in my attempts to educate my collaborators and the reviewers of funding proposals. The exposition of topics results from many interactions with students, who on occasion have made abundantly clear to me what doesn't work. This has led to a "cyclical" approach, where I return to topics for more detailed discussions and reiterate explanations in different contexts. This does lead to some redundancy, but also allows chapters to be read individually, although they do build on one another.

Since new detectors will be always be stretching the envelope, it is important to understand the basic principles underlying the technology to see how far one can push, so I emphasize physical principles and how they are applied. Although the big picture is important, details are crucial, so in important areas I go into some detail, especially in the sections on noise, signal processing and devices. What I don't do, is delve deeply into the intricacies of detector technology and circuit design. Engineers have developed very powerful analysis techniques and technological tools that are essential for a working design. Fortunately, knowledge of sensor and IC fabrication details, complex frequency space, and Laplace transforms is not needed to understand the key principles of the systems and then go to the experts with the right questions. This is not a cookbook. Technology progresses continually, so specific examples are included to illustrate concepts, rather than prescriptions for designs. I emphasize the key mechanisms and their interplay, with the goal of helping readers focus the analysis of their own designs, as the usual curricula tend to teach how to calculate, rather than what to calculate.

The foundations provided by a good undergraduate physics education should be sufficient to follow the discussions in this book. The book is designed to be self-contained. Key elements are derived, to make clear their origins and limits. Some of the derivations can be found readily in the more specialized literature, but I've included them since it allows me to emphasize those aspects that are important for this specific application, for example in the diode equation and the treatment of the bipolar transistor. Whenever the derivations would disrupt the main thread of the discussion, I've moved them to the appendices. For those who are not familiar with equivalent circuits or with complex notation to describe phase shifts, brief tutorial descriptions are also included as appendices.

Chapter 1 gives an overview and summarizes key aspects of semiconductor detector arrays to provide context for the subsequent more detailed discussions. It can also serve as an executive summary for those who wish to appear knowledgeable without going to the trouble of understanding too much. Chapters 2 and 3 cover signal formation and electronic noise and lead to Chapter 4, which discusses signal processing and optimization of signal-to-noise ratio. Chapter 5 discusses digitization techniques, including their flaws, which provides some key insights needed for a brief introduction to digital signal processing. Up to this point the discussions are largely technology-independent, *i.e.* these principles can be applied to either old or new technology and remain valid as technology progresses. Technology limits enter into Chapter 6, Transistors and Amplifiers,

which discusses how technology affects the device parameters that are critical in detector systems. This gives insight into practical noise limits and also provides the basis for Chapter 7, which discusses radiation effects and mitigation techniques. This area shows clearly how degradation of individual device properties can be mitigated by system architecture and appropriate choice of system parameters. Chapter 8 selects some specific detector designs to illustrate conflicts and trade-offs, and shows how different solutions address the same goals. This chapter also includes some comments on design and assembly techniques, reliability issues, and testing. In closing, Chapter 9 turns to what is probably the most important problem, why things don't work. Although not exhaustive, it discusses many of the interference sources and design flaws that cause systems to perform more poorly than expected, or even be unusable. This is a complex topic, rife with simple recipes, which tend to be wrong.

In the past two decades semiconductor detector arrays have come a long way. We have encountered many difficulties that were not foreseen, but silicon technology is mature and highly developed, so the power of the technology has always provided the flexibility to find solutions. Although the next generation of detector arrays is still under construction, future accelerator upgrades now under discussion will tax detector capabilities even more. Furthermore, we see developments coming full circle. After being driven by tracking applications in high-energy physics, we now see increasing interest in applying array technology to x-ray spectroscopy for astrophysics, materials science, and medical imaging. Even though the requirements are daunting and solutions not always obvious, we can be assured that semiconductor detector arrays will be key components in frontier experiments for years to come.

ACKNOWLEDGEMENTS

This book is the result of countless interactions with friends, colleagues, and collaborators too numerous to list individually. However, some bear more responsibility than others. Through persistant questioning over the past two decades, Bob Ely and Carl Haber prompted me to rethink and develop significant portions of the material presented here. Unfortunately, the responsibility for any mistakes or inaccuracies is mine.

At Lawrence Berkeley National Laboratory my work is supported by the Director, Office of Science, Office of High Energy and Nuclear Physics, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, but a project-driven environment with inadequate budgets is hardly conducive to probing beyond what is absolutely necessary for detector design and construction. As a result, the lengthy process of understanding the physics and technology of semiconductor sensors and electronics, developing (and discarding) concepts, and finally the preparation of this book were very much a night-time and weekend effort. Foremost, I thank my wife Sigrid for wholeheartedly supporting this effort and understanding my faraway gazes at the breakfast and dinner table.

Helmuth Spieler

Pinole, California
July, 2005

CONTENTS

1 Detector systems overview 1
1.1 Sensor 2
1.2 Preamplifier 3
1.3 Pulse shaper 3
1.4 Digitizer 5
1.5 Electro–mechanical integration 6
1.6 Sensor structures I 8
1.6.1 Basic sensor 8
1.6.2 Position sensing 9
1.6.3 Pixel devices 11
1.7 Sensor physics 12
1.7.1 Signal charge 12
1.7.2 Sensor volume 13
1.7.3 Charge collection 16
1.7.4 Energy resolution 19
1.7.5 Position resolution 19
1.8 Sensor structures II – monolithic pixel devices 24
1.8.1 Charge coupled devices 24
1.8.2 Silicon drift chambers 25
1.8.3 Monolithic active pixel sensors 26
1.9 Electronics 29
1.10 Detection limits and resolution 29
1.10.1 Electronic noise 31
1.10.2 Amplitude measurements 33
1.10.3 Timing measurements 35
1.11 Subsystems 36
1.11.1 Circuit integration and bussing 36
1.11.2 Detector modules, services, and supports 38
1.11.3 Data acquisition 40
1.12 Further reading 40
References 40

2 Signal formation and acquisition 43
2.1 The signal 43
2.2 Detector sensitivity 48
2.2.1 Low energy quanta ($E \approx E_g$) 48
2.2.2 High energy quanta ($E \gg E_g$) 51
2.2.3 Fluctuations in signal charge – the Fano factor 52
2.3 Signal formation 55
2.3.1 Formation of a high-field region 55

2.3.2 Doping 56
2.3.3 The *pn*-junction 59
2.3.4 The reverse-biased diode 61
2.3.5 Strip and pixel detectors 66
2.4 Charge collection 67
2.5 Time dependence of the signal current 71
2.5.1 Induced charge – Ramo's theorem 73
2.5.2 Parallel plate geometry with uniform field 75
2.5.3 Double-sided strip detector 78
2.6 Charge collection in the presence of trapping 82
2.7 Semiconductor detector materials 83
2.8 Photodiodes 86
2.9 Signal acquisition 91
2.9.1 Voltage-sensitive amplifier 91
2.9.2 Current-sensitive amplifier 91
2.9.3 Voltage and current mode with capacitive sources 92
2.9.4 Feedback amplifiers – the "charge-sensitive amplifier" 93
2.9.5 Realistic charge-sensitive amplifiers 95
2.9.6 Input impedance of a charge-sensitive amplifier 100
References 102

3 Electronic noise 105
3.1 Electronic noise and resolution 105
3.2 Electronic noise 107
3.3 Some general properties of noise 107
3.3.1 Thermal (Johnson) noise 109
3.3.2 Shot noise 109
3.3.3 Low frequency ("$1/f$") noise 109
3.4 Derivation of spectral densities 110
3.4.1 Spectral density of thermal noise 110
3.4.2 Spectral density of shot noise 111
3.4.3 Spectral density of low-frequency noise 113
3.5 "Noiseless" resistances 114
3.5.1 Dynamic resistances 114
3.5.2 Active resistances 114
3.5.3 Radiation resistance of an antenna 114
3.6 Correlated noise 115
3.7 Signal equivalent noise measures 116
3.7.1 Noise equivalent power 116
3.7.2 Equivalent noise charge 116
3.8 Noise in Amplifiers 117
3.8.1 Amplifier noise model 117
3.8.2 Noise bandwidth *vs.* signal bandwidth 120
3.9 Amplifier noise matching 121

3.9.1 Resistive sources 121
3.9.2 Noise matching with a transformer 122
3.10 Capacitive sources 123
3.10.1 Noise *vs.* capacitance in a charge-sensitive amplifier 123
3.10.2 S/N *vs.* input time constant 125
3.11 Complex sensors 127
3.11.1 Cross-coupled noise 129
3.11.2 Backside readout 131
3.12 Quantum noise limits in amplifiers 132
References 133

4 Signal processing 134
4.1 Simple pulse shapers 134
4.1.1 Effect of relative time constants 135
4.2 Evaluation of equivalent noise charge 138
4.2.1 Experiment 139
4.2.2 Numerical simulation (*e.g.* SPICE) 141
4.2.3 Analytical simulation 141
4.3 Noise analysis of a detector and front-end amplifier 142
4.3.1 Detector bias current 143
4.3.2 Parallel resistance 144
4.3.3 Series resistance 145
4.3.4 Amplifier input noise 145
4.3.5 Cumulative input noise voltage 145
4.3.6 Equivalent noise charge 146
4.4 Examples 148
4.4.1 Photodiode readout 148
4.4.2 High-rate x-ray spectroscopy 151
4.5 Noise analysis in the time domain 153
4.5.1 Principles of noise analysis in the time domain 154
4.5.2 The weighting function 156
4.5.3 Time-variant shapers 158
4.5.4 Noise analysis of a correlated-double sample pulse shaper 160
4.6 Detector noise summary 166
4.7 Threshold discriminator systems 169
4.7.1 Noise rate 171
4.7.2 Noise occupancy 173
4.7.3 Measurement of noise in a threshold discriminator system 174
4.8 Some other aspects of pulse shaping 175
4.8.1 Baseline restoration 175
4.8.2 Tail (pole–zero) cancellation 177
4.8.3 Bipolar *vs.* unipolar shaping 178
4.9 Timing measurements 179
4.9.1 Pulse shaping in timing systems 180

4.9.2 Choice of rise time in a timing system 181
4.9.3 Time walk 182
4.9.4 Lowest practical threshold in leading edge triggering 183
4.9.5 Zero-crossing timing 184
4.9.6 Constant fraction timing 185
4.9.7 Fast timing – some results 187
References 189

5 Elements of digital electronics and signal processing 191
5.1 Digital circuit elements 191
5.1.1 Logic elements 191
5.1.2 Propagation delays and power dissipation 194
5.1.3 Logic arrays 195
5.2 Digitization of pulse height and time 196
5.2.1 ADC parameters 197
5.2.2 Analog-to-digital conversion techniques 203
5.3 Time-to-digital converters (TDCs) 209
5.3.1 Counter 209
5.3.2 Analog ramp 209
5.3.3 Digitizers with clock interpolation 209
5.4 Digital signal processing 210
References 216

6 Transistors and amplifiers 217
6.1 Bipolar transistors 217
6.1.1 Bipolar transistors in amplifiers 222
6.2 Field effect transistors 229
6.2.1 Junction field effect transistors 230
6.2.2 Metal-oxide-semiconductor field effect transistors 236
6.2.3 MOSFET types 241
6.2.4 MOS Transistors in Amplifiers 242
6.3 Noise in transistors 243
6.3.1 Noise in field effect transistors 243
6.3.2 Low-frequency excess noise ("$1/f$ noise") 248
6.3.3 Noise in bipolar transistors 248
6.3.4 Comparison between bipolar and field effect transistors 251
6.3.5 Noise optimization – capacitive matching revisited 252
6.4 Composite amplifiers 256
6.5 Overall noise of a detector module 265
6.6 Optimization for low power 266
6.6.1 Optimum operating current 267
6.6.2 Technology improvements 271
6.7 Power dissipation of an active pixel array *vs.* strip readout 274
References 275

7 Radiation effects 277
7.1 Radiation damage mechanisms 278
7.1.1 Displacement damage 279
7.1.2 Ionization damage 282
7.2 Radiation damage in diodes 283
7.2.1 Contributions to N_{eff} 286
7.2.2 Trapping 289
7.2.3 Ionization effects 292
7.3 Radiation damage in transistors and integrated circuits 292
7.3.1 Bipolar transistors 292
7.3.2 Junction field effect transistors (JFETs) 295
7.3.3 Metal-oxide-silicon field effect transistors (MOSFETs) 296
7.3.4 Radiation effects in integrated circuit structures 302
7.4 Dosimetry 303
7.5 Mitigation techniques 304
7.5.1 Detectors 304
7.5.2 Electronics 306
7.5.3 Summary 309
References 309

8 Detector systems 315
8.1 Conflicts and compromises 315
8.2 Design considerations 316
8.2.1 Detector geometry 316
8.2.2 Efficiency 316
8.2.3 Event rate 316
8.2.4 Readout 317
8.2.5 Support structures, cooling, and cabling 317
8.2.6 Cost 317
8.3 Segmentation 318
8.4 Tracking and vertex detectors at e^+e^- colliders 319
8.4.1 Layout and detector geometry 319
8.4.2 Electronics 323
8.4.3 "Common mode noise" 326
8.4.4 Noise limits in long strip detectors 327
8.4.5 CCD detectors at e^+e^- colliders 330
8.5 Vertex and tracking detectors at hadron colliders 337
8.5.1 CDF and D∅ 337
8.6 Silicon trackers at the Large Hadron Collider 342
8.6.1 Coping with high rates 343
8.6.2 Radiation damage 344
8.6.3 Layout 345
8.6.4 Readout electronics 348
8.6.5 Detector modules 353

8.6.6 Pixel detectors 357
8.6.7 ATLAS pixel detector 357
8.7 Monolithic active pixel devices 363
8.7.1 CMOS imagers 363
8.7.2 DEPFET pixel detectors 364
8.8 Astronomical imaging 366
8.9 Emerging applications 367
8.9.1 Space applications 367
8.9.2 X-ray imaging and spectroscopy 369
8.10 Design, assembly and test 372
8.10.1 Design 372
8.10.2 Assembly 374
8.10.3 Testing 375
8.11 Summary 377
References 378

9 Why things don't work 386
9.1 Reflections on transmission lines 386
9.2 Common pickup mechanisms 389
9.2.1 Noisy detector bias supplies 389
9.2.2 Light pickup 389
9.2.3 Microphonics 390
9.2.4 RF pickup 391
9.3 Pickup reduction techniques 392
9.3.1 Shielding 392
9.3.2 "Field line pinning" 394
9.3.3 "Self-shielding" structures 395
9.3.4 Inductive coupling 396
9.3.5 "Self-shielding" cables 397
9.3.6 Shielding summary 397
9.4 Shared current paths – grounding and the power of myth 398
9.4.1 Shared current paths ("ground loops") 398
9.4.2 Remedial techniques 400
9.4.3 Potential distribution on ground planes 403
9.4.4 Connections in multi-stage circuits 405
9.5 Breaking parasitic current paths 405
9.5.1 Isolate sensitive loops 406
9.5.2 Differential signal transmission 406
9.5.3 Blocking Common Mode Currents 408
9.5.4 Isolating parasitic ground connections by series resistors 409
9.5.5 Directing the current flow away from sensitive nodes 410
9.5.6 The folded cascode 412
9.6 Capacitors 414
9.7 System considerations 415

9.7.1 Choice of shaper 415
9.7.2 Local referencing 416

A Semiconductor device technology 418
A.1 Bulk material 418
A.2 Introduction of dopants 419
A.3 Deposition 420
A.4 Patterning 421
A.5 Surface passivation 422
A.6 Detector fabrication 422
A.7 Detector process flow 423
A.8 Strip detector structures 426
A.9 CMOS devices 428
References 429

B Phasors and complex algebra in electrical circuits 432

C Equivalent circuits 434

D Feedback amplifiers 438
D.1 Gain of a feedback amplifier 438
D.2 Linearity 439
D.3 Bandwidth 439
D.4 Series and shunt feedback 440
D.5 Input and output impedance 440
D.5.1 Series feedback 441
D.5.2 Shunt feedback 441
D.5.3 Output impedance 442
D.6 Loop gain 443
D.7 Stability 444
References 446

E The diode equation 447
E.1 Carrier concentrations in pure semiconductors 447
E.2 Carrier concentrations in doped crystals 450
E.3 *pn*-junctions 451
E.4 The forward-biased *pn*-junction 453
References 458

F Electrical effects of impurities and defects 459
F.1 Emission and capture processes 459
F.1.1 Electron capture 460
F.1.2 Electron emission 460
F.1.3 Hole capture and emission 460
F.1.4 Emission probabilities 461
F.2 Recombination 462
F.2.1 Band-to-band recombination 462

F.2.2 Recombination via intermediate states 463
F.3 Carrier generation 465
F.3.1 Generation in the depletion region 465
F.3.2 Generation in the neutral region 466
F.4 The origin of recombination and generation centers 467
F.5 The diode equation revisited 468
F.5.1 Reverse Current 468
F.5.2 Forward current 470
F.5.3 Comments 470
References 471

G Bipolar transistor equations 472
References 477

Index 478

1

DETECTOR SYSTEMS OVERVIEW

All semiconductor detector systems include the same basic functions. The signal from each sensor or sensor channel in a detector array must be amplified and processed for storage and analysis. Some functions are clearly associated with individual circuit blocks, but frequently circuit blocks perform multiple functions.

Figure 1.1 compares a "traditional" silicon detector system for charged particle spectroscopy with an integrated detector module. The left panel shows a room-temperature silicon detector, removed from the vacuum chamber in which it is operated. The detector is connected to a preamplifier through a vacuum feedthrough mounted on a vacuum flange. The pulse shaper and detector bias

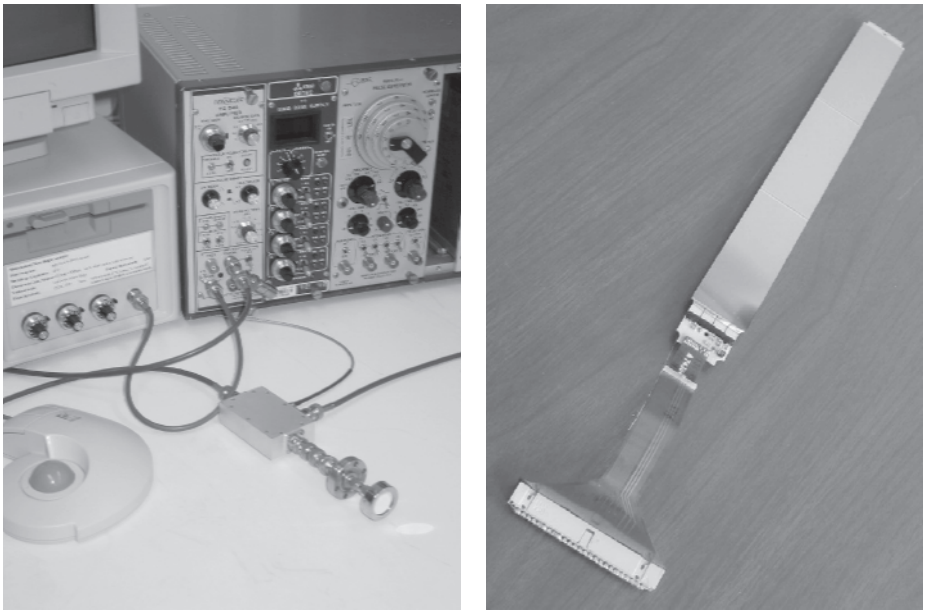

FIG. 1.1. Left: A "traditional" silicon detector system showing a single readout channel. The silicon sensor is the cylindrical object at the lower right. Right: A 512-channel detector module used for particle tracking. Three 2.5 cm wide × 6 cm long sensors are ganged together and read out by four integrated circuits with 128 channels each. A low-mass ribbon cable provides data and power connections to the external readout electronics.

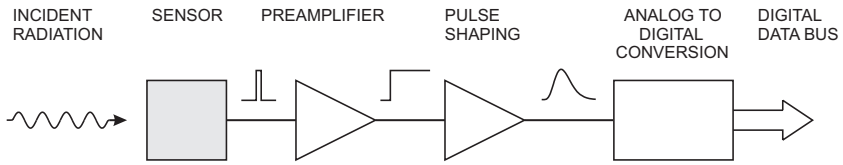

FIG. 1.2. Basic detector functions: Radiation is absorbed in the sensor and converted into an electrical signal. This low-level signal is integrated in a preamplifier, fed to a pulse shaper, and then digitized for subsequent storage and analysis.

supply reside in a NIM bin and the data acquisition system is a plug-in card in a PC, which also provides data storage and data display. In single-channel systems digitization and data storage are often combined in a single unit, a multichannel analyzer, whereas more complicated systems utilize a bank of external digitizers read out through a data bus (CAMAC, VME, VXI, PCI, *etc.*) and fed to a computer. Such systems are still in widespread use for high-resolution x-ray and gamma spectroscopy.

In contrast, the right panel of Figure 1.1 shows a 512-channel detector module from a high-energy physics experiment, CDF at FermiLab. The primary function of this detector is position sensing. Multiple layers of these detectors provide space points to reconstruct particle trajectories. The silicon sensor, the preamplifier, pulse shaper, digital readout control, and signal bussing are combined in one integrated unit, a detector module. The 512 channels of analog and digital electronics are accommodated in four integrated circuits (ICs), each about 6 mm in size (Kleinfelder *et al.* 1988).

Here the term detector becomes ambiguous, especially in experiments where the "detector" consists of several detector subsystems – tracking, calorimetry, muon detection – which in turn consist of many individual detector modules. Whenever ambiguities might arise we'll refer to the device that translates the presence of a particle to an electrical signal as a sensor.

The sequence of detector functions is illustrated in Figure 1.2 and described below.

1.1 Sensor

The sensor converts the energy deposited by a particle (or photon) to an electrical signal. This can be achieved in a variety of ways, but in this context energy is absorbed in a semiconductor, for example silicon, which produces mobile charge carriers – electron–hole pairs. An electric field applied to the sensor sweeps the charge carriers to electrodes, inducing an electrical current. The number of electron–hole pairs is proportional to the absorbed energy, so by integrating the signal current one obtains the signal charge, which is proportional to energy. As will be shown below, the sensor pulses can be quite short (of order nanoseconds

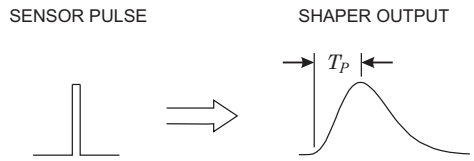

FIG. 1.3. In energy measurements a pulse processor typically transforms a short sensor current pulse to a broader pulse with a peaking time T_P.

or less) and the spatial extent of the charge cloud is small (of order microns), so semiconductor sensors can handle very high particle rates.

1.2 Preamplifier

The signal charge can be quite small, about 50 aC ($5 \cdot 10^{-17}$ C) for 1 keV x-rays and 4 fC ($4 \cdot 10^{-15}$ C) in a typical high-energy tracking detector, so the sensor signal must be amplified. The magnitude of the sensor signal is subject to statistical fluctuations, and electronic noise further "smears" the signal. These fluctuations will be discussed in detail in Chapters 2 and 3, but at this point we note that the sensor and preamplifier must be designed carefully to minimize electronic noise. A critical parameter is the total capacitance in parallel with the input, *i.e.* the sensor capacitance and input capacitance of the amplifier. The signal-to-noise ratio increases with decreasing capacitance. The contribution of electronic noise also relies critically on the next stage, the pulse shaper.

1.3 Pulse shaper

In semiconductor detector systems the primary function of the pulse shaper is to improve the signal-to-noise ratio. Although we are considering signal pulses, *i.e.* time-varying signals, the signal power is also distributed in frequency space, quantified by the pulse's Fourier transform. The frequency spectra of the signal and the noise differ, so one can improve the signal-to-noise ratio by applying a filter that tailors the frequency response to favor the signal, while attenuating the noise. Changing the frequency response also changes the time response, the pulse shape, so this function is called pulse shaping. As will be shown below, improving the signal-to-noise ratio commonly implies reducing the bandwidth, which increases the duration of the pulse (Figure 1.3).

Usually, we are not interested in measuring just one pulse, but many pulses in succession and often at a very high rate. Too large a pulse width will lead to pile-up of successive pulses, as shown in Figure 1.4 (left). A system that measures the peak amplitude will give an erroneous result for the second pulse. Pile-up can be ameliorated by reducing the pulse width, as shown in the second panel of Figure 1.4.

Figure 1.5 shows how the pulse transformation shown in Figure 1.3 can be accomplished. The preamplifier is configured as an integrator, which converts

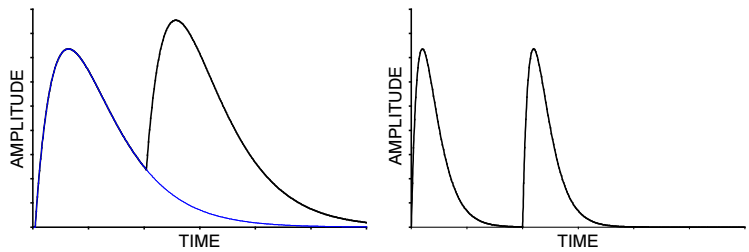

FIG. 1.4. Amplitude pile-up occurs when two pulses overlap (left). Reducing the shaping time allows the first pulse to return to the baseline before the second pulse arrives.

the narrow current pulse from the sensor into a step impulse with a long decay time. A subsequent CR high-pass filter introduces the desired decay time and an RC low-pass filter limits the bandwidth and sets the rise time. This will be discussed in more detail in Chapter 4. Shapers can be much more complex, using multiple integrators to improve pulse symmetry, for example. However, common to all shapers are operations that constrain the upper frequency bound, which sets the rise time, and the lower frequency bound, which determines the pulse duration. When designing a system it is necessary to find a balance between the conflicting requirements of reducing noise and increasing speed. Sometimes minimum noise is crucial, sometimes rate capability is paramount, but usually a compromise between the two must be found.

Although the primary measure of the signal energy is the charge, when the pulse shape is the same for all signal magnitudes, the pulse amplitude or "pulse height" is equivalent (hence the frequently used term "pulse height analysis"). The pulse height spectrum is the energy spectrum. This is convenient, since

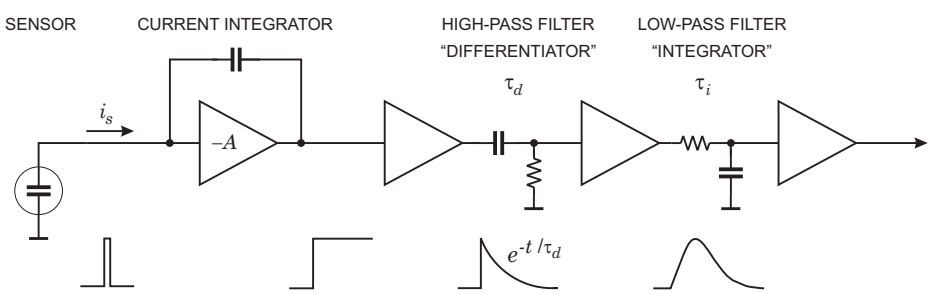

FIG. 1.5. Components of a pulse shaping system. The signal current from the sensor is integrated to form a step impulse with a long decay. A subsequent high-pass filter ("differentiator") limits the pulse width and the low-pass filter ("integrator") increases the rise-time to form a pulse with a smooth cusp.

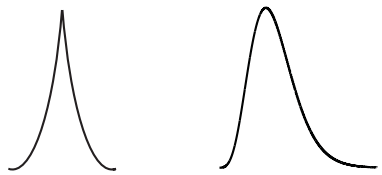

FIG. 1.6. The theoretical "optimum" shaper output (left) and a practical pulse (right), which attains its maximum for a measurable time.

analog-to-digital converters (ADCs) measure voltage or current amplitude. However, this imposes an additional requirement on the pulse shaper; the pulse shape should be compatible with the digitizer. Since the digitizer has a finite response time, the maximum signal amplitude should be maintained for a commensurate time, so the shaper output should have a smooth maximum. This is worth remembering, since the filter that theoretically "optimizes" signal-to-noise ratio for many detectors is a cusp, where the peak amplitude is attained for only an infinitesimally short time, as shown in Figure 1.6. Clearly, determining the amplitude of this pulse in a realistic system is fraught with uncertainties.

Sometimes the shaper is hidden; "charge sensing" ADCs are often used to digitize short pulses from photomultiplier tubes. Internally, the input stage integrates the input pulse and translates the signal charge to a voltage level, which is held for the duration of the digitization. This is also a form of pulse shaping. Very sophisticated shapers have been developed to optimize noise and rate capability, and also to reduce sensitivity to variations in sensor pulse shape. However, in many applications, shapers can be quite simple. Since all amplifiers have a limited bandwidth, every amplifier is a pulse shaper. Frequently, rather sophisticated pulse shaping can be implemented by tailoring the bandwidths of the amplifiers needed anyway to increase the signal level.

1.4 Digitizer

Analog-to-digital conversion translates a continuously varying amplitude to discrete steps, each corresponding to a unique output bit pattern. First developed for use in radiation detection, analog-to-digital conversion today is a mainstream technique and ADCs with a wide range of characteristics are available. A conceptually simple ADC is shown in Figure 1.7. The signal is fed in parallel to a bank of comparators with monotonically increasing thresholds, provided by a resistor voltage divider. When the pulse height exceeds a certain threshold, all comparators with lower thresholds fire and a decoder translates the hit pattern to a more convenient (*e.g.* binary) form. This technique is very fast, but requires many comparators, as the number of comparators determines the resolution. For example, 256 comparators can provide a full scale range of 1 V with 3.9 mV resolution. In the age of vacuum tubes or discrete transistors this technique was not very practical, as the space required for many precision comparators was pro-

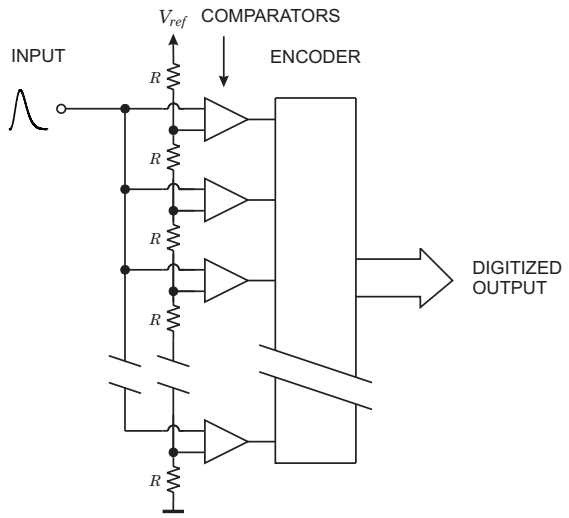

FIG. 1.7. A conceptually simple technique for analog-to-digital conversion utilizes a bank of comparators with increasing threshold levels. The address of the highest level comparator responding to a signal is encoded to provide a binary output.

hibitive. However, in monolithically integrated circuits it is quite feasible, but in practice power dissipation and chip size constrain the obtainable resolution. Generally, increasing circuit speed requires more power, so ADCs trade off resolution *vs.* speed. More sophisticated conversion techniques have been developed to provide high resolution (as high as 24 bits) with fewer circuit elements, but at the expense of conversion time. Generally, speed and resolution are opposing parameters, as are speed and power. Although a bit pattern appears unambiguous, ADCs are not perfect. Analog-to-digital conversion techniques with their strengths and flaws are discussed in Chapter 5.

1.5 Electro–mechanical integration

The ability to combine many sensor channels in a small volume brings with it the need to implement many connections, both within a detector module and also to connect modules to the "outside world". One must remove the heat due to electrical power dissipation, control "cross-talk" (unwanted coupling between different channels), provide precise mechanical positioning, and deal with a host of other problems that straddle the realms of electronic and mechanical design.

To illustrate some of these problems, consider vertex detection in high-energy physics. A powerful tool in identifying interesting events is the detection of secondary vertices. A particle formed in the primary collision, a B meson, for example, decays after a brief time to form new particles, whose tracks form a vertex displaced from the primary collision point. The formation of the initial particle

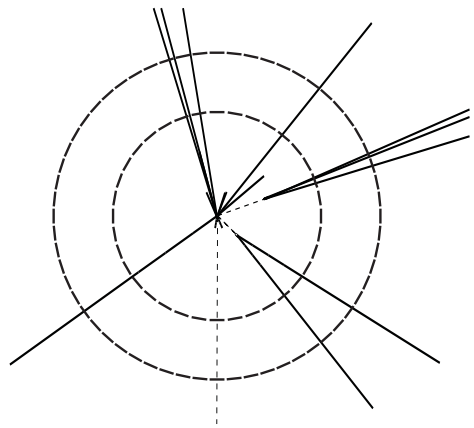

FIG. 1.8. Axial view of a collider event. Most tracks originate from the primary interaction point in the center. A reconstructed neutrino is shown as a dashed track, as it is not directly detected. Two B mesons are emitted toward the right and decay in flight. The decay products originate from displaced vertices, which are a few mm distant from the primary vertex. Concentric arrays of position-sensitive detectors, shown schematically and not to scale, provide track coordinates at two radii.

is inferred by reconstructing the trajectories of the secondaries and detecting the displaced vertex.

Figure 1.8 illustrates the detection of displaced vertices. Segmentation of the concentric detector layers provides both position resolution and the ability to separate adjacent tracks. When the track density is not too high, high resolution in the $r\varphi$ plane alone is sufficient for pattern recognition and track reconstruction. Basic requirements for vertex detection can be derived from this simple tracking system with two layers at radii r_1 and r_2 and resolutions of σ_1 at r_1 and σ_2 at r_2. The impact parameter resolution

$$\sigma_b^2 \approx \left(\frac{\sigma_1 r_2}{r_2 - r_1}\right)^2 + \left(\frac{\sigma_2 r_1}{r_2 - r_1}\right)^2 = \frac{1}{(r_2 - r_1)^2}\left[(\sigma_1 r_2)^2 + (\sigma_2 r_1)^2\right] \ . \tag{1.1}$$

The position resolution at the inner radius is weighted by the outer radius, so precision at the inner radius is paramount. If the two layers have equal resolution $\sigma_1 = \sigma_2 = \sigma$, this result can be rewritten as

$$\left(\frac{\sigma_b}{\sigma}\right)^2 \approx \left(\frac{1}{1 - r_1/r_2}\right)^2 + \left(\frac{1}{r_2/r_1 - 1}\right)^2 \ . \tag{1.2}$$

The geometrical impact parameter resolution is limited by the ratio of the outer to inner radius, so it is desirable to measure the first space point at as small a

radius as possible. The obtainable impact parameter resolution improves rapidly from $\sigma_b/\sigma = 7.8$ at $r_2/r_1 = 1.2$ to $\sigma_b/\sigma = 2.2$ at $r_2/r_1 = 2$ and attains values $<$ 1.3 at $r_2/r_1 > 5$. For $\sigma = 10$ μm and $r_2/r_1 = 2$, $\sigma_b \approx 20\mu$m. Thus, the inner layer requires a high-resolution detector, which also implies a high-density electronic readout with associated cabling and cooling, mounted on a precision support structure. All of this adds material, which imposes an additional constraint.

The obtainable vertex resolution is affected by angular deflection due to multiple scattering from material in the detector volume. The scattering angle

$$\Theta_{rms} = \frac{0.0136[\mathrm{GeV/c}]}{p_\perp}\sqrt{\frac{x}{X_0}}\left[1 + 0.038 \cdot \ln\left(\frac{x}{X_0}\right)\right] , \tag{1.3}$$

where $p_\perp$ is the particle momentum, x the thickness of the material, and X_0 the radiation length (see Particle Data Group 2004 for a concise summary). As noted above, the position resolution at inner radii is critical, so it is important to minimize material close to the interaction. Typically, the first layer of material is the beam pipe.

Consider a Be beam pipe of $x = 1$ mm thickness and $R = 5$ cm radius. The radiation length of Be is $X_0 = 35.3$ cm, so $x/X_0 = 2.8 \cdot 10^{-3}$ and at $p_\perp = 1$ GeV/c the scattering angle $\Theta_{rms} = 0.56$ mrad. This corresponds to $\sigma_b\Theta_{rms} = 28\,\mu$m, which in this example would dominate the obtainable resolution. Clearly, any material between the interaction and the measurement point should be minimized and the first measurement should be at as small a radius as possible. This exercise shows how experimental requirements drive the first detector layers to small radii, which increases the particle flux (hits per unit area) and radiation damage.

The need to reduce material imposes severe constraints on the sensor and electronics, the support structures, and the power dissipation, which determines the material in the cooling systems and power cabling. Since large-scale arrays combine both analog and digital functions in the detector module, special techniques must be applied to reduce pickup from digital switching without utilizing massive shielding. Similar constraints apply in other applications, x-ray imagers, for example, where Compton scattering blurs the image.

Subsequent chapters will provide detailed discussions of the relevant physics and design parameters. In the spirit of a "road map" the remainder of this chapter summarizes the key aspects of semiconductor detector systems.

1.6 Sensor structures I

1.6.1 *Basic sensor*

Semiconductor detectors are basically ionization chambers. In the simplest configuration an absorbing medium is subtended by a pair of electrodes with an applied voltage, as illustrated in Figure 1.9. Absorbed radiation liberates charge pairs, which move under the influence of an applied field and induce an electrical current in the external circuit.

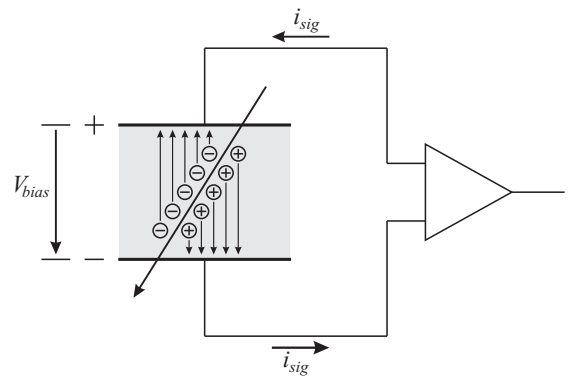

FIG. 1.9. Charge collection in a simple ionization chamber.

1.6.2 *Position sensing*

The electrodes of the sensor can be segmented to provide position information. Now the magnitude of the signal measured on a given electrode depends on its position relative to the sites of charge formation, as shown in Figure 1.10. Segmenting one electrode into strips, as shown in the left panel of Figure 1.11, provides position information in one dimension. Angled tracks will deposit charge on two or more strips. Evaluating the ratio of charge deposition allows interpolation to provide position resolution better than expected from the electrode pitch alone. We'll return to this later. A second orthogonal set of strips on the opposite face gives two-dimensional position readout, shown in the second panel of Figure 1.11.

In a colliding-beam experiment the strip pitch (center-to-center distance) is typically $25 - 100\,\mu$m and lengths range from centimeters to tens of centimeters, usually aligned parallel to the beam axis to provide $r\varphi$ coordinates. The maximum strip length per sensor is limited by wafer size ($10 - 15$ cm for detector-grade Si), so multiple sensors are ganged to form longer electrodes. Practical detectors

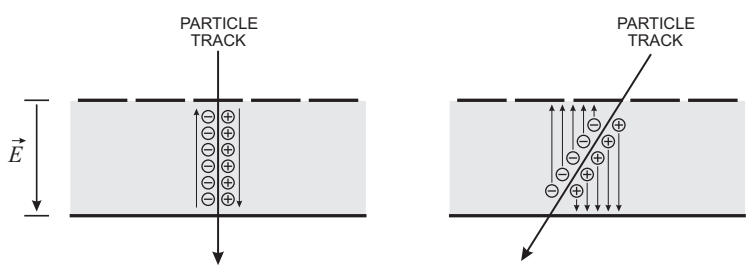

FIG. 1.10. Segmenting the sensor electrode provides position information. Angled tracks deposit charge on two or more electrodes.

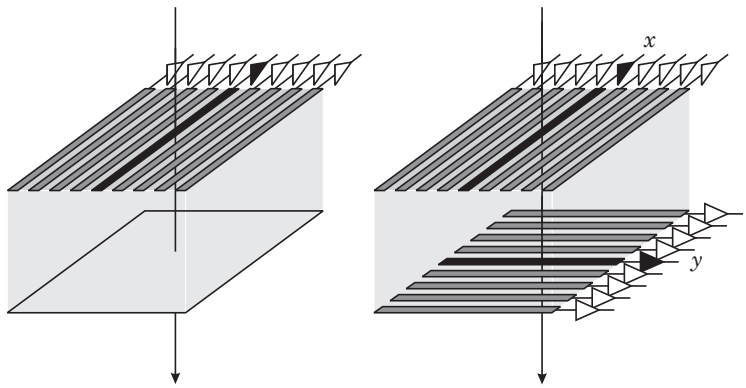

FIG. 1.11. Subdividing an electrode into strips provides one-dimensional position sensing (left). Subdividing both electrodes to form orthogonal strips provides two-dimensional imaging.

have used strips as long as 30 or 40 cm, limited by electronic noise and the hit rate per strip.

Two-dimensional position sensing using crossed strips is simple, but has problems at high hit densities. Each hit generates an x- and a y-coordinate. However, n tracks generate n x-coordinates and n y-coordinates, simulating n^2 hits of which $n^2 - n$ are fake. The "ghosts" can only be exorcised with additional information to eliminate coordinates not consistent with tracks, clearly a formidable task in a mixture of stiff and soft tracks with low-momentum particles looping in a magnetic field. A compromise solution that is often adequate utilizes "small-angle stereo", where the strips subtend a small angle, rather than 90°.

Small-angle stereo is illustrated in Figure 1.12. The area subtended by two sensing elements (strips) of length L_1 and L_2 arranged at an angle 90° is $A = L_1 L_2$, so a hit in a given strip can form combinations with hits on all of the transverse strips – the probability of "ghosting" is maximal. However, if the

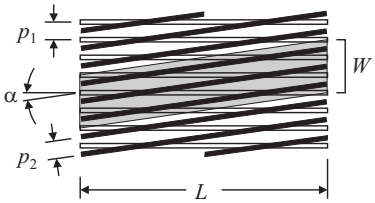

FIG. 1.12. Small-angle stereo reduces the area subtended by strips that could provide a coincident signal. The width W of the shaded area subject to confusion is $L(p_2/p_1)\tan\alpha + p_2$.

angle α subtended by the two strip arrays is small (and their lengths L are approximately equal), the capture area

$$A \approx L^2 \frac{p_2}{p_1} \tan\alpha + Lp_2 \ . \tag{1.4}$$

Consider a given horizontal strip struck by a particle. To determine the longitudinal coordinate, all angled strips that cross the primary strip must be checked and every hit that deposits charge on these strips adds a coordinate that must be considered in conjunction with the coordinate defined by the horizontal strip. Since each strip captures charge from a width equal to the strip pitch, the exact width of the capture area is an integer multiple of the strip pitch. The probability of multiple hits within the acceptance area, and hence the number of "ghosts", is reduced as α is made smaller, but at the expense of resolution in the longitudinal coordinate.

1.6.3 *Pixel devices*

To obtain unambiguous two-dimensional information the sensor must provide fine segmentation in both dimensions, which can be achieved either by geometrical or electronic segmentation. Charge coupled devices (CCDs), random access pixel devices, and silicon drift chambers represent different approaches to obtaining nonprojective two-dimensional information. The conceptually simplest implementation is shown in Figure 1.13. The sensor electrodes are patterned as a checkerboard and a matching two-dimensional array of readout electronics is connected via a two-dimensional array of contacts, for example solder bumps. In this scheme the pixel size is limited by the area required by each electronic readout cell. Pixel sizes of 30 – 100 μm are practical today, depending on the complexity of the circuitry required in each pixel. Figure 1.13 also shows that the readout IC requires more area than the pixel array to accommodate the

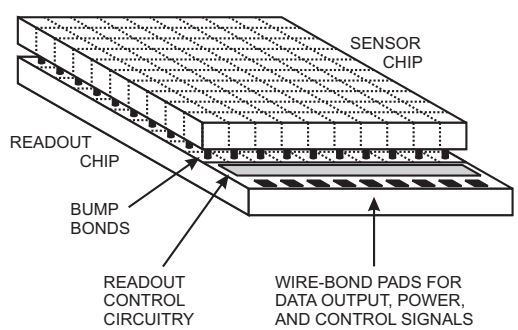

FIG. 1.13. Schematic view of a hybrid pixel detector. A pixellated sensor chip is connected to a matching array of readout amplifiers by a two-dimensional array of solder bumps. The readout chip extends beyond the sensor chip to accommodate readout and control circuitry in addition to wire bonds for external connections.

readout control and driver circuitry and additional bond pads for the external connections. Since multiple readout ICs are needed to cover more than several cm^2, this additional area constrains designs that require full coverage. Examples for integrating multiple readout ICs and sensors are discussed in Chapter 8.

Implementing this structure monolithically would be a great simplification and some work has proceeded in this direction. Before describing these structures, it is useful to discuss some basics of semiconductor detectors.

1.7 Sensor physics

1.7.1 *Signal charge*

All of the configurations discussed above differ only in the structures at the surface of the sensor. Common to them is that the charge pairs are formed in the sensitive volume and the average signal charge

$$Q_s = \frac{E}{E_i} e \ , \tag{1.5}$$

where E is the absorbed energy, E_i the energy required to form a charge pair, and e the electronic charge. In solids the absorbed energy must exceed the bandgap to form mobile charge carriers. In Si the gap energy is 1.12 eV, so photons with greater energy, *i.e.* wavelengths less than 1.1 μm, can be detected. At higher energies (> 50 eV) the additional constraint of momentum conservation becomes significant and the ionization energy $E_i = 3.6$ eV. As will be discussed in Chapter 2, the ionization energy E_i is proportional to the bandgap, so higher bandgap materials yield less signal charge.

For a charged particle track traversing the sensor, the energy loss E – and hence the signal charge Q_s – will increase with sensor thickness. Minimum ionizing particles average about 80 electron–hole pairs per μm path length in silicon. For x-rays absorbed by the photoelectric effect, the deposited energy is fixed, but the sensor must be sufficiently thick to provide good efficiency. For gamma-rays above 100 keV Compton interactions dominate, so the sensor volume must be sufficiently large to accommodate multiple sequential interactions (for a discussion see Knoll 2000).

When a low-energy x-ray is absorbed by the photoelectric effect, the charge deposition is localized, with a charge cloud whose extent is determined by the range of the ejected photoelectron. A charged particle traversing the sensor forms charge pairs along the track with a radial extent of order μm. The signal is formed when the liberated charge carriers move, which changes the induced charge on the sensor electrodes. This will be treated quantitatively in Chapter 2, so at this point we'll simply note that when all signal charges have reached their respective electrodes, the change in induced charge, *i.e.* the integrated signal current, is Q_s.

To establish the electric field a potential is applied between the electrodes to accelerate the charge carriers. As the carriers move through the medium they scatter. After a short equilibration time (of order ps in Si) carrier transport

becomes nonballistic and the velocity does not depend on the duration of acceleration, but only on the magnitude of the local electric field (see Sze 1981). Thus, the velocity of carriers at position x depends only on the local electric field $E(x)$, regardless of where they originated and how long they have moved. The carrier velocity

$$\vec{v}(x) = \mu \vec{E}(x) \ , \tag{1.6}$$

where μ is the mobility. For example, in Si the mobility is $1350\,\mathrm{V/cm \cdot s^2}$ for electrons and $450\,\mathrm{V/cm \cdot s^2}$ for holes. As an estimate to set the scale, applying 30 V across a 300 μm thick absorber yields an average field of 10^3 V/cm, so the velocity of electrons is about $1.4 \cdot 10^6$ cm/s and it will take about 20 ns for an electron to traverse the detector thickness. A hole takes three times as long.

1.7.2 *Sensor volume*

To establish a high field with a small quiescent current, the conductivity of the absorber must be low. Signal currents are typically of order μA, so if in the above example the quiescent current is to be small compared to the signal current, the resistance between the electrodes should be $\gg 30\,\mathrm{M\Omega}$. In an ideal solid the resistivity depends exponentially on the bandgap. Increasing the bandgap reduces the signal charge, so the range of suitable materials is limited. Diamond is an excellent insulator, but the ionization energy E_i is about 6 eV and the range of available thickness is limited. In semiconductors the ionization energy is smaller, 2.9 eV in Ge and 3.6 eV in Si. Si material can be grown with resistivities of order $10^4\,\Omega\,\mathrm{cm}$, which is too low; a 300 μm thick sensor with $1\,\mathrm{cm}^2$ area would have a resistance of $300\,\Omega$, so 30 V would lead to a current flow of 100 mA and a power dissipation of 3 W. On the other hand, high-quality single crystals of Si and Ge can be grown economically with suitably large volumes, so to mitigate the effect of resistivity one resorts to reverse-biased diode structures.

The conductivity of semiconductors is controlled by introducing dilute concentrations of impurities into the crystal, a process called doping. Let the semiconductor be of atomic number Z. If the dopant is of atomic number $Z+1$, one of the shell electrons is only lightly bound and can be thermally excited into the conduction band, so electrons are available as mobile charge carriers. If the atomic number of the dopant is $Z-1$, one of the bonds lacks an electron, but only little energy is needed to "borrow" an electron from a nearby atom. Thus, the unfilled bond moves and acts like a positive mobile charge, a "hole". To form a diode, one can start with material doped to provide mobile electrons, "n-type" material. By introducing a $Z-1$ dopant from the surface, a region can be formed with holes as mobile carriers, "p-type" material. This forms a "pn-junction". When a voltage is applied with positive polarity on the n-side and negative on the p-side (reverse bias), the electrons on the n-side and the holes on the p-side are drawn away from the junction. Thus, the region adjacent to the pn-junction is depleted of mobile charge and forms an insulator, over which the applied voltage builds up the desired electric field, as illustrated in Figure 1.14.

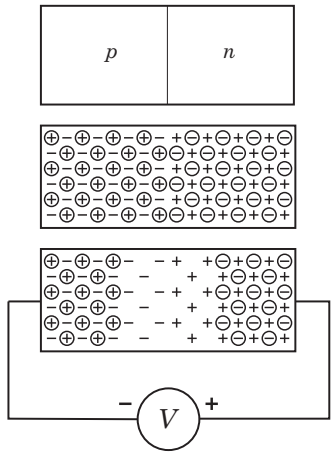

FIG. 1.14. Adjoining regions of p- and n-type doping form a pn-junction (top). The charge of the mobile electrons and holes (circled) is balanced by the charge of the atomic cores, so charge neutrality is maintained. When an external potential is applied with positive polarity on the n-side and negative polarity on the p-side (bottom), the mobile charges are drawn away from the junction. This leaves a net space charge from the atomic cores, which builds up a linear electric field in the junction. This is treated analytically in Chapter 2.

Note that even in the absence of an externally applied voltage, thermal diffusion forms a depletion region. As electrons and holes diffuse from their original host atoms, a space charge region is formed and the resulting field limits the extent of thermal diffusion. As a result, every pn-junction starts off with a nonzero depletion width and a potential difference between the p- and n-sides, the "built-in" potential V_{bi}.

Figure 1.15 shows the cross-section of a typical detector diode. The pn-junction is formed by introducing the dopant at the upper surface. The detector junction is in the middle. Similarly doped regions to the left and right indicate a guard ring, which surrounds the detector diode to isolate it from the edge of the wafer. Mechanical damage at the edge leads to very large leakage currents. The guard ring, biased at the same potential as the detector electrode, captures the edge currents and also forms a well-defined electrical boundary for the detector diode (the active area ends midway between the detector electrode and the guard ring). Metallization layers (typically aluminum) deposited on the electrodes provide electrical contact. The intermediate silicon surface is protected by a layer of SiO_2 that provides a well-controlled interface to the silicon lattice. In detectors the surface side of the junction is usually much more heavily doped than the substrate material, so the resulting asymmetric junction depletes into the bulk. Appendix A provides more details on detector structures and fabrication.

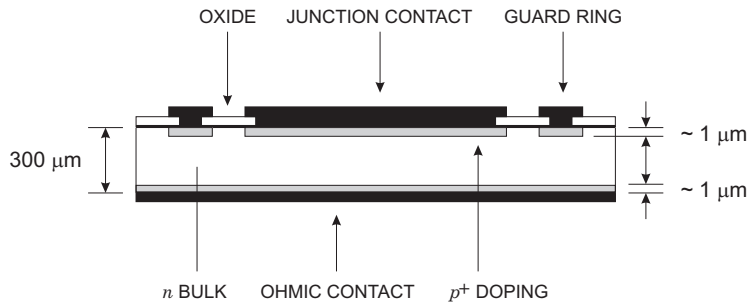

FIG. 1.15. Detector diodes are usually asymmetric, with a highly doped layer at the surface and a lightly doped bulk. With reverse bias the junction depletes into the bulk. SiO_2 layers protect the silicon surface and a guard ring isolates the diode from the edges of the chip.

A reverse bias voltage V_b yields the depletion width

$$w_d = \sqrt{\frac{2\varepsilon(V_b + V_{bi})}{Ne}} \ , \tag{1.7}$$

where N is the dopant concentration in the bulk and ε the dielectric constant ($11.9\varepsilon_0$ for Si). The "built-in" junction potential V_{bi} in detector diodes is typically about 0.5 V. When the depletion width is less then the silicon thickness, the diode is "partially depleted". When w_d extends to the back contact the diode is "fully depleted".

The depleted junction volume is free of mobile charge and thus forms a capacitor, bounded by the conducting p- and n-type semiconductor on each side. The capacitance

$$C = \varepsilon \frac{A}{w_d} = A\sqrt{\frac{\varepsilon e N}{2(V_b + V_{bi})}} \ . \tag{1.8}$$

For bias voltages $V_b \gg V_{bi}$

$$C \propto \frac{1}{\sqrt{V_b}} \ . \tag{1.9}$$

In technical units

$$\frac{C}{A} = \frac{\varepsilon}{w_d} \approx 1 \left[\frac{\mathrm{pF}}{\mathrm{cm}}\right] \frac{1}{W} \ . \tag{1.10}$$

A Si diode with 100 μm thickness has a capacitance of about 1 pF/mm^2. This applies to a detector whose electrodes are large compared to the depletion thickness. In strip and pixel detectors the fringing capacitance to neighboring electrodes usually dominates. The interstrip capacitance depends on the ratio of electrode

width w to strip pitch p. For typical geometries the interstrip capacitance C_s per cm length l follows the relationship (Demaria *et al.* 2000)

$$\frac{C_s}{l} = \left(0.03 + 1.62\frac{w + 20\mu\mathrm{m}}{p}\right)\left[\frac{\mathrm{pF}}{\mathrm{cm}}\right] \ . \tag{1.11}$$

Typically, the interstrip capacitance is about 1 pF/cm. The backplane capacitance

$$C_b \approx \varepsilon\varepsilon_0\frac{pl}{w} \ . \tag{1.12}$$

Since the adjacent strips confine the fringing field lines to the interstrip boundaries, the strip appears as an electrode with a width equal to the strip pitch. Corrections apply at large strip widths (Barberis *et al.* 1994).

Ideally, reverse bias removes all mobile carriers from the junction volume, so no current can flow. However, thermal excitation can promote electrons across the bandgap, so a current flows even in the absence of radiation, hence the term "dark current". The probability of electrons surmounting the bandgap is increased strongly by the presence of impurities in the lattice, as they introduce intermediate energy states in the gap that serve as "stepping stones". As derived in Appendix F the reverse bias current depends exponentially on temperature T,

$$I_R \propto T^2 \exp\left(-\frac{E_g}{2kT}\right) \ , \tag{1.13}$$

where E_g is the bandgap energy and k the Boltzmann constant, so cooling the detector can reduce leakage substantially. The ratio of leakage currents at temperatures T_1 and T_2

$$\frac{I_R(T_2)}{I_R(T_1)} = \left(\frac{T_2}{T_1}\right)^2 \exp\left[-\frac{E_g}{2k}\left(\frac{T_1 - T_2}{T_1 T_2}\right)\right] \ . \tag{1.14}$$

In Si (E_g =1.12 eV) this yields a ten-fold reduction in leakage current when the temperature is lowered by 14 °C from room temperature.

1.7.3 *Charge collection*

How quickly electrons and holes are swept from the depletion region is determined by the electric field. To simplify the following equations we'll set $V \equiv V_b + V_{bi}$. At low reverse bias the field in the depletion region initially has a triangular profile

$$|E(x)| = \frac{eN}{\varepsilon}(w_d - x) = \sqrt{\frac{2Ne}{\varepsilon}V} \cdot \left(1 - \frac{x}{w_d}\right) \equiv E_{max} \cdot \left(1 - \frac{x}{w_d}\right) \tag{1.15}$$

up to the voltage where the depletion width w_d equals the thickness of the bulk d, corresponding to the depletion voltage

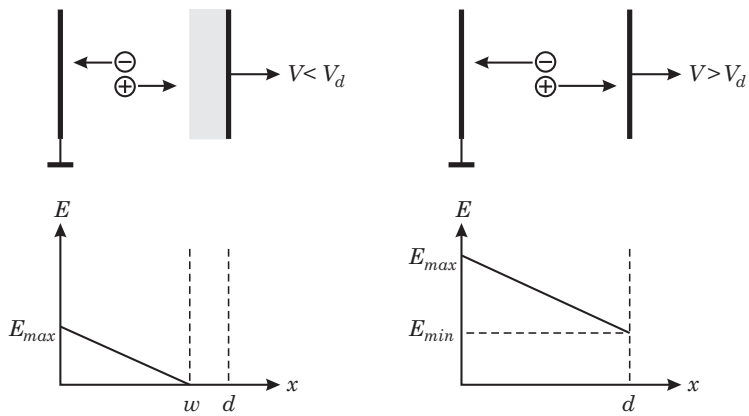

FIG. 1.16. Electric field distributions in a partially depleted detector (left) and a detector operated with overbias (right).

$$V_d = \frac{Ned^2}{2\varepsilon} \ . \tag{1.16}$$

Increasing the bias voltage V beyond this point ("overbias", often called "overdepletion") increases the field uniformly and evens out the field profile

$$|E(x)| = \frac{2V_d}{d}\left(1 - \frac{x}{d}\right) + \frac{V - V_d}{d} \ . \tag{1.17}$$

Then the maximum field is $(V + V_d)/d$ and the minimum field $(V - V_d)/d$. Figure 1.16 illustrates the electric field distributions in partial depletion and with overbias.

When radiation forms electron–hole pairs, they drift under the influence of the field with a velocity $v = \mu E$. The time required for a carrier to traverse the full detector thickness, the collection time, is

$$t_c = \frac{d^2}{2\mu V} \log\left(\frac{V + V_d + 2V_{bi}}{V - V_d}\right) \ , \tag{1.18}$$

where the collection time for electrons or holes is obtained by using the appropriate mobility. At full depletion or beyond, the collection time can be estimated by using the average field $\overline{E} = V/d$, so

$$t_c \approx \frac{d}{v} = \frac{d}{\mu\overline{E}} = \frac{d^2}{\mu V} \tag{1.19}$$

and charge collection can be sped up by increasing the bias voltage. In partial depletion, however, the collection time is independent of bias voltage and determined by the doping concentration alone, as d^2/V remains constant. This is discussed in Chapter 2.

In practice the dopant concentration N of silicon wafers is expressed as the resistivity of the material $\rho = (e\mu N)^{-1}$, as this is readily measurable. Using this parameter and introducing technical units yields the depletion voltage

$$V_{dn} = 4 \left[\frac{\Omega \cdot \mathrm{cm}}{(\mu\mathrm{m})^2} \right] \cdot \frac{d^2}{\rho_n} - V_{bi} \tag{1.20}$$

for n-type material and

$$V_{dp} = 11 \left[\frac{\Omega \cdot \mathrm{cm}}{(\mu\mathrm{m})^2} \right] \cdot \frac{d^2}{\rho_p} - V_{bi} \tag{1.21}$$

for p-type material. The resistivity of silicon suitable for tracking detectors (or more precisely, the highest resistivity available economically) is $5 - 10\,\mathrm{k}\Omega\,\mathrm{cm}$. Note that in $10\,\mathrm{k}\Omega\,\mathrm{cm}$ n-type Si the built-in voltage by itself depletes $45\,\mu\mathrm{m}$ of material. Detector wafers are typically $300\,\mu\mathrm{m}$ thick. Hence, the depletion voltage in n-type material is $35 - 70\,\mathrm{V}$ for the resistivity range given above. Assuming $6\,\mathrm{k}\Omega\,\mathrm{cm}$ material ($V_d = 60\,\mathrm{V}$) and an operating voltage of $90\,\mathrm{V}$, the collection times for electrons and holes are $8\,\mathrm{ns}$ and $27\,\mathrm{ns}$, respectively. Electron collection times tend to be somewhat longer then given by eqn 1.15 since the electron mobility decreases at fields $> 10^3\,\mathrm{V/cm}$ (see Chapter 2 and Sze 1981) and eventually the drift velocity saturates at $10^7\,\mathrm{cm/s}$. At saturation velocity the collection time is $10\,\mathrm{ps}/\mu\mathrm{m}$. In partial depletion, as noted above, the collection time is independent of voltage and depends on resistivity alone. For electrons the collection time constant

$$\tau_{cn} = \rho\varepsilon = 1.05 \left[\frac{\mathrm{ns}}{\mathrm{k}\,\Omega \cdot \mathrm{cm}} \right] \cdot \rho \ . \tag{1.22}$$

To increase the depletion width or speed up the charge collection one can increase the voltage, but ultimately this is limited by the onset of avalanching. At sufficiently high fields (greater than about $10^5\,\mathrm{V/cm}$ in Si) electrons acquire enough energy between collisions that secondary electrons are ejected. At even higher fields holes can eject secondary electrons, which in turn can eject new secondaries, and a self-sustaining charge avalanche forms (see Chapter 2). This phenomenon is called "breakdown" and can cause permanent damage to the sensor. In practice avalanching often occurs at voltages much lower than predicted by eqn 1.17, since high fields can build up at the relatively sharp edges of the doping distribution or electrode structures. When controlled, charge avalanching can be used to increase the signal charge, as discussed in Chapter 2. In detecting visible light, the primary signal charge is quite small, so this technique is most often applied in photodiodes to provide internal gain and bring the signal above the electronic noise level (avalanche photodiodes or APDs). APDs must be designed carefully to prevent breakdown and also to reduce additional signal fluctuations introduced by the avalanche process. Bias voltage and temperature both affect the gain strongly, so they must be kept stable.

1.7.4 *Energy resolution*

The minimum detectable signal and the precision of the amplitude measurement are limited by fluctuations. The signal formed in the sensor fluctuates, even for a fixed energy absorption. Generally, sensors convert absorbed energy into signal quanta. In a scintillation detector absorbed energy is converted into a number of scintillation photons. In an ionization chamber energy is converted into a number of charge pairs (electrons and ions in gases or electrons and holes in solids). The absorbed energy divided by the excitation energy yields the average number of signal quanta

$$N = \frac{E}{E_i} \ . \tag{1.23}$$

This number fluctuates statistically, so the relative resolution

$$\frac{\Delta E}{E} = \frac{\Delta N}{N} = \frac{\sqrt{FN}}{N} = \sqrt{\frac{FE_i}{E}} \ . \tag{1.24}$$

The resolution improves with the square root of energy. F is the Fano factor, which comes about because multiple excitation mechanisms can come into play and reduce the overall statistical spread. For example, in a semiconductor absorbed energy forms electron–hole pairs, but also excites lattice vibrations – quantized as phonons – whose excitation energy is much smaller (meV *vs.* eV). Thus, many more excitations are involved than apparent from the charge signal alone and this reduces the statistical fluctuations of the charge signal. For example, in Si the Fano factor is 0.1. The Fano factor is explained in Chapter 2.

In most applications, the intrinsic energy resolution of semiconductor sensors is so good that external contributions determine the overall fluctuations. However, for low-energy x-rays signal charge fluctuations are significant, whereas in gamma-ray detectors electronic noise tends to determine the obtainable energy resolution. For minimum ionizing charged particles, it is the statistics of energy loss. Since the energy deposited by minimum ionizing particles varies according to a Landau–Vavilov distribution (Figure 1.17) with $\sigma_Q/Q_s \approx 0.2$ in $300\,\mu$m of Si, the inherent energy resolution of the detector is negligible. Nevertheless, electronic noise is still important in determining the minimum detectable signal, *i.e.* the detection efficiency.

1.7.5 *Position resolution*

The position resolution of the detector is determined to first order by the electrode geometry. The size and shape of the electrodes is limited by the size of a wafer, on the one hand (10 or 15 cm diameter for detector grade material), and the resolution capability of IC fabrication technology on the the other ($\sim 1\,\mu$m). In practice the lower bound is set by the readout electronics, which in the smallest dimension tend to require $20 - 50\,\mu$m overall width. Most commonly, sensors for tracking applications have strip electrodes. The strips are usually $8 - 12\,\mu$m wide, placed on a pitch of $25 - 50\,\mu$m, and $6 - 12$ cm long. Frequently, multiple sensor wafers are ganged to form longer electrodes.

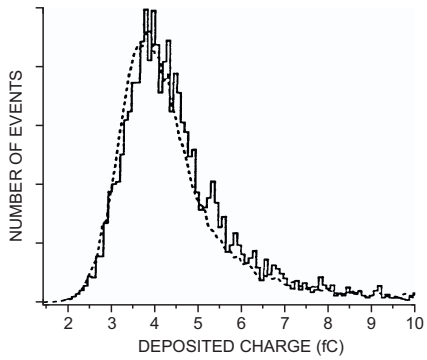

FIG. 1.17. Measured energy loss distribution of 1.5 MeV/c electrons in a silicon detector. The dashed line is a Vavilov theory calculation. (Wood *et al.* 1991. Figure courtesy of P. Skubic)

It is important to note that despite the gaps between electrodes, the detectors still remain 100% efficient. The field lines remain parallel in the detector until near the surface, where they bend along the surface and end on the electrode. Hence, the electrical segmentation is determined by the electrode pitch, rather than the width. Since the response function is essentially box-like, the position resolution of a single detector is equal to the strip pitch p. However, for tracks randomly aligned with respect to a strip, the differences between the measured and the true positions have a Gaussian distribution with the standard deviation

$$\sigma^2 = \int_{-p/2}^{p/2} \frac{x^2}{p} dx = \frac{p^2}{12} \ , \tag{1.25}$$

so the root mean square (rms) resolution is the strip pitch divided by $\sqrt{12}$. The same mechanism leads to "sampling noise" in image processing or when digitizing an analog waveform, and is discussed in Section 5.2.

To first order the electrons and holes simply follow the field lines on which they originated and end on a certain electrode. In reality, however, they are also subject to thermal diffusion, which spreads the charge cloud transversely as the charges drift through the detector, with an rms width

$$\sigma_y = \sqrt{2Dt} \ . \tag{1.26}$$

Since the diffusion constant is linked to the mobility by the Einstein relation

$$D = \frac{kT}{e}\mu \tag{1.27}$$

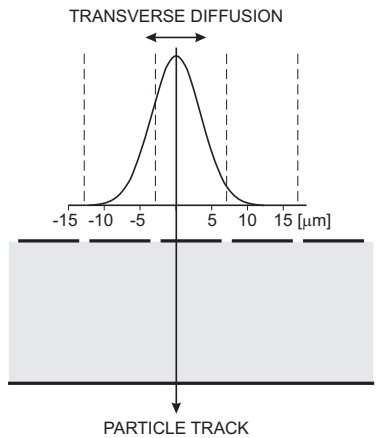

FIG. 1.18. Transverse diffusion distributes charge on multiple strip electrodes, here shown on a $20\,\mu$m pitch. The charge division boundaries are indicated by dashed lines. By evaluating the distribution of charge on the electrodes, the position resolution can be improved beyond the geometric resolution $\sigma = p/\sqrt{12}$.

and the collection time is inversely proportional to the carrier mobility, the transverse diffusion is the same for electrons and holes. Using the average field approximation $\overline{E} = V/d$ the transverse diffusion

$$\sigma_y = \sqrt{2Dt} \approx \sqrt{2\frac{kT}{e}\frac{d^2}{V_b}}\ , \tag{1.28}$$

which is independent of mobility, giving the same result for electrons and holes. For $d = 300\,\mu$m, $T = 300\,$K and $V_b = 100\,$V the transverse diffusion $\sigma_y \approx 7\,\mu$m.

At first glance this might seem to degrade the obtainable position resolution, but it can, in fact, be turned to advantage, since transverse diffusion spreads charge to neighboring strips. As illustrated in Figure 1.18, one can evaluate the charge distribution over a central strip and its neighbors to improve the position resolution beyond the limit given by strip geometry. Since the fractional charge terminating on the neighboring strip is determined by superimposed Gaussian distributions (Lüth 1990), whose integral falls off rapidly for deviations beyond several standard deviations, this technique is practical only for a rather limited range of strip pitches. In the interpolation regime the position resolution is inversely proportional to signal-to-noise ratio. Kenney *et al.* (1993) have applied a weighted interpolation algorithm to rectangular pixels of $34\,\mu\text{m} \times 125\,\mu$m with $S/N = 55$. In the direction of the $34\,\mu$m pitch the measured resolution was $2.2\,\mu$m, whereas in the direction of the $125\,\mu$m pitch the interpolation could only be applied to the outer $25\,\mu$m regions to yield $5.3\,\mu$m resolution. The resolution in the central region was $75/\sqrt{12}\,\mu$m.

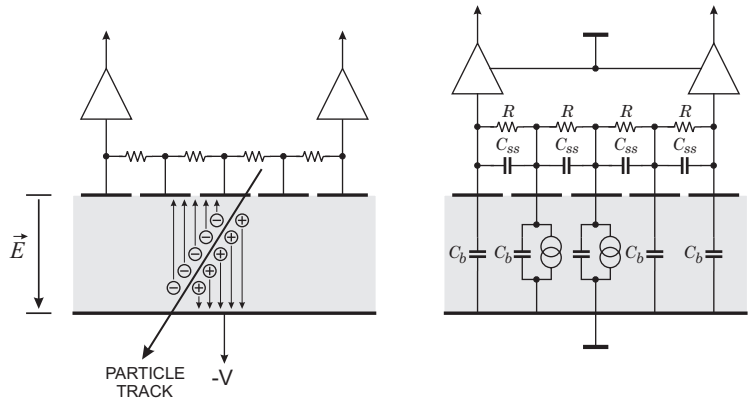

FIG. 1.19. Intermediate "floating" electrodes can be used to reduce the effective readout pitch (left). The equivalent circuit (right) shows how the signal current induced on the "floating" electrodes is transferred to the readout amplifiers through the intermediate capacitive dividers formed by the strip-to-strip capacitance C_{ss} and the backplane capacitance C_b.

The range of charge interpolation can be extended by introducing intermediate strips that are not connected to readout channels (Figure 1.19). The bias resistors keep all strips at the same quiescent potential, but the time constant formed by the bias resistance and the strip capacitance is made so large that the potential of a "floating" strip can change individually in response to signal charge. The charge induced on the "floating" strips is coupled capacitively to its neighbors. The readout amplifiers must have a low input impedance, so that the signal current from a given electrode will divide inversely proportional to the effective coupling capacitance. It is crucial that all electrodes be at the same quiescent potential, to ensure uniform charge collection efficiency. Connecting the bias resistors as shown in Figure 1.19 ensures that each electrode is biased at the input voltage of the amplifiers. The biasing resistors must be large to reduce electronic noise, as will be discussed in Chapter 4, but still small enough that the detector leakage current does not alter the electrode voltages significantly.

For simplicity, first assume that the backplane capacitance C_b is zero. Then the capacitances coupling the central strip to the two amplifiers are formed by two interstrip capacitances C_{ss} in series on each side. Thus, the signal current will divide equally. For the strip to the left of center, the coupling capacitance is C_{ss} to the left-hand amplifier and $C_{ss}/3$ to the right-hand amplifier, so the left-hand amplifier will receive 3/4 of the signal.

This technique can also be used to reduce the number of readout channels. However, a portion of the signal charge remains on the backplane capacitance C_b of each strip in the signal path. For charge induced on the central strip, the charge transferred to one of the two amplifiers

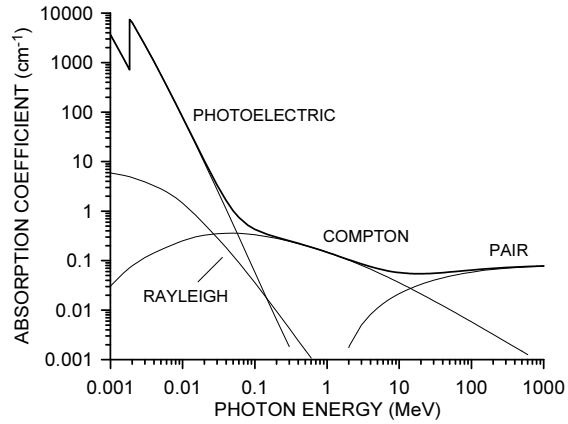

FIG. 1.20. The photon absorption coefficient μ *vs.* energy in silicon. At low energies photoelectric absorption dominates. Above 100 keV Compton scattering takes over and at high energies pair production dominates. The fraction of photons that have interacted within a distance x is $1 - \exp(\mu x)$.

$$\frac{Q_a}{Q_s} = \frac{1}{2} \cdot \frac{1}{1 + 2(C_b/C_{ss}) + \frac{1}{2}\left(C_b/C_{ss}\right)^2} \approx \frac{1}{2} \cdot \frac{1}{1 + 2\left(C_b/C_{ss}\right)} \ . \tag{1.29}$$

For $C_b/C_{ss} = 0.1$, a typical ratio for strip detectors, the exact expression yields $Q_{1a}/Q_s = 0.41$, so the backplane capacitance incurs an 18% loss in the summed signal from both amplifiers. However, the capacitance presented to the amplifier inputs is also smaller than in a fully instrumented system, so interpolation via floating strips can reduce overall power dissipation with respect to a fully instrumented readout. Double-track resolution, however, will be determined by the readout pitch.

In attempting to optimize position resolution, other effects must also be considered. The energy deposited by minimum ionizing particles fluctuates along the track, so tracks impinging at the same position will show varying centroids in the induced charge. A further limit on the position resolution is imposed by energetic delta-electrons formed along the track trajectory, which can skew the charge centroid appreciably (Bedeschi *et al.* 1989).

In x-ray imaging applications where photoelectric absorption dominates, the resolution is limited by the range of the emitted photoelectron, as it deposits energy along its path. The binding energy $E_b \approx 2\,\text{keV}$, so the photoelectron's energy

$$E_k = E_{photon} - E_b \ . \tag{1.30}$$

For a 20 keV photon the photoelectron's range is about 5 μm, whereas for 100 keV the range is about 80 μm. As can be seen in Figure 1.20 Compton scattering is about equally probable at 100 keV and dominates up to about 10 MeV.

However, given the maximum practical silicon detector thickness of several mm reasonable efficiency only obtains to 30 or 40 keV. Materials with higher atomic number are necessary for higher energies. Alternative materials are discussed briefly in Chapter 2. At high energies pair production can be used to determine the direction of gamma rays by detecting the emitted electrons and positrons in a silicon strip tracker (Chapter 8).

1.8 Sensor structures II – monolithic pixel devices

In the early years of large-scale semiconductor detectors the monolithic integration of large scale sensors with electronics was viewed as the "holy grail". Clearly, it is an appealing concept to have a $6 \times 6\,\mathrm{cm}^2$ detector tile that combines a strip detector and 1200 channels of readout electronics with only the power and data readout as external connections. The problem was perceived at the time to be the incompatibility between IC and detector fabrication processes (see Appendix A). Development of an IC-compatible detector process allowed the monolithic integration of high-quality electronics and full depletion silicon sensors without degrading sensor performance (Holland and Spieler 1990), subsequently extended to full CMOS circuitry (Holland 1992). Nevertheless, a simple yield estimate shows that this isn't practical. In the conventional scheme reading out ~ 1200 channels with a $50\,\mu$m readout pitch requires 10 ICs with 128 channels each. These devices are complex, so their yield is not 100%. Even when assuming 90% functional yield per 128-channel array, the probability of ten adjacent arrays on the wafer being functional is prohibitively small. The integration techniques are applicable, however, to simpler circuitry and have been utilized in monolithic pixel detectors (Snoeys *et al.* 1992). The oldest and most widespread wafer-scale monolithic imaging device is the charge coupled device.

1.8.1 *Charge coupled devices*

The classic high-resolution pixel array is the charge coupled device (CCD), which combines the charge readout with the sensors. Figure 1.21 illustrates the principle. In signal acquisition mode the pixels function as small ionization chambers. To transfer the signal charge to the readout amplifier, additional electrodes are appropriately biased to shift the charge to the adjacent electrode. By applying the appropriate sequence of pulses, the signal charge is sequentially transferred to the output electrode, which in turn is connected to a readout amplifier. This structure allows small pixel sizes, about $10\,\mu$m, and provides full coverage. The drawback is that the readout is sequential, so larger arrays require more readout time. Since charge is commonly transferred over thousands of pixels, the charge transfer efficiency η from one pixel to the next must be very close to unity. After transferring through n pixels the signal arriving at the output node is attenuated by η^n, but modern fabrication techniques provide practically 100% charge transfer over $\sim 10^4$ pixels. Pixels are read out sequentially, column by column as shown in Figure 1.21. Typically, a single amplifier reads out the entire array. Low noise militates against fast clocking, so readout times are long. This is discussed

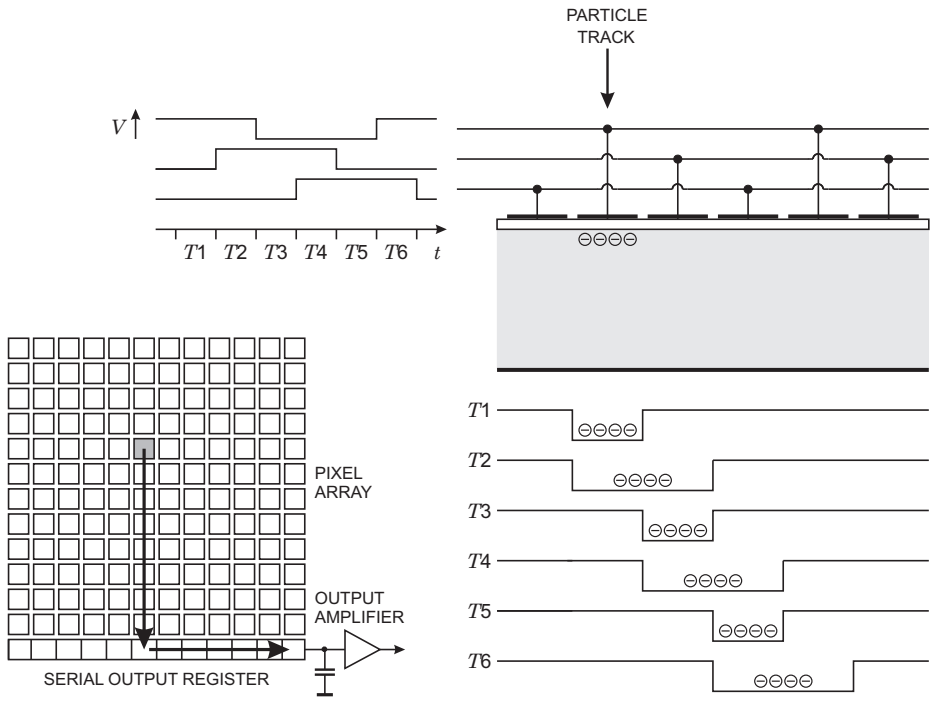

FIG. 1.21. Upper right: schematic cross-sectional view of a CCD. Voltages are applied to the electrodes according to the timing diagram at the upper left. The potential sequence shifts the charge from the track to the right. Three electrodes comprise one pixel, but all charge from the track subtended by the pixel is drawn to the pixel's left-most electrode. Six clock periods shift the charge to the neighboring pixel. The pixels are read out sequentially (bottom left). Charge is transferred down the column and then horizontally. The charge is deposited on a storage capacitor and transferred to the readout line by the output amplifier.

in Chapter 4. Large arrays are commercially available; the SLD detector (Abe *et al.* 1997) used $16 \times 80\,\mathrm{mm}^2$ devices with $20\,\mu$m pixels. The sensitive depth was $20\,\mu$m, so minimum ionizing particles yielded a broad charge distribution peaking at about $1200\,e$. Electronic noise was $100\,e$. The thin depletion depth reduces the signal, but limits transverse diffusion and provides excellent position resolution. The readout rate was 5 MHz and four readout amplifiers were used to speed up the readout. CCDs are in widespread use, but high-energy physics and x-ray detection require specialized devices.

1.8.2 *Silicon drift chambers*

An ingenious structure that provides the functionality of a CCD without discrete transfer steps is the silicon drift chamber (Gatti and Rehak 1984, Rehak *et*

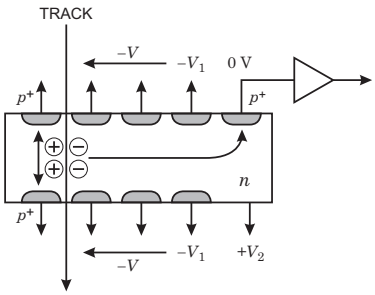

FIG. 1.22. Principle of a silicon drift chamber. The n-type bulk is depleted from both surfaces by a series of p^+ electrodes, biased to provide a positive potential gradient along the center axis of the detector. Holes drift to the p electrodes, whereas electrons are transported parallel to the surface and then attracted to the collection electrode, where the signal is read out.

al. 1985). In this device a potential trough is established in the bulk, so that the signal charge is collected in the trough and then drifts towards the readout electrode (Figure 1.22). The position is derived from the time it takes for a signal charge to move to the output, so the detector requires a time reference. When a pulsed accelerator or pulsed x-ray tube is used, the start time is readily available. With random rates, as with radioactive sources, the time reference must be derived from the sensor. This will be discussed in Chapters 3 and 4. Although originally proposed as a position-sensing device, the Si drift chamber's other useful application is energy spectroscopy. Since this structure collects charge from a large area onto a small collection electrode, the capacitance presented to the readout amplifier is small (order 10 – 100 fF), so the electronic noise can be very low. This can be exploited in x-ray detection and in photodiodes. Various drift detector topologies are described by Lutz (1999).

1.8.3 *Monolithic active pixel sensors*

Neither CCDs nor silicon drift devices can be fabricated using standard IC fabrication processes. The doping levels required for diode depletion widths of 100 μm or more are much lower than used in commercial integrated circuits. The process complexity and yield requirements of the readout electronics needed for most application dictate the use of industry-standard fabrication processes. In contrast to detectors, where the entire thickness of the silicon wafer is utilized for charge collection, integrated electronics utilize only a thin layer, of order μm, at the surface of the silicon. The remainder of the typically 500 μm thick wafer provides mechanical support, but also serves to capture deleterious impurities, through gettering processes described in Appendix A. IC substrate material – typically grown by the Czochralski method – has both crystalline defects and impurities, whereas detector grade material utilizes float-zone material, which is

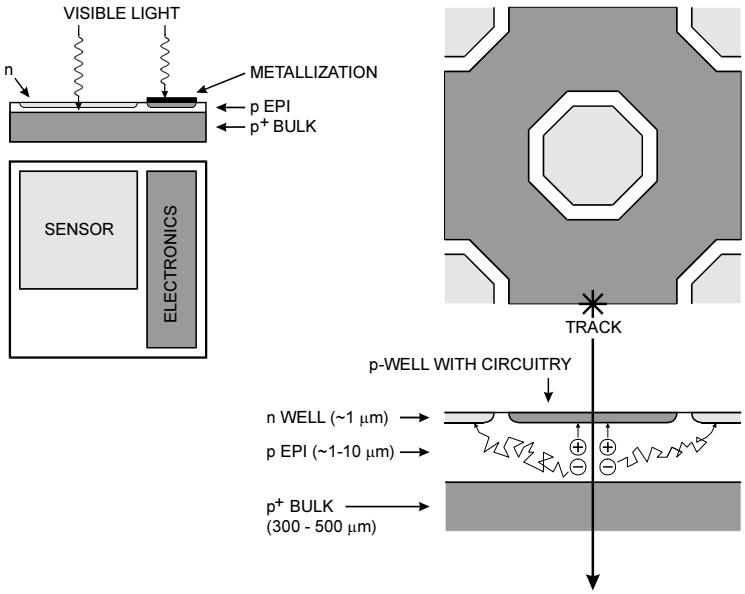

FIG. 1.23. An active pixel sensor that integrates sensors and electronics monolithically. The left shows an implementation for visible light. Electrons are collected directly by the sensor electrode formed by the n-well. Light penetrates only a short distance, so the portion of the epi-layer covered by the electronics is insensitive. The right shows an alternative layout for high-energy particles. Electrons formed by a track traversing the electronics diffuse towards the n-well electrodes. This structure provides 100% sensitive area, but diffusion transport leads to long collection times.

dislocation-free and achieves the very low impurity levels needed for high resistivity. High-quality transistors also require low defect densities, but much higher doping levels than detectors, so a thin high quality layer is epitaxially grown on the Czochralski substrate (referred to as the "epi-layer"). The doping levels are still quite high, typically corresponding to $1 - 10\,\Omega\,\mathrm{cm}$ resistivity (10^3 times lower than in typical detectors), so breakdown limits depletion widths to several μm or less.

Visible light ($400 - 700\,\mathrm{nm}$) in silicon is practically fully absorbed in a thickness of $0.5 - 7\,\mu$m, so thin depletion layers are usable, also because diffusion from non-depleted silicon also adds to the charge signal. Driven by the digital camera market and other commercial applications, there is widespread activity in the application of conventional IC fabrication processes to optical imaging (Fossum 1997). These devices, called CMOS imagers or active pixel sensors, utilize a portion of the pixel cell as a sensor, as illustrated in Figure 1.23. Each pixel includes an active region (the sensor) with adjacent amplifier and readout circuitry. Metallization layers provide connections between the sensor and the electronics, and

between the components in the electronics cell. Light impinging on the sensor is detected, but is blocked by the metallization. Typical "fill factors", the ratio of light-sensitive area to pixel area, are 20 – 30%, but this can be improved as smaller feature sizes shrink the electronics.

High-energy particles, on the other hand, traverse both the metallization and the transistors, so devices have been developed that seek to utilize the epi-layer in the entire pixel as the sensor region, as shown in the right panel of Figure 1.23 (Deptuch *et al.* 2003, Turchetta *et al.* 2003, Kleinfelder *et al.* 2004). Since the depletion layers are thin, this device relies on diffusion for a substantial portion of the recovered charge. Thus, the recovered signal charge is much smaller than in fully depleted detectors, about $1000\,e$ compared to $22\,000\,e$ for a $300\,\mu$m thick sensor. Diffusion is channeled laterally by the potential well formed in the epi-layer, as it is lightly doped with respect to the substrate and the p-wells that accommodate the electronics. Since the pixel capacitance is small, electronic noise levels can also be low.

However, relying on diffusion increases the collection time to about 100 ns, which can still provide the time resolution required at high-luminosity colliders, but radiation damage will degrade the carrier lifetime (see Appendix F) to order 1 – 10 ns after a relatively short time at high luminosity and the small radii where pixel detectors are needed. Incomplete charge collection also limits the usability of these devices in applications that require good energy resolution, *e.g.* x-ray spectroscopy, although they may be acceptable for counting measurements.

The conceptually simplest form of an active pixel array is a matrix of transistors. During image acquisition all transistors are inactive and signal charge is stored on their input capacitance. Control electrodes are bussed by row and outputs by column. During readout each transistor is addressed individually by selecting the appropriate row and all columns read out simultaneously. This structure has been implemented by monolithically integrating the transistors (called DEPFETs) in a high-resistivity substrate (Kemmer and Lutz 1987). This arrangement allows the readout of individual pixels, but unlike more complex active pixel devices can't signal which pixels to read out. When reading out full image frames the performance of this structure is comparable to a CCD with a fully parallel readout. Understanding the limits of this technique requires some additional background, so we'll return to it in Chapter 8.

Some active pixel designs replicate the fully sequential readout used for CCDs. This is a good match to digital photography, where every pixel carries information. The electronic circuitry in each pixel cell is quite simple and readout can be slow, so the circuitry does not occupy much area. Slow readout applied to charged particle detection also allows simple circuitry and facilitates low electronic noise. However, in sparse data environments with high event rates, such as high luminosity hadron colliders, "smart pixels" that signal the presence of a hit and then allow the selective readout of struck pixels sorted by time stamp are necessary. This requires both fast response, to allow time stamping, and local threshold discrimination, buffering, and readout logic. This drives up circuit com-

plexity substantially, so the "real estate" occupied by the electronics increases, both in the pixel and in the common control and readout circuitry. This will be illustrated in Chapter 8. Comparison of various technologies requires careful scrutiny that the adopted architecture and circuit design match the intended purpose and not some simpler situation.

1.9 Electronics

Electronics are a key component of all modern detector systems. Although experiments and their associated electronics can take very different forms, the same basic principles of the electronic readout and the optimization of signal-to-noise ratio apply to all.

The purpose of pulse processing and analysis systems is to

1. Acquire an electrical signal from the sensor. Typically this is a short current pulse.
2. Tailor the time response of the system to optimize
   (a) the minimum detectable signal (detect hit/no hit),
   (b) energy measurement,
   (c) event rate,
   (d) time of arrival (timing measurement),
   (e) insensitivity to sensor pulse shape,
   or some combination of the above.
3. Digitize the signal and store for subsequent analysis.

Position-sensitive detectors utilize the presence of a hit, amplitude measurement, or timing, so these detectors pose the same set of requirements. Generally, these properties cannot be optimized simultaneously, so compromises are necessary.

In addition to these primary functions of an electronic readout system, other considerations can be equally or even more important, for example radiation resistance, low power (portable systems, large detector arrays, satellite systems), robustness, and – last, but not least – cost.

1.10 Detection limits and resolution

In addition to signal fluctuations originating in the sensor, the minimum detection limit and energy resolution are subject to fluctuations introduced by the electronics. The gain can be controlled very precisely, but electronic noise introduces baseline fluctuations, which are superimposed on the signal and alter the peak amplitude. Figure 1.24 (left) shows a typical noise waveform. Both the amplitude and time distributions are random.

When superimposed on a signal, the noise alters both the amplitude and time dependence. Figure 1.24 (right) shows the noise waveform superimposed on a small signal. As can be seen, the noise level determines the minimum signal whose presence can be discerned.

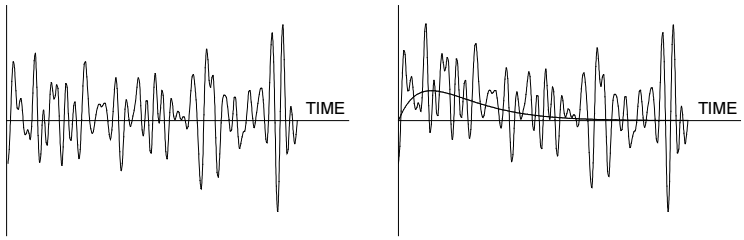

FIG. 1.24. Waveforms of random noise (left) and signal + noise (right), where the peak signal is equal to the rms noise level ($S/N = 1$). The noiseless signal is shown for comparison.

In an optimized system, the time-scale of the fluctuations is comparable to that of the signal, so the peak amplitude fluctuates randomly above and below the average value. This is illustrated in Figure 1.25, which shows the same signal viewed at four different times. The fluctuations in peak amplitude are obvious, but the effect of noise on timing measurements can also be seen. If the timing signal is derived from a threshold discriminator, where the output fires when the signal crosses a fixed threshold, amplitude fluctuations in the leading edge translate into time shifts. If one derives the time of arrival from a centroid analysis, the timing signal also shifts (compare the top and bottom right figures). From this

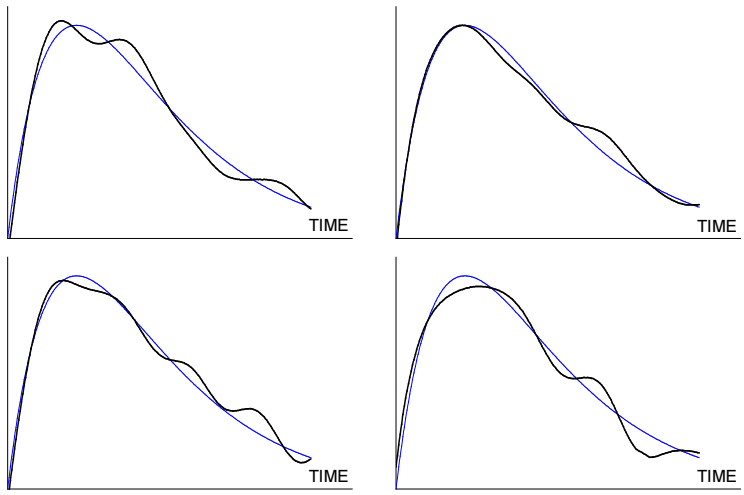

FIG. 1.25. Signal plus noise at four different times, shown for a signal-to-noise ratio of about 20. The noiseless signal is superimposed for comparison.

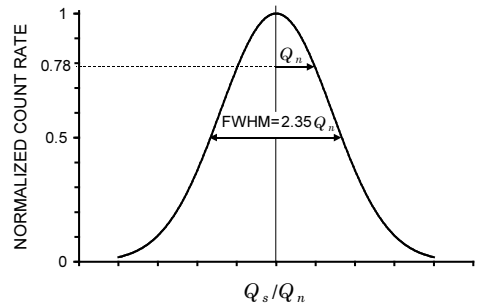

FIG. 1.26. Repetitive measurements of the signal charge yield a Gaussian distribution whose standard deviation equals the rms noise level Q_n. Often the width is expressed as the full width at half maximum (FWHM), which is 2.35 times the standard deviation.

one sees that signal-to-noise ratio is important for all measurements – sensing the presence of a signal or the measurement of energy, timing, or position.

1.10.1 *Electronic noise*

Electronic noise originates as both velocity or number fluctuations. This is discussed in detail in Chapter 3. Velocity fluctuations arise from thermal excitation. The spectral density of the noise power can be derived directly as the long wavelength limit in Planck's theory of black-body radiation (see Chapter 3). At the frequencies of interest here the spectral density is independent of frequency; the spectrum is "white". Number fluctuations occur when charge carriers are injected into a sample independently of one another. Thermionic emission or current flow through a semiconductor *pn*-junction are common examples. This is called "shot noise" and also has a white spectrum.

In electronic circuits the noise sources can be modeled either as voltage or current sources. Generally, the frequency spectra of the signal and the noise are different. Typically, the noise spectra extend over a greater frequency band than the signal, so by shaping the frequency response of the system one can optimize the signal-to-noise ratio. The amplitude distribution of the noise is Gaussian, so superimposing a constant amplitude signal on a noisy baseline will yield a Gaussian amplitude distribution whose width equals the noise level (Figure 1.26).

To analyze the contributions of electronic noise, let's consider a typical detector front-end as shown in Figure 1.27. The sensor is represented by a capacitance C_d, a relevant model for most detectors. Bias voltage is applied through resistor R_b and the signal is coupled to the preamplifier through a blocking capacitor C_c. The series resistance R_s represents the sum of all resistances present in the input signal path, *e.g.* the electrode resistance, any input protection networks, and parasitic resistances in the input transistor. The preamplifier provides gain and feeds a pulse shaper, which tailors the overall frequency response to optimize

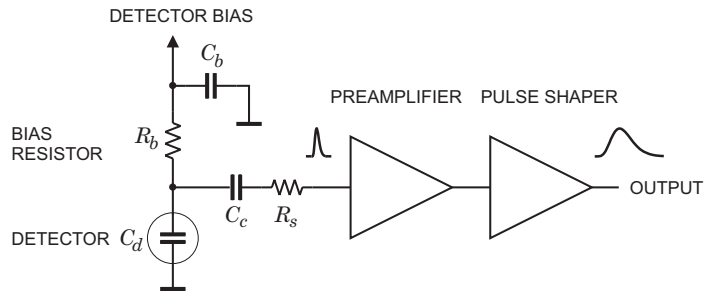

FIG. 1.27. Typical detector front-end circuit.

signal-to-noise ratio while limiting the duration of the signal pulse to accommodate the signal pulse rate. Even if not explicitly stated, all amplifiers provide some form of pulse shaping, due to their limited frequency response.

The equivalent circuit for the noise analysis (Figure 1.28) includes both current and voltage noise sources. The leakage current of a semiconductor detector, for example, fluctuates due to electron emission statistics. This "shot noise" i_{nd} is represented by a current noise generator in parallel with the detector. Resistors exhibit noise due to thermal velocity fluctuations of the charge carriers. This noise source can be modeled either as a voltage or current generator. Generally, resistors shunting the input act as noise current sources and resistors in series with the input act as noise voltage sources (which is why some in the detector community refer to current and voltage noise as "parallel" and "series" noise). Since the bias resistor effectively shunts the input, as the capacitor C_b passes current fluctuations to ground, it acts as a current generator i_{nb} and its noise current has the same effect as the shot noise current from the detector. Any other shunt resistances can be incorporated in the same way. Conversely, the series resistor R_s acts as a voltage generator. The amplifier's noise is described

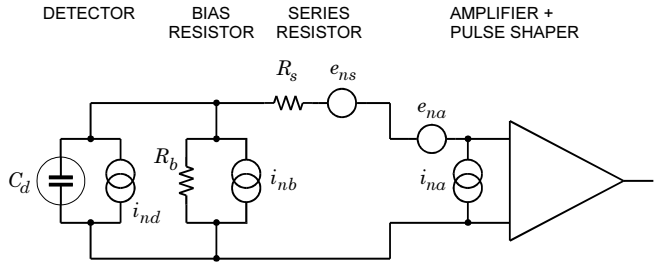

FIG. 1.28. Equivalent circuit for noise analysis of the detector front-end in Figure 1.27.

fully by a combination of voltage and current sources at its input, shown as e_{na} and i_{na}.

Shot noise and thermal noise have a "white" frequency distribution, *i.e.* the spectral densities are constant with the magnitudes

$$\begin{aligned} i_{nd}^2 &= 2eI_d \\ i_{nb}^2 &= 4kT/R_b \\ e_{ns}^2 &= 4kTR_s \end{aligned}$$

where e is the electronic charge, I_d the detector bias current, k the Boltzmann constant, and T the temperature. Typical amplifier noise parameters e_{na} and i_{na} are of order $\mathrm{nV}/\sqrt{\mathrm{Hz}}$ and fA to $\mathrm{pA}/\sqrt{\mathrm{Hz}}$. Trapping and detrapping processes in resistors, dielectrics and semiconductors can introduce additional fluctuations whose noise power frequently exhibits a $1/f$ spectrum. The spectral density of the $1/f$ noise voltage is

$$e_{nf}^2 = \frac{A_f}{f} \ , \tag{1.31}$$

where the noise coefficient A_f is device specific and of order $10^{10} - 10^{12}\,\mathrm{V}^2$.

A portion of the noise currents flows through the detector capacitance, resulting in a frequency-dependent noise voltage $i_n/\omega C_d$, which is added to the noise voltages in the input circuit. Since the individual noise contributions are random and uncorrelated, they add in quadrature. The total noise at the output of the pulse shaper is obtained by integrating over the full bandwidth of the system.

1.10.2 *Amplitude measurements*

Since radiation detectors are typically used to measure charge, the system's noise level is conveniently expressed as an equivalent noise charge Q_n, which is equal to the detector signal that yields a signal-to-noise ratio of one. The equivalent noise charge is commonly expressed in Coulombs, the corresponding number of electrons, or the equivalent deposited energy (eV). For a capacitive sensor

$$Q_n^2 = i_n^2 F_i T_S + e_n^2 F_v \frac{C^2}{T_S} + F_{vf} A_f C^2 \ , \tag{1.32}$$

where C is the sum of all capacitances shunting the input. Note that the voltage noise contributions increase with capacitance. The shape factors F_i, F_v, and F_{vf} depend on the shape of the pulse determined by the shaper. T_S is a characteristic time, for example the peaking time of a semi-Gaussian pulse (Figure 1.3) or the prefilter integration time in a correlated double sampler (discussed in Chapter 4). The shape factors F_i, F_v are easily calculated,

$$F_i = \frac{1}{2T_S} \int_{-\infty}^{\infty} [W(t)]^2 \, dt \quad \text{and} \quad F_v = \frac{T_S}{2} \int_{-\infty}^{\infty} \left[\frac{dW(t)}{dt} \right]^2 dt \ , \tag{1.33}$$

where for time invariant pulse shaping $W(t)$ is simply the system's impulse response (the output signal seen on an oscilloscope) with the peak output signal

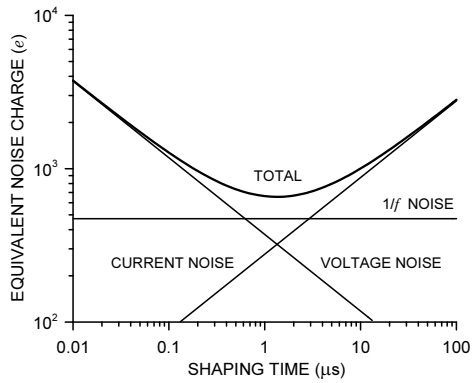

FIG. 1.29. Equivalent noise charge *vs.* shaping time of a typical detector system.

normalized to unity. For more details see the papers by Goulding (1972), Radeka (1968, 1974), and Goulding and Landis (1982).

A pulse shaper formed by a single differentiator and integrator with equal time constants $\tau_d = \tau_i = \tau \equiv T_S$ as in Figure 1.5 has $F_i = F_v = 0.9$ and $F_{vf} = 4$, independent of the shaping time constant. The overall noise bandwidth, however, depends on the time constant, *i.e.* the characteristic time T_S. The contribution from noise currents increases with shaping time, *i.e.* pulse duration, whereas the voltage noise decreases with increasing shaping time. Noise with a $1/f$ spectrum depends only on the ratio of upper to lower cutoff frequencies (integrator to differentiator time constants), so for a given shaper topology the $1/f$ contribution to Q_n is independent of T_S . Increased detector capacitance shifts the voltage noise contribution upward and the noise minimum to longer shaping times. Pulse shapers can be designed to reduce the effect of current noise, *e.g.* mitigate radiation damage. Increasing pulse symmetry tends to decrease F_i and increase F_v (*e.g.* to 0.45 and 1.0 for a shaper with one CR differentiator and four cascaded integrators).

For the circuit shown in Figures 1.27 and 1.28

$$Q_n^2 = \left(2eI_d + \frac{4kT}{R_b} + i_{na}^2\right) F_i T_S + \left(4kTR_s + e_{na}^2\right) F_v \frac{C_d^2}{T_S} + F_{vf} A_f C_d^2 \ . \quad (1.34)$$

As the shaping time T_S is changed, the total noise goes through a minimum, where the current and voltage contributions are equal. Figure 1.29 shows a typical example. At short shaping times the voltage noise dominates, whereas at long shaping times the current noise takes over. The noise minimum is flattened by the presence of $1/f$ noise. Increasing the detector capacitance will increase the voltage noise and shift the noise minimum to longer shaping times (Figure 4.29).

For quick estimates one can use the following equation, which assumes a field effect transistor (FET) amplifier (negligible i_{na}) and a simple *CR-RC* shaper with time constants τ (equal to the peaking time).

$$Q_n^2 = 12\left[\frac{e^2}{\mathrm{nA}\cdot\mathrm{ns}}\right] I_d\tau + 6\cdot 10^5\left[\frac{e^2\,\mathrm{k}\Omega}{\mathrm{ns}}\right]\frac{\tau}{R_b} + 3.6\cdot 10^4\left[\frac{e^2\,\mathrm{ns}}{(\mathrm{pF})^2(\mathrm{nV})^2/\mathrm{Hz}}\right] e_n^2\frac{C^2}{\tau}$$

For a given amplifier (*i.e.* e_n), noise is improved by reducing the detector capacitance and leakage current, judiciously selecting all resistances in the input circuit, and choosing the optimum shaping time constant.

The noise parameters of the amplifier depend primarily on the input device. Chapter 6 treats this in detail. In field effect transistors the noise current contribution is very small, so reducing the detector leakage current and increasing the bias resistance will allow long shaping times with correspondingly lower noise. In bipolar transistors the base current sets a lower bound on the noise current, so these devices are best at short shaping times. In special cases where the noise of a transistor scales with geometry, *i.e.* decreasing noise voltage with increasing input capacitance, the lowest noise is obtained when the input capacitance of the transistor is equal to the detector capacitance, albeit at the expense of power dissipation. Capacitive matching is useful with FETs, but not bipolar transistors, as discussed in Chapter 6. In bipolar transistors the minimum obtainable noise is independent of shaping time, but only at the optimum collector current I_C, which does depend on shaping time:

$$Q_{n,min}^2 = 4kT\frac{C}{\sqrt{\beta_{DC}}}\sqrt{F_iF_v} \quad \text{at} \quad I_C = \frac{kT}{e}C\sqrt{\beta_{DC}}\sqrt{\frac{F_v}{F_i}}\frac{1}{T_S}\ , \tag{1.35}$$

where β_{DC} is the direct current gain. For a *CR-RC* shaper and $\beta_{DC} = 100$,

$$Q_{n,min} \approx 250\left[\frac{e}{\sqrt{\mathrm{pF}}}\right]\cdot\sqrt{C} \quad \text{at} \quad I_C = 260\left[\frac{\mu\mathrm{A}\cdot\mathrm{ns}}{\mathrm{pF}}\right]\cdot\frac{C}{T_S}\ . \tag{1.36}$$

Practical noise levels range from $< 1\,e$ for CCDs at long shaping times to $\approx 10^4\,e$ in high-capacitance liquid argon calorimeters. Silicon strip detectors typically operate at $\approx 10^3\,e$, whereas pixel detectors with fast readout can provide noise of order $100\,e$.

1.10.3 *Timing measurements*

In timing measurements the slope-to-noise ratio must be optimized, rather than the signal-to-noise ratio alone, so the rise time t_r of the pulse is important. The "jitter" σ_t of the timing distribution

$$\sigma_t = \frac{\sigma_n}{(dS/dt)_{S_T}} \approx \frac{t_r}{S/N}\ , \tag{1.37}$$

where σ_n is the rms noise and the derivative of the signal dS/dt is evaluated at the trigger level S_T. To increase dS/dt without incurring excessive noise the

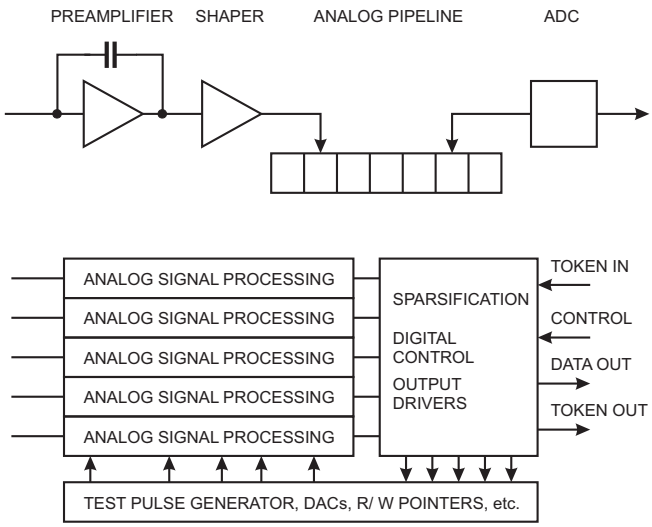

FIG. 1.30. Circuit blocks in a representative readout IC. The analog processing chain is shown at the top. Control is passed from chip to chip by token passing.

amplifier bandwidth should match the rise-time of the detector signal. The 10 – 90% rise time of an amplifier with bandwidth f_u is

$$t_r = 2.2\tau = \frac{2.2}{2\pi f_u} = \frac{0.35}{f_u} \ . \tag{1.38}$$

For example, an oscilloscope with 350 MHz bandwidth has a 1 ns rise time. When amplifiers are cascaded, which is invariably necessary, the individual rise times add in quadrature

$$t_r \approx \sqrt{t_{r1}^2 + t_{r2}^2 + \ldots + t_{rn}^2} \ . \tag{1.39}$$

Thus, reducing the risetime of the electronics beyond the risetime of the sensor signal will increase the electronic noise more rapidly than improve the signal risetime. Time resolution improves with signal-to-noise ratio, so minimizing the total capacitance at the input is also important. At high signal-to-noise ratios the time jitter can be much smaller than the rise time. When a simple threshold discriminator is used the timing signal will shift with pulse amplitude ("walk"), but this can be corrected by various means, either in hardware or software. Timing measurements are discussed in Chapter 4 and by Spieler (1982).

1.11 Subsystems

1.11.1 *Circuit integration and bussing*

A detector array combines the sensor and the analog signal processing circuitry together with a readout system. Figure 1.30 shows the circuit blocks in a representative readout IC. Individual sensor electrodes connect to parallel channels of

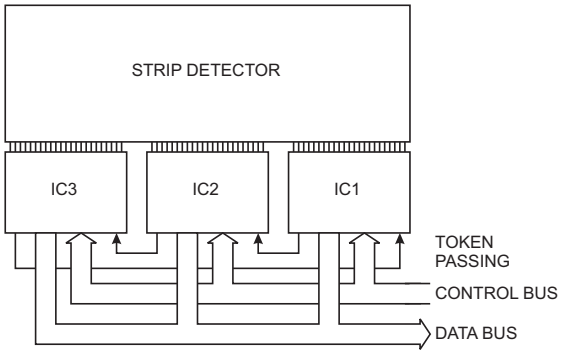

FIG. 1.31. Multiple ICs are ganged to read out a strip detector. The right-most chip IC1 is the master. A command on the control bus initiates the readout. When IC1 has written all of its data it passes the token to IC2. When IC2 has finished it passes the token to IC3, which in turn returns the token to the master IC1.

analog signal processing circuitry. Data are stored in an analog pipeline pending a readout command. Variable write and read pointers are used to allow simultaneous read and write. The signal in the time slot of interest is digitized, compared with a digital threshold and read out. Circuitry is included to generate test pulses that are injected into the input to simulate a detector signal. This is a very useful feature in setting up the system and is also a key function in chip testing prior to assembly. Analog control levels are set by digital-to-analog converters (DACs). Multiple ICs are connected to a common control and data output bus, as shown in Figure 1.31. Each IC is assigned a unique address, which is used in issuing control commands for setup and *in situ* testing. Sequential readout is controlled by token passing. IC1 is the master, whose readout is initiated by a command (trigger) on the control bus. When it has finished writing data it passes the token to IC2, which in turn passes the token to IC3. When the last chip has completed its readout the token is returned to the master IC, which is then ready for the next cycle. The readout bit stream begins with a header, which uniquely identifies a new frame. Data from individual ICs are labeled with a chip identifier and channel identifiers. Many variations on this scheme are possible. As shown, the readout is event oriented, *i.e.* all hits occurring within an externally set exposure time (*e.g.* time slice in the analog buffer in Figure 1.30) are read out together.

In colliding-beam experiments only a small fraction of beam crossings yields interesting events. The time required to assess whether an event is potentially interesting is typically of order microseconds, so hits from multiple beam crossings must be stored on-chip, identified by beam crossing or time-stamp. Upon receipt of a trigger the interesting data are digitized and read out. This allows use of a digitizer that is slower than the collision rate. It is also possible to read out analog signals and digitize them externally. Then the output stream is a se-

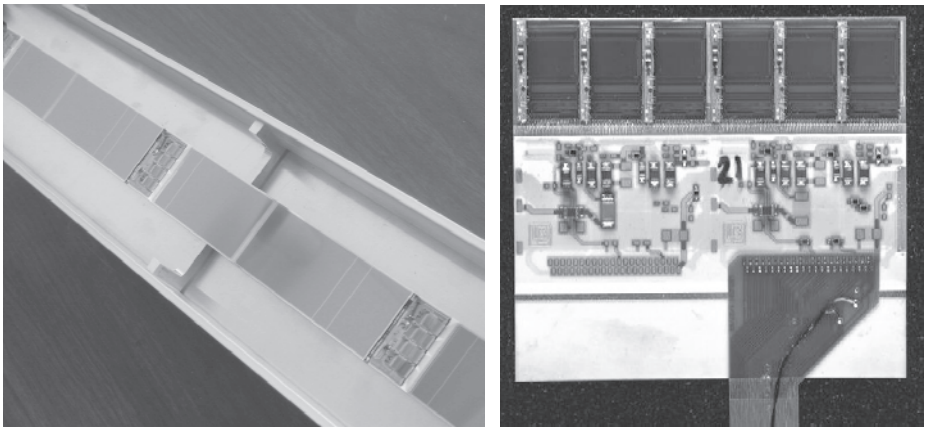

FIG. 1.32. Detector modules combining silicon strip sensors and electronics, shown in a protective enclosure. Multiple modules are mounted on a stave that also incorporates the signal and power busses. The right-hand panel shows the ceramic hybrid that combines six readout ICs (top row) with the power supply and data bussing. Bypass capacitors are visible as small rectangles beneath the ICs. A flex ribbon cable connects the hybrid to the control and data acquisition system and also provides power. (Photographs courtesy of M. Garcia-Sciveres and C. Haber.)

quence of digital headers and analog pulses. An alternative scheme only records the presence of hits. The output of a threshold comparator signifies the presence of a signal and is recorded in a digital pipeline that retains the crossing number.

When reading out pulse height information, either in analog or digitized form, the "smearing" of pulse height information by electronic noise will be clearly visible in the output data. In a "binary" readout the presence of noise is not so obvious. With a large signal-to-noise ratio the threshold can be set high, so predominantly true hits will appear in the output stream. However, when sensitivity is of the essence, the threshold will be set as low as possible. If set too low, the comparator will fire predominantly on noise pulses. If set too high, noise hits are suppressed, but efficiency for desired events will suffer. Thus, a compromise threshold is chosen that will provide high efficiency with an acceptable rate of noise hits. In any case, since the "tails" of the noise distribution extend to infinity, the output of every binary system is contaminated by noise pulses. Only the ratio of noise hits to true hits will be different and depend on the signal-to-noise ratio. This is discussed quantitatively in Chapter 4.

1.11.2 *Detector modules, services, and supports*

The outputs of the ICs must be transferred to the off-detector electronics. To provide this interface the readout ICs are mounted on a substrate, which accommodates the signal bussing between ICs, control signals, and power supply busses.

Important components are bypass capacitors and filter networks to block external interference from the readout ICs, but also to keep digital switching spikes from propagating through the power supply lines. Figure 1.32 shows an assembly of multiple detectors with readout circuitry, mounted on a stave that also integrates the data and power busses. The right-hand panel of Figure 1.32 shows the electronics unit (called a "hybrid", as it combines multiple technologies). Each integrated circuit includes 128 channels of front-end circuitry, analog pipeline, analog-to-digital conversion, and readout logic and driver with on-chip zero suppression (sparsification), so that only struck channels are read out. This hybrid utilizes a multilayer ceramic substrate to integrate the readout ICs, associated capacitors, and interconnects. Power, control, and data lines are implemented as polyimide ribbon cables. Figure 1.33 shows a closeup of ICs mounted on a hybrid using a flexible kapton substrate (Kondo *et al.* 2002), described in Chapter 8 (Section 8.6.5). The wire bonds connecting the IC to the hybrid are clearly visible. Channels on the IC are laid out on a $\sim 50\,\mu$m pitch and pitch adapters fan out to match the $80\,\mu$m pitch of the strip detector. The space between chips accommodates bypass capacitors and connections for control busses carrying signals from chip to chip. Other examples are discussed in Chapter 8.

In large systems optical links are often chosen to eliminate cross-coupling from other lines, but properly designed differential cable drivers and receivers can also provide high noise immunity. Optical links require additional interface

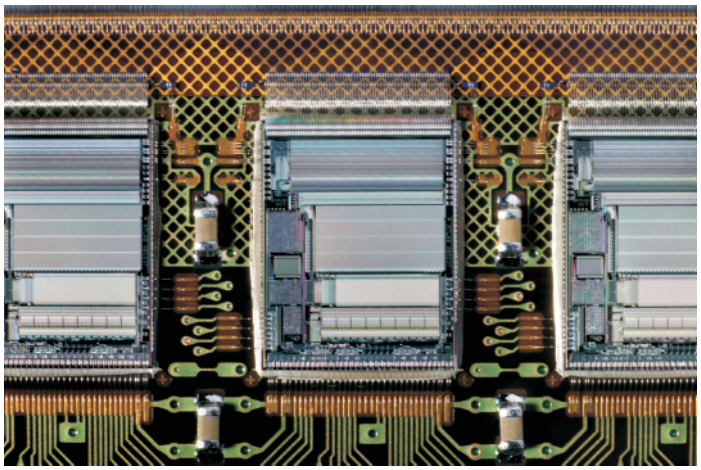

FIG. 1.33. Close-up of ICs mounted on a hybrid utilizing a flexible polyimide substrate (Kondo *et al.* 2002). The high-density wire bonds at the upper edges connect via pitch adapters to the $80\,\mu$m pitch of the silicon strip detector. The ground plane is patterned as a diamond grid to reduce material. (Photograph courtesy of A. Ciocio.)

ICs, as silicon is not well-suited for optical emitters. In some designs all drivers are accommodated on the detector module, whereas in others multiple detector modules interface through a common driver module.

1.11.3 *Data acquisition*

Signals are bussed from the detector modules to a readout module, which includes data buffering. Depending on the complexity of the systems, several alternatives exist for interfacing to the data acquisition computer. Small systems can interface directly via a plug-in card in a PC. In large systems VME or PCI bus interfaces are frequently used. In the past the nuclear instrumentation community designed interfaces such as CAMAC or FastBus, but now the availability of suitable standard industry interfaces has displaced these community-specific interfaces. The interface modules that accept the data from the detector module and transfer it to the computer data acquisition bus are usually custom designed and typically contain buffer memory, FPGA-based logic, and local processors to preprocess the data. Digital interfacing utilizes standard techniques known to many engineers and scientists, so it will not be covered in this book. For an overview of data acquisition systems in high-energy physics see Butler (2003).

1.12 Further reading

The following chapters go into detail on the topics laid out in this introduction, but emphasize aspects relevant to large scale semiconductor systems. Books by G. Knoll (2000) and C. Grupen (1996) provide excellent general introductions to radiation detection and techniques. S.M. Sze (1981) gives a concise description of semiconductor physics and a comprehensive treatment of semiconductor devices. The following chapters and appendices include references to more specialized texts.

References

Abe, K. *et al.* (1997). Design and performance of the SLD vertex detector: a 307 Mpixel tracking system. *Nucl. Instrum. and Meth.* **A400** (1997) 287–343

Barberis, E. *et al.* (1994). Capacitances in silicon microstrip detectors. *Nucl. Instr. and Meth.* **A342** (1994) 90–95

Bedeschi, F. *et al.* (1989). CDF silicon detector prototype test beam results. *IEEE Trans. Nucl Sci.* **NS-36** (1989) 35–39

Butler, J. (2003). Triggering and data acquisition general considerations, in *Instrumentation in Elementary Particle Physics, AIP Conf. Proc.* **674** (2003) 101–129

Demaria, N. *et al.* (2000). New results on silicon microstrip detectors of CMS tracker. *Nucl. Instr. and Meth.* **A447** (2000) 142–150

Deptuch, G. *et al.* 2003. Development of monolithic active pixel sensors for charged particle tracking. *Nucl. Instr. and Meth.* **A511** (2003) 240–249

Fossum, E.R. (1997). CMOS image sensors: electronic camera on a chip. IEEE. Trans. Electron Dev. **ED-44/10** (1997) 1689–1698

Gatti, E. and Rehak, P. (1984). Semiconductor drift chamber – an application of a novel charge transport scheme. *Nucl. Instr. and Meth.* **225** (1984) 608–614

Goulding, F.S. (1972). Pulse shaping in low-noise nuclear amplifiers: a physical approach to noise analysis. *Nucl. Instr. and Meth.* **100** (1972) 493–504

Goulding, F.S. and Landis, D.A. (1982). Signal processing for semiconductor detectors. *IEEE Trans. Nucl. Sci.* **NS-29/3** (1982) 1125–1141

Grupen, C. (1996). *Particle Detectors.* Cambridge University Press, Cambridge. ISBN 0-521-55216-8, QC787.G6G78

Holland, S. and Spieler, H. (1990). A monolithically integrated detector–preamplifier on high-resistivity silicon. *IEEE Trans. Nucl. Sci.* **NS-37** (1990) 463–468

Holland, S. (1992). Properties of CMOS Devices and circuits fabricated on high-resistivity, detector-grade Silicon. *IEEE Trans. Nucl. Sci.* **NS-39/4** (1992) 809–813

Kemmer, J. and Lutz, G. (1987). New detector concepts. *Nucl. Instr. and Meth.* **A253** (1987) 365–377

Kenney, C. *et al.* (1993). Performance of a monolithic pixel detector. *Nucl. Phys. B (Proc. Suppl.)* **32** (1993) 460 and First test beam results from a monolithic silicon pixel detector. *Nucl. Instrum. and Meth.* **A326** (1993) 144–149

Kleinfelder, S.A. *et al.* (1988). A flexible 128 channel silicon strip detector instrumentation integrated circuit with sparse data readout. *IEEE Trans. Nucl. Sci.* **NS-35** (1988) 171–175

Kleinfelder, S.A. *et al.* (2004). Novel integrated CMOS sensor circuits. *IEEE Trans. Nucl. Sci.* **NS-51/5** (2004) 2328–2336

Knoll, G.F. (2000). *Radiation Detection and Measurement*(3rd edn). Wiley, New York, ISBN 0-471-07338-5, QC787.C6K56 1999

Kondo, T. *et al.* (2002). Construction and performance of the ATLAS silicon microstrip barrel modules. *Nucl. Instr. and Meth.* **A485** (2002) 27–42

Lüth, V. (1990). *Spatial Resolution of Silicon Detectors.* SLAC BaBar Note #54 9-7-90

Lutz, G. (1999). *Semiconductor Radiation Detectors.* Springer Verlag, Berlin, 1999. ISBN 3-5406-4859-3

Particle Data Group (2004). Review of Particle Physics. *Phys. Lett.* **B592** (2004) 1–1109 and at http://pdg.lbl.gov

Radeka, V. (1968). Optimum signal-processing for pulse-amplitude spectrometry in the presence of high-rate effects and noise. *IEEE Trans. Nucl. Sci.* **NS-15/3** (1968) 455–470

Radeka, V. (1974). Signal, noise and resolution in position-sensitive detectors. *IEEE Trans. Nucl. Sci.* **NS-21** (1974) 51–64

Rehak, P. *et al.* (1985). Semiconductor drift chambers for position and energy measurements. *Nucl. Instrum. and Meth.* **A235** (1985) 224–234

Snoeys, W. *et al.* (1992). A new integrated pixel detector for high energy physics. *IEEE Trans. Nucl. Sci.* **39** (1992) 1263–1269

Spieler, H. (1982). Fast timing methods for semiconductor detectors. *IEEE Trans. Nucl. Sci.* **NS-29** (1982) 1142–1158

Sze, S.M. (1981). *Physics of Semiconductor Devices*(2nd edn). Wiley, New York 1981. ISBN 0-471-05661-8, TK7871.85.S988

Turchetta, R. *et al.* (2003). Monolithic active pixel sensors (MAPS) in a VLSI CMOS technology. *Nucl. Instr. and Meth.* **A501** (2003) 251–259

Wood, M.L. *et al.* (1991). Charge correlation measurements of double-sided direct-coupled silicon microstripe detectors. *Supercollider 3 (J. Nonte, ed). Proc. 3rd Annual International Industrial Symposium on the Super Collider, Atlanta, Georgia, Mar 13–15, 1991.* pp. 903–926. Plenum Press, New York. ISBN 0-3064-4037-7, QC787.P7.I57

2

SIGNAL FORMATION AND ACQUISITION

Semiconductor detectors, regardless of their electrode structure, are basically ionization chambers. Figure 2.1 shows a detector with amplifier. As will be discussed later in this chapter, the measured signal depends critically on the combined response of the detector and amplifier. However, first we consider signal formation in the detector, whose equivalent circuit, a current source representing the signal current $i_s(t)$ in parallel with the detector capacitance C_d, is shown in the second panel of Figure 2.1. The capacitance C_d is the capacitance formed by the two detector electrodes. To complete the equivalent circuit we must include the magnitude and time structure of the signal. The combined response of the detector and amplifier will be discussed towards the end of this chapter.

2.1 The signal

In semiconductor detectors the electrical signal is formed directly by ionization. Incident radiation quanta impart sufficient energy to individual atomic electrons to form electron–hole pairs. This is in contrast to other detection mechanisms such as the excitation of optical states (as in scintillators), lattice vibrations (as in cryogenic bolometers), the breakup of Cooper pairs in superconductors, or the formation of superheated droplets in superfluid He, just to name a few examples. Typical excitation energies are listed in Table 2.1. The measured signal, *i.e.* the average number of signal quanta, is the absorbed energy divided by the excitation energy.

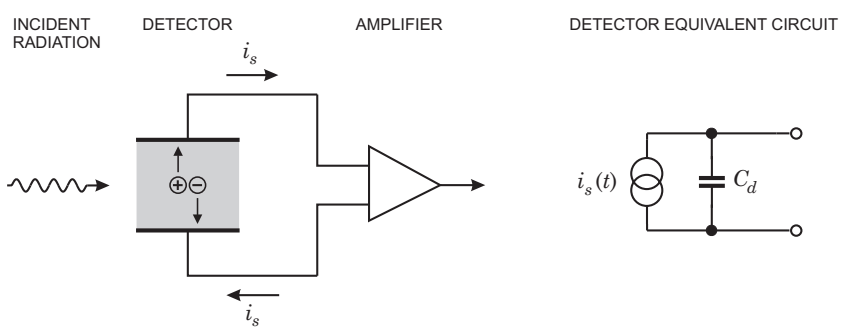

FIG. 2.1. The detector and amplifier (left) and the equivalent circuit of the detector (right).

Table 2.1 Excitation energies for some representative detector media.

Ionization in gases	30 eV
Ionization in semiconductors	1 – 5 eV
Scintillation	10 – 200 eV
Phonons	meV
Breakup of Cooper pairs	meV

The small excitation energies associated with phonon production or the breakup of Cooper pairs can only be exploited at very low temperatures. Among room-temperature media – gases, semiconductors, and scintillators – semiconductors have the lowest excitation energies, so they yield the largest signal.

In semiconductors the energy of elementary excitations is determined by the periodicity of the crystal lattice. Si and Ge have a "diamond" lattice, illustrated in Figure 2.2. The dimension a is the lattice constant, which is 3.56 Å in diamond, 5.65 Å in Ge and 5.43 Å in Si. A wafer can be cut at different orientations relative to the lattice, specified by crystal indices $[hkl]$ (see Kittel 1996, Sze 2001, or other texts). Common orientations are [100] (parallel to the face of the cube) and [111] (the diagonal plane passing through three nonadjacent corners).

Si and Ge are group 4 elements in the periodic table, so they have four valence electrons, shown as "pegs" on the corner atoms in Figure 2.2. Although the bonds are shown as discrete objects, the wavefunctions of the valence electrons extend over distances of 1 – 2 Å, as illustrated in Figure 2.3. These combine with neighbors to form covalent bonds and close the outer shells. Wavefunctions

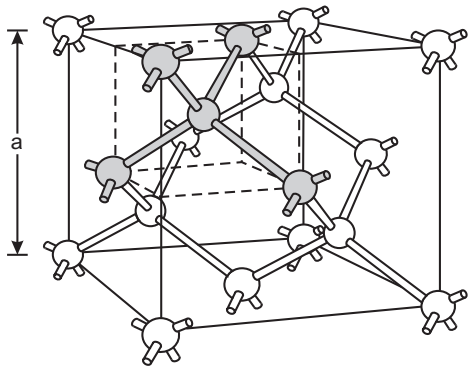

FIG. 2.2. Lattice structure of Si and Ge. The shaded atoms form the basic building block of the lattice, a central atom bonded to four equidistant neighbors.

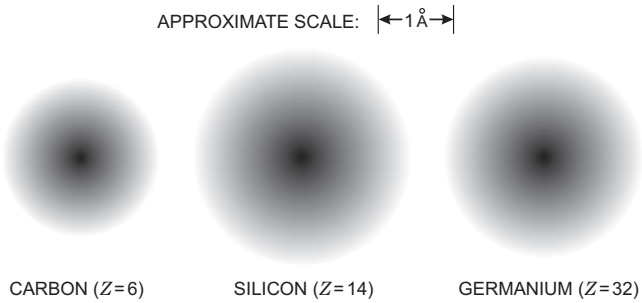

FIG. 2.3. Visualization of the probability density of electrons surrounding carbon, silicon, and germanium atoms. (Following Shockley 1950.)

of individual atoms can merge to form bonding states or antibonding states. The latter have a vanishing occupancy at the midpoint between atoms. At large distances, two of the Si bonding electrons in are in s- and two in p-states. All possible s states are occupied, whereas only two of the six possible p-states are filled. When combined in a lattice the discrete energy states of the atomic shell broaden to form bands, as shown in Figure 2.4. In a metal the antibonding states are partially filled, so states are available for the small incremental increases in energy required for conduction. As the interatomic spacing is reduced further the bands cross, forming a forbidden gap with no available states. Below the forbidden gap the bonding states form the valence band. The number of bonding states equals the number of electrons, so in the absence of any additional excitation, such as heat, no higher energy states are occupied. Then the solid is an insulator, as setting electrons in motion requires an increase in energy. However, no free

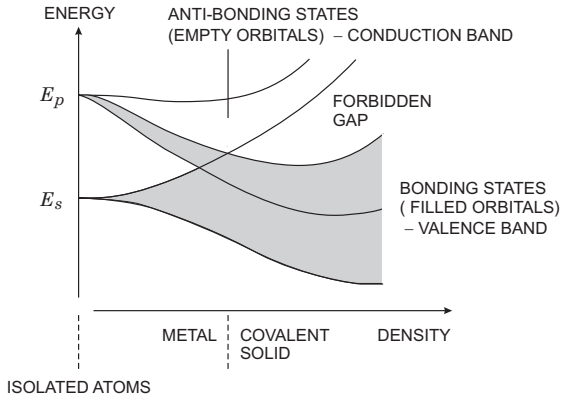

FIG. 2.4. As atoms are moved closer together the bonding and antibonding states spread to form the valence band, a forbidden gap, and a conduction band.

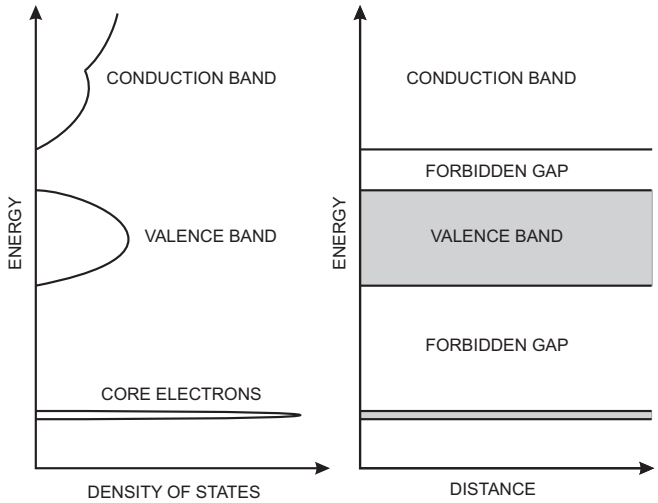

FIG. 2.5. Density of states (energy levels per unit volume) and band structure. Although the band structure is commonly shown in simplified form, as in the right–hand panel, the density of states within the band varies greatly with energy (left). (Following Shockley 1950.)

states are available in the valence band. This requires promoting electrons across the forbidden gap to the empty orbitals above, so this set of states is called the conduction band.

Each atom in the lattice contributes its quantum states to each band, so that the number of quantum states in the band is equal to the number of states from which the band was formed, *i.e.* at least one for each atom in the lattice. Since the density of atoms in Si is $5.0 \cdot 10^{22}$ cm^{-3}, many states can be available. Although they comprise energy levels originally associated with individual atoms, the bands are extended states, *i.e.* the state contributed by an individual atom extends throughout the crystal.

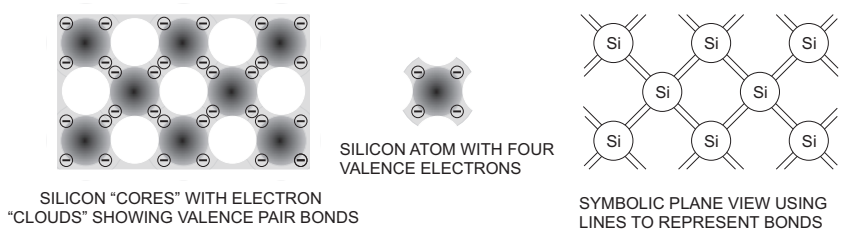

FIG. 2.6. Bonds in a diamond lattice shown schematically. (Following Shockley 1950.)

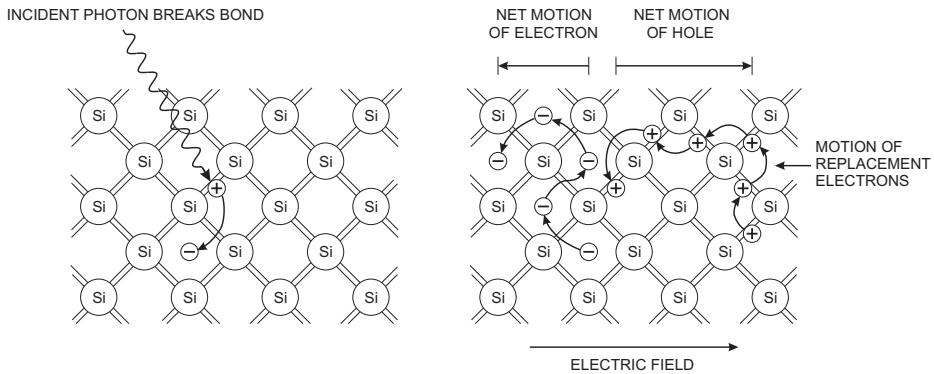

FIG. 2.7. An incident particle can break a bond, promoting an electron into the conduction band, so it can move freely. The vacant bond with positive net charge can also move by successively "borrowing" electrons from neighboring bonds. (Following Shockley 1950.)

The density of states is not uniform; it depends on energy, as shown in the left side of Figure 2.5. In space, the bands extend uniformly throughout the crystal, as shown in the right side of Figure 2.5. Typical widths of the forbidden band are 0.7 eV in Ge, 1.1 eV in Si, 1.4 eV in GaAs and 5.5 eV in diamond.

Figure 2.6 shows the bonds in a diamond lattice schematically. At 0 K all electrons occupy bonding states, completely filling the valence band and, as noted above, no electrical conduction is possible, as no states are available without imparting the substantial energy to surmount the bandgap.

However, if energy is imparted to a bond by incident radiation, for example a photon, the bond can be broken. This excites an electron into the conduction band and leaves back a vacant state in the valence band, a "hole", as illustrated in Figure 2.7. The electron can move freely in its extended state in the conduction band. However, although the hole appears in the valence band, it can also move, albeit by an indirect mechanism. Since the vacant state is completely equivalent to those contributed other atoms, the hole can be filled by an electron from a nearby atom, thereby moving to another position.

The motion of the electron and hole can be directed by an electric field. Holes can be treated as positive charge carriers just like the electrons, although they tend to move more slowly as hole transport involves sequential transition probabilities (the wavefunction overlap of the hole and its replacement electron). Thus, the absorbing volume acts as an ionization chamber. The minimum detectable quantum of energy is set by the bandgap. Larger absorbed energies can promote multiple electrons into the conduction band, yielding a correspondingly larger signal.

Although it is intuitively obvious that a discrete energy deposition can excite an electron from the valence into the conduction band, it is not so obvious what

the effect of thermal excitation is. After all, at room temperature, thermal energy is about 1/40 eV, which is much smaller than the bandgap.

In a pure semiconductor electrons in the conduction band can only originate through thermal excitation from the valence band, so the concentration of electrons and holes must be equal. Since electrons are fermions, the probability of occupying an energy state is given by the Fermi–Dirac distribution

$$f_e(E) = \frac{1}{e^{(E-E_F)/kT} + 1} \ , \tag{2.1}$$

where E_F is the Fermi level (equal to the chemical potential in thermodynamics). The probability of a hole state not being occupied, *i.e.* a valence state being empty, is

$$f_h(E) = 1 - f_e(E) = \frac{1}{e^{(E_F-E)/kT} + 1} \ , \tag{2.2}$$

so equal concentrations of electrons and holes place the Fermi level in the middle of the bandgap $E_F = E_g/2$, the "intrinsic" level. In silicon the bandgap energy $E_g = 1.12$ eV. For a thermal energy of 26 meV the probability of an electron occupying a state in the conduction band is $4.4 \cdot 10^{-10}$. Nevertheless, because of the high density of states, the intrinsic carrier concentration $n_i = 1.45 \cdot 10^{10}$ cm^{-3}, corresponding to a resistivity of about $3 \cdot 10^5\,\Omega\,$cm.

From this we see what distinguishes insulators from semiconductors. At 0 K semiconductors are insulators, but at higher temperatures they have substantial conductivity, depending on the magnitude of the bandgap. Insulators have sufficiently large bandgaps so that the concentration of carriers in the conduction band is negligible at all temperatures of interest.

In a radiation sensor, thermal excitation leads to a continuous current flow, from which the presence of any radiation signal must be distinguished. Furthermore, as was shown in Chapter 1, random fluctuations in the quiescent current flow limit the precision with which the background can be measured, which also sets a minimum signal threshold. A large bandgap greatly reduces the thermally excited current, but reduces the number of charge pairs due to the desired signal. Conversely, a small bandgap increases the signal, but leads to an exponential increase in background current. Furthermore, in applications that require timing information or high pulse rates the signal charge should be swept rapidly from the sensitive volume, so it is important to establish a high electric field. Thus, the conductivity of the sensor material must be low, to allow application of an appropriate voltage without excessive current flow. As a result, the range of bandgaps suitable for practical radiation sensors is quite limited. These considerations are summarized in Table 2.2.

2.2 Detector sensitivity

2.2.1 *Low energy quanta ($E \approx E_g$)*

The ionization energy in solids is proportional to the bandgap, so the bandgap sets the minimum detection threshold. At incident energies below the gap energy,

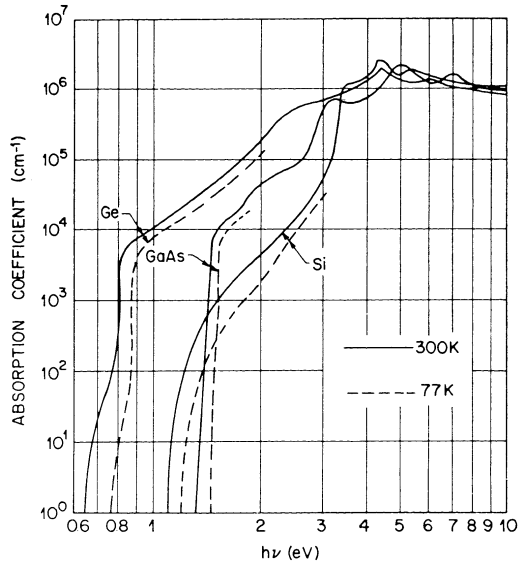

FIG. 2.8. Photon absorption coefficient of Ge, GaAs, and Si at energies near the bandgap. (From Sze 1981. ©John Wiley & Sons, reproduced with permission.)

no energy is transferred, so the absorption is very low. Figure 2.8 shows the absorption coefficient of Ge, GaAs and Si at temperatures of 77 K and 300 K. The rapid onset of absorption at the gap energy is clearly visible, but the curves also show additional structure. At energies well above the bandgap, this is due

Table 2.2 The conducting properties of materials depend on the magnitude of the bandgap.

Bandgap	Type	Properties	Examples
small	conductor	small electric field DC current $\gg$ signal current	Al, Ag, Au, Cu
large	insulator	high electric field small signal charge small DC current	glass, diamond, ceramics
moderate	semiconductor	high electric field "large" signal charge small DC current, but "*pn*-junction" required	Si, Ge, GaAs

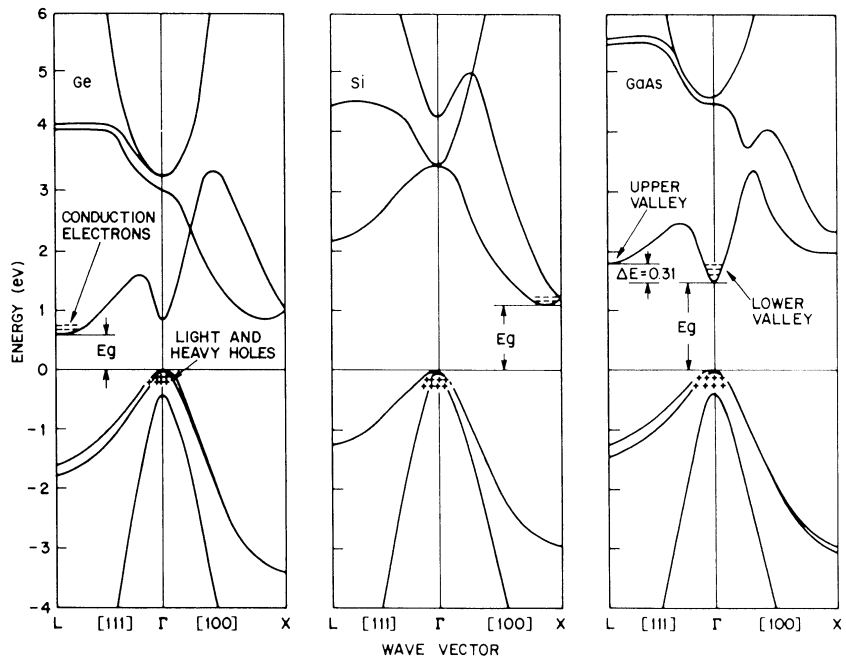

FIG. 2.9. The band structure *vs.* wavevector shows multiple peaks and valleys and depends on orientation. In Ge and Si the minimum bandgap is associated with a non-zero wavevector (momentum), whereas in GaAs the transition occurs with zero momentum (right). (From Sze 1981. ©John Wiley & Sons, reproduced with permission.)

to the density of states in the conduction band (see Figure 2.5). However, when comparing the onset of absorption between GaAs and Ge or Si, GaAs shows a steep rise, whereas Ge and Si show a more gradual transition. This is because the structure of the bands is not just a set of parallel boundaries. In reality the magnitude of the bandgap also depends on momentum. Figure 2.9 shows the band structure *vs.* wavevector (or momentum). The details of the band structure are discussed in solid state physics texts, but in this context the important feature is that the minimum of the conduction band and the maximum of the valence band are offset in Ge and Si, whereas they are aligned at zero wavevector in GaAs. Thus, excitation of an electron in Ge and Si to the conduction band requires simultaneous transfer of both energy and momentum. Momentum is transferred to lattice vibrations, which are quantized as phonons. The density of phonon states depends on both energy and propagation direction in the crystal. Thus, the onset of absorption in Ge and Si in Figure 2.8 is more gradual than in GaAs.

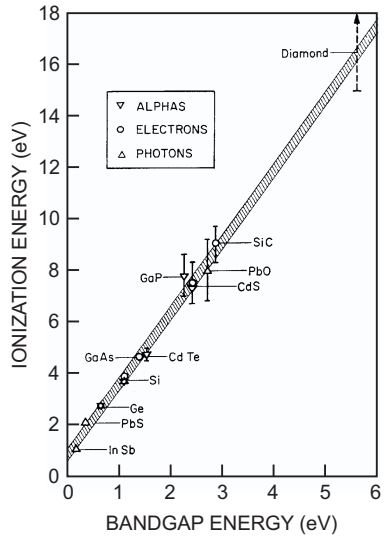

FIG. 2.10. Energy required to form an electron–hole pair *vs.* bandgap. (Adapted from Klein 1968. ©American Institute of Physics, reproduced with permission)

Measurements on silicon photodiodes show that for photon energies below 4 eV one electron–hole pair is formed per incident photon (Scholze *et al.* 2000). The ionization energy E_i attains a maximum of 4.4 eV at photon energies around 6 eV and assumes a constant value above 1.5 keV.

2.2.2 *High energy quanta ($E \gg E_g$)*

For high energy quanta the role of momentum transfer is even more pronounced. The absorption process must conserve both energy and momentum. For photons the momentum $p = E/c$, so for 1 eV photons the momentum that must be absorbed by the lattice is very small. However, for x-rays, gamma-rays and charged particles this is not the case, so a significant portion of the absorbed energy must go into excitations that carry momentum. This explains the experimental observation that the energy required to form an electron–hole pair exceeds the bandgap. This is shown in Figure 2.10. The energy required to form an electron–hole pair is roughly proportional to the bandgap, yielding a good fit to the expression (Owens 2004)

$$E_i \approx 2.8 E_g + 0.6\,\mathrm{eV}\ . \tag{2.3}$$

In Si the ionization energy is about 3.6 eV. The bandgap is 1.12 eV, so about 70% of the ionization energy goes into phonon excitation, or put differently, only about 30% of the absorbed energy goes into signal charge. Ionization energies for a variety of semiconductors are tabulated together with other properties in

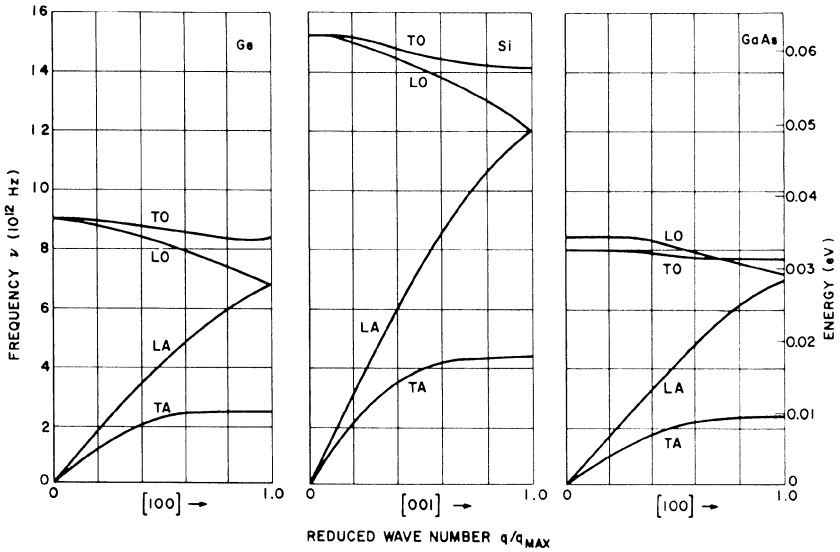

FIG. 2.11. Phonon energy expressed as frequency $\nu = E/h$ *vs.* wave number (momentum) in Ge, Si, and GaAs. (From Sze 1981. ©John Wiley & Sons, reproduced with permission.)

Section 2.7. Note that the ionization energies of some alloys yield a smaller constant term in eqn 2.3 (Owens 2004). Examples are HgI_2, TlBr, 4H-SiC, and diamond.

The relationship between momentum and energy of phonons is complicated, as illustrated in Figure 2.11. The "optical" branches TO and LO show a rather weak dependence of energy *vs.* momentum, whereas the "acoustic" branches LA and TA show a more intuitive behavior at low momenta, *i.e.* energy is roughly proportional to momentum. In this context the important feature is that many combinations of energy and momentum are allowed. This is true in all semiconductors, so on the average the relative energy associated with momentum transfer is roughly the same and we find the same proportionality factor between ionization energy and bandgap.

2.2.3 *Fluctuations in signal charge – the Fano factor*

A key characteristic of signal sensors is not just the magnitude of the signal, but also the fluctuations of the signal for a given absorbed energy. Both determine the minimum signal threshold and the relative resolution $\Delta E/E$. One of the remarkable features of the signal fluctuations in semiconductor sensors is that they are smaller than the simple statistical variance $\sigma_Q = \sqrt{N_Q}$. A detailed calculation of this phenomenon is quite complicated, so the following derivation introduces some simplifications to bring out the basic mechanism.

Two mechanisms contribute to the mean ionization energy. First, conservation of momentum requires excitation of lattice vibrations, and second, many modes are available for momentum and energy transfer with an excitation energy less than the bandgap.

Energy can be absorbed by either lattice excitation, *i.e.* phonon production (with no formation of mobile charge), or ionization, *i.e.* formation of a mobile charge pair. Assume that in the course of energy deposition N_x excitations produce N_P phonons and N_{ion} ionization interactions form N_Q charge pairs. The sum of the energies going into excitation and ionization is equal to the energy deposited by the incident radiation

$$E_0 = E_{ion} N_{ion} + E_x N_x \ , \tag{2.4}$$

where E_{ion} and E_x are the energies required for a single excitation or ionization. In a semiconductor E_{ion} is the bandgap and E_x is the average phonon energy. Assuming Gaussian statistics, the variance in the number of excitations $\sigma_x = \sqrt{N_x}$ and the variance in the number of ionizations $\sigma_{ion} = \sqrt{N_{ion}}$.

For a single event, the energy E_0 deposited in the detector is fixed (although this may vary from one event to the next). If the energy required for excitation E_x is much smaller than required for ionization E_{ion}, sufficient degrees of freedom will exist for some combination of ionization and excitation processes to dissipate precisely the deposited energy. Hence, for a given energy deposited in the sample a fluctuation in excitation must be balanced by an equivalent fluctuation in ionization:

$$E_x \Delta N_x + E_{ion} \Delta N_{ion} = 0 \ . \tag{2.5}$$

If for a given event more energy goes into charge formation, less energy will be available for excitation. Averaging over many events this means that the variances in the energy allocated to the two types of processes must be equal $E_{ion}\sigma_{ion} = E_x \sigma_x$, so

$$\sigma_{ion} = \frac{E_x}{E_{ion}} \sqrt{N_x} \ . \tag{2.6}$$

From the total energy $E_0 = E_{ion} N_{ion} + E_x N_x$ (eqn 2.4) we can extract

$$N_x = \frac{E_0 - E_{ion} N_{ion}}{E_x} \tag{2.7}$$

and insert this into the preceding equation 2.6 to obtain

$$\sigma_i = \frac{E_x}{E_{ion}} \sqrt{\frac{E_0}{E_x} - \frac{E_{ion}}{E_x} N_{ion}} \ . \tag{2.8}$$

Overall, the number N_Q of charge pairs formed is the total deposited energy E_0 divided by the average energy deposition E_i required to produce a charge pair. Since each ionization forms a charge pair that contributes to the signal,

$$N_{ion} = N_Q = \frac{E_0}{E_i} \ . \tag{2.9}$$

Thus, the variance in ionization processes

$$\sigma_{ion} = \frac{E_x}{E_{ion}} \sqrt{\frac{E_0}{E_x} - \frac{E_{ion}}{E_x}\frac{E_0}{E_i}} \ , \tag{2.10}$$

which can be rewritten as

$$\sigma_{ion} = \sqrt{\frac{E_0}{E_i}} \cdot \sqrt{\frac{E_x}{E_{ion}} \left(\frac{E_i}{E_{ion}} - 1 \right)} \ . \tag{2.11}$$

The second factor on the right-hand side is called the Fano factor F. Since σ_{ion} is proportional to the variance in signal charge Q and the number of charge pairs is $N_Q = E_0/E_i$,

$$\sigma_Q = \sqrt{F N_Q} \ . \tag{2.12}$$

In silicon $E_x = 0.037\,\mathrm{eV}$, $E_{ion} = E_g = 1.1\,\mathrm{eV}$, and $E_i = 3.6\,\mathrm{eV}$ for which the above expression yields $F = 0.08$, in reasonable agreement with the measured value $F = 0.1$. Thus, the variance of the signal charge is smaller than naively expected, $\sigma_Q \approx 0.3\sqrt{N_Q}$.

A similar treatment can be applied if the degrees of freedom are much more limited and Poisson statistics are necessary. However, when applying Poisson statistics to the situation of a fixed energy deposition, which imposes an upper bound on the variance, one cannot use the usual expression for the variance $\mathrm{var}N = \overline{N}$. Instead, the variance is $\overline{(N - \overline{N})^2} = F\overline{N}$, as shown by Fano (1947) in the original paper. An accurate calculation of the Fano factor requires a detailed accounting of the energy dependent cross sections and the density of states of the

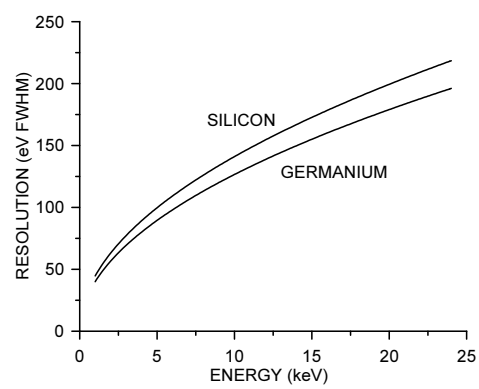

FIG. 2.12. Intrinsic resolution of silicon and germanium detectors *vs.* energy.

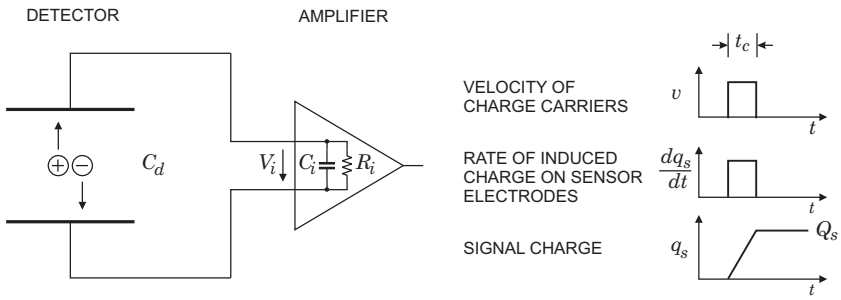

FIG. 2.13. Charge collection and signal integration in an ionization chamber.

phonon modes. This was discussed by Alkhazov *et al.* (1967) and van Roosbroeck (1963).

We can use this result to calculate the intrinsic resolution of semiconductor detectors

$$\Delta E_{FWHM} = 2.35 \cdot E_i \sqrt{FN_Q} = 2.35 \cdot E_i \sqrt{F\frac{E}{E_i}} = 2.35 \cdot \sqrt{FEE_i} \ , \qquad (2.13)$$

where in Si $E_i = 3.6\,\text{eV}$ and in Ge $E_i = 2.9\,\text{eV}$. For both the Fano factor $F = 0.1$. The energy resolution *vs.* energy is shown in Figure 2.12. Detectors with good efficiency at x-ray energies typically have sufficiently small capacitance to allow electronic noise of $\sim 100\,\text{eV}$ FWHM, so the variance of the detector signal is a significant contribution. At energies $> 100\,\text{keV}$ the detector sizes required tend to increase the electronic noise to dominant levels.

2.3 Signal formation

Semiconductor detectors are ionization chambers. Particles deposit energy in the detection volume, forming positive and negative charge carriers. Under an applied electric field the charge carriers move and induce a change in induced charge on the electrodes. The duration of the induced signal depends on the carriers' velocity, which depends on the electric field. This is illustrated in Figure 2.13.

A high field in the detection volume is desirable for fast response, but also for improved charge collection efficiency. Crystal lattices are not perfect; irregularities in the crystal structure and impurities can form trapping sites for the charge carriers. This is discussed in more detail in Appendix F and Chapter 7. One result is that trapping leads to a carrier lifetime, so if the carriers are swept more rapidly from the crystal, the trapping probability is reduced.

2.3.1 *Formation of a high-field region*

As already noted above, the conduction band is only empty at 0 K. As the temperature is increased, thermal excitation can promote electrons across the

bandgap into the conduction band. In pure Si the carrier concentration is $\sim 10^{10}\,\mathrm{cm}^{-3}$ at 300 K, corresponding to a resistivity $\rho \approx 400\,\mathrm{k}\Omega\,\mathrm{cm}$. Since the Si lattice includes $5 \cdot 10^{22}$ atoms/cm^3, many states are available in the conduction band to allow carrier motion. In reality, crystal imperfections and minute impurity concentrations limit Si carrier concentrations to $\sim 10^{11}\ \mathrm{cm}^{-3}$ at 300 K, corresponding to a resistivity $\rho \approx 40\,\mathrm{k}\Omega\,\mathrm{cm}$. In practice, resistivities up to $20\,\mathrm{k}\Omega\,\mathrm{cm}$ are available, with mass production ranging from 5 to $10\,\mathrm{k}\Omega\,\mathrm{cm}$.

As already noted in Chapter 1, these resistivities are too low for use in a simple crystal detector. However, a high-field region with low leakage current can be established by using a reverse-biased *pn*-junction. The key to this technology is the deliberate introduction of impurities to control the conductivity. This process is called doping.

2.3.2 *Doping*

The conductivity of semiconductors can be controlled by introducing special impurities. Required concentrations are in the range 10^{12} – $10^{18}\,\mathrm{cm}^{-3}$, where the former is typical in radiation detectors. In semiconductors the conductivity can be provided by either electrons (*n*-type) or holes (*p*-type).

2.3.2.1 *n-type doping* Replacing a silicon atom (group 4 in the periodic table, *i.e.* four valence electrons) by an atom with five valence electrons (*e.g.* P, As, Sb) leaves one valence electron without a partner (Figure 2.14). Since the impurity contributes an excess electron to the lattice, it is called a donor.

The donor electron cannot be accommodated in the valence band, but it is lightly bound to the impurity atom. As illustrated in Figure 2.15 the wavefunction of the dopant atom extends over many neighbors, so one can, at least as an approximation, utilize lattice properties such as the dielectric constant. Thus, the Coulomb force that binds the electron to the donor atom is reduced by the

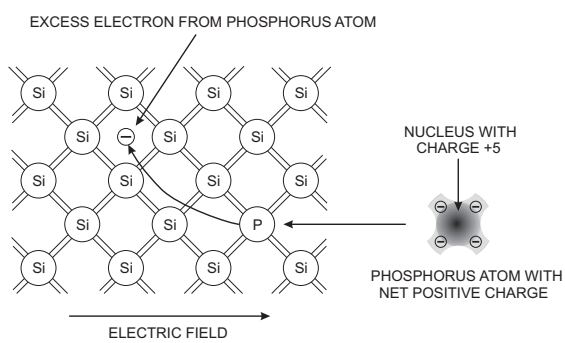

FIG. 2.14. Introducing a group 5 impurity (*e.g.* P or As) introduces a lightly bound electron that can move freely under the influence of an electric field. (Following Shockley 1950.)

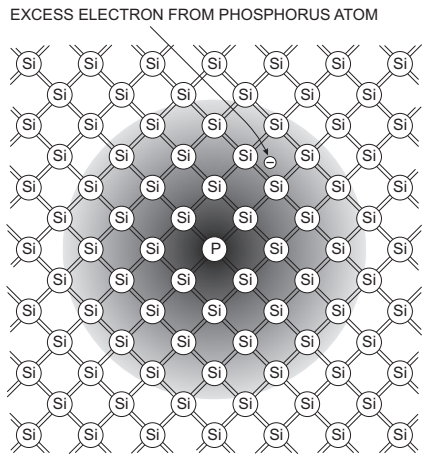

FIG. 2.15. The donor wavefunction extends over many lattice sites. (Following Shockley 1950.)

dielectric constant ε of the medium ($\varepsilon = 11.9$ in Si), so the binding energy

$$E_b(lattice) \approx \frac{E_b(atom)}{\varepsilon^2} \ .$$

The bound level of this unpaired electron is of order 0.01 eV below the conduction band (*e.g.* for P in Si: $E_c - 0.045\,\text{eV}$), as illustrated in Figure 2.16. As a result, at room temperature ($E = 0.026\,\text{eV}$) the probability of ionization is substantial and mobile electrons are introduced into the conduction band. Energy levels of impurity states are shown in Figure F.3.

2.3.2.2 *p-type doping* Introducing a group 3 atom (B, Al, Ga, In) into a lattice site provides bonds for all Si valence electrons, but leaves one impurity valence electron without a partner. This is illustrated in Figure 2.17 for a boron impurity.

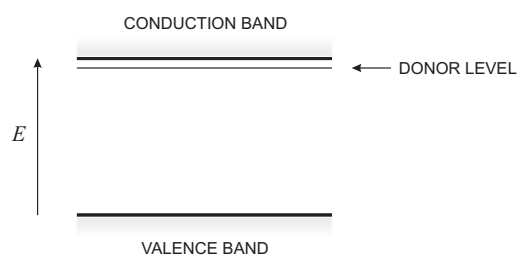

FIG. 2.16. The donor level lies in the forbidden gap close to the conduction band edge, so thermal excitation can promote electrons into the conduction band.

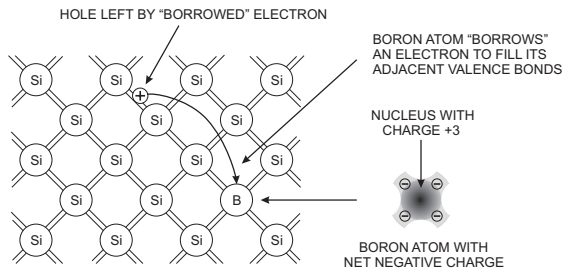

FIG. 2.17. Introducing a group 3 impurity leaves an unpaired silicon bond, which can attract a neighboring electron. As other electrons are "borrowed" to fill the unpaired bond, the resulting vacancy, a "hole" moves through the lattice. (Following Shockley 1950.)

To close its shell the B atom "borrows" an electron from a lattice atom in the vicinity. This type of dopant is called an "acceptor". The "borrowed" electron is bound, but somewhat less than other valence electrons since the B nucleus only has charge three. This introduces a bound state close to the valence band, also of order 0.01 eV from the band edge (Figure 2.18).

For example, a B atom in Si forms a state at $E_v + 0.045\,\text{eV}$. Again, as this energy is comparable to kT at room temperature, electrons from the valence band can be excited to fill a substantial fraction of these states. The electrons missing from the valence band form mobile positive charge states called "holes", which behave similarly to an electron in the conduction band, *i.e.* they can move freely throughout the crystal.

Since the charge carriers in the donor region are electrons, *i.e.* negative, it is called "n-type". Conversely, as the charge carriers in the acceptor region are holes, *i.e.* positive, it is called "p-type" (actually, these designations were coined before the conduction mechanism was understood, but still turned out to be correct).

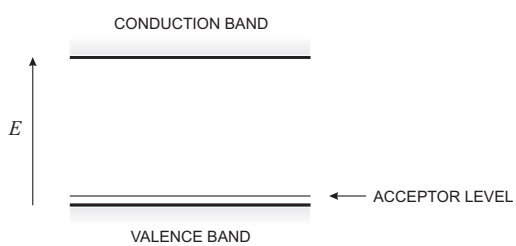

FIG. 2.18. The acceptor level lies in the forbidden gap just above valence band edge. Thermal excitation can promote electrons from the valence band to the fill the acceptor state, leaving a "hole" in the valence band.

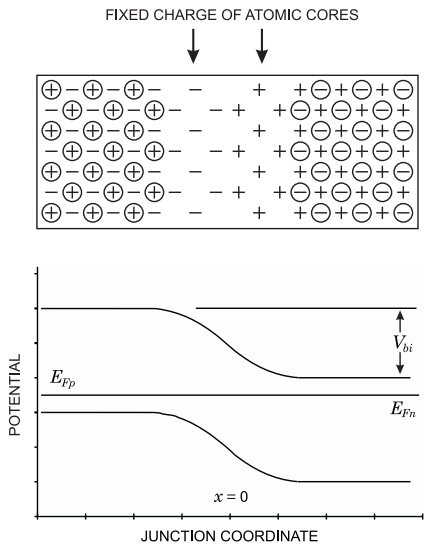

FIG. 2.19. Diffusion of electrons and holes across the junction forms a depletion zone with a resulting potential between the p- and n-regions.

2.3.3 *The pn-junction*

Consider a crystal suitably doped that donor and acceptor regions adjoin, a "*pn*-junction". Initially, the p- and n-regions are electrically neutral, but thermal diffusion will drive holes and electrons across the junction. As electrons diffuse from the n- to the p-region, they uncover their respective donor atoms, leaving a net positive charge in the n-region. This positive space charge exerts a restraining force on the electrons that diffused into the p-region, *i.e.* diffusion of electrons into the p-region builds up a potential. The diffusion depth is limited when the space charge potential exceeds the available energy for thermal diffusion. The corresponding process also limits the diffusion of holes into the n-region. Figure 2.19 shows the resulting potential distribution establishing a "built-in" potential V_{bi} between the p- and n-regions.

The diffusion of holes and electrons across the junction leads to a region free of mobile carriers – the "depletion region", bounded by conductive regions, which are n- and p-doped, respectively. Strictly speaking, the depletion region is not completely devoid of mobile carriers, as the diffusion profile is a gradual transition. Nevertheless, since the carrier concentration is substantially reduced, it is convenient to treat the depletion zone as an abrupt transition between bulk and zero carrier concentration.

The formation of the two adjacent space charge regions builds up a potential barrier between the n- and p-regions, which impedes the further diffusion of charge. The magnitude of this potential barrier depends on the doping levels. As

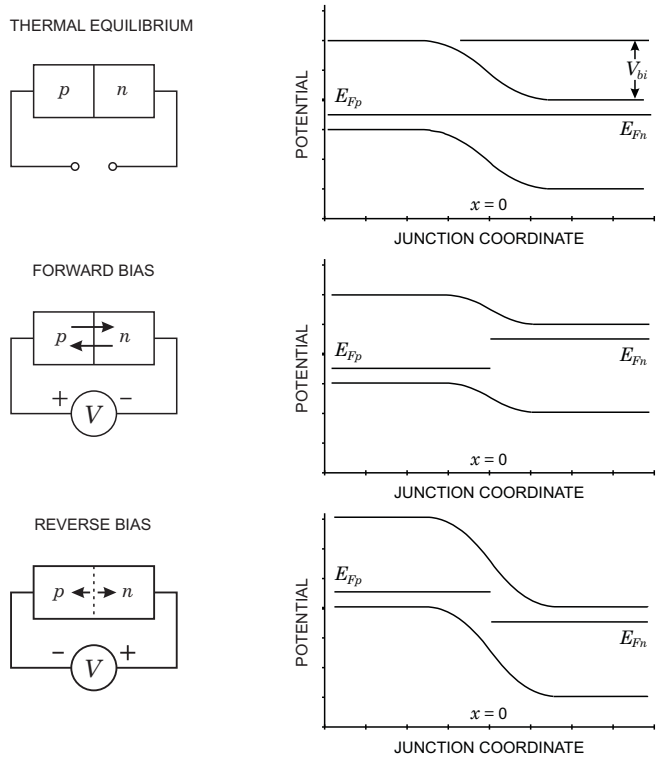

FIG. 2.20. In a *pn*-diode's quiescent state (top) the valence and conduction bands bend so that the Fermi level is constant throughout the device. Applying forward bias (middle) lowers the potential difference and increases the flow of electrons and holes across the junction. Reverse bias (bottom) raises the potential barrier, which reduces the electron and hole concentrations at the *pn*-junction and widens the depletion region.

shown by eqns 2.1 and 2.2 the Fermi level in a pure semiconductor is at mid-gap. If donor impurities are introduced, the electron concentration is increased, so the Fermi level shifts closer to the conduction band. Conversely, acceptors shift the Fermi level closer to the valence band.

In isolation the Fermi levels in the *p*- and *n*-regions are different. However, in thermal equilibrium the Fermi level must be constant throughout the device, so in a *pn*-junction the bands are offset and make a gradual transition from the *p*- to the *n*-regions, following the potential distribution. This is illustrated in the bottom panel of Figure 2.19. The potential difference between the *p*- and *n*-regions is the built-in potential, equal to the difference between the respective Fermi levels $V_{bi} = E_{Fn} - E_{Fp}$.

For a typical doping level in detector-grade silicon of $10^{12}\,\mathrm{cm}^{-3}$ the Fermi level $E_F = 0.1\,\mathrm{eV}$, measured with respect to the middle of the bandgap. A p-electrode doped at 10^{16} cm^{-3} has $E_F \approx -0.4\,\mathrm{eV}$, so the built-in voltage $V_{bi} \approx 0.5\,\mathrm{V}$. The built-in voltage depends logarithmically on doping level (Sze 1981 and Appendix E),

$$V_{bi} = \frac{kT}{e} \log\left(\frac{N_a N_d}{n_i^2}\right) , \tag{2.14}$$

where N_a and N_d are the acceptor and donor concentrations and n_i is the intrinsic carrier concentration, which in Si at 300 K is $1.45 \cdot 10^{10}$ cm^{-3}.

When an external potential is applied, thermal equilibrium no longer holds. With positive potential applied to the p-region and negative to the n-region, the potential barrier is reduced and the flow of electrons and holes across the junction increases (forward bias). When the opposite polarity is applied, *i.e.* negative potential to the p-region and positive to the n-region, the potential barrier is increased and the width of the depletion grows. Forward and reverse bias are illustrated in the middle and bottom panels of Figure 2.20.

The pn-junction is asymmetric with respect to current flow. The dependence of diode current *vs.* voltage is given by the "Shockley equation"

$$I = I_0(e^{eV/kT} - 1) , \tag{2.15}$$

which is derived in Appendix E. For positive bias voltages V the exponential term dominates and the current increases rapidly with voltage. For large negative bias the exponential term becomes negligible and $I = -I_0$, the reverse bias current in saturation. Figure 2.21 shows the current *vs.* voltage (I–V curve) of a semiconductor diode. The current under forward bias rises rapidly, attaining $20I_0$ at a voltage of $3kT/e$ (about 80 mV at room temperature) and $150I_0$ at $5kT/e$. The reverse bias current saturates rather quickly, attaining 95% of the saturation current I_0 at a reverse bias voltage of $3kT/e$. This large asymmetry in current allows the pn-diode to be used as a rectifier. Note that the bandgap does not appear explicitly in the diode equation, although it enters indirectly through the reverse saturation current. The reverse saturation current is strongly affected by impurities and defects, which can increase it by orders of magnitude. Then the reverse diode current increases with the depletion width. This is discussed in Appendix F. Not all defects are electrically active and some are useful (Queisser and Haller 1998). One example is gettering, where defects "capture" harmful impurities (Appendix A).

2.3.4 *The reverse-biased diode*

The reverse-biased diode is of special interest for radiation detection. Since the depletion region is a volume devoid of mobile carriers it forms a capacitor, where the undepleted p- and n-regions are the electrodes and the depletion region is the dielectric. The electric field in the depletion region will sweep mobile carriers to the electrodes, so the diode forms an ionization chamber. The depletion region

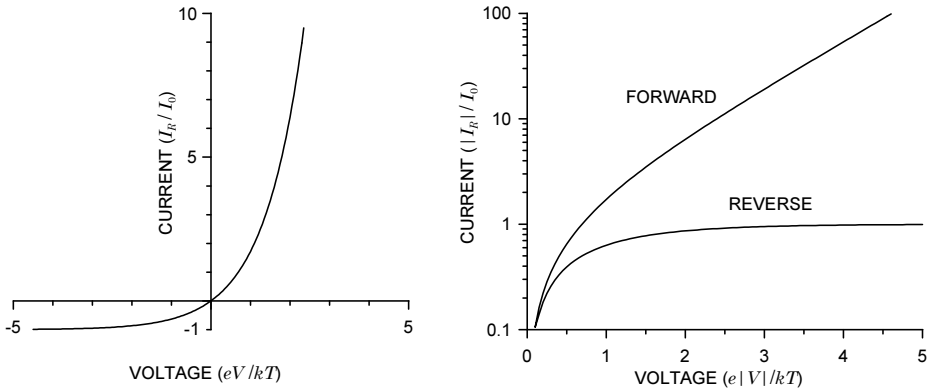

FIG. 2.21. Current *vs.* voltage of a semiconductor diode, plotted in units of the reverse saturation current I_S *vs.* thermal voltage kT/e. These curves are the same for all materials.

formed by thermal diffusion is rather thin, of order μm in typical diodes used in electronic circuitry. However, the width of the depletion region can be increased by applying a reverse bias voltage.

2.3.4.1 *Depletion width and electric field in a pn-junction* Assume a diode that is reverse-biased by an external potential V_b and where the potential V in the depletion region changes only in the direction perpendicular to the $n-p$ interface. The potential distribution is determined by the space charge due to the dopant atoms minus the conduction electrons or holes, which have been swept from the depletion region. The potential is described by Poisson's equation

$$\frac{d^2V}{dx^2} + \frac{Ne}{\varepsilon} = 0 \ , \tag{2.16}$$

where N is the dopant concentration, e the electronic charge, and ε the dielectric constant. For simplicity assume an abrupt junction, where the charge densities on the n and p sides are $N_d e$ and $N_a e$, respectively. First consider the n-side. Setting the limit of the depletion region to x_n, after two successive integrations one obtains

$$\frac{dV}{dx} = -\frac{eN_d}{\varepsilon}(x - x_n) \tag{2.17}$$

and

$$V = -\frac{eN_d}{\varepsilon}\frac{x^2}{2} + \frac{eN_d x x_n}{\varepsilon} + V_j \ , \tag{2.18}$$

where V_j is the potential at the metallurgical junction, the interface where the n- and p-regions join. At the boundary of the depletion region $x = x_n$ the potential

$$V(x_n) = V_b = \frac{eN_d x_n^2}{2\varepsilon} + V_j \tag{2.19}$$

and the contribution of the n-region to the total reverse bias potential becomes

$$V_b - V_j = \frac{eN_d x_n^2}{2\varepsilon} \ . \tag{2.20}$$

Correspondingly, in the p-region

$$V_j = \frac{eN_a x_p^2}{2\varepsilon} \tag{2.21}$$

and the total potential becomes

$$V_b = \frac{e}{2\varepsilon}(N_d x_n^2 + N_a x_p^2) \ , \tag{2.22}$$

where V_b is the applied reverse bias voltage. Since overall charge neutrality must be maintained

$$N_d x_n = N_a x_p \ , \tag{2.23}$$

so

$$V_b = \frac{e}{2\varepsilon} \left(1 + \frac{N_a}{N_d}\right) N_a x_p^2 = \frac{e}{2\varepsilon} \left(1 + \frac{N_d}{N_a}\right) N_d x_n^2 \ . \tag{2.24}$$

The depletion widths on the n- and p-side of the junction are

$$\begin{aligned} x_n &= \sqrt{\frac{2\varepsilon V_b}{eN_d(1 + N_d/N_a)}} \\ x_p &= \sqrt{\frac{2\varepsilon V_b}{eN_a(1 + N_a/N_d)}} \end{aligned} \tag{2.25}$$

and the total depletion width

$$w = x_n + x_p = \sqrt{\frac{2\varepsilon V_b}{e} \ \frac{N_a + N_d}{N_a N_d}} \ . \tag{2.26}$$

Combining 2.21, 2.23, and 2.25 yields the junction potential

$$V_j = \left(\frac{N_d}{N_a}\right) \frac{V_b}{(1 + N_d/N_a)} \ . \tag{2.27}$$

For an asymmetrical junction with $N_d \ll N_a$ the junction potential

$$V_j \approx \frac{N_d}{N_a} V_b \tag{2.28}$$

and the junction potential is practically equal to the potential of the p contact, so all of the bias voltage develops across the lightly doped n-region of the depletion width.

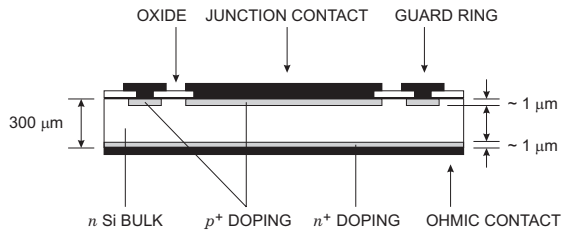

FIG. 2.22. Detector diodes are commonly formed by introducing a highly doped surface layer into a lightly doped bulk. The depletion zone then extends into the bulk. Metallization layers (black) provide electrical contacts to the doped p^+ and n^+ layers that form the junction and the back electrode (ohmic contact).

The depletion width increases with the square root of the reverse bias voltage, so increasing the bias voltage will increase the sensitive volume and reduce the capacitance. For a charged particle traversing the detector this increases the signal charge and reduces the electronic noise, so it is very beneficial. However, the maximum voltage one can apply is limited. At fields $> 10^5$ V/cm electrons acquire sufficient energy to form secondary electron–hole pairs, ultimately leading to a destructive avalanche, called "breakdown". This is discussed in more detail in Section 2.8. Given this limit, the width of the depletion region can be increased by reducing the dopant concentration. Ultimately, this is limited by the minimum residual impurity levels in the crystal. In practice, lightly doped semiconductors include both donor and acceptor impurities and the net doping is $N_d - N_a$.

As already noted in Chapter 1, detector diodes are usually asymmetrically doped, as shown in Figure 2.22. The starting material (bulk) is lightly doped and the junction is formed by diffusing or ion-implanting a highly doped p^+ layer into the n-type bulk. The depletion region then extends predominantly into the lightly doped bulk. The back contact is a highly doped layer of the same type as the bulk, forming a nonrectifying "ohmic" contact.

In addition to the basic diode, Figure 2.22 shows a guard ring, which isolates the wafer edge (saw cut) from the active region. The guard ring is biased at the same potential as the adjacent electrode, so the boundary of the detector's sensitive volume is midway between the detector electrode and the guard ring. In the gap between the detector electrode and the guard ring it is critical to provide a neutral interface at the silicon surface to prevent formation of a conductive path, as discussed in Chapter 6. This is best accomplished by oxide passivation (SiO_2, Appendix A).

When as discussed above $N_a \gg N_d$, the depletion region extends predominantly into the n- side and the total depletion width is

$$w \approx x_n = \sqrt{\frac{2\varepsilon V_b}{eN_d}} \ . \tag{2.29}$$

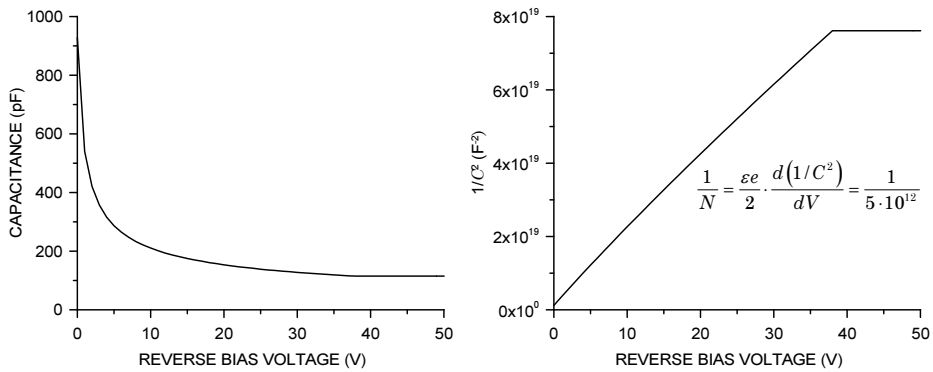

FIG. 2.23. The diode capacitance decreases as the reverse bias voltage is raised until the diode is fully depleted (left). This is more apparent in a plot of $1/C^2$ *vs.* V (right), which also yields the doping level.

The doping concentration is commonly expressed in terms of resistivity

$$\rho = \frac{1}{\mu e N} \ , \tag{2.30}$$

because this is a readily measurable quantity. The mobility μ describes the relationship between the applied field and carrier velocity (to be discussed below). Using resistivity the depletion width becomes

$$w = \sqrt{2\varepsilon\mu_n\rho_n V_b} \ . \tag{2.31}$$

Note that this introduces an artificial distinction between the n- and p-regions, because the mobilities μ for electrons and holes are different. Since the mobility of holes is approximately 1/3 that of electrons, p-type material will have three times the resistivity of n-type material with the same doping concentration.

As discussed earlier, even in the absence of an external voltage electrons and holes diffuse across the junction, establishing a "built-in" reverse bias voltage V_{bi}. If we take this inherent bias voltage into account, the bias voltage V_b in the above equations becomes $V_b + V_{bi}$, and the depletion width of the one-sided junction

$$w \approx x_1 = \sqrt{\frac{2\varepsilon(V_b + V_{bi})}{eN_d}} = \sqrt{2\varepsilon\mu_n\rho_n(V_b + V_{bi})} \ . \tag{2.32}$$

The depleted junction volume is free of mobile charge and thus forms a capacitor, bounded by the conducting p- and n-type semiconductor on each side. The capacitance

$$C = \varepsilon\frac{A}{w} = A\sqrt{\frac{\varepsilon e N}{2(V_b + V_{bi})}}. \tag{2.33}$$

In technical units (V_b in volts and ρ in $\Omega \cdot$ cm) the depletion width in n-type silicon

$$w = 0.5 \ [\mu\text{m}] \times \sqrt{\rho(V_b + V_{bi})}$$

and in p-type material

$$w = 0.3 \ [\mu\text{m}] \times \sqrt{\rho(V_b + V_{bi})} \ .$$

For bias voltages $V_b \gg V_{bi}$ the depletion width increases with the square root of bias voltage $w \propto \sqrt{V_b}$.

The capacitance per unit area

$$\frac{C}{A} = \frac{\varepsilon}{w} \approx 1 \left[\frac{\text{pF}}{\text{cm}}\right] \cdot \frac{1}{w} \ ,$$

so a Si diode with $100\,\mu$m thickness has about $1\,\text{pF/mm}^2$. The capacitance *vs.* voltage characteristic of a diode can be used to determine the doping concentration of the detector material. From eqn 2.33

$$\frac{1}{N} = \frac{\varepsilon e}{2} \cdot \frac{d\left(1/C^2\right)}{dV} \ . \tag{2.34}$$

In a plot of $(A/C)^2$ *vs.* the detector bias voltage V_b the slope of the voltage dependent portion yields the doping concentration N. Figure 2.23 illustrates capacitance *vs.* voltage curves.

2.3.5 *Strip and pixel detectors*

The detector electrodes can be segmented to form strips or pixels. Figure 2.24 shows the cross-section of a typical strip detector on an n-type substrate. The highly doped p^+ electrodes are introduced by ion implantation through a mask to form the strips (see Appendix A). Each strip forms a pn-diode. The gaps between strips must be electrically controlled to maintain isolation between adjacent diodes. A layer of thermally grown oxide (see Appendix A) terminates the "dangling" bonds at the silicon surface and also provides a protective layer that is impermeable to many contaminants. An aluminum layer deposited on the electrodes provides a low-resistance signal path to the readout electronics at the end of the detector.

Double-sided detectors also pattern the ohmic contact. The strips and substrate comprise an n^+-n-n^+ structure that forms a conducting path unless additional isolation structures are introduced. These can be implemented as intermediate p-regions ("p-stops"), as shown in Figure 2.24, as a contiguous "p spray" (Richter *et al.* 1996), or as field plates appropriately biased to deplete the surface of electrons (Avset *et al.* 1990, Chabaud *et al.* 1996). The latter technique is adopted from MOS transistors, discussed in Chapter 6.

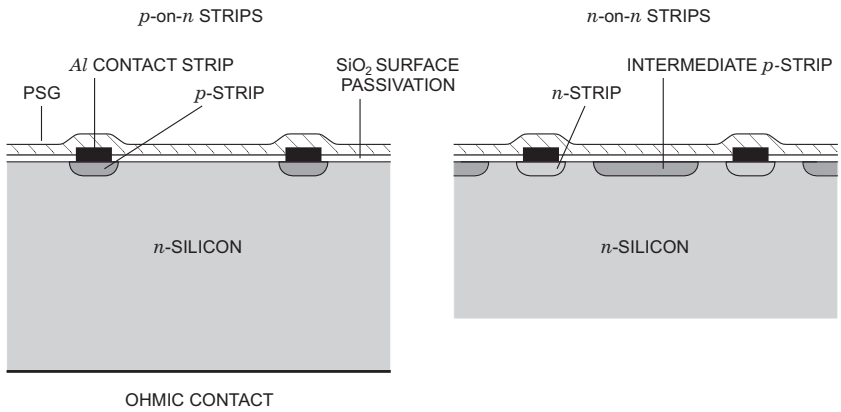

FIG. 2.24. Cross-section of the electrode structure in strip detectors. The left shows the junction side of a single-sided detector. In double-sided detectors the electrodes on the ohmic side require additional isolation structures, for example an intermediate p-region as shown at the right. A layer of phosphosilicate glass (PSG) protects the structure.

2.4 Charge collection

Carrier transport can proceed through diffusion or drift. Diffusion is driven by a concentration gradient. Thermal energy causes carriers to move in random directions, but in the presence of a concentration gradient they collide more frequently in the direction of higher concentration, so the net motion is in the opposite direction. The concentration profile spreads out with time forming a Gaussian distribution with the variance

$$\sigma = \sqrt{Dt} \ , \tag{2.35}$$

where D is a material-dependent diffusion constant.

In the presence of an electric field, the carriers move parallel to the field (drift). However, the velocity does not depend on the time during which the charge carrier is accelerated, as in normal ballistic motion, since the charge carrier also interacts with the crystal lattice, exciting lattice vibrations (phonons). The characteristic times for phonon excitation are much smaller than the transport times, so the carrier is always in equilibrium with the lattice and the velocity is only a function of the electric field

$$\vec{v} = \mu \vec{E} \ , \tag{2.36}$$

where μ is the mobility. The mobility is linked to the diffusion constant through the Einstein relation

$$\mu = \frac{e}{kT} D \ . \tag{2.37}$$

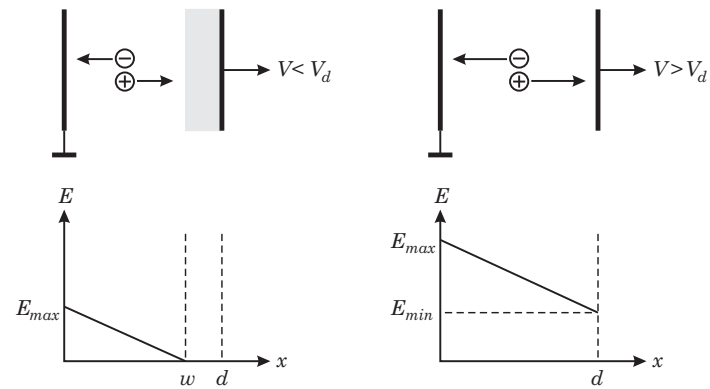

FIG. 2.25. Electric field in a reverse-biased diode in partial depletion (left) and with overbias (right)

Electrons and holes have different mobilities, 1350 and 450 $\mathrm{cm}^2/\mathrm{Vs}$ in Si, respectively (Beadle, Tsai, and Plummer 1984). Thus, in a field of 10^3 V/cm the electron velocity is $1.35 \cdot 10^6$ cm/s. For comparison, the thermal velocity of an electron in Si at room temperature is about 10^7 cm/s, so the carrier motion is the superposition of a substantial random thermal motion and the drift due to the electric field.

Radiation absorbed in the detector's sensitive region forms mobile electrons and holes, which move under the influence of the electric field. Although electrons and holes move in opposite directions, their contribution to the signal current is of the same polarity since they have opposite charge. The time required for a charge carrier to traverse the sensitive volume is called the collection time.

Using the depletion width eqn 2.25 one can rewrite eqn 2.17 for the electric field

$$E(x) = \frac{2(V_b + V_{bi})}{w} \left(\frac{x}{w} - 1\right) . \tag{2.38}$$

The detector bulk is completely depleted of mobile charge when the depletion width equals the thickness of the detector $W = d$. This occurs at the externally applied depletion voltage

$$V_d = \frac{eN_d w^2}{2\varepsilon} - V_{bi} . \tag{2.39}$$

The field drops linearly from its maximum value at the junction to zero at the opposite contact. Increasing the bias voltage beyond this value adds a uniform field due to the voltage beyond depletion, yielding a distribution

$$E(x) = \frac{2V_{di}}{d} \left(1 - \frac{x}{d}\right) + \frac{V_b - V_{di}}{d} , \tag{2.40}$$

where $V_{di} \equiv V_d + V_{bi}$ has been defined as the internal depletion voltage. Figure 2.25 shows the field distribution in partial depletion and with overbias.

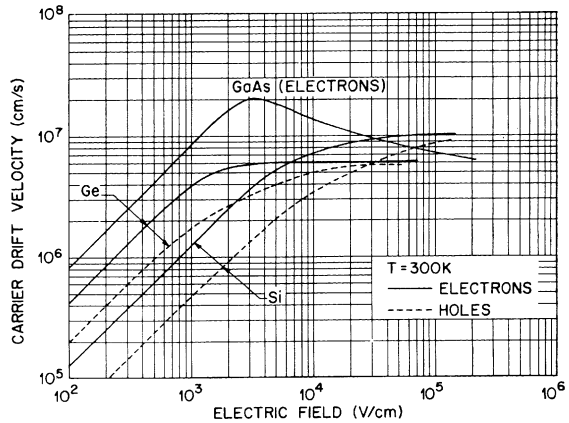

FIG. 2.26. Drift velocity *vs.* electric field in Ge, GaAs and Si. (From Sze 1981. ©John Wiley & Sons, reproduced with permission.)

To calculate the collection time first consider a detector operated at partial depletion $V_b < V_d$. The electric field

$$E(x) = -\frac{eN_d}{\varepsilon}(w - x) \equiv E_0(w - x) \tag{2.41}$$

and the local velocity of a charge carrier

$$v(x) = \mu E(x) = \mu E_0(w - x) \ . \tag{2.42}$$

In Si at 300 K the mobility at low fields is 1350 $\mathrm{cm}^2/\mathrm{Vs}$ for electrons and 480 $\mathrm{cm}^2/\mathrm{Vs}$ for holes. The mobility is constant up to about 10^4 V/cm, but then increased phonon emission reduces the energy going into electron motion, so the mobility decreases. At high fields $E > 10^5$ V/cm the mobility $\mu \propto 1/E$ and carriers attain a constant drift velocity of 10^7 cm/s, as shown in Figure 2.26. The following calculations assume constant mobility; in numerical calculations it is straightforward to include a field-dependent mobility.

The time required for a charge originating at x_0 to reach a point x is

$$t(x) = \int_{x_0}^{x} \frac{1}{v(x)} dx = \frac{1}{\mu E_0} \int_{x_0}^{x} \frac{1}{w - x} dx = -\frac{1}{\mu E_0} \left[\log(w - x)\right]_{x_0}^{x}$$

$$t(x) = -\frac{1}{\mu E_0} \ \log \frac{w - x}{w - x_0} = \frac{\varepsilon}{\mu e N_d} \ \log \frac{w - x}{w - x_0} \ . \tag{2.43}$$

Consider a hole drifting toward the high-field region and collected at the p-electrode $x = 0$. Using the hole mobility μ_p, eqn 2.43 yields

$$t(x_0) = -\frac{1}{\mu_p E_0} \ \log \frac{w}{w - x_0} = \frac{\varepsilon}{\mu_p e N_d} \ \log \frac{w}{w - x_0} \ . \tag{2.44}$$

If we define a characteristic collection time

$$\tau_p \equiv \frac{\varepsilon}{\mu_p e N_d} \ , \tag{2.45}$$

then

$$t(x_0) = \tau_p \ \log \frac{w}{w - x_0} \ . \tag{2.46}$$

For example, $t(x_0 = 0.5w) = 0.7\tau_p$ and $t(x_0 = 0.95w) = 3.0\tau_p$.

For the electrons drifting toward the low-field electrode $x = w$, eqn 2.43 does not yield a solution. However, it can be rewritten to yield the position as a function of time

$$x(t) = w - (w - x_0)\ e^{-t/\ \tau_n} \ , \tag{2.47}$$

where τ_n has been defined analogously to τ_p. For a charge originating at the metallurgical junction $x_0 = 0$ and drifting toward $x = w$

$$x(t) = w(1 - \ e^{-t/\ \tau_n}) \ . \tag{2.48}$$

In this simple picture, a charge carrier drifting toward the low-field region is never collected (in reality this is accomplished by diffusion), although after a time $t = 3\tau_n$ the carrier will have traversed 95% of the detector. Note that in a partially depleted detector the collection time constants τ_n and τ_p are independent of the applied bias voltage (and depletion thickness), but determined only by the doping concentration of the bulk material and the carrier mobility. τ_n is numerically equal to the dielectric relaxation time of the n-type bulk

$$\tau = \rho\varepsilon = \varepsilon_{Si}\ \varepsilon_0\ \rho = 1.05\ \left[\frac{\text{ns}}{\text{k}\Omega \cdot \text{cm}}\right]\ \ \rho \ ,$$

so the resistivity gives a quick estimate of the collection time. In n-type silicon of $10\,\text{k}\Omega\,\text{cm}$ resistivity $\tau_n = 10.5\,\text{ns}$ and $\tau_p = 31.5\,\text{ns}$, so collection times are about 30 and 90 ns, respectively.

The collection time can be reduced by operating the detector at bias voltages exceeding the depletion voltage (overbias, often referred to as "overdepletion"). The field distribution was given in eqn 2.40, which can be rewritten as

$$E(x) = E_0\ \left(1 - \frac{x}{d}\right) + E_1 \ . \tag{2.49}$$

This yields a collection time

$$t(x) = \int_{x_0}^{x} \frac{1}{v(x)} dx = \frac{1}{\mu} \int_{x_0}^{x} \frac{1}{E_0\left(1 - \frac{x}{d}\right) + E_1} dx = -\frac{d}{\mu E_0}\left[\log\left(E_0 + E_1 - E_0 \frac{x}{d}\right)\right]_{x_0}^{x}$$

$$t(x) = \frac{d}{\mu E_0} \log \frac{E_0 + E_1 - E_0 \dfrac{x}{d}}{E_0 + E_1 - E_0 \dfrac{x_0}{d}} \ . \tag{2.50}$$

For holes originating at $x_0 = w$ and drifting to the p-electrode $x = 0$

$$t_{cp} = \frac{d}{\mu_p E_0} \log \left(1 + \frac{E_0}{E_1}\right) \ . \tag{2.51}$$

The corresponding result obtains for electrons originating at $x_0 = 0$ and drifting to the n-electrode $x = w$

$$t_{cn} = \frac{d}{\mu_n E_0} \log \left(1 + \frac{E_0}{E_1}\right) \ . \tag{2.52}$$

For large overbias $E_1 \gg E_0$,

$$\log \left(1 + \frac{E_0}{E_1}\right) \approx \frac{E_0}{E_1} \tag{2.53}$$

and

$$t_{cp} = \frac{d}{\mu_p E_1} \ , \tag{2.54}$$

as expected for a uniform field.

Rewritten in terms of voltages, eqns 2.52 and 2.53 become

$$t_{cp} = \frac{d^2}{2\mu_p V_{di}} \log \left(\frac{V_b + V_{di}}{V_b - V_{di}}\right) \tag{2.55}$$

and

$$t_{cn} = \frac{d^2}{2\mu_n V_{di}} \log \left(\frac{V_b + V_{di}}{V_b - V_{di}}\right) \ , \tag{2.56}$$

where $V_{di} \equiv V_d + V_{bi}$.

For example, consider a sensor made of n-type silicon with $10\,\mathrm{k\Omega\,cm}$ resistivity and a thickness of $300\,\mu\mathrm{m}$. The depletion voltage is $30\,\mathrm{V}$. When operated at twice the depletion voltage $V_b = 60\,\mathrm{V}$ (*i.e.* $E_0 = 2 \cdot 10^3$ and $E_1 = 10^3\,\mathrm{V/cm}$), the collection times are 12 ns for electrons and 36 ns for holes. These are substantially smaller than in the partially depleted device, where collection times are 30 ns for electrons and 90 ns for holes.

2.5 Time dependence of the signal current

As illustrated in Figure 2.27, charge moving in the sensitive volume of the sensor gives rise to a signal current, as indicated in the accompanying equivalent circuit. At this point we need to determine $i_s(t)$. When does the signal current begin? When the charge reaches the electrode or when the charge begins to move?

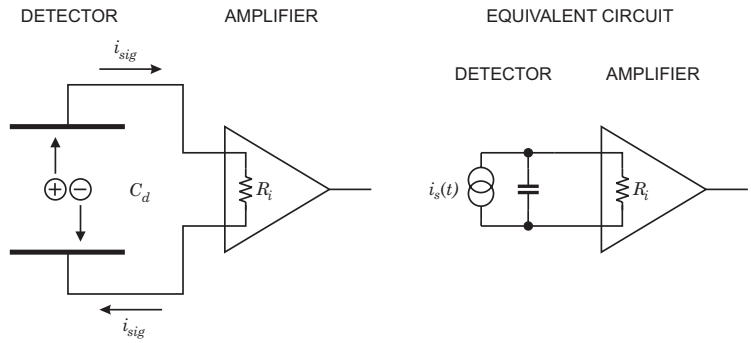

FIG. 2.27. Charge moving in the detector volume induces a signal current in the external circuit (left). The detector's equivalent circuit is shown at the right.

Although the first answer is quite popular (encouraged by the phrase "charge collection"), the second is correct; current flow begins instantaneously.

To understand the physics of induced charge, first consider a charge q near a single, infinitely large electrode. All electric field lines from the charge terminate on the electrode. Integrating the field on a Gaussian surface S surrounding the charge yields

$$\oint_S \vec{E} d\vec{a} = q \ .$$

Correspondingly, integrating over a Gaussian surface enclosing only the electrode yields the charge $-q$. Since the direction of the field lines is opposite relative to the first integral, this charge – the "induced charge" – has the opposite sign. Next, add a second electrode, as shown in Figure 2.28. If the charge is positioned

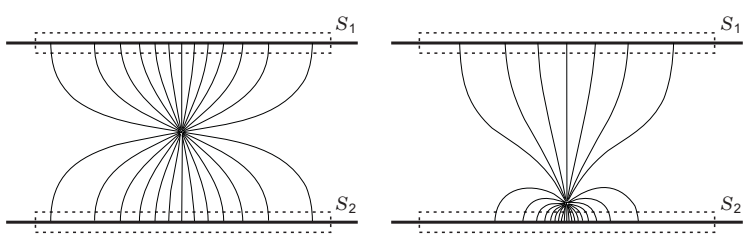

FIG. 2.28. A charge q positioned midway between two parallel plates induces equal charge on each plate (left). Integrating over the Gaussian surface S_1 or S_2 yields the induced charge $-q/2$. When positioned close to the bottom plate (right panel) more field lines terminate on the lower than on the upper plate, so the charge enclosed by S_2 is larger than the charge enclosed by S_1, *i.e.* the induced charge on the lower plate has increased.

midway between the two electrodes, half of the field lines will terminate on the upper and the other half on the lower electrode. Integrating over a Gaussian surface S_1 enclosing the upper electrode yields $-q/2$, as does integration around the lower electrode. If the charge is moved very close to the lower electrode, as in the right panel of Figure 2.28, most of the field strength will terminate there, and the induced charge will be correspondingly higher. Thus, a charge moving from the upper to the lower electrode will initially induce most of its charge on the upper electrode, with an increasing proportion shifting to the lower electrode as the charge moves toward it.

We cannot observe the induced charge directly, but we can measure its change. If the two electrodes are connected to form a closed circuit, the change in induced charge manifests itself as a current. Integrating the induced current as the charge traverses the distance from the top to the bottom electrode yields the difference in induced charge q.

Quantitatively, this is described by Ramo's theorem (Ramo 1939). The same problem has been solved by others, *e.g.* by Shockley (1938), but Ramo presented a particularly elegant formulation of the solution. The following discussion applies to *all* structures that register the effect of charges moving in an ensemble of electrodes, *i.e.* not just semiconductor or gas-filled ionization chambers, but also resistors, capacitors, photoconductors, vacuum tubes, *etc.*

Frequently, the detector signal is calculated using an energy balance approach, where the energy gained by the charge moving along the electric field is set equal to the change in energy of the capacitor

$$dU = eEdx = CV_b dV \tag{2.57}$$

to calculate the change in voltage across the detector. However, this neglects the additional energy expended in interactions with the lattice, *i.e.* phonon excitation. More generally, this objection applies to all collision-limited transport, whether in solids, liquids, or gases. Thus, there are very few detector structures where the energy balance approach is valid. Although in certain configurations the energy balance sometimes gives the correct signal magnitude, it tends to predict the wrong pulse shape, as will be illustrated in Section 2.5.2. The induced charge formalism described below applies the correct physics to obtain generally valid results.

2.5.1 *Induced charge – Ramo's theorem*

Consider a mobile charge in the presence of any number of grounded electrodes. Surround the charge q with a small equipotential sphere. Then, if V is the potential of the electrostatic field, in the region between conductors

$$\nabla^2 V = 0 \ . \tag{2.58}$$

Call V_q the potential of the small sphere and note that $V = 0$ on the conductors. Applying Gauss' law yields

$$\int\limits_{\substack{\text{sphere's}\\ \text{surface}}} \frac{\partial V}{\partial n} ds = 4\pi q \ . \tag{2.59}$$

Next, consider the charge removed and one conductor A raised to unit potential. Call the potential V_1, so that

$$\nabla^2 V_1 = 0 \tag{2.60}$$

in the space between the conductors, including the site where the charge was situated. Call the new potential at this point V_{q1}. Green's theorem states that

$$\int\limits_{\substack{\text{volume}\\ \text{between}\\ \text{boundaries}}} (V_1 \nabla^2 V - V \nabla^2 V_1) dv = -\int\limits_{\substack{\text{boundary}\\ \text{surfaces}}} \left[V_1 \frac{\partial V}{\partial n} - V \frac{\partial V_1}{\partial n} \right] ds \ . \tag{2.61}$$

Choose the volume to be bounded by the conductors and the tiny sphere. Then the left-hand side is 0 and the right-hand side may be divided into three integrals:

1. Over the surfaces of all conductors except A. This integral is 0 since on these surfaces $V = V_1 = 0$.
2. Over the surface of A. As $V_1 = 1$ and $V = 0$ this reduces to

$$-\int\limits_{\text{surfaceA}} \frac{\partial V}{\partial n} ds$$

3. Over the surface of the sphere.

$$-V_{q1} \int\limits_{\substack{\text{sphere's}\\ \text{surface}}} \frac{\partial V}{\partial n} ds + V_q \int\limits_{\substack{\text{sphere's}\\ \text{surface}}} \frac{\partial V_1}{\partial n} ds$$

The second integral is zero by Gauss' law, since in this case the charge is removed. Combining these three integrals yields

$$0 = -\int\limits_{\text{surfaceA}} \frac{\partial V}{\partial n} ds - V_{q1} \int\limits_{\substack{\text{sphere's}\\ \text{surface}}} \frac{\partial V}{\partial n} ds = 4\pi Q_A - 4\pi q V_{q1} \tag{2.62}$$

or

$$Q_A = q V_{q1} \ . \tag{2.63}$$

If the charge q moves in direction x, the current on electrode A is

$$i_A = \frac{dQ_A}{dt} = q \frac{dV_{q1}}{dt} = q \left(\frac{\partial V_{q1}}{\partial x} \frac{dx}{dt} \right) \ . \tag{2.64}$$

Since the charge's velocity

$$v_x = \frac{dx}{dt} \ ,$$

the induced current on electrode A is

$$i_A = qv_x \frac{\partial V_{q1}}{dx} \equiv qv_x \frac{\partial \Phi}{dx} \ , \tag{2.65}$$

where Φ is the "weighting potential" that describes the coupling of a charge at any position to electrode A. The weighting potential applies to a specific electrode and is obtained by setting the potential of the electrode to 1 and setting all other electrodes to potential 0.

Summary of results:

- If a charge q moves along any path s from position 1 to position 2, the net induced charge on electrode k is

$$\Delta Q_k = q(V_{q1}(2) - V_{q1}(1)) \equiv q\left(\Phi_k(2) - \Phi_k(1)\right) \ . \tag{2.66}$$

- The instantaneous current can be expressed in terms of a weighting field

$$i_k = -q\overrightarrow{v} \cdot \overrightarrow{E_Q} \ . \tag{2.67}$$

The weighting field is determined by applying unit potential to the measurement electrode and zero to all others. *Note that the electric field and the weighting field are distinctly different.*

- The electric field determines the charge trajectory and velocity.
- The weighting field depends only on geometry and determines how charge motion couples to a specific electrode.
- Only in two-electrode configurations are the electric field and the weighting field of the same form.

2.5.2 *Parallel plate geometry with uniform field*

A semiconductor detector with very large overbias can be approximated by a uniform field. The bias voltage V_b is applied across the electrode spacing d. The electric field

$$E = \frac{V_b}{d} \tag{2.68}$$

determines the motion of a charge carrier in the detector. The carrier's velocity

$$v = \mu E = \mu \frac{V_b}{d} \ . \tag{2.69}$$

The weighting field is obtained by applying unit potential to the collection electrode and grounding the other:

$$E_Q = \frac{1}{d} \ , \tag{2.70}$$

so the induced current

$$i = qvE_Q = q\mu\frac{V_b}{d}\frac{1}{d} = q\mu\frac{V_b}{d^2} \ . \tag{2.71}$$

Since both the electric field and the weighting field are uniform throughout the detector, the current is constant until the charge reaches its terminal electrode.

Assume that the charge is created at the opposite electrode and traverses the detector thickness d. The required collection time, *i.e.* the time required to traverse the distance d

$$t_c = \frac{d}{v} = \frac{d}{\mu\frac{V_b}{d}} = \frac{d^2}{\mu V_b} \ . \tag{2.72}$$

The induced charge

$$Q = it_c = q\mu\frac{V_b}{d^2}\frac{d^2}{\mu V_b} = q \ . \tag{2.73}$$

Next, assume an electron–hole pair formed at coordinate x from the positive electrode. The collection time for the electron

$$t_{ce} = \frac{x}{v_e} = \frac{xd}{\mu_e V_b} \tag{2.74}$$

and the collection time for the hole

$$t_{ch} = \frac{d-x}{v_h} = \frac{(d-x)d}{\mu_h V_b} \ . \tag{2.75}$$

Since electrons and holes move in opposite directions, they induce current of the same sign at a given electrode, despite their opposite charge. The induced charge due to the motion of the electron

$$Q_e = e\mu_e\frac{V_b}{d^2}\frac{xd}{\mu_e V_b} = e\frac{x}{d} \ . \tag{2.76}$$

Correspondingly, the hole contributes

$$Q_h = e\mu_h\frac{V_b}{d^2}\frac{(d-x)d}{\mu_h V_b} = e\left(1-\frac{x}{d}\right) \ . \tag{2.77}$$

Assume that $x = d/2$. After the collection time for the electron

$$t_{ce} = \frac{d^2}{2\mu_e V_b} \tag{2.78}$$

the induced charge is $e/2$. At this time the hole, due to its lower mobility $\mu_h \approx \mu_e/3$, has induced $e/6$, yielding a cumulative induced charge of $2e/3$. After the additional time for the hole collection, the remaining charge $e/3$ is induced, yielding the total charge e. The measured charge depends on the integration

time. Integration times larger than the collection time of all charge carriers yield the full charge. A shorter integration time yields a fractional charge.

Equation 2.76 illustrates the difference between the induced charge and the erroneous energy balance approach, which predicts the incremental signal charge $dQ = e\ dV/V$ instead of $dQ = e\ dx/d$. Superficially, it appears that the energy balance gives a different result because the potential distribution in a parallel plate semiconductor detector is not linear due to space charge. However, as discussed by Cavalleri *et al.* (1972) the validity of Ramo's theorem is not affected by the presence of space charge. Fundamentally, the discrepancy arises because we are not dealing with ballistic transport and the energy balance approach invokes a conservation law without considering the total energy of the system.

In the parallel plate configuration electrons and holes contribute equally to the currents on both electrodes and the instantaneous current at any time is the same on both electrodes, although of opposite sign. The continuity equation (Kirchhoff's law) must be satisfied:

$$\sum_k i_k = 0 \ . \tag{2.79}$$

With only two electrodes, $i_1 = -i_2$ and the currents observed on the n and p electrodes differ only in their polarity. In the presence of multiple electrodes the instantaneous current from one electrode must balance the sum of the currents from the others, so all signal currents can be different. This is the situation in strip detectors.

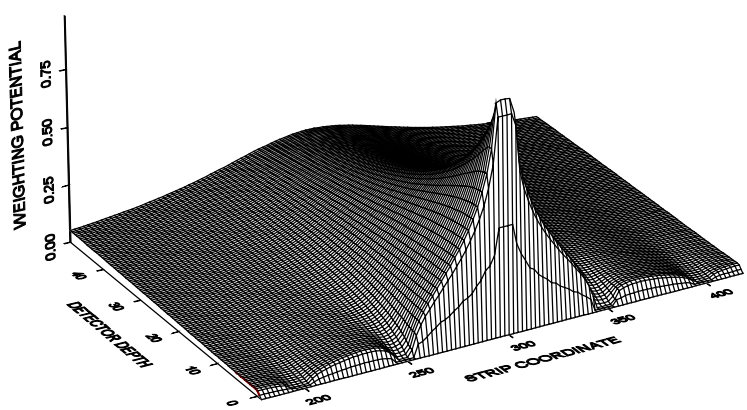

FIG. 2.29. Weighting potential for a 300 μm thick strip detector with strips on a pitch of 50 μm. The central strip is at unit potential and the others at zero. Only 50 μm of depth are shown.

2.5.3 *Double-sided strip detector*

The detector has strip electrodes on both faces. The strip pitch is assumed to be small compared to the thickness. The electric field is similar to a parallel-plate geometry, except in the immediate vicinity of the strips. The signal weighting potential, however is very different, as shown in Figure 2.29.

Figure 2.30 shows cuts through the weighting potential and the weighting field along the center of the signal strip (left) and the neighbor strip (right). Consider an electron–hole pair q_n, q_p originating at a point x_0 on the center-line of a strip. The motion of the electron towards the n-electrode at coordinate x_n is equivalent to the motion of a hole in the opposite direction to the p-electrode at x_p. The total induced charge on electrode k after the charges have traversed the detector is

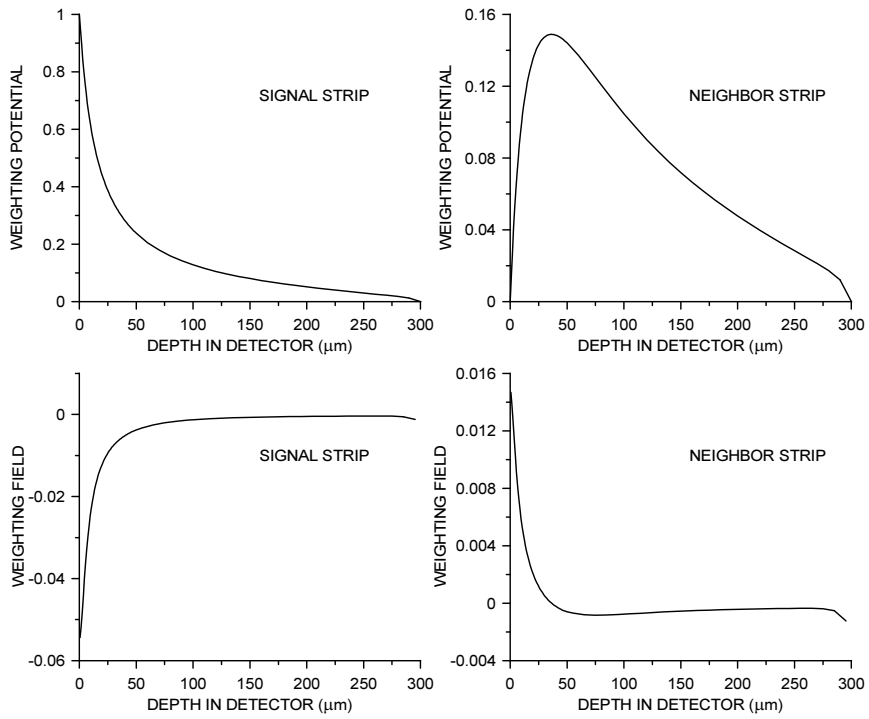

FIG. 2.30. Cuts of the weighting field through the center of a strip electrode and through the neighboring electrode. Along the axis of a measurement electrode the weighting field is monotonic, so the signal charge increases as the carrier approaches the electrode. The weighting field through the neighbor electrode changes sign at about 40 μm depth, so the induced current inverts and the signal on this strip integrates to zero.

$$Q_k = q_p[\Phi_{Qk}(x_p) - \Phi_{Qk}(x_0)] + q_n[\Phi_{Qk}(x_n) - \Phi_{Qk}(x_0)] \ . \tag{2.80}$$

Since the hole charge $q_p = e$ and $q_n = -e$,

$$Q_k = e[\Phi_{Qk}(x_p) - \Phi_{Qk}(x_0)] - e[\Phi_{Qk}(x_n) - \Phi_{Qk}(x_0)] \ . \tag{2.81}$$

If the signal is measured on the p-electrode, collecting the holes,

$$Q_k = e[\Phi_{Qk}(x_p) - \Phi_{Qk}(x_n)] \ . \tag{2.82}$$

Then $\Phi_{Qk}(x_p) = 1$, $\Phi_{Qk}(x_n) = 0$, and $Q_k = e$. If, however, the charge is collected on the neighboring strip $k+1$, then $\Phi_{Qk+1}(x_p) = 0$, $\Phi_{Qk+1}(x_n) = 0$, and $Q_{k+1} = 0$.

In general, if a moving charge does not terminate on the measurement electrode, signal current will be induced, but the current changes sign and integrates to zero. This is illustrated in Figure 2.31. The plots of the weighting field in Figure 2.30 show that the induced current on both the signal and neighbor strips is strongly peaked near the strip. However, the weighting field of the neighbor strip changes sign at about $40\,\mu$m distance from the strip electrode, so the induced charge initially builds up as a carrier drifts $260\,\mu$m from the far side and then rapidly reduces to zero as the carrier reaches the signal strip.

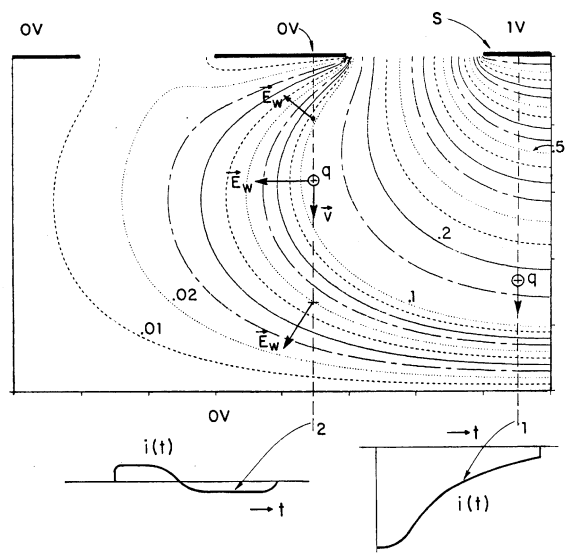

FIG. 2.31. The weighting field in a strip detector. The measurement electrode is the right-most strip. The induced current is shown for a charge terminating on the measurement electrode (right) and the neighbor electrode (left), showing the change in polarity. (From Radeka 1988. ©Annual Reviews www.annualreviews.org, reprinted with permission)

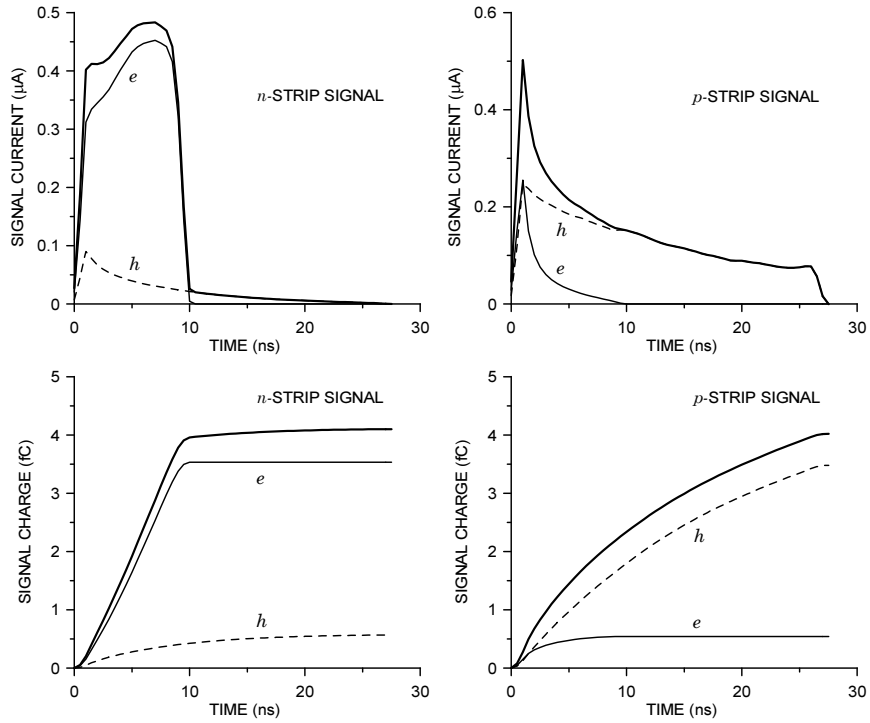

FIG. 2.32. Strip detector signals for an n-bulk device with 60 V depletion voltage operated at a bias voltage of 90 V. The electron (e) and hole (h) components are shown together with the total signal (bold). The top row shows the signal current and the bottom row shows the integrated current. Despite marked differences in shape, the total charge is the same on both sides.

Note, however, that in general charge cancellation on "non-collecting" electrodes relies on the motion of both electrons and holes. Assume, for example, that the holes are stationary, so they don't induce a signal. Then the first term of eqn 2.80 vanishes, which leaves a residual charge

$$Q_k = e[\Phi_{Qk}(x_0) - \Phi_{Qk}(x_n)] \ , \tag{2.83}$$

since for any coordinate not on an electrode $Q_k(x_0) \neq 0$, although it may be very small.

An important consequence of this analysis is that one cannot simply derive pulse shapes by analogy with a detector with contiguous electrodes (*i.e.* a parallel plate detector of the same overall dimensions as a strip detector). Specifically,

1. The shape of the current pulses can be quite different.

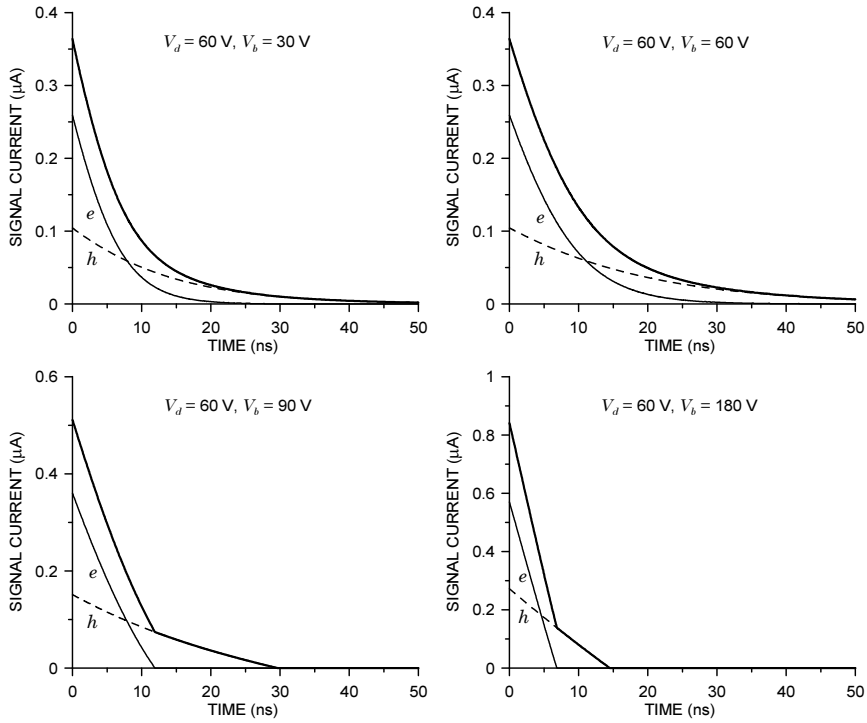

FIG. 2.33. Current signals for tracks traversing a detector with parallel-plate electrodes. The upper two plots are at partial and full depletion. The collection times are practically the same, with only minor differences due to the field-dependent mobility. The lower two plots show the faster charge collection with overbias, at 90 and 180 V, respectively.

2. The signals seen on opposite strips of a double-sided detector have different shapes.
3. The net induced charge on the p- or n-side is not split evenly between electrons and holes.
   - Because the weighting potential is strongly peaked near the signal electrode, most of the charge is induced when the moving charge is near the signal electrode.
   - As a result, most of the signal charge is due to the charge terminating on the signal electrode.

Figure 2.32 shows the current pulses at the p- and n- electrodes of a double-sided strip detector. The pulse durations are determined by the collection time, but the pulse shapes are very different on the two sides. Nevertheless, when integrating over the collection time one obtains the same signal charge on both sides,

as also shown in Figure 2.32. Since the weighting function is so strongly peaked near the strip electrodes, nearly all of the signal charge on the n-side is due to electrons and on the p-side due to holes. For comparison, Figure 2.33 shows the signal currents of a detector with a simple parallel-plate electrode structure and the same depletion voltage. The pulse shapes are strikingly different from those shown for a strip detector. At partial and full depletion the pulse shapes are practically unchanged, except for minor differences due to the field dependent mobility. The speed-up in charge collection with overbias is also apparent. Overbias also removes the drawn-out "tails" due to the vanishing field as carriers approach the ohmic electrode. For the same depletion and bias voltages the pulse durations are the same as in strip detectors, but the shapes of the current differ greatly because of the different weighting field.

2.6 Charge collection in the presence of trapping

Practical semiconductor crystals suffer from imperfections introduced during crystal growth, during device fabrication, or by radiation damage. Defects in the crystal such as impurity atoms, vacancies, and structural irregularities (*e.g.* dislocations) introduce states into the crystal that can trap charge.

Extremely minute trap concentrations lead to significant effects. As a carrier drifts, superimposed is an isotropic random motion due to its much higher thermal velocity. As the particle "scans" the crystal the probability of encountering a trap is proportional to the elapsed time. As a consequence, charge trapping is characterized by a carrier lifetime τ, the time a charge carrier can "survive" in a crystal before trapping or recombination with a hole. This is discussed quantitatively in Appendix F.

Trapping removes mobile charge available for signal formation. Depending on the nature of the trap, thermal excitation or the externally applied field can release the carrier from the trap, leading to delayed charge collection.

Given a lifetime τ, a packet of charge Q_0 will decay so that as a function of time the remaining charge

$$Q(t) = Q_0 e^{-t/\tau} \ . \tag{2.84}$$

In an electric field the charge drifts with a velocity $v = \mu E$. The time required to traverse a distance x

$$t = \frac{x}{v} = \frac{x}{\mu E} \tag{2.85}$$

after which the remaining charge

$$Q(x) = Q_0 e^{-x/\mu E \tau} \equiv Q_0 e^{-x/L} \ . \tag{2.86}$$

Since the drift length L is proportional to the mobility–lifetime product, $\mu\tau$ is often used as a figure of merit.

Assume a detector with a simple parallel-plate geometry. For a charge traversing the increment dx of the detector thickness d, the induced signal charge

$$dQ_s = Q(x)\frac{dx}{d} \ , \tag{2.87}$$

so the total induced charge

$$Q_s = \frac{1}{d}\int_0^d Q(x)dx = \frac{1}{d}\int_0^d Q_0 e^{-x/L}dx$$

$$Q_s = Q_0\frac{L}{d}\left(1 - e^{-d/L}\right) \ . \tag{2.88}$$

If the thickness of the detector is much greater than the drift length $d \gg L$, the measured fraction of the signal charge

$$\frac{Q_s}{Q_0} \approx \frac{L}{d} \ . \tag{2.89}$$

For $> 95\%$ charge yield the detector thickness must be greater than $3L$. The $\mu\tau$ product differs for electrons and holes, so the two contributions must be evaluated separately. For a charge deposition at coordinate x_0 the induced charge is given by

$$\frac{Q_s}{Q_o} = \frac{L_e}{d}\left[1 - \exp\left(\frac{d - x_o}{L_e}\right)\right] + \frac{L_h}{d}\left[1 - \exp\left(\frac{x_o}{L_h}\right)\right] \ , \tag{2.90}$$

where $L_e = (\mu\tau)_e E$ and $L_h = (\mu\tau)_h E$. This expression is often referred to as the Hecht equation (Hecht 1932).

In high quality silicon detectors $\tau \approx 10\,\mathrm{ms}$. Since the electron mobility $\mu_e = 1350\,\mathrm{V/cm \cdot s^2}$, the drift length at a field $E = 10^4\,\mathrm{V/cm}$ is about $10^4\,\mathrm{cm}$. In amorphous silicon (short lifetime, low mobility) typical drift lengths are of order $10\,\mu\mathrm{m}$. In high quality deposited diamond layers drift lengths are in the range of hundreds of μm.

In tracking devices where particles traverse the detector the charge loss reduces the average signal. In x-ray or gamma spectroscopy charge loss smears the signal distribution to lower energies. Even small amounts of trapping can lead to significant low energy tails that adversely affect the ability to separate adjacent lines.

2.7 Semiconductor detector materials

The most commonly used solid state detector materials are silicon and germanium. Both materials provide excellent energy resolution and large volume single crystals with good electric properties can be grown. Germanium detectors with several hundred cm^3 volume are common and carrier lifetimes are ample to provide excellent charge collection efficiency. Germanium crystals can be grown with extremely high purity (Haller, Hansen, and Goulding 1981), so large volumes can be depleted with fields well below breakdown levels. Silicon dominates

in charged particle and x-ray spectroscopy. In high-resolution x-ray detectors it must be cooled to minimize the shot noise contribution of the reverse bias current. Furthermore, photoelectric absorption limits its use to energies less than about 100 keV. The higher atomic number of germanium extends the energy range to 10 MeV or so. The lower bandgap of Ge (0.66 eV *vs.* 1.12 eV for Si) increases the signal charge yield, but greatly increases the reverse bias current, so Ge detectors are typically operated at liquid nitrogen temperature (77 K). Silicon has a high-quality and robust native oxide that – when properly grown – provides a well-controlled interface to the underlying silicon, while protecting the bulk from environmental contamination. This property of silicon is unique among semiconductors and it is a key ingredient in high-density integrated circuits and finely patterned detectors.

Nevertheless, many applications would benefit from a better combination of bandgap and absorption. For example, materials with absorption similar to Ge but with a larger bandgap would allow room temperature operation of high-resolution gamma-ray detectors. Because of the exponential dependence of reverse bias current on temperature, even small increases make a big difference. Conversely, a material with the bandgap of silicon, but with higher atomic number would extend absorption efficiency to higher energies. GaAs for example, has a slightly larger bandgap but much higher stopping power than Si, so detectors can cover a larger energy range with good resolution (for example see Owens 1999). GaAs has suffered from poor lifetimes, as the commonly used "semi-insulating" material owes its high resistivity to the presence of mid-gap states. High purity GaAs has shown better results (Owens 1999). Many interesting materials are not available as single crystals of sufficient size. Furthermore, direct bandgap materials tend to be more susceptible to recombination.

In high-energy physics radiation damage is the main driver in the search for new materials. Cost is sometimes cited, but in the total cost of a detector system, the cost of raw silicon material is minor and perceived cost benefits in sensor material are easily outweighed by increased electronic requirements. For example, the use of amorphous silicon was promoted with the argument that as a disordered material it is not susceptible to displacement damage. However, the charge yield is so low that the required electronic power dissipation increases substantially to maintain an adequate signal-to-noise ratio for minimum ionizing particles. Crystalline silicon must be irradiated well beyond fluences of $10^{15}\,\mathrm{cm}^{-2}$ before its characteristics degrade to the level of amorphous silicon. The advantage of crystalline material is that it provides much superior performance over the major part of its operational lifetime. This does not mean that amorphous silicon is useless; it has been successfully applied in certain x-ray imaging applications where charge yield and electronic noise are less important.

Unfortunately, the ideal detector material does not exist, as desirable qualities are often at odds. For example, increasing charge yield requires a smaller bandgap, which increases the reverse bias current. Another consideration is the dielectric constant, which affects the capacitance and thus the electronic noise.

If the noise current I_n is made negligible, by cooling the detector for example, the equivalent noise charge (Chapter 3)

$$Q_n \propto e_n C_d = e_n \varepsilon \frac{A}{d} \ . \tag{2.91}$$

Here e_n is the amplifier's equivalent input noise voltage, C_d the detector capacitance, and A and d are the sensor's active area and thickness. The energy resolution is the product of the noise charge Q_n and the energy required to form an electron–hole pair E_i,

$$\Delta E = E_i Q_n \propto e_n E_i \varepsilon \frac{A}{d} \ . \tag{2.92}$$

Thus, for a given pulse shaper and sensor geometry, constant noise requires that the product of the amplifier's input noise voltage and the sensor's ionization energy and dielectric constant remains constant. As will be shown in Chapter 6, at best the power dissipation scales inversely with the square of the required noise charge $P \propto 1/Q_n^2$, so the desire for wide bandgap materials tends to carry a substantial penalty in front-end power. To some degree this can be alleviated by segmentation, *i.e.* reducing the electrode area A per channel. Nevertheless, the product of ionization energy and dielectric constant $E_i\varepsilon$ is an important figure of merit for sensor materials.

Table 2.3 summarizes the properties of some representative materials. As noted above, the bandgap affects the reverse bias current. Linked with this is

Table 2.3 Representative detector materials. Mobilities μ are in units of $\mathrm{cm^2V^{-1}s^{-1}}$ and $\mu\tau$ products in $\mathrm{cm^2V^{-1}}$.

Material	E_g (eV)	E_i (eV)	ε	μ_e	μ_h	$(\mu\tau)_e$	$(\mu\tau)_h$	ρ	$\langle Z \rangle$
Si	1.12	3.6	11.7	1350	450	> 1	> 1	2.33	14
Ge	0.67	2.96	16	3900	1900	> 1	> 1	5.33	32
GaAs	1.43	4.2	12.8	8000	400	$8 \cdot 10^{-5}$	$4 \cdot 10^{-6}$	5.32	31.5
Diamond	5.5	13	5.7	1800	1200			3.52	6
4H-SiC	3.26	8	9.7	1000	115	$4 \cdot 10^{-4}$	$8 \cdot 10^{-5}$	3.21	10
GaN	3.39	8 – 10		1000	30			6.15	19
InP	1.35	4.2	12.4	4600	150	$5 \cdot 10^{-6}$	$< 10^{-5}$	4.78	32
CdTe	1.44	4.43	10.9	1100	100	$3 \cdot 10^{-3}$	$2 \cdot 10^{-4}$	5.85	50
$Cd_{0.9}Zn_{0.1}Te$	1.572	4.64	10	1000	120	$4 \cdot 10^{-3}$	$1.2 \cdot 10^{-4}$	5.78	49.1
HgI_2	2.15	4.2	8.8	100	4	$3 \cdot 10^{-4}$	$4 \cdot 10^{-5}$	6.4	62
TlBr	2.68	6.5	30	30	4	$5 \cdot 10^{-4}$	$2 \cdot 10^{-6}$	7.56	58
a-Si	1.9	6	12	1 – 4	0.05	$2 \cdot 10^{-7}$	$3 \cdot 10^{-8}$	2.3	14

the ionization energy E_i, which sets the charge yield. The electronic noise is set by both the reverse bias current and the detector capacitance, which depends on the dielectric constant ε. The electron and hole mobilities μ_e and μ_h determine the collection time, whereas the mobility–lifetime products $(\mu\tau)_e$ and $(\mu\tau)_h$ indicate the maximum useful detector thickness. The density gives the energy loss for minimum ionizing particles in high energy particle tracking and the atomic number is key in setting the absorption in x-ray and gamma detection. The data for diamond and amorphous silicon are for films deposited by chemical vapor deposition, so the material is polycrystalline with properties strongly dependent on the growth conditions. Papers on diamond detectors quote a "charge collection distance", which specifies how far electrons and holes are separated before recombination. A typical value for high-quality films is $200\,\mu$m at a field of 10^4 V/cm (Edwards 2004). The size of the microcrystals in diamond becomes visible in precision measurements of the position resolution (Lari 2005).

For more details on materials for x-ray spectroscopy see Owens (2004). Moll (2003) and the RD50 website give information on radiation resistance. Compound semiconductors are of great interest in this context, as they allow the bandgap to be tuned by changing the composition ("bandgap engineering"). For example, in ZnTe the bandgap $E_g = 2\,\mathrm{eV}$ and in CdTe it is $1.5\,\mathrm{eV}$. Varying the composition x of $\mathrm{Cd}_{(1-x)}\mathrm{Zn}_x\mathrm{Te}$ changes the bandgap between the two extremes. Empirically it has been found that (Olega 1985)

$$E_g(x) = 1.510 + 0.606x + 0.139x^2\,\mathrm{eV}\ . \tag{2.93}$$

In compound semiconductors the hole mobility tends to be much smaller than for electrons. If the integration time of the electronics is tailored to the electron collection time, the hole contribution may be negligible, so on the average only half the signal charge is available. For charged particles traversing the detector the signal will be the same, but for photons the induced electron signal will depend on where the photon was absorbed. As shown in Section 2.5.3 the charge induced on small-pitch strip or pixel electrodes is preferentially due to the carrier type collected on the respective electrode. Furthermore, the charge induced on non-collecting electrodes only integrates to zero after both carrier types have reached their respective electrodes. This can be exploited in coplanar strip arrays to mitigate the effect of incomplete hole collection. By ganging every other strip and subtracting the signals from the two sets the net electron signal is measured (Luke 1994, 1995, Amman and Luke 1999).

Work on alternative materials to complement Si and Ge has been pursued for decades. Progress has been made and new materials have been applied in niche areas. Clearly, many properties play together and developments in materials growth technology can change the picture significantly.

2.8 Photodiodes

Photodiodes differ from charged particle and x-ray detectors in the very small absorption depth of visible light. As the photon energies are close to the bandgap,

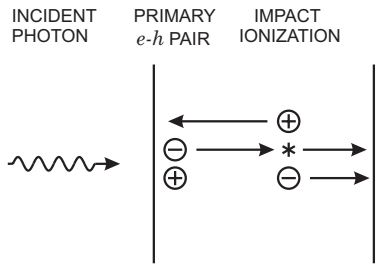

FIG. 2.34. At sufficiently high fields an electron gains sufficient energy for impact ionization, which in this example doubles the signal charge.

the absorption length in silicon changes strongly with wavelength (Figure 2.8). At a typical scintillator emission wavelength of 400 nm the absorption length is of order 100 nm, whereas at 700 nm wavelength it is about 5 μm. To obtain good quantum efficiency at short wavelengths the "dead layers" at the entrance surface must be significantly thinner than typical metallization or doping layers. Unlike photomultiplier tubes whose quantum efficiency is in the 10 – 30% range, optimized silicon photodiodes achieve quantum efficiencies > 80% at wavelengths from 400 nm to nearly 1 μm. UV-extended photodiodes have useful efficiency down to 200 nm.

At best the signal charge is one electron per photon, so low noise is crucial. Examples will be presented in Chapters 4 and 8. Small pixels, notably CCDs, allow single photon sensitivity, but large area devices as might be used in scintillation detectors are severely limited by electronic noise. The signal charge can be increased by incorporating internal gain in the detector. The mechanism is impact ionization. At sufficiently high fields electrons gain sufficient energy between interactions with the lattice that they can eject electrons from lattice atoms (Figure 2.34). The gain within a length d is

$$G_n = e^{\alpha_n d} \ . \tag{2.94}$$

The electron ionization coefficient α_n depends strongly on the electric field

$$\alpha_n = \alpha_{n0} e^{-E_n/|E|} \ . \tag{2.95}$$

In silicon $\alpha_n = 3.36 \cdot 10^6\,\mathrm{cm}^{-1}$ and $E_n = 1.76 \cdot 10^6\,\mathrm{V/cm}$ (Lee 1964). The ionization coefficient is also strongly temperature dependent.

The secondary hole can also ionize and form additional electron–hole pairs. Since the hole mobility is less than the electron mobility, higher fields are required than for same electron ionization. For example, at an electric field of $2 \cdot 10^5\,\mathrm{V/cm}$ the ionization coefficient $\alpha_n \approx 2 \cdot 10^5\,\mathrm{cm}^{-1}$ for electrons and $\alpha_p \approx 80$ for holes. Increasing the field by 25% increases α_n to $7 \cdot 10^3$ and α_p tenfold to 800.

Table 2.4 The breakdown gain depends strongly on the field in the avalanche region. The calculations are for a uniform field.

Field (V/cm)	Max. gain	Detector thickness (μm)	Bias voltage (V)
$2 \cdot 10^5$	$2.2 \cdot 10^3$	520	$1 \cdot 10^4$
$3 \cdot 10^5$	50	5	150
$4 \cdot 10^5$	6.5	0.5	20
$5 \cdot 10^5$	2.8	0.1	5

The higher probability of ionization by electrons is fortunate, as the formation of secondary holes leads to a positive feedback process. When the partial gain due to holes

$$G_p \geq 2 \tag{2.96}$$

the combined multiplication of electrons and holes leads to a sustained avalanche, *i.e.* breakdown. Only in silicon is the ratio of electron to hole ionization coefficients significantly greater than one, but it is field dependent and decreases with increasing field

$$\frac{\alpha_n}{\alpha_p} = 0.15 \cdot \exp\left(\frac{1.15 \cdot 10^6}{|E|}\right) \ . \tag{2.97}$$

This leads to the limits of gain and detector thickness *vs.* electric field shown in Table 2.4.

Lower fields allow higher gains, favored by the larger ratio α_n/α_p. High voltage operation is limited by local high-field regions and several techniques have been developed to limit radii of curvature and edge effects, for example (for an overview see Baliga 1987). Detectors should be operated well below breakdown. Operation at low fields reduces the sensitivity to variations in voltage and detector thickness, so an optimum design balances the length of the avalanche region and operating voltage for the desired gain.

The avalanche process also introduces noise, which is exacerbated by the positive feedback due to holes. The field profile must be optimized to achieve a given gain with a minimum hole component (McIntyre 1966, Webb *et al.* 1974, McIntyre 1999). The expressions given above are only valid in geometries with extended avalanche regions and must be modified for peaked fields or short devices (McIntyre 1999, Yuan 1999).

The optimum structure of an avalanche photodiode (APD) is the reach-through structure shown schematically in Figure 2.35 (Webb *et al.* 1974) or its variant, the reverse reach-through APD (McIntyre *et al.* 1996). Photons incident on the p^+-contact form electron–hole pairs. Since the absorption layer is very thin, the primary holes are collected with practically zero probability of avalanching. The electrons drift to the localized high-field region established

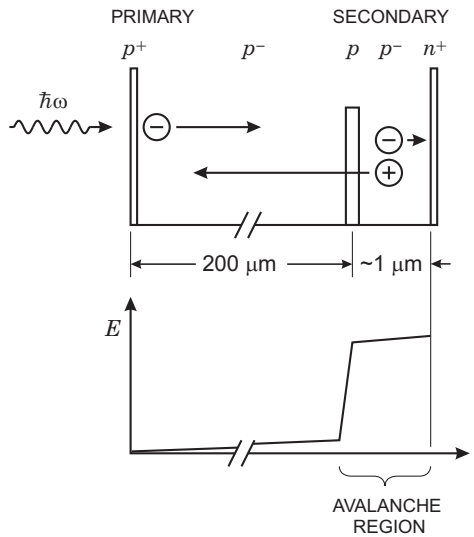

FIG. 2.35. Schematic doping structure of a reach through APD and the field distribution. Incident photons form primary signal electrons, which drift into the high-field region and release secondary electrons and holes.

by the internal p-layer. Here secondary electrons and holes are created by impact ionization. The secondary holes drift through the low-field region to the p^+-contact and contribute the major portion of the total induced charge.

This structure has several advantages over other types of APDs:

1. Only those primary carriers with the higher ionization coefficient are transferred to the avalanche region, as is desired for low avalanche noise and good stability.
2. The field in the avalanche region is primarily determined by the charge in the intermediate p-layer. This can be relatively unaffected by doping variations in the bulk material.
3. The avalanche field profile can be quite flat, even with realistic doping profiles. This allows a lower field for a given gain, which further improves stability.
4. The detector can be made much thicker than the avalanche region to reduce capacitance.

As shown in eqn 2.94 the gain is an exponential function of the width of the avalanche region times the ionization coefficient. The ionization coefficient in turn is an exponential function of field. Thus, both the width and the field must be precisely controlled over the full area of the device to obtain good gain uniformity.

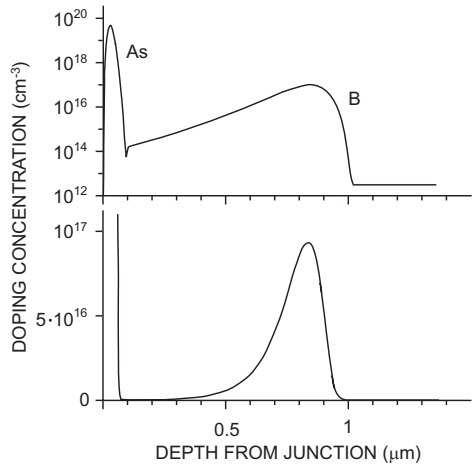

FIG. 2.36. Simulated net doping profile of the avalanche region of a reach-through APD. A shallow arsenic implant forms the *pn*-junction with the *p*-type bulk. A deep boron implant sets the boundary of the high-field region. Note that the orientation is reversed with respect to Figure 2.35

Figure 2.36 shows a practical doping profile, calculated with a full process simulator. The intermediate *p*-layer is formed by ion implantation of boron at 400 keV. This yields the desired flat field profile; to the left of the doping peak the field is "flat" to 10%. At the depletion voltage of $\sim$ 500 V the calculated gain is 35. If the doping is uniform to 0.5%, the variation in gain $\Delta G/G = 10\%$ at $G = 10$ and $\approx 35\%$ at a gain of 50. This is not presented as an "optimum" design, but to illustrate the demands on process control. In addition the strong dependence of gain on temperature and bias voltage place stringent demands on large systems. Nevertheless, large APD arrays are used successfully in high energy physics for the readout of scintillating tiles in calorimeters. For a summary of silicon photodiode design and technology see Webb *et al.* (1974).

The development of photodetectors is a very active field and many device structures have been developed. One example is the VLPC (Petroff 1987, 1989) used in the fiber tracker of *D*∅ (Wayne *et al.* 1997). These devices provide single photon sensitivity, but must operate at cryogenic temperatures (typically 5 – 7 K). Another class of devices utilizes multilayer structures of compound semiconductors utilizing bandgap engineering with graded doping profiles to yield low noise amplification at high gains ("staircase" photomultiplier, Capasso 1983). However, the fabrication of these devices is much more complicated (and costly) than of the silicon devices described above. Complexity notwithstanding, heterojunction superlattice structures offer many very interesting possibilities and one should monitor future developments. The desire to replace photomultiplier tubes by APDs does not require that they have the same gain. Since low-noise

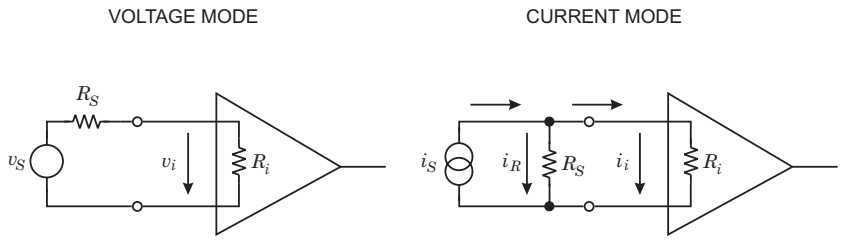

FIG. 2.37. Amplifiers can operate either in voltage mode (left) or current mode (right), depending on the ratio of amplifier input resistance R_I to source resistance R_S.

high-bandwidth amplifiers are inexpensive, moderate gains of order 10 – 100 are often sufficient to achieve the required signal level.

2.9 Signal acquisition

The preceding analysis yielded the current pulse shapes provided by the sensor. However, to be utilized the pulses must be amplified. Several types of amplifiers can be used, all of which affect the measured pulse shape. To illustrate the different modes, we'll first consider resistive sources.

2.9.1 *Voltage-sensitive amplifier*

A voltage-sensitive amplifier is designed to minimize loss of signal voltage at the amplifier input. The equivalent circuit is shown in Figure 2.37 (left). The voltage generator has zero source resistance, so the series resistor R_S represents the actual source resistance. The signal voltage at the amplifier input

$$v_i = \frac{R_i}{R_S + R_i} v_S \ . \tag{2.98}$$

If the signal voltage at the amplifier input is to be approximately equal to the signal voltage $v_i \approx v_S$, the input resistance $R_i \gg R_S$. To operate in the voltage-sensitive mode, the amplifier's input resistance (or impedance) must be large compared to the source resistance, or in general, the source impedance.

In ideal voltage amplifiers one sets $R_i = \infty$. Although this is never true in reality, it can be fulfilled to a good approximation. To provide a voltage output, the amplifier should have a low output resistance, *i.e.* its output resistance should be small compared to the input resistance of the following stage.

2.9.2 *Current-sensitive amplifier*

The right-hand panel of Figure 2.37 shows the equivalent circuit of a current-sensitive amplifier. The signal source is represented by a current generator, which by definition has infinite source resistance. The finite source resistance of a real source is represented by the shunt resistance R_S. The signal current divides into

the source resistance and the amplifier's input resistance. The fraction of current flowing into the amplifier

$$i_i = \frac{R_s}{R_s + R_i} i_S \ . \tag{2.99}$$

If the current flowing into the amplifier is to be approximately equal to the signal current $i_i \approx i_S$, then $R_i \ll R_S$. To operate in the current-sensitive mode, the amplifier's input resistance (or impedance) must be small compared to the source resistance (impedance).

One can also model a current source as a voltage source with a series resistance. For the signal current to be unaffected by the amplifier input resistance, the input resistance must be small compared to the source resistance, as derived above. At the output, to provide current drive the output resistance should be high, *i.e.* large compared to the input resistance of the next stage. This corresponds to the condition found for the input impedance; for optimum current transfer the source resistance must be large compared to the load.

- Whether a specific amplifier operates in the current or voltage mode depends on the ratio of source resistance to amplifier input resistance.
- Amplifiers can be configured as current mode input and voltage mode output or, conversely, as voltage mode input and current mode output. The gain is then expressed as V/A (transresistance) or A/V (transconductance).
- Since the mode of operation depends only on the ratio of source to input resistance, a voltage amplifier can function in either voltage or current mode, depending on the source resistance. In current mode the voltage at the amplifier input is $v_i = i_i R_i$ and the output voltage $v_o = A_v v_i$, where A_v is the voltage gain.

2.9.3 *Voltage and current mode with capacitive sources*

The preceding examples used resistive sources to illustrate the criteria for voltage and current mode amplification. A similar reasoning can be applied to capacitive sources, as appropriate for detectors. Figure 2.38 shows the equivalent circuit. The sensor signal is a current pulse of magnitude i_s and duration t_c , so the signal charge $Q_s = \int i_s(t)dt = i_s t_c$. As in the preceding examples, we consider the voltage gain of the amplifier, so that the output voltage

$$v_o = \text{voltage gain } A_v \times \text{input voltage } v_s \ .$$

Whether the amplifier operates in current or voltage mode depends on the charge collection time t_c of the sensor and the input time constant $R_i C_d$:

1. $R_i C_d \ll t_c$: The sensor capacitance discharges rapidly, so the output voltage is proportional to the instantaneous current $v_o \propto i_s(t)$. The combined sensor–amplifier system is operating in current mode.

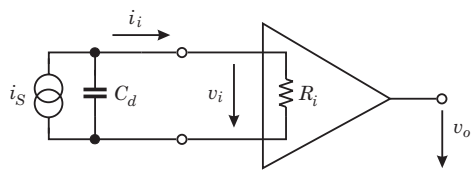

FIG. 2.38. With a capacitive source current or voltage mode are determined by the input time constant $\tau_i = R_i C_d$.

2. $R_i C_d \gg t_c$: The detector capacitance discharges slowly, so the signal current induced in the sensor is initially integrated on the sensor capacitance before discharging through the input resistance. The output voltage $v_o = V_o \exp(-t/R_i C_d)$, where $V_o = Q_s/C_d \propto \int i_s(t)dt$. In this case the signal manifests itself as a voltage developed across the sensor, so the system is operating in voltage mode.

In both cases the amplifier is providing voltage gain, so the output signal voltage is directly proportional to the input voltage. The difference is that the shape of the input voltage pulse is determined either by the instantaneous current (current mode) or by the integrated current and the decay time constant (voltage mode). This is discussed further in Chapter 3.

2.9.4 *Feedback amplifiers – the "charge-sensitive amplifier"*

Feedback can be used in amplifiers to control the gain and input resistance, as described in Appendix D. However, feedback can also be used to perform special functions. A very useful configuration for sensor readout is the charge-sensitive amplifier, shown in Figure 2.39.

The basic building block is an inverting voltage amplifier with a high input resistance. For simplicity assume an infinite input resistance, so that no signal current can flow into the amplifier. Since the amplifier inverts, the voltage gain $dv_o/dv_i = -A$, so $v_o = -Av_i$. A feedback capacitor C_f is connected from the output to the input. If an input signal produces a voltage v_i at the amplifier input, the voltage at the amplifier output is $-Av_i$. Thus, the voltage difference

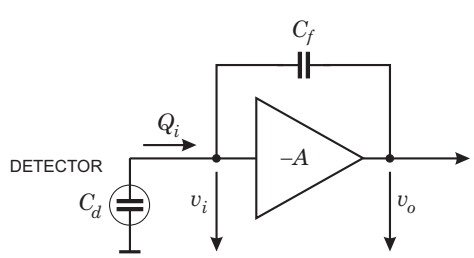

FIG. 2.39. Principle of a charge-sensitive amplifier

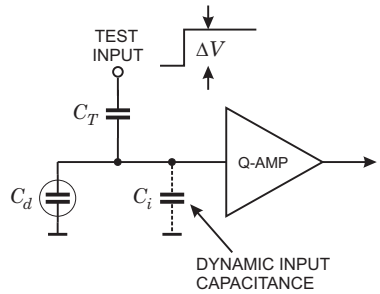

FIG. 2.40. Adding a test input to a charge-sensitive amplifier provides a simple means of absolute charge calibration.

across the feedback capacitor $v_f = (A+1)v_i$ and the charge deposited on C_f is $Q_f = C_f v_f = C_f(A+1)v_i$. Since no current can flow into the amplifier, all of the signal current must charge up the feedback capacitance, so $Q_f = Q_i$. The amplifier input appears as a "dynamic" input capacitance

$$C_i = \frac{Q_i}{v_i} = C_f(A+1) \ . \tag{2.100}$$

The enhanced input capacitance corresponds to a reduction in input impedance $1/\omega C_i$, as expected for a shunt feedback amplifier (see Appendix D).

The voltage output per unit input charge

$$A_Q = \frac{v_o}{Q_i} = \frac{Av_i}{C_i v_i} = \frac{A}{C_i} = \frac{A}{A+1} \cdot \frac{1}{C_f} \approx \frac{1}{C_f} \quad (A \gg 1) \ , \tag{2.101}$$

so the charge gain is determined by a well-controlled component, the feedback capacitor.

The signal charge Q_S will be distributed between the sensor capacitance C_d and the dynamic input capacitance C_i. The ratio of measured charge to signal charge

$$\frac{Q_i}{Q_s} = \frac{Q_i}{Q_d + Q_{s,amp}} = \frac{C_i}{C_d + C_i} = \frac{1}{1 + \dfrac{C_d}{C_i}} \ , \tag{2.102}$$

so the dynamic input capacitance must be large compared to the sensor capacitance.

Another very useful feature of the integrating amplifier is the ease of charge calibration. By adding a test capacitor as shown in Figure 2.40, a voltage step injects a well-defined charge into the input node. If the dynamic input capacitance C_i is much larger than the test capacitance C_T , the voltage step ΔV at the test input will be applied nearly completely across the test capacitance C_T, thus injecting a charge $C_T \Delta V$ into the input. More precisely, the injected charge

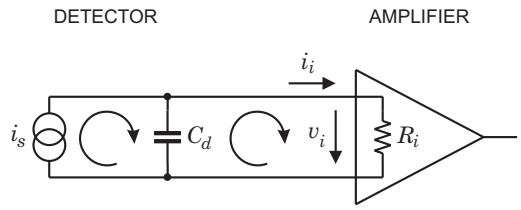

FIG. 2.41. Charge integration in a realistic charge-sensitive amplifier. First, charge is integrated on the sensor capacitance and subsequently transferred to the charge-sensitive loop, as it becomes active.

$$Q_T = \frac{C_T}{1 + \dfrac{C_T}{C_i + C_d}} \cdot \Delta V \approx C_T \left(1 - \frac{C_T}{C_i + C_d}\right) \Delta V \ , \tag{2.103}$$

so for the best accuracy the system should be calibrated with the detector connected.

2.9.5 *Realistic charge-sensitive amplifiers*

The preceding discussion assumed that the amplifiers are infinitely fast, that is that they respond instantaneously to the applied signal. In reality this is not the case; charge-sensitive amplifiers often respond much more slowly than the time duration of the current pulse from the sensor. However, as shown in Figure 2.41, this does not obviate the basic principle. Initially, signal charge is integrated on the sensor capacitance, as indicated by the left-hand current loop. Subsequently, as the amplifier responds the signal charge is transferred to the amplifier. Thus, the signal charge is preserved and the full signal appears at the amplifier output, even if the amplifier is much slower than the collection time.

Nevertheless, the time response of the amplifier does affect the measured pulse shape. First, consider a simple amplifier as shown in Figure 2.42.

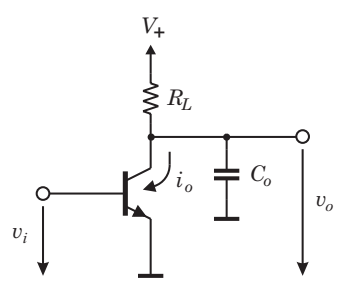

FIG. 2.42. A simple amplifier demonstrating the general features of any single-stage gain stage, whether it uses a bipolar transistor (shown) or an FET.

The gain element shown is a bipolar transistor, but it could also be a field effect transistor (JFET or MOSFET) or even a vacuum tube. Transistors are discussed in Chapter 6, but here it suffices to know that the transistor's output current changes as the input voltage is varied. For small signals the current i_o increases proportionally to the input voltage v_i, so the output voltage v_o decreases when the input voltage increases. Thus, the amplifier inverts and the voltage gain

$$A_v = -\frac{dv_o}{dv_i} = -\frac{di_o}{dv_i} \cdot Z_L \equiv -g_m Z_L \ , \tag{2.104}$$

where the minus sign indicates inverting gain. The parameter g_m is the transconductance, a key parameter that determines gain, bandwidth and noise of transistors. The load impedance Z_L is the parallel combination of the load resistance R_L and the output capacitance C_o. This capacitance is unavoidable; every gain device has an output capacitance, the following stage has an input capacitance, and in addition the connections and additional components introduce stray capacitance. The load impedance is given by

$$\frac{1}{Z_L} = \frac{1}{R_L} + \mathbf{i}\omega C_o \ , \tag{2.105}$$

where the imaginary $\mathbf{i}$ indicates the phase shift associated with the capacitance (an explanation of this notation is given in Appendix B). The voltage gain

$$A_v = -g_m \left(\frac{1}{R_L} + \mathbf{i}\omega C_o \right)^{-1} \ . \tag{2.106}$$

Figure 2.43 shows the frequency and phase response of a representative amplifier. Since amplifiers with a single cutoff frequency f_u show the same frequency response whether inverting or noninverting, the gain and phase are shown for a noninverting amplifier, for which the low-frequency phase shift is zero. The phase response shows the *change* in phase from the low-frequency response, so for an inverting amplifier as in Figure 2.42, the low-frequency phase shift is 180° and at high frequencies the phase shifts from 180° to 90°.

At low frequencies where the second term of eqn 2.106 is negligible, the gain is constant $A_v = -g_m R_L$. However, at high frequencies the second term dominates and the gain falls off linearly in frequency with an additional 90° phase shift. Figure 2.44 shows the frequency response in a simplified form, again for a noninverting amplifier. For an inverting amplifier a minus sign is applied. This "shorthand" representation is useful when considering the asymptotic response.

The cutoff frequency, where the asymptotic low and high frequency responses intersect, is determined by the output time constant $R_L C_o$, so the cutoff frequency

$$f_u = \frac{1}{2\pi R_L C_o} \ . \tag{2.107}$$

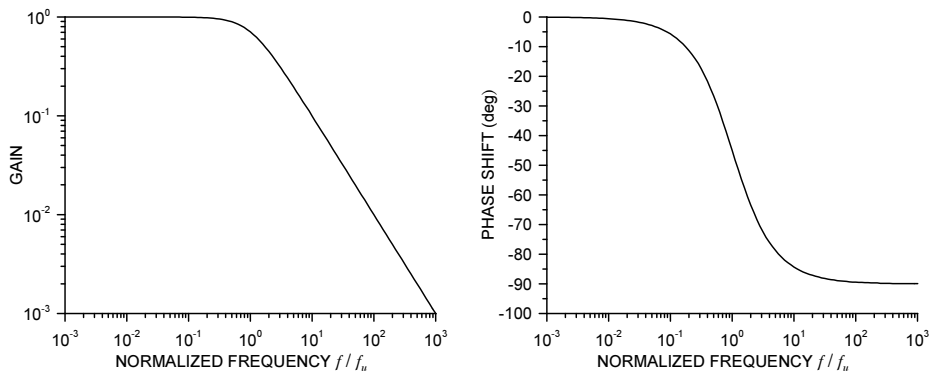

FIG. 2.43. Frequency (left) and phase (right) response of a "single pole" amplifier as shown in Figure 2.42. At the upper cutoff frequency f_u the gain drops off by $1/\sqrt{2}$ and the phase shift is 45°. The right-hand plot shows the change in phase, relative to the low-frequency response, so for an inverting amplifier the phase changes from 180 to 90°.

In the regime where the gain drops linearly with frequency the product of gain and frequency is constant, so the amplifier can be characterized by its gain–bandwidth product, which is equal to the frequency where the gain is one, the unity gain frequency f_0.

The amplifier gain–bandwidth product is independent of the low frequency gain, as increasing the load resistance R_L increases the gain, while decreasing the cutoff frequency. The product of low-frequency gain and the cutoff frequency

$$|A_v|\, f_u = f_0 = g_m R_L \frac{1}{2\pi R_L C_o} = \frac{1}{2\pi} \cdot \frac{g_m}{C_o} \tag{2.108}$$

depends only on the transconductance and the load capacitance. Increasing the gain–bandwidth product, *i.e.* increasing transconductance and reducing the load capacitance, raises the obtainable gain at any frequency $> f_u$.

As apparent from Figure 2.43 a logarithmic plot of gain *vs.* frequency is quite simple, so gain is commonly expressed in decibels (dB)

$$A_{dB} = 20 \log_{10} |A_v| \ . \tag{2.109}$$

Strictly formulated this should include the input and load resistances R_i and R_L, as the decibel is defined in terms of power

$$A_{dB} = 10 \log_{10} \frac{P_o}{P_i} = 10 \log_{10} \left(\frac{{v_o}^2}{{v_i}^2} \frac{R_i}{R_L} \right) = 20 \log_{10} |A_v| + 10 \log_{10} \frac{R_i}{R_L} \ .$$

The resistance term is commonly ignored when specifying amplifier voltage gains, so one must bear in mind that eqn 2.109 does not translate directly into a power gain.

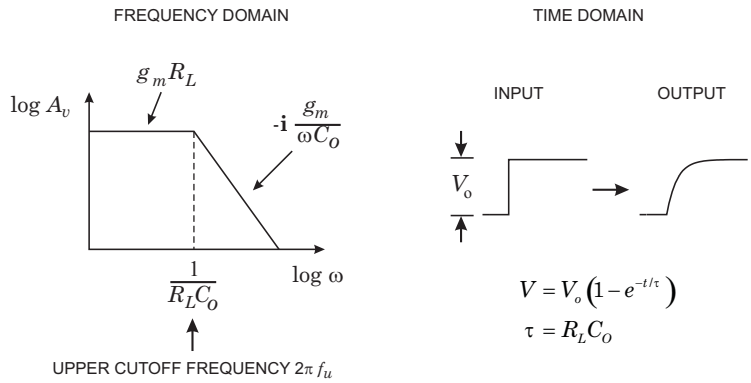

FIG. 2.44. The time constants of an amplifier affect both the frequency and the time response. Both are fully equivalent representations.

The frequency response translates into a time response. If a voltage step is applied to the input of the amplifier, the output does not respond instantaneously, as the output capacitance must first charge up. This is shown in the second panel of Figure 2.44, which illustrates that the response of an amplifier can be described in either the frequency or time domain. The response to a signal step of magnitude V_0 is

$$v_o(t) = V_0\left(1 - e^{-t/\tau}\right) \ , \tag{2.110}$$

where the time constant $\tau = R_L C_o$. Both representations are fully equivalent, but the choice of parameter space can greatly simplify analyses, so they should be considered in concert.

In practice, amplifiers utilize multiple stages, all of which contribute to the frequency response. Then additional corner frequencies appear, as illustrated in Figure 2.45, which shows two corner frequencies. At low frequencies the additional phase shift is zero, becoming 90° above the first corner frequency f_{u1}. Beyond the second corner frequency f_{u2} the phase shift becomes 180°, so what is an inverting amplifier at low frequencies becomes noninverting. This is critical when applying feedback. If the magnitude of the in-phase frequency components is too high, the system will oscillate. In Figure 2.39 the fedback signal is attenuated by the capacitive divider

$$\frac{X(C_d)}{X(C_d) + X(C_f)} = \frac{1/\omega C_d}{(1/\omega C_d) + (1/\omega C_f)} = \frac{C_f}{C_d + C_f} \approx \frac{C_f}{C_d} \ . \tag{2.111}$$

If at frequencies in the 180° phase shift regime the forward gain of the amplifier $A(f)$ is greater than the attenuation of the feedback network, then the amplifier will oscillate. This puts a limit on the amount of feedback, which for a given amplifier and feedback capacitance C_f depends on the detector capacitance C_d. As a consequence, an amplifier that is stable when operating with a detector

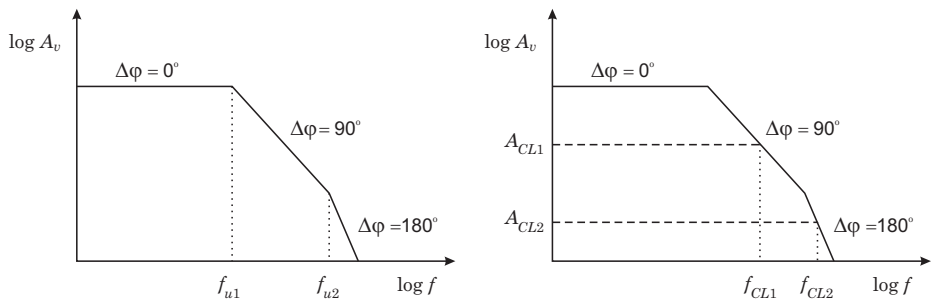

FIG. 2.45. Multiple gain stages introduce additional corner frequencies and additional phase shift. This plot shows two corner frequencies. At low frequencies the additional phase shift is zero, becoming 90° above the first corner frequency f_{u1} and 180° beyond the second corner frequency f_{u2}. Applying feedback sets the gain and corner frequency as shown in the second panel for two values of fedback gain A_{CL1} and A_{CL2}. The former is stable, whereas the latter is unstable.

can break into oscillation when the detector is disconnected, because of a faulty wire-bond, for example.

Practical amplifiers suitable for fedback systems always have multiple corner frequencies, which are commonly called "poles", since they appear as singularities in the Laplace transforms commonly used in feedback analysis. As a rule of thumb, a phase margin of 45° (an additional phase shift of 135°) ensures stable operation. Since the additional phase shift introduced by each pole is 45° at the corner frequency (see Figure 2.43), stable operation obtains when

$$\frac{C_d + C_f}{C_f} > |A|\,(f_{u2}) \ . \tag{2.112}$$

Figure 2.45 also shows two values of closed loop gain A_{CL1} and A_{CL2}. The former attains a corner frequency within the 90° phase shift regime, so it results in stable operation. Increasing feedback to yield A_{CL2} moves the corner frequency into the 180° phase shift regime, so the amplifier is unstable.

A clear discussion of feedback theory with more precise stability criteria was given by Nyquist (1932). There is a well-defined relationship between the slope of the gain *vs.* frequency and the phase shift (Bode 1940), so frequency-dependent gain and phase shift are coupled. Feedback amplifiers are discussed in numerous more recent texts (for example, Gray 2001, Mancini 2003) and summarized in Appendix D. Feedback amplifiers are designed with a dominant pole at a rather low frequency and all other poles at much higher frequencies, to ensure that the forward gain of the amplifier is low in the regime of critical phase shifts.

2.9.6 *Input impedance of a charge-sensitive amplifier*

We can now use the frequency response to calculate the input impedance and time response of a charge-sensitive amplifier. Applying the same reasoning as in Section 2.9.4 (and derived in Appendix D), the input impedance of a shunt feedback amplifier with gain A_v and a generalized feedback impedance Z_f is

$$Z_i = \frac{Z_f}{1 - A_v} \approx -\frac{Z_f}{A_v} \quad (|A_v| \gg 1) \ . \tag{2.113}$$

Since the amplifier inverts, A_v is negative. At low frequencies the gain is constant with no additional phase shift, so the input impedance is of the same nature as the feedback impedance, but reduced by $1/|A_v|$. At high frequencies well beyond the amplifier's cutoff frequency f_u the gain drops linearly with frequency with an additional 90° phase shift, so the gain (eqn 2.106) can be expressed as

$$A_v = \mathbf{i}\frac{\omega_0}{\omega} \ . \tag{2.114}$$

In a charge-sensitive amplifier the feedback impedance

$$Z_f = -\mathbf{i}\frac{1}{\omega C_f} \ , \tag{2.115}$$

so the input impedance eqn 2.113 becomes

$$Z_i \approx -\frac{Z_f}{A_v} = -\frac{-\mathbf{i}/(\omega C_f)}{\mathbf{i}(\omega_0/\omega)} = \frac{1}{\omega_0 C_f} \ . \tag{2.116}$$

The imaginary component vanishes, so the input impedance is real. In other words, it appears as a resistance R_i. Thus, at low frequencies $f \ll f_u$ the input of a charge-sensitive amplifier appears capacitive, whereas at high frequencies $f \gg f_u$ it appears resistive.

Suitable amplifiers invariably have corner frequencies well below the frequencies of interest for radiation detectors, so the input impedance is resistive. This allows a simple calculation of the time response. The sensor capacitance is discharged by the resistive input impedance of the fedback amplifier with the time constant

$$\tau_i = R_i C_d = \frac{1}{\omega_0 C_f} \cdot C_d \ . \tag{2.117}$$

From this we see that the rise time of the charge-sensitive amplifier increases with sensor capacitance. For reasons that will become apparent later, the feedback capacitance should be much smaller than the sensor capacitance. If $C_f = C_d/100$, the amplifier's gain–bandwidth product must be $100/\tau_i$, so for a rise time constant of 10 ns the gain–bandwidth product must be 10^{10} radians = 1.6 GHz. The same result can be obtained using conventional operational amplifier feedback theory.

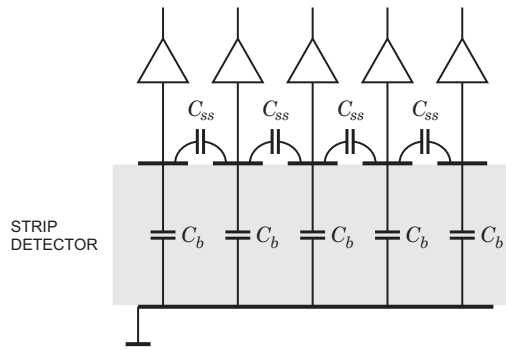

FIG. 2.46. To preserve the position resolution of strip detectors the readout amplifiers must have a low input impedance to prevent spreading of signal charge to the neighboring electrodes.

In acquiring the signal the ohmic input impedance at high frequencies and the dynamic input capacitance at low frequencies play complementary roles. In a simple approximation the charge transferred to the amplifier

$$Q(t) \approx \frac{Q_s}{1 + \dfrac{C_d}{C_i}} \left(1 - e^{-t/\tau_i}\right) \ . \tag{2.118}$$

The input time constant $\tau_i = R_i C_d$ determines how quickly charge is transferred from the detector and the dynamic input capacitance C_i determines what fraction of the signal charge Q_s can be acquired.

The mechanism of reducing the input impedance through shunt feedback leads to the concept of the "virtual ground". If the gain is infinite, the input impedance is zero. Although very high gains (of order 10^5 – 10^6) are achievable in the kHz range, at the frequencies relevant for detector signals the gain is much smaller. The input impedance of typical charge-sensitive amplifiers in strip detector systems is of order kΩ. Fast amplifiers designed to optimize power dissipation achieve input impedances of $100 - 500\,\Omega$ (Kipnis 1994). None of these qualify as a "virtual ground", so this concept should be applied with caution.

Apart from determining the signal rise time, the input impedance is critical in position-sensitive detectors. Figure 2.46 illustrates a silicon-strip sensor read out by a bank of amplifiers. Each strip electrode has a capacitance C_b to the backplane and a fringing capacitance C_{ss} to the neighboring strips. If the amplifier has an infinite input impedance, charge induced on one strip will capacitively couple to the neighbors and the signal will be distributed over many strips (determined by C_{ss}/C_b). If, on the other hand, the input impedance of the amplifier is low compared to interstrip impedance, practically all of the charge will flow into the amplifier and the neighbors will show only a small signal. The relevant frequencies are set by the peaking time T_P of the pulse shaper.

For strip pitches that are much smaller than the bulk thickness the capacitance is dominated by the fringing capacitance to the neighboring strips C_{ss}. Typically, the fringing capacitance is $1-2\,\mathrm{pF/cm}$ for strip pitches of $25-100\,\mu\mathrm{m}$ on Si. The backplane capacitance C_b is typically 20% of the strip-to-strip capacitance. If the real part of the amplifier's input impedance is R_i, cross-coupling will be negligible at times greater than 2 to 3 times R_iC_{ss} and if $C_i \gg C_{ss}$.

References

Alkhazov, G.D. *et al.* (1967). Ionization fluctuations and resolution of ionization chambers and semiconductor detectors. *Nucl. Instr. and Meth.* **48** (1967) 1–12

Amman, M. and Luke, P.N. (1999). Optimization criteria for coplanar-grid detectors. *IEEE Trans. Nucl. Sci.* **NS-46/3** 205–212

Avset, B.S. *et al.* (1990). A new microstrip detector with double-sided readout. *IEEE Trans. Nucl. Sci.* **37/3** (1990) 1153–1161

Baliga, B. Jayant (1987). *Modern Power Devices.* Chapter 3. Wiley, New York. ISBN 0-471-81986-7, TK7881.15.B35

Beadle, W.C. , Tsai, J.C.C., and Plummer, R.D. (1984). *Quick Reference Manual for Silicon Integrated Circuit Technology.* Wiley-Interscience, New York, ISBN 0-471-81588-8, TK7874.Q38 1984

Bode, H.W. (1940). Relations between attenuation and phase in feedback amplifier design. *Bell System Tech. Journal* **19** (1940) 421–454

Cavalleri, G. *et al.* (1971). Extension of Ramo's theorem to induced charge in semiconductor detectors. *Nucl. Instr. and Meth.* **92** (1971) 137–140

Chabaud, V. *et al.* (1996). The DELPHI silicon strip microvertex detector with double sided readout. *Nucl. Instr. and Meth.* **A368** (1996) 314–332

Conradi, J. (1972). The distribution of gains in uniformly multiplying avalanche photodiodes: experimental. *IEEE Trans. Electron Dev.* **ED-19/6** (1972) 713–718

Edwards, A.J. *et al.* (2004). Radiation monitoring with diamond sensors in BaBar. *IEEE Trans. Nucl. Sci.* **NS51/4** (2004) 1808–1811

Fano, U. (1947). Ionization yield of radiations. II. The fluctuations of the number of ions. *Phys. Rev.* **72** (1947) 26–29

Gray, Paul R. *et al.* (2001). *Analysis and Design of Analog Integrated Circuits.* (4th edn). Wiley, New York, ISBN 0-471-32168-0

Haller, E.E., Hansen, W.L., and Goulding, F.S. (1981). Physics of ultra-pure germanium. *Advances in Physics* **30(1)** (1981) 93–138

Hecht, K. (1932). Zum Mechanismus des lichtelektrischen Primärstroms in isolierenden Kristallen. *Z. Physik* **77** (1932) 235–245

Kittel, C. (1996). *Introduction to Solid State Physics.* Wiley, New York, ISBN 0-471-11181-3, QC176.K5 1996.

Klein, C.A. (1968). Bandgap dependence and related features of radiation ionization energies in semiconductors. *J. Applied Physics* **39** (1968) 2029–2038

Lari, T. *et al.* (2005). Characterization and modeling of non-uniform charge collection in CVD diamond pixel detectors. *Nucl. Instr. and Meth.* **A537** (2005) 581–593

Lee, C.A. *et al.* (1964) Ionization rates of holes and electrons in silicon. *Phys. Rev.* **134/3A** (1964) A761–A773

Luke, P.N. (1994). Single-polarity charge sensing in ionization detectors using coplanar electrodes. *Appl. Phys. Lett.* **65** (1994) 2884–2886

Luke, P.N. (1995). Unipolar charge sensing with coplanar electrodes – Application to Semiconductor Detectors. *IEEE Trans. Nucl. Sci.* **NS-42/4** (1995) 207–213

Mancini, R. (2003). *Op Amps for Everyone* (2nd edn), Newnes, Amsterdam, ISBN 0-750-67701-5

McIntyre, R.J. (1966). Multiplication noise in uniform avalanche diodes. *IEEE Trans. Electron Dev.* **ED-13/1** (1966) 164–168

McIntyre, R.J. (1966). The distribution of gains in uniformly multiplying avalanche photodiodes: theory. *IEEE Trans. Electron Dev.* **ED-19/6** (1972) 703–713

McIntyre, R.J. *et al.* (1996). A short-wavelength selective reach-through avalanche photodiode. *IEEE Trans. Nucl. Sci.* **NS-43/3** (1996) 1341–1346

McIntyre, R.J. (1999). A new look at impact ionization – Part I: A theory of gain, noise, breakdown probability, and frequency response. *IEEE Trans. Electron Dev.* **ED-46/8** (1999) 1623–1631

Nyquist, H. (1932). Regeneration theory. *Bell System Tech. Journal* **11** (1932) 126–147

Owens, A. *et al.*(1999). Synchrotron characterization of deep depletion epitaxial GaAs detectors. *J. Appl. Phys.* **86** (1999) 4341–4347

Owens, A. and Peacock, A. (2004). Compound semiconductor radiation detectors. *Nucl. Instr. and Meth.* **A531** (2004) 18–37

Petroff, M.D. *et al.* (1987). Detection of individual 0.4 – 28 μm wavelength photons via impurity-impact ionization in a solid-state photomultiplier. *Appl. Phys. Lett.* **51** (1987) 406–408

Petroff, M.D. and Stapelbroek, M.G. (1989). Photon-counting solid-state photomultiplier. *Nucl. Instr. and Meth.* **NS-36/1** (1989) 158–162

Queisser, H.J. and Haller, E.E. (1998). Defects in semiconductors: some fatal, some vital. *Science* **281** (1998) 945–950

Radeka, V. (1988). Low noise techniques in detectors. *Ann. Rev. Nucl. Part. Sci.* **38**(1988) 217–277

Ramo, S. (1939). Currents induced by electron motion. *Proc. IRE* **27** (1939) 584–585

Richter, R.H. *et al.* (1996). Strip detector design for ATLAS and HERA-B using two-dimensional device simulation. *Nucl. Instr. and Meth.* **A377** (1996) 412–421

Scholze, F. *et al.* (2000). Determination of the electron–hole pair creation energy for semiconductors from the spectral responsivity of photodiodes. *Nucl. Instr. and Meth.* **A439** (2000) 208–215

Shockley, W. (1938). Currents to conductors induced by a moving point charge. *J. Appl. Phys.* **9** (1938) 635–636

Shockley, W. (1950). *Electrons and Holes in Semiconductors*. van Nostrand, Princeton

Sze, S.M. (1981). *Physics of Semiconductor Devices* (2nd edn). Wiley, New York. ISBN 0-471-05661-8, TK7871.85.S988

Sze, S.M. (2001). *Semiconductor Devices – Physics and Technology* (2nd edn). Wiley, New York. ISBN 0-471-33372-7, TK7871.85.S9883

van Roosbroeck, W. (1963). Theory of the yield and Fano factor of electron–hole pairs generated in semiconductors by high-energy particles. *Phys. Rev.* **139** (1963) A1702

Wayne, M.R. *et al.* (1997). Visible light photon counters and the DØ scintillating fiber tracker. *Nucl. Instr. and Meth.* **A387** (1997) 278–281

Webb, P.P., McIntyre, R.J. and Conradi, J. (1972). Properties of avalanche photodiodes. *RCA Review* **25** (1974) 234–278

Yuan, P. *et al.* (1999). A new Look at impact ionization – Part II: Gain and noise in short avalanche photodiodes. *IEEE Trans. Electron Dev.* **ED-46/8** (1999) 1623–1631

3

ELECTRONIC NOISE

3.1 Electronic noise and resolution

Electronic noise places a lower bound on the detectable signal level and also determines the ability to distinguish signal levels or measure them precisely. Figure 3.1 shows two examples. The left panel compares gamma-ray spectra taken with a scintillator and a semiconductor detector. Where the NaI(Tl) scintillator shows broad bumps, the Ge detector resolves a multitude of discrete energy peaks. The right panel of Figure 3.1 shows how resolution affects sensitivity. Higher resolution, or smaller line width, distributes the signal counts over a narrower background range, so the signal becomes more pronounced.

As noted in Chapter 1, one objective of signal processing is to improve the signal-to-noise ratio by tailoring the spectral distributions of the signal and the

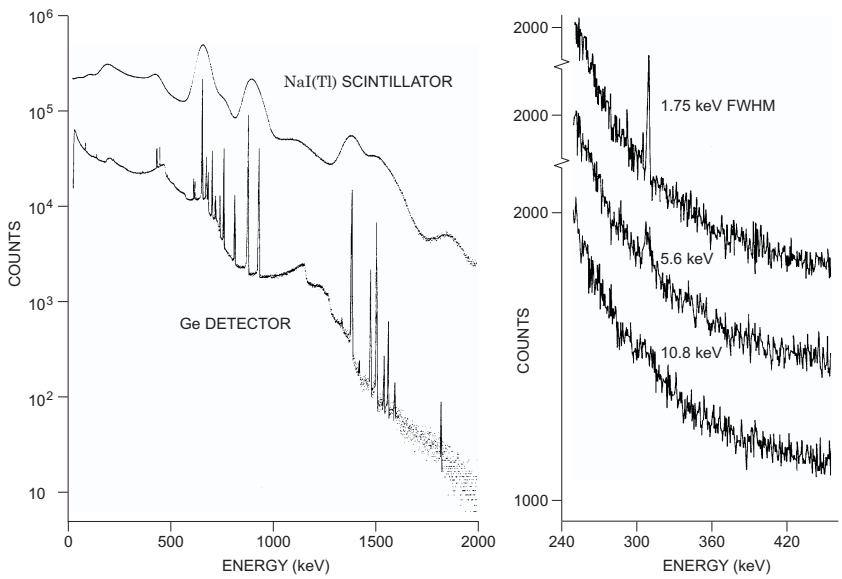

FIG. 3.1. The comparison of gamma-ray spectra taken with a scintillator and a semiconductor detector (left) clearly shows how improved resolution reveals detailed structure (adapted from Philippot 1970). Higher resolution also improves the signal-to-noise ratio (right) (adapted from Armantrout *et al.* 1972). Figures ©IEEE, reprinted with permission.

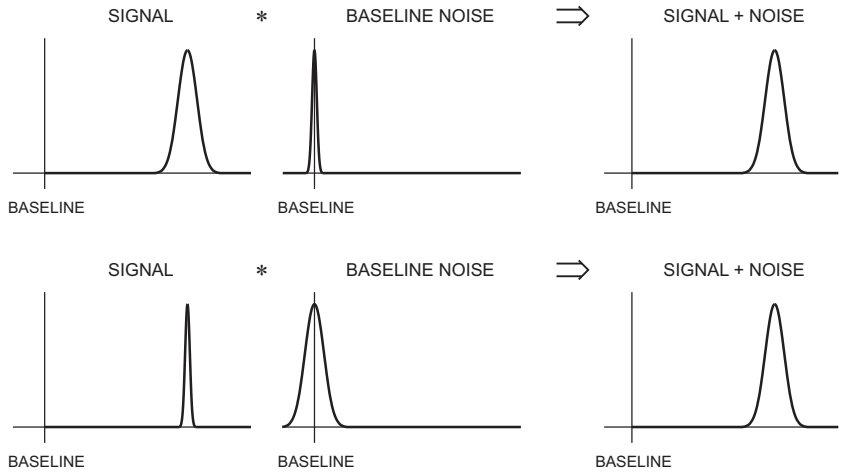

FIG. 3.2. Signal and baseline fluctuations add in quadrature. For large signal variance (top) as in scintillation detectors or proportional chambers the baseline noise is usually negligible, whereas for small signal variance as in semiconductor detectors or liquid Ar ionization chambers, baseline noise is critical.

electronic noise. However, in many detectors electronic noise does not determine the resolution. For example, in a NaI(Tl) scintillation detector measuring 511 keV gamma-rays, as in a positron-emission tomography system, 25 000 scintillation photons are produced. Because of reflective losses, about 15 000 reach the photocathode. This translates to about 3000 electrons reaching the first dynode. The gain of the electron multiplier will yield about $3 \cdot 10^9$ electrons at the anode. However, despite the magnitude of the output signal, its statistical spread is determined by the smallest number of electrons in the chain, *i.e.* the 3000 electrons reaching the first dynode. Thus, the resolution $\Delta E/E = 1/\sqrt{3000} = 2\%$, which at the anode corresponds to about $5 \cdot 10^4$ electrons. This is much larger than electronic noise in any reasonably designed system. This situation is illustrated in the upper panel of Figure 3.2 (top). Under these circumstances, signal acquisition and count rate capability may be the prime objectives of the pulse processing system.

The bottom panel of Figure 3.2 shows the situation for high resolution sensors providing small signals. Semiconductor detectors, photodiodes or ionization chambers are typical examples. Here baseline fluctuations dominate the overall resolution, so low noise is critical. The baseline fluctuations can have many origins, external interference, artifacts due to imperfect electronics, *etc.*, but the fundamental limit is electronic noise.

3.2 Electronic noise

Consider a current flowing through a sample bounded by two electrodes, *i.e.* n electrons moving with velocity v. The induced current depends on the spacing s between the electrodes (see "Ramo's theorem" in Chapter 2), so

$$i = \frac{nev}{s} \ . \tag{3.1}$$

The fluctuation of this current is given by the total differential

$$\langle di \rangle^2 = \left(\frac{ne}{s} \langle dv \rangle \right)^2 + \left(\frac{ev}{s} \langle dn \rangle \right)^2 \ , \tag{3.2}$$

where the two terms add in quadrature, as they are statistically uncorrelated. From this one sees that two mechanisms contribute to the total noise, velocity and number fluctuations.

To evaluate the overall effect of these noise sources in a measurement system with a frequency response $A(f)$ we need to know the spectral distribution of these various types of noise. However, before deriving their frequency distributions we'll simply summarize the results and discuss some general features.

Velocity fluctuations originate from thermal motion. Superimposed on the average drift velocity are random velocity fluctuations due to thermal excitation. This "thermal noise" is described by the long wavelength limit of Planck's black body spectrum where the spectral density, *i.e.* the power per unit bandwidth, is constant ("white" noise).

Number fluctuations occur in many circumstances. One source is carrier flow that is limited by emission over a potential barrier. Examples are thermionic emission or current flow in a semiconductor diode. The probability of a carrier crossing the barrier is independent of any other carrier being emitted, so the individual emissions are random and not correlated. In a reverse-biased diode the current is determined by statistically independent generation and recombination processes (Appendix F). This is called "shot noise", which also has a "white" spectrum. Another source of number fluctuations is carrier trapping. Impurities or imperfections in a crystal lattice can trap charge carriers and release them after a characteristic lifetime. As shown below, this leads to a frequency-dependent power spectrum $dP_n/df = 1/f^\alpha$, where α is typically in the range of $0.5 - 2$.

3.3 Some general properties of noise

Fundamentally, the spectral distribution of noise is described as a power density dP_n/df, or in other words, the power in a narrow slice of frequency space. However, in analyzing electronic noise we need to describe the noise in terms of voltage and current spectral densities $dv_n/\sqrt{df}$ and $di_n/\sqrt{df}$. In circuit design literature and data sheets these are commonly abbreviated as $dv_n/\sqrt{df} \equiv e_n$ and $di_n/\sqrt{df} \equiv i_n$, so we'll follow that convention. This does lead to inconsistencies; i_s might represent a signal current with the unit A, whereas i_n represents a

spectral density $\mathrm{A}/\sqrt{\mathrm{Hz}}$. In the following just bear in mind that e_n and i_n have this special connotation.

The total noise is obtained by integrating the noise power over the relevant frequency range of the system, the bandwidth. Since power is proportional to either the voltage or current squared, the output noise of an amplifier with a frequency-dependent gain $A(f)$ is

$$v_{no}^2 = \int_0^\infty e_n^2 A^2(f)df \quad \text{or} \quad i_{no}^2 = \int_0^\infty i_n^2 A^2(f)df \ . \tag{3.3}$$

The total noise v_{no} or i_{no} increases with the square root of bandwidth. Since small bandwidths correspond to large rise times, increasing the speed of a pulse measurement system will increase the noise. The effect of reducing bandwidth on signal-to-noise ratio is illustrated in Figure 3.3.

The amplitude distribution of the noise is Gaussian, so noise fluctuations superimposed on the signal also yield a Gaussian distribution (Fig 1.26). Thus, by measuring the width of the amplitude spectrum of a well-defined signal, one can determine the noise level.

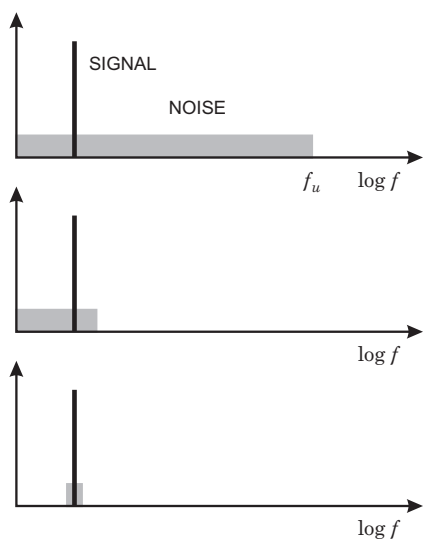

FIG. 3.3. The signal-to-noise ratio compares the signal power to the total noise power. If the system spans a large frequency range, as determined by the high-frequency cutoff f_u in an amplifier (top) the integrated power will include more noise than in the middle panel, which shows a lower cutoff frequency. Introducing a bandpass filter that is just wide enough to pass the signal components will optimize the signal-to-noise ratio (bottom).

First we'll describe the properties of various types of noise and then derive the power spectra. More details may be found in books by van der Ziel (1986), for a more theoretical treatment, and by Motchenbacher and Connelly (1993), who take a more practical approach.

3.3.1 *Thermal (Johnson) noise*

The most common example of noise due to velocity fluctuations is the noise of resistors. The spectral noise density *vs.* frequency

$$\frac{dP_n}{df} = 4kT \ , \tag{3.4}$$

where k is the Boltzmann constant and T the absolute temperature. Since the power in a resistance R is

$$P = \frac{v^2}{R} = i^2 R \ , \tag{3.5}$$

the spectral voltage noise density

$$\frac{dv_n^2}{df} \equiv e_n^2 = 4kTR \tag{3.6}$$

and the spectral current noise density

$$\frac{di_n^2}{df} \equiv i_n^2 = \frac{4kT}{R} \ . \tag{3.7}$$

3.3.2 *Shot noise*

The spectral noise density of shot noise

$$i_n^2 = 2eI \ , \tag{3.8}$$

where I is the average current and e the electronic charge. Note that the criterion for shot noise is that carriers are injected independently of one another, as in thermionic or semiconductor diodes. Current flow determined by an ohmic conductor ($I = V/R$) does not carry shot noise. Any local fluctuation in electron density relative to the stationary positive charge of the host atoms will set up an electric field that can easily draw in additional carriers to equalize the disturbance.

3.3.3 *Low frequency ("$1/f$") noise*

The noise spectrum becomes nonuniform whenever the fluctuations are not purely random in time, for example when carriers are trapped and then released with a time constant τ. With an infinite number of uniformly distributed time constants the spectral power density assumes a pure $1/f$ distribution. However, as shown below, with as few as three time constants spread over one or two

decades, the spectrum shows a nearly perfect $1/f$ distribution over a limited frequency range, so this form of noise is very common.

Assume a spectral power density $P_{nf} = S_f/f$ and a corresponding voltage density $e_{nf}^2 = A_f/f$. Then the total noise in a frequency band extending from f_1 to f_2 is

$$v_{nf}^2 = \int_{f_1}^{f_2} \frac{A_f}{f} df = A_f \log\left(\frac{f_2}{f_1}\right) \ . \tag{3.9}$$

Thus, for a $1/f$ spectrum the total noise depends on the ratio of the upper to lower cutoff frequencies, rather than the absolute bandwidth. Since the $1/f$ distribution refers to the power spectrum, the associated voltage or current spectral density changes ten-fold over a 100-fold span in frequency.

3.4 Derivation of spectral densities

3.4.1 *Spectral density of thermal noise*

Two approaches can be used to derive the spectral distribution of thermal noise:

1. The thermal velocity distribution of the charge carriers is used to calculate the time dependence of the induced current, which is then transformed into the frequency domain.
2. Application of Planck's theory of black-body radiation.

The first approach clearly shows the underlying physics, whereas the second "hides" the physics by applying a general result of statistical mechanics. However, the first requires mathematical tools that go well beyond the standard curriculum, so the "black-body" approach will be used here.

In Planck's theory of black body radiation the energy per mode

$$\overline{E} = \frac{hf}{e^{hf/kT} - 1} \tag{3.10}$$

and the spectral density of the radiated power

$$\frac{dP}{df} = \frac{hf}{e^{hf/kT} - 1} \ . \tag{3.11}$$

This is the power that can be extracted in equilibrium. At low frequencies $hf \ll kT$ the spectral power density

$$\frac{dP}{df} \approx \frac{hf}{\left(1 + \frac{hf}{kT}\right) - 1} = kT \ , \tag{3.12}$$

so the spectral density is independent of frequency and for a total bandwidth B the noise power that can be transferred to an external device is $P_n = kTB$.

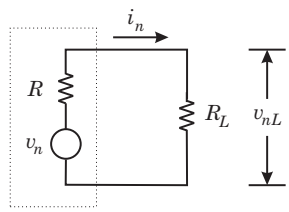

FIG. 3.4. Equivalent circuit to determine how much thermal noise power can be extracted from a resistor.

To apply this result to the noise of a resistor, consider a resistance R whose thermal noise gives rise to a noise voltage v_n. To determine the power transferred to an external device consider the circuit shown in Figure 3.4. The power dissipated in the load resistor R_L is

$$\frac{v_{nL}^2}{R_L} = i_n^2 R_L = \frac{v_n^2 R_L}{(R+R_L)^2} \ . \tag{3.13}$$

Maximum power transfers when the load resistance equals the source resistance $R_L = R$, so

$$v_{nL}^2 = \frac{v_n^2}{4} \ . \tag{3.14}$$

Since the power transferred to R_L is kTB,

$$\frac{v_{nL}^2}{R} = \frac{v_n^2}{4R} = kTB \ , \tag{3.15}$$

$$P_n = \frac{v_n^2}{R} = 4kTB \ . \tag{3.16}$$

and the spectral density of the noise power

$$\frac{dP_n}{df} = 4kT \ . \tag{3.17}$$

3.4.2 *Spectral density of shot noise*

When an excess electron is injected into a device, it forms a current pulse of duration t. In a thermionic diode t is the transit time from cathode to anode, for example. In a forward-biased semiconductor diode t is the recombination time (see Appendix E). If these times are short with respect to the periods of interest $t \ll 1/f$, the current pulse can be represented by a delta pulse. The Fourier transform of a delta pulse yields a "white" spectrum, *i.e.* the amplitude distribution in frequency is uniform

$$\frac{di_{n,pk}}{\sqrt{df}} = 2e \ . \tag{3.18}$$

Within an infinitesimally narrow frequency band each individual spectral component is a pure sinusoid, so its rms value

$$i_n \equiv \frac{di_n}{\sqrt{df}} = \frac{2e}{\sqrt{2}} = \sqrt{2}e \ . \tag{3.19}$$

If N electrons are emitted at the same average rate, but at different times, they will have the same spectral distribution, but the coefficients will differ in phase. For example, for two currents i_p and i_q with a relative phase φ the total rms current is given by

$$\left\langle i^2 \right\rangle = \left(i_p + i_q e^{i\varphi}\right)\left(i_p + i_q e^{-i\varphi}\right) = i_p^2 + i_q^2 + 2 i_p i_q \cos\varphi \ . \tag{3.20}$$

For a random phase the third term averages to zero and

$$\left\langle i^2 \right\rangle = i_p^2 + i_q^2 \ , \tag{3.21}$$

so if N electrons are randomly emitted per unit time, the individual spectral components simply add in quadrature

$$i_n^2 = 2Ne^2 \ . \tag{3.22}$$

The average current $I = Ne$, so the spectral noise density

$$i_n^2 \equiv \frac{di_n^2}{df} = 2eI \ . \tag{3.23}$$

This result can also be obtained by applying Carson's theorem (van der Ziel 1986). If a single pulse has the amplitude $A(t)$ and its Fourier transform

$$P(f) = \int_{-\infty}^{\infty} A(t) \exp(-\mathbf{i}\omega t) dt \ , \tag{3.24}$$

then a random sequence of pulses occurring at a rate r has the spectral power distribution

$$S(f) = 2r \left|P(f)\right|^2 \ . \tag{3.25}$$

Shot noise can be considered as a sequence of delta pulses, which have a white frequency spectrum, so the pulse sequence also has a white spectrum. Since the rate $r = I/e$ and the integral $P(f) = e$, the spectral density of shot noise

$$i_n^2 = 2eI \ , \tag{3.26}$$

as derived above in a less general manner. This derivation also demonstrates that a direct current formed by a random sequence of individual pulses retains the spectral distribution of the individual pulses. In other words, the spectral distribution of a DC signal carries information of the signal's origin.

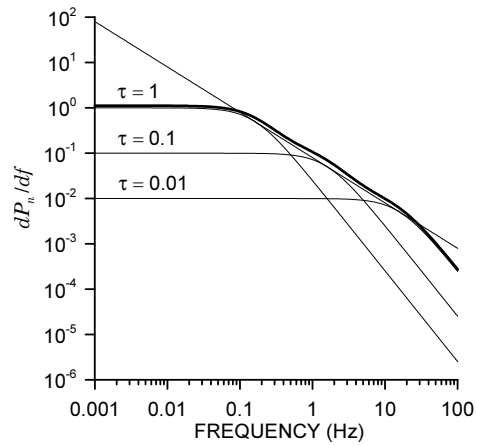

FIG. 3.5. Traps with three time constants of 0.01, 0.1 and 1 s yield a $1/f$ distribution over two decades in frequency, although individual traps introduce a $1/f^2$ spectrum. A line with a pure $1/f$ dependence is shown for comparison.

3.4.3 *Spectral density of low-frequency noise*

Assume a conductor where charge carriers get trapped and subsequently released. This corresponds to a change in conductance ΔG. These events occur independently in time in a random sequence. Because characteristic times are involved, the shot noise spectrum deviates from white noise in certain frequency ranges. First, let's consider that all traps have the same lifetime τ. Since this is a form of shot noise, it depends on the magnitude of the current I. The magnitude of the excess noise also depends on the conductance G and the number N of traps in the sample. The spectral density

$$i_{nf}^2 = 4NI^2 \left(\frac{\Delta G}{G}\right)^2 \frac{\tau}{1+(\omega\tau)^2} \ . \tag{3.27}$$

At frequencies well below $\omega = 2\pi f = 1/\tau$ the spectrum is white and at high frequencies well above the corner frequency the spectral density falls off with $1/f^2$. It is not apparent how this can yield a $1/f$ distribution.

However, it is rare that only a single time constant is involved. With just a few time constants spread over a decade or two, the envelope of the composite spectral density has a $1/f$ distribution over the corresponding frequency range. This is shown in Figure 3.5, where only three time constants distributed over two decades in time yield a nearly ideal $1/f$ response over two decades in frequency. By extension, an infinite number of time constants uniformly distributed yields a $1/f$ spectrum over the full frequency range. Figure 3.5 also illustrates why in practice $1/f$ spectra do not extend down to zero frequency. Another unrealistic assumption is that all traps occur at the same concentration. Clearly, reality

doesn't always follow this pattern, so low frequency noise spectra commonly show various regimes with $dP_n/df = 1/f^\alpha$, where α is typically in the range of 0.5 – 2. Also bear in mind that this is the power density, so when measuring voltage or current, the magnitude of $1/f$ noise increases ten-fold over *two* decades in frequency.

3.5 "Noiseless" resistances

3.5.1 *Dynamic resistances*

In many instances a resistance is formed by the slope of a device's current–voltage characteristic, rather than by a static ensemble of electrons agitated by thermal energy. A common example is a forward-biased semiconductor diode. The diode current *vs.* voltage (Appendix E)

$$I = I_0(e^{eV/kT} - 1) \ . \tag{3.28}$$

The differential resistance

$$r_d = \frac{dV}{dI} = \frac{kT}{eI} \ , \tag{3.29}$$

so at a given current the diode presents a resistance, *e.g.* $26\,\Omega$ at $I = 1\,\mathrm{mA}$ and $T = 300\,\mathrm{K}$.

Note that two diodes can have different charge carrier concentrations, but will still exhibit the same dynamic resistance at a given current, so the dynamic resistance is not uniquely determined by the number of carriers, as in a resistor. There is no thermal noise associated with this "dynamic" resistance, although the current flow carries shot noise.

3.5.2 *Active resistances*

In the previous chapter it was shown that the input of an amplifier with capacitive feedback can appear resistive. The input resistance is the result of amplifier gain, phase shift, and a feedback capacitor. The feedback capacitor is noiseless for all practical purposes, since we can ignore vacuum fluctuations. If the amplifier is noiseless (impossible in reality, but useful as a "thought experiment"), then the input will appear as a noiseless resistance. In practice, the amplifier noise determines the equivalent noise of the input resistance R_i, but this can be much smaller than $4kTR_i$, so it is equivalent to a "cooled" resistance (Radeka 1974). This technique can provide terminations for transmission lines without incurring the full thermal noise.

3.5.3 *Radiation resistance of an antenna*

Consider a receiving antenna with the normalized power pattern $P_n(\theta, \phi)$ pointing at a brightness distribution $B(\theta, \phi)$ in the sky. The power per unit bandwidth received by the antenna

$$\frac{dP}{df} = \frac{A_e}{2} \iint B(\theta, \phi) P_n(\theta, \phi) d\Omega \tag{3.30}$$

where A_e is the effective aperture, *i.e.* the "capture area" of the antenna. For a given field strength E, the captured power $P \propto EA_e$ (Kraus 1986).

If the brightness distribution is from a black-body radiator and we're measuring in the Rayleigh-Jeans regime,

$$B(\theta, \phi) = \frac{2kT}{\lambda^2} \tag{3.31}$$

and the power received by the antenna

$$\frac{dP}{df} = \frac{kT}{\lambda^2} A_e \Omega_A \ . \tag{3.32}$$

Ω_A is the beam solid angle of the antenna (measured in rad^2), *i.e.* the angle through which all the power would flow if the antenna pattern were uniform over its beamwidth. Since $A_e \Omega_A = \lambda^2$ (see antenna textbooks, for example Kraus 1988), the received power

$$\frac{dP}{df} = kT \ . \tag{3.33}$$

The received power is independent of the radiation resistance, as would be expected for thermal noise. However, it is not determined by the temperature of the antenna, but by the temperature of the sky the antenna pattern is subtending.

For example, in a region dominated by the cosmic microwave background, the measured power corresponds to a resistor at a temperature of $\sim 3\,\mathrm{K}$, though the antenna may be at $300\,\mathrm{K}$. For a more detailed discussion see Kraus (1986).

3.6 Correlated noise

Generally, noise power is additive:

$$P_{n,tot} = P_{n1} + P_{n1} + \ldots \tag{3.34}$$

However, in a coherent system (*i.e.* a system that preserves phase), the power often results from the sum of voltages or currents, which is sensitive to relative phase.

For two correlated noise sources N_1 and N_2 the total noise

$$N^2 = N_1^2 + N_2^2 + 2CN_1N_2 \ , \tag{3.35}$$

where the correlation coefficient C can range from -1 (anticorrelated, *i.e.* identical, but $180°$ out of phase) to $+1$ (fully correlated). For uncorrelated noise components $C = 0$ and then individual current or voltage noise contributions add in quadrature,

$$v_{n,tot} = \sqrt{\sum_i v_{ni}^2} \ . \tag{3.36}$$

3.7 Signal equivalent noise measures

The preceding discussion has expressed noise in terms of power and voltage or current. Rather than specifying the noise in absolute terms, it is often more useful to express it in terms of the quantity to be measured. The noise level is then specified as the signal level for which the signal-to-noise ratio equals one.

3.7.1 *Noise equivalent power*

For example, in a system that measures power, one can express the noise in terms of noise equivalent power (NEP), which is equal to the signal input power for which the signal-to-noise ratio is one. This measure is commonly used in infrared and mm-wave measurements. If the signal-to-noise ratio S/N is known for a given input power P_{signal}, the noise equivalent power

$$\mathrm{NEP} = \frac{P_{signal}}{(S/N)} \tag{3.37}$$

or, if the noise current and the responsivity (signal current per unit signal power) are known,

$$\mathrm{NEP} = \frac{\text{Noise Current}\left[\mathrm{A}/\sqrt{\mathrm{Hz}}\right]}{\text{Current Responsivity}\,[\mathrm{A/W}]} \ . \tag{3.38}$$

3.7.2 *Equivalent noise charge*

Similarly, detector readout systems that measure signal charge can be characterized in terms of equivalent noise charge (ENC), *i.e.* the signal charge that yields a signal-to-noise ratio of one. If absorbed energy yields a signal charge Q_S and a signal-to-noise ratio S/N, the equivalent noise charge

$$\mathrm{ENC} \equiv Q_n = \frac{Q_S}{S/N} \ . \tag{3.39}$$

ENC is commonly expressed in fC or units of the electronic charge $e = 1.602 \cdot 10^{-19}$ C.

For a given detector material, the signal charge can be translated into absorbed energy, so the noise can be expressed in terms of energy, *i.e.* eV or keV. For an ionization energy E_i the energy resolution

$$\Delta E_n = E_i \cdot \mathrm{ENC} \ . \tag{3.40}$$

Equivalent noise charge is the most convenient measure of system noise when designing semiconductor detector systems. However, in analyzing the individual noise contributions, the basic noise parameters – voltage and current – are more useful. The combination of individual voltage and current noise contributions together with the system bandwidth yields the ENC, so it is a derived, rather than a primary quantity.

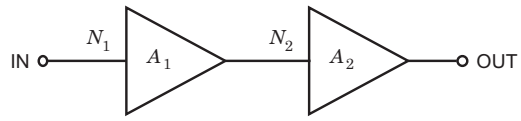

FIG. 3.6. In cascaded amplifiers the equivalent input noise of the first amplifier amplified by the first stage's gain can override the noise of the second stage.

3.8 Noise in Amplifiers

Consider a chain of two amplifiers (or amplifying devices), with gains A_1 and A_2, and input noise levels N_1 and N_2, as shown in Figure 3.6. When a signal S is applied to the first amplifier, the input signal-to-noise ratio is S/N_1. At the output of the first amplifier the signal is A_1S and the noise A_1N.

Both the signal and the noise are amplified by the second amplifier, but in addition the second amplifier contributes its noise, so the signal-to-noise ratio at the output of the second amplifier

$$\left(\frac{S}{N}\right)^2 = \frac{(SA_1A_2)^2}{(N_1A_1A_2)^2 + (N_2A_2)^2} = \frac{S^2}{N_1^2 + \left(\dfrac{N_2}{A_1}\right)^2}$$

$$\left(\frac{S}{N}\right)^2 = \left(\frac{S}{N_1}\right)^2 \frac{1}{1 + \left(\dfrac{N_2}{A_1N_1}\right)^2} \tag{3.41}$$

The overall signal-to-noise ratio is reduced, but the noise contribution from the second-stage can be negligible, provided the gain of the first stage is sufficiently high. In a well-designed system the noise is dominated by the first gain stage.

3.8.1 *Amplifier noise model*

The noise properties of any amplifier can be described fully in terms of a voltage noise source and a current noise source at the amplifier input, as shown in Figure 3.7. Typical magnitudes are $\mathrm{nV}/\sqrt{\mathrm{Hz}}$ and fA to $\mathrm{pA}/\sqrt{\mathrm{Hz}}$. Rather than specifying the total noise over the full bandwidth, the magnitude of each noise source

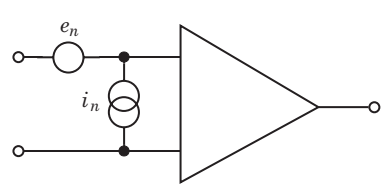

FIG. 3.7. An amplifier's noise characteristics are fully specified by equivalent input noise voltage and current generators.

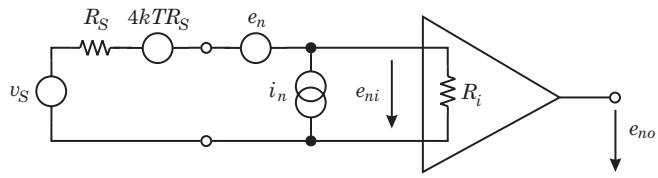

FIG. 3.8. The amplifier's noise equivalent circuit including a resistive signal source.

is characterized by its spectral density. This is convenient because the effects of frequency-dependent impedances in the input circuit and of the amplifier bandwidth can be assessed separately.

The noise sources do not have to physically present at the input. Noise also originates within the amplifier. Assume that at the output the combined contribution of all internal noise sources has the spectral density e_{no}. If the amplifier has a voltage gain A_v, this is equivalent to a voltage noise source at the input $e_n = e_{no}/A_v$.

Next we assess the interaction of the amplifier noise sources with a signal source, shown in Figure 3.8. A sensor with resistance R_S is connected to an amplifier with voltage gain A_v. First assume an infinite input resistance $R_i = \infty$, so no current flows into the amplifier.

The spectral noise densities e_n and i_n can be treated as voltages or currents at a discrete frequency. Thus, the input noise current i_n flows through the source resistance R_S to yield a noise voltage $i_n R_S$, which adds to the thermal noise of the source resistance and the input noise voltage of the amplifier. All terms add in quadrature, since they are not correlated. The total noise voltage at the input of the amplifier

$$e_{ni}^2 = 4kTR_S + e_n^2 + (i_n R_S)^2 \tag{3.42}$$

and at the output of the amplifier

$$e_{no}^2 = (A_v e_{ni})^2 = A_v^2 \left[4kTR_S + e_n^2 + (i_n R_S)^2\right] \ . \tag{3.43}$$

The signal-to-noise ratio at the amplifier output

$$\left(\frac{S}{N}\right)^2 = \frac{A_v^2 v_S^2}{A_v^2 \left[4kTR_S + e_n^2 + (i_n R_S)^2\right]} \tag{3.44}$$

is independent of the amplifier gain and equal to the input S/N, as both the input noise and the signal are amplified by the same amount.

In the preceding example the amplifier had an infinite input resistance, so no current flowed into the amplifier. Is the signal-to-noise ratio affected by a finite input resistance? The signal at the input of the amplifier

$$v_{Si} = v_S \frac{R_i}{R_S + R_i} \ . \tag{3.45}$$

The noise voltage at the input of the amplifier

$$e_{ni}^2 = \left(4kTR_S + e_n^2\right)\left(\frac{R_i}{R_i + R_S}\right)^2 + i_n^2\left(\frac{R_i R_S}{R_i + R_S}\right)^2 \ , \tag{3.46}$$

where the bracket in the i_n^2 term is the parallel combination of R_i and R_S. The signal-to-noise ratio at the output of the amplifier

$$\left(\frac{S}{N}\right)^2 = \frac{A_v^2 v_{Si}^2}{A_v^2 e_{ni}^2} = \frac{v_S^2\left(\dfrac{R_i}{R_i + R_S}\right)^2}{\left(4kTR_S + e_n^2\right)\left(\dfrac{R_i}{R_i + R_S}\right)^2 + i_n^2\left(\dfrac{R_i R_S}{R_i + R_S}\right)^2}$$

$$\left(\frac{S}{N}\right)^2 = \frac{v_S^2}{(4kTR_S + e_n^2) + i_n^2 R_S^2} \tag{3.47}$$

is the same as for an infinite input resistance. This result also holds for a complex input impedance, *i.e.* a combination of resistive and capacitive or inductive components. Since S/N is independent of amplifier input impedance, we can perform valid noise analyses using "idealized" amplifiers with infinite input impedance.

As noted above, noise sources can be correlated. In the above example, if the input noise voltage and current are correlated, the input noise voltage

$$e_{ni}^2 = 4kTR_S + e_n^2 + i_n^2 + 2Ce_n i_n R_S \ . \tag{3.48}$$

The total noise at the output is obtained by integrating over the spectral noise power $P_n(f) \propto e_{no}^2(f)$. The frequency distribution of the noise is determined both by the spectral distribution of the input noise voltage and current and by the frequency response of the amplifier:

$$v_{no}^2 = \int_0^\infty e_{no}^2(f)df = \int_0^\infty e_{ni}^2(f)\left|A_v\right|^2 df \ . \tag{3.49}$$

The amplifier gain factor is shown as magnitude squared, as in general the amplifier has a frequency-dependent gain and phase, so it is a complex number.

For noise whose spectral density is constant over the amplifier bandwidth, the total noise voltage (or current) increases with the square root of bandwidth. For $1/f$ noise, however,

$$v_n^2 = \int_{f_l}^{f_u} \frac{A_f}{f} df = A_f \log\frac{f_u}{f_l} \ , \tag{3.50}$$

so the total noise depends only on the ratio of the upper to lower cutoff frequencies. For a decade span $f_u = 10f_l$ the total noise $v_n^2 = 2.3A_f$. This is a useful

relationship because $1/f$ noise is often specified as the total noise over one or two decades of bandwidth.

Amplifiers commonly exhibit "$1/f$" noise at low frequencies and white noise at high frequencies. The "corner frequency" is the frequency where the $1/f$ noise equals the white noise, so the $1/f$ noise is $1/\sqrt{2}$ of the total noise. At $f = f_c/100$ the $1/f$ noise voltage is 10 times larger. The corner frequency alone is not sufficient to specify $1/f$ noise as it depends on the magnitude of the white noise. For the same $1/f$ noise, a higher white noise level will yield a lower corner frequency.

3.8.2 *Noise bandwidth vs. signal bandwidth*

Consider an amplifier with the frequency response $A(f)$. This can be rewritten

$$A(f) \equiv A_0 G(f) \tag{3.51}$$

where A_0 is the maximum gain and $G(f)$ describes the frequency response. For example, in the simple amplifier described in Chapter 2 the gain

$$A_v = g_m \left(\frac{1}{R_L} + \mathbf{i}\omega C_o \right)^{-1} = g_m R_L \frac{1}{1 + \mathbf{i}\omega R_L C_o} \tag{3.52}$$

and using the above convention

$$A_0 \equiv g_m R_L \quad \text{and} \quad G(f) \equiv \frac{1}{1 + \mathbf{i}(2\pi f R_L C_o)} \ . \tag{3.53}$$

If a "white" noise source with spectral density e_{ni} is present at the input, the total noise voltage at the output is

$$v_{no} = \sqrt{\int_0^\infty e_{ni}^2 \left| A_0 G(f) \right|^2 df} = e_{ni} A_0 \sqrt{\int_0^\infty G^2(f) df} \equiv e_{ni} A_0 \sqrt{\Delta f_n} \ . \tag{3.54}$$

Δf_n is the "noise bandwidth".

Note that, in general, the noise bandwidth and the signal bandwidth are not the same. If the upper cutoff frequency is determined by a single RC time constant, as in the "simple amplifier" discussed in Chapter 2, the signal bandwidth

$$\Delta f_s = f_u = \frac{1}{2\pi RC} \tag{3.55}$$

and the noise bandwidth

$$\Delta f_n = \frac{1}{4RC} = \frac{\pi}{2} f_u \ . \tag{3.56}$$

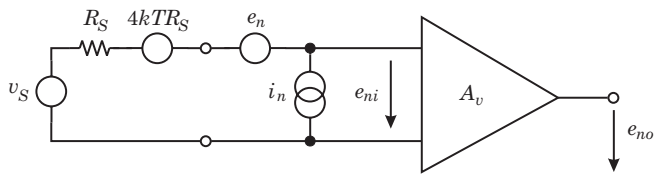

FIG. 3.9. Equivalent circuit for determining the total noise with a resistive signal source.

3.9 Amplifier noise matching

3.9.1 *Resistive sources*

The concept of noise matching is often misapplied in detector systems, so a discussion is appropriate to clarify where matching applies and where not. First consider a resistive source, as shown in Figure 3.9. The resistance of the signal source contributes thermal noise, but also determines the contribution of the amplifier's current noise contribution

$$e_{ni}^2 = 4kTR_S + e_n^2 + (i_n R_S)^2 \ . \tag{3.57}$$

Consider the total noise power in the input circuit. The source resistance contributes $4kT\Delta f_n$ and the power due to the amplifier's input noise voltage and current depends on the source resistance,

$$P_n = \left(4kT + \frac{e_n^2}{R_S} + i_n^2 R_S\right)\Delta f_n \ . \tag{3.58}$$

The total power attains a minimum for

$$R_S = \frac{e_n}{i_n} \ . \tag{3.59}$$

A common measure of amplifier noise is the "noise factor" F, which is the ratio of the total noise to the thermal noise of the sensor.

$$F = \frac{e_{ni}^2}{4kTR_S} = 1 + \frac{e_n^2 + (i_n R_S)^2}{4kTR_S} = 1 + \frac{1}{4kT}\left(\frac{e_n^2}{R_S} + i_n^2 R_S\right) \ . \tag{3.60}$$

For a noiseless amplifier $F = 1$. In a matched system with a resistive source

$$F_{opt} = 1 + \frac{e_n i_n}{2kT} \ . \tag{3.61}$$

The noise factor is frequently expressed in dB as the "noise figure" NF $= 10\log_{10} F$.

This principle of "noise matching" must be applied with caution:

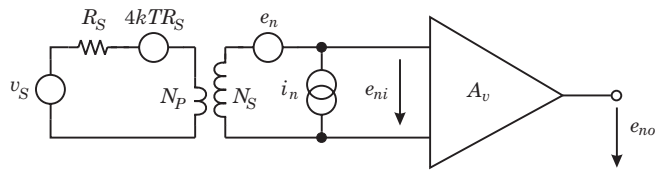

FIG. 3.10. Noise matching with a transformer.

1. Power is not always the relevant measure. Sometimes the noise voltage is most important. Minimum noise voltage e_{ni} always obtains with $R_S = 0$.
2. Merely increasing the source resistance will increase the total input noise without improving the signal-to-noise ratio. The advantage of noise matching only obtains when both the signal and the effective source resistance are modified simultaneously, for example by a transformer.

3.9.2 *Noise matching with a transformer*

The sensor is coupled to the amplifier through a transformer with the turns ratio $N = N_S/N_P$, as shown in Figure 3.10. Assume unity coupling in the transformer. Then the sensor voltage appearing at the secondary

$$v_{SS} = N v_S \ . \tag{3.62}$$

The thermal noise of the sensor at the secondary

$$e_{nSS}^2 = N^2 4kTR_S \ . \tag{3.63}$$

Because the transformer also converts impedances, the source resistance appears at the secondary as

$$R_{SS} = N^2 R_S \ . \tag{3.64}$$

Thus, as the signal voltage is increased, so is the noise contribution due to the input noise current

$$e_{ni}^2 = 4kTR_S N^2 + e_n^2 + R_S^2 N^4 i_n^2 \ . \tag{3.65}$$

The signal-to-noise ratio

$$\left(\frac{S}{N}\right)^2 = \frac{v_S^2 N^2}{4kTR_S N^2 + e_n^2 + R_S^2 N^4 i_n^2} = \frac{v_S^2}{4kTR_S + \frac{e_n^2}{N^2} + N^2 R_S^2 i_n^2} \ , \tag{3.66}$$

which attains a maximum for

$$R_S N^2 = \frac{e_n}{i_n} \ . \tag{3.67}$$

Minimum noise obtains when the transformer secondary presents the optimum source resistance to the amplifier.

3.10 Capacitive sources

The noise measures and derivations applied above do not apply to capacitive sources, because

1. The capacitance is not a noise source, so matching the amplifier noise to the sensor noise is meaningless.
2. The noise power in the input circuit is zero, as the current and voltage in any reactance (capacitance or inductance) are 90° out of phase.

Capacitive sources have a different matching criterion, which depends on the frequency range used in the measurement, *i.e.* the shaping time. This will be developed in the next chapter. There are also optimization criteria for various types of amplifying devices, which will be described in Chapter 6. Transformers can be useful in system with capacitive sources, but they are bulky, so they are not well-matched to microelectronics, and they must be designed carefully to maintain signal integrity and minimize additional noise.

3.10.1 *Noise vs. capacitance in a charge-sensitive amplifier*

In a voltage-sensitive amplifier with negligible input noise current the noise voltage at the output is essentially independent of detector capacitance C_d, *i.e.* the *equivalent* input noise voltage $v_{ni} = e_{no}\sqrt{\Delta f_n}/A_v$. Since for a given input charge Q_s the input voltage $v_s = Q_s/C$, the input signal decreases with increasing total input capacitance C (the sum of detector and amplifier input capacitance), so the signal-to-noise ratio depends on detector capacitance.

In a charge-sensitive amplifier, the signal at the amplifier output is independent of detector capacitance (if $C_i \gg C_d$). In this case the mechanism that determines noise *vs.* sensor capacitance is quite different. Noise appearing at the output of the amplifier is fed back to the input with opposite phase, decreasing the output noise from the open-loop value $e_{no} = e_{ni}A_v$. The magnitude of the fedback signal depends on the shunt impedance at the input, *i.e.* the sensor capacitance. Thus, the signal-to-noise ratio at the amplifier output depends on the amount of feedback, which changes with detector capacitance.

Note, that although specified as an input noise, the dominant noise sources are typically internal to the amplifier. Only in a fedback configuration is some of this noise actually present at the input. In other words, the primary noise signal is not a physical charge (or voltage) at the amplifier input, to which the loop responds in the same manner as to a detector signal.

To analyze the noise consider Figure 3.11. Start with an output noise voltage e_{no}, which is fed back to the input through the capacitive voltage divider $C_f - C_d$. The voltage divider formed by the feedback capacitor and the sensor capacitance sets the condition

$$e_{no} = e_{nf}\frac{X_{C_f} + X_{C_d}}{X_{C_d}} = e_{nf}\frac{(1/\omega C_f) + (1/\omega C_d)}{1/\omega C_d} = e_{nf}\left(1 + \frac{C_d}{C_f}\right) \quad . \tag{3.68}$$

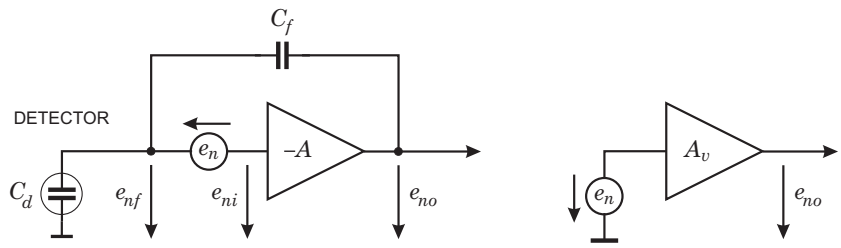

FIG. 3.11. Circuit for the noise analysis of a charge-sensitive amplifier. The convention for setting the phase of e_n is shown at the right.

Next we establish the relationship between e_{nf} and the equivalent input noise of the amplifier e_n. Inspection of Figure 3.11 shows that e_{nf}, e_n, and e_{ni} form a loop. Since the sum of the potentials in a loop must be zero,

$$e_{nf} + e_n - e_{ni} = 0 \ . \tag{3.69}$$

The noise voltages all have the same origin, so they are fully correlated and add algebraically. The phase of e_n has been set to conform to the general gain relationship for the amplifier without feedback $e_{no} = A_v e_n$, as shown in the second panel of Figure 3.11. In this case $A_v = -A$, which regardless of feedback equals sets the ratio e_{no}/e_{ni}. Introducing this relationship and eqn 3.68 into eqn 3.69 yields

$$e_{nf} = -e_n \frac{1}{1 + \dfrac{1}{A} + \dfrac{C_d}{AC_f}} \ . \tag{3.70}$$

If the open loop gain of the amplifier is sufficiently large such that $A \gg 1$ and $AC_f \gg C_d$, the noise voltage at the amplifier input becomes negligible $|e_{ni}| \ll |e_{nf}|$, so $|e_{nf}| \approx |e_n|$.

To calculate the equivalent input noise charge, assume that the charge-sensitive amplifier feeds a pulse shaper that provides a signal gain of unity and whose noise bandwidth Δf_n yields an equivalent noise charge Q_{ni}. Using the charge sensitivity eqn 2.101 derived in Chapter 2

$$\begin{aligned} Q_{ni} &= F_S \frac{e_{no}}{A_Q} = F_S e_{no} C_f \\ Q_{ni} &= F_S e_{ni} \left(C_D + C_f\right) \approx F_S e_n \left(C_D + C_f\right) \ , \end{aligned} \tag{3.71}$$

where the factor $F_S = A_{VS}\sqrt{\Delta f_n}$ combines the noise bandwidth and gain that characterize the pulse shaper. This will be discussed quantitatively in the next chapter. Thus, the signal-to-noise ratio

$$\frac{Q_s}{Q_{ni}} = \frac{1}{F_S} \frac{Q_s}{e_{nf}(C_d + C_f)} \approx \frac{1}{F_S} \frac{1}{C_d} \frac{Q_s}{e_n} \quad (C_d \gg C_f) \ . \tag{3.72}$$

This is the same result as for a voltage-sensitive amplifier, *but here the signal is constant and the noise grows with increasing* C_d. However, note that the additional feedback capacitor adds to the detector capacitance in determining the noise.

Since the amplifier inverts, e_{no} is 180° out of phase with e_n and so is e_{nf}. Thus, the net noise voltage at the amplifier input $e_{ni} = e_n - |e_{nf}|$ is diminished with respect to the open loop condition shown in the second panel of Figure 3.11, as necessary for negative feedback. As a consequence, the noise at the amplifier output will increase with detector capacitance, as this reduces the fedback voltage e_{nf}.

Equation 3.70 also shows that the requirement $AC_f \gg C_d$ is important for low noise as well as for signal acquisition, as derived in Chapter 2. Note that the frequencies of interest usually lie well above the open loop corner frequency of the amplifier, so A is much smaller than the DC gain, which is what is often quoted. The frequency dependence $A(f)$ must always be taken into account.

As was also shown in Chapter 2, the pulse rise time at the amplifier output also increases with total capacitive input load, because of reduced feedback. In contrast, the rise time of a voltage sensitive amplifier is not affected by the input capacitance, although the equivalent noise charge increases with C_d just as for the charge-sensitive amplifier.

This discussion illustrates a general feature. The optimum S/N is independent of whether the voltage, current, or charge signal is sensed. Furthermore, under optimum conditions S/N is not affected by feedback. The more rigorous statement is that S/N *cannot be improved* by feedback. The feedback circuit components can introduce additional noise. Note in eqn 3.71 that the feedback capacitor adds to the effective capacitive load at the input and hence increases the equivalent noise charge. However, this contribution can be negligible, since a design goal is $C_f \ll C_d$. Resistive components in the feedback circuit will add thermal noise. A common noise source is a resistor in parallel to the feedback capacitor to provide a discharge path.

3.10.2 *S/N vs. input time constant*

The preceding analysis implies that the signal-to-noise ratio could increase arbitrarily by reducing the detector capacitance. The following analysis shows the limits. We make use of the fact that the signal-to-noise ratio is the same in any amplifier configuration and analyze the voltage signal developed at the amplifier input in Figure 3.12.

In the analysis the signal current $i_s(t)$ and the detector capacitance C_d are kept constant, but the input resistance of the amplifier is changed while keeping its noise parameters constant. The total input noise voltage is constant, so the signal-to-noise ratio depends only on the magnitude of the voltage signal $v_s(t)$.

The pulse shape and the peak voltage registered by the amplifier depend on the input time constant R_iC_d. For a rectangular detector current pulse of duration T and peak magnitude I_s the input current to the amplifier is

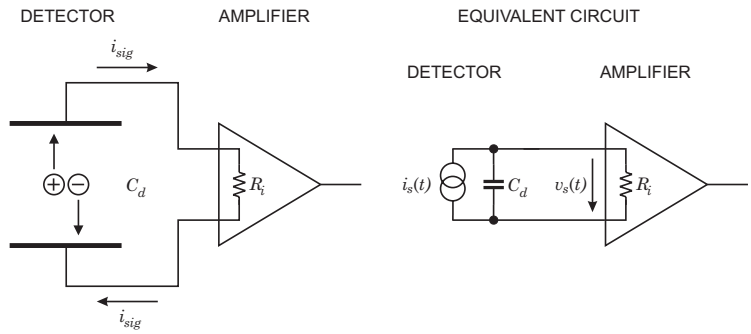

FIG. 3.12. Circuit to analyze the signal-to-noise ratio *vs.* input time constant R_iC_d.

$$\begin{aligned} 0 \le t < T: \quad & i_{in}(t) = I_s\left(1 - e^{-t/RC}\right) \\ T \le t \le \infty: \quad & i_{in}(t) = I_s\left(e^{T/RC} - 1\right) \cdot e^{-t/RC} \end{aligned} \tag{3.73}$$

The time dependence of the current flowing into the amplifier is shown in Figure 3.13 for input time constants ranging from $0.01T$ to 10^3T. At short time constants $R_iC_d \ll T$ the amplifier pulse approximately follows the detector current pulse. As the input time constant R_iC_d increases, the amplifier signal becomes longer and the peak amplitude decreases, although the integral, *i.e.* the signal charge, remains the same.

At long time constants the detector signal current is integrated on the detector capacitance and the resulting voltage sensed by the amplifier

$$V_{in} = \frac{Q_s}{C} = \frac{\int i_s dt}{C} \ . \tag{3.74}$$

Then the peak amplifier signal is inversely proportional to the *total capacitance at the input*, *i.e.* the sum of the sensor capacitance, the input capacitance of the amplifier, and any stray capacitances present at the input.

Figure 3.14 shows the peak signal *vs.* input time constant, which depends directly on input capacitance. At small time constants the amplifier signal approximates the detector current pulse and is independent of capacitance. At large input time constants ($RC/T > 5$) the maximum signal falls linearly with capacitance.

In practically all situations it is valid to assume that the signal-to-noise ratio increases linearly with decreasing capacitance. However, one must exercise caution when extrapolating to very small capacitances. If $S/N = 1$ at $R_iC_d/T = 100$, decreasing the capacitance to 1/10 of its original value ($R_iC_d/T = 10$), increases S/N to 10. However, if initially $R_iC_d/T = 1$, the same 10-fold reduction in capacitance (to $R_iC_d/T = 0.1$) only yields $S/N = 1.6$, reflecting the transition from voltage to current mode operation.

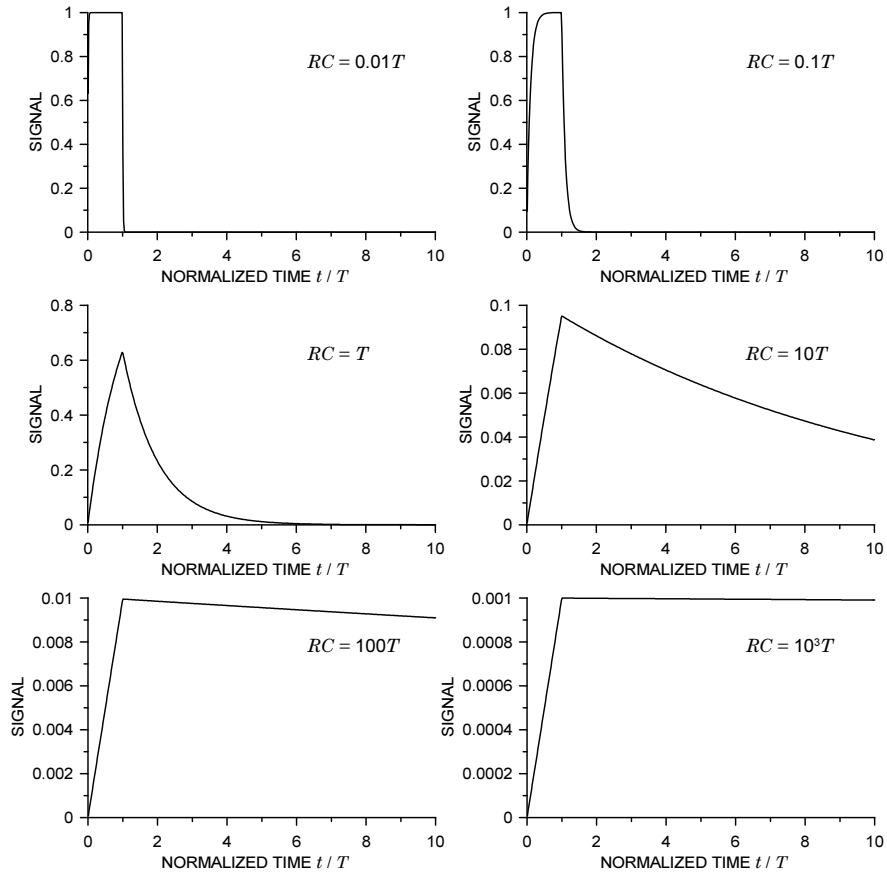

FIG. 3.13. Pulse amplitudes *vs.* time for input time constants RC ranging from $0.01T$ to 10^3T, where T is the duration of the detector current pulse.

3.11 Complex sensors

Up to now the sensor was modeled as a simple capacitance. This is valid for simple "pad" sensors, but not for all configurations. One example that requires a more complex model is a strip detector. Figure 3.15 shows the detector model for noise simulations. Noise is evaluated in the center channel and includes the strip impedance and noise coupled from the neighbor amplifiers.

Individual strip electrodes are modeled as distributed RLC elements, where the R, L, and C are the strip electrode's resistance, inductance, and capacitance per unit length. The capacitance includes the strip-to-strip and strip-to-backplane capacitances C_{ss} and C_b. In most applications the inductive reactance is sufficiently small that the strip can be treated as purely resistive. The strip resistance is determined by the thickness and width of the metallization.

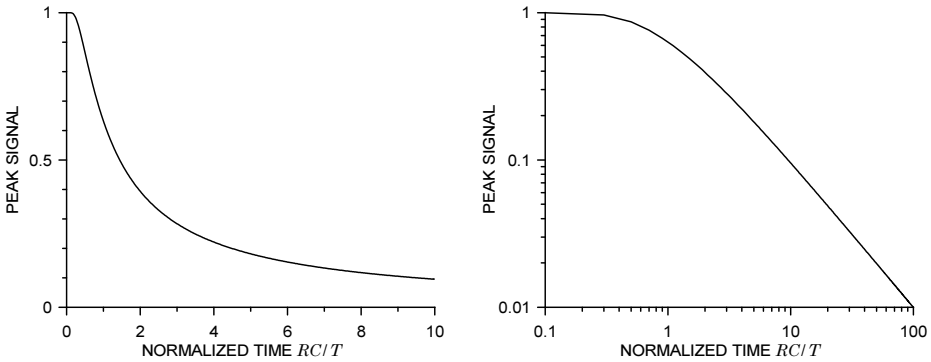

FIG. 3.14. Maximum signal *vs.* input time constant, plotted on linear and logarithmic scales.

The material is commonly a sputtered Al-Si alloy with a typical resistivity of $4 \cdot 10^{-6}\,\Omega\,\mathrm{cm}$. Film thicknesses of $0.5 - 1\,\mu$m are typical with a practical upper limit of about $2\,\mu$m. A width of $10\,\mu$m and $1\,\mu$m thickness give a strip resistance of $40\,\Omega/\mathrm{cm}$.

A signal current is injected at the desired position of the center electrode. Rather then using a simple rectangular pulse, realistic pulse shapes as shown in the previous chapter are used. Multiple amplifiers are included to assess the effect of amplifier input impedance and time response, but also to include the injection of noise from the neighbor amplifiers. This model can be included in a full SPICE simulation together with the full amplifier circuit to assess the equivalent noise charge. Each strip acts as a low-pass filter, so the thermal noise

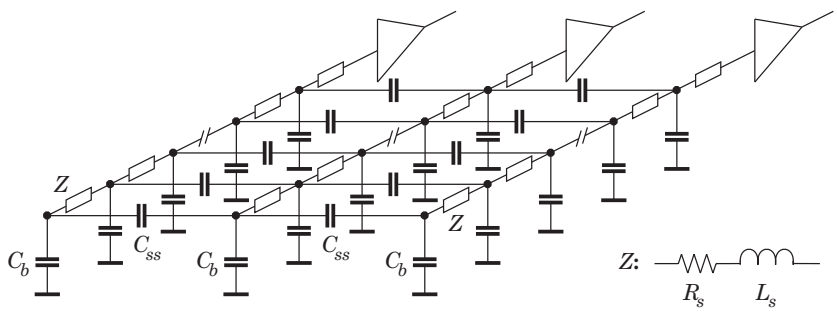

FIG. 3.15. Equivalent circuit of a strip detector for noise simulations. The strip electrodes are modeled as distributed RLC networks. The impedance Z is the strip electrode's inductance and resistance per unit length, C_{ss} the capacitance to the neighbor strip, and C_b the strip's capacitance to the backplane. In most applications the strip inductance can be neglected, so $Z \approx R$.

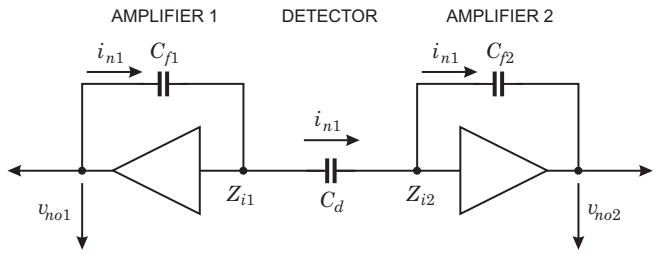

FIG. 3.16. Cross coupling of amplifier noise through a common capacitance. The circuit represents two amplifiers connected to opposite electrodes of a simple parallel-plate detector, for example.

of the strip resistance is attenuated at high frequencies. The cutoff frequency of the low-pass filter $f_u = (2\pi R_{strip} C_{strip})^{-1}$, where R_{strip} and C_{strip} are the total strip resistance and capacitance. At low frequencies the strip resistance contributes $e_{nR}^2 = 4ktR_{strip}$ to the input noise voltage. At frequencies beyond f_u the noise is attenuated, but the rise time of the signal pulse also increases. A second contribution is the noise coupled through the strip-to-strip capacitance from the neighbor amplifiers.

3.11.1 *Cross-coupled noise*

The mechanism for cross-coupling of amplifier is illustrated in Figure 3.16. To assess the contribution of amplifier 1 to amplifier 2, we first assume that amplifier 2 is noiseless and that the noise voltage v_{no1} is present at the output of amplifier 1. This causes a current i_{n1} to flow through C_{f1} and the detector capacitance C_d into the input of amplifier 2. Note that for a signal originating at the output of amplifier 1, the impedance Z_{i1} is high (infinite for an idealized amplifier), so all of i_{n1} flows through C_d and then into the input of amplifier 2. For this current the input of amplifier 2 does present a low impedance, so its magnitude

$$i_{n1} = \frac{v_{no1}}{\dfrac{1}{\omega C_{f1}} + \dfrac{1}{\omega C_d}} \ . \tag{3.75}$$

Since the output voltage of amplifier 2 is the product of the input current and feedback impedance,

$$v_{no12} = \frac{i_{n1}}{\omega C_{f2}} = \frac{v_{no1}}{\dfrac{1}{\omega C_{f1}} + \dfrac{1}{\omega C_d}} \frac{1}{\omega C_{f2}} = \frac{v_{no1}}{\dfrac{C_{f2}}{C_{f1}} + \dfrac{C_{f2}}{C_d}} \ . \tag{3.76}$$

If the two amplifiers are the same, $C_{f1} = C_{f2}$. Furthermore, $C_{f2} \ll C_d$, so the additional noise at the output of amplifier 2 due to amplifier 1 is $v_{no12} = v_{no1}$, which adds in quadrature to v_{no2}. Since both amplifiers are the same, $v_{no1} = v_{no2}$ and the noise increases by a factor $\sqrt{2}$.

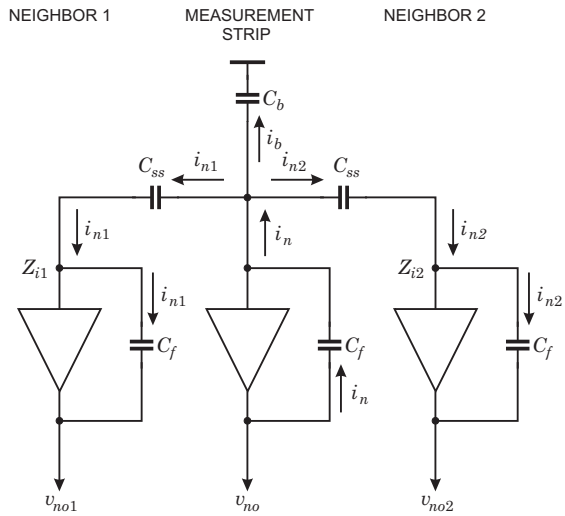

FIG. 3.17. Noise cross-coupling in a strip detector.

In a strip detector the cross-coupling capacitance is the strip-to-strip capacitance C_{ss} and the current originating in the center amplifier is split between the two neighbors and the backplane capacitance C_b, as shown in Figure 3.17. Thus, the current injected from the center amplifier

$$i_n = \frac{v_{no}}{\dfrac{1}{\omega C_f} + \dfrac{1}{\omega\left(2C_{ss} + C_b\right)}} \ , \tag{3.77}$$

which divides into the two currents to the neighbors

$$i_{n1} = i_{n2} = \frac{i_n}{2}\left(1 - \frac{1}{1 + C_{ss}/2C_b}\right) \ . \tag{3.78}$$

Thus, the additional noise voltage at the outputs of the two neighbor amplifiers

$$v_{no1} = v_{no2} = \frac{i_{n1}}{\omega C_f} = \frac{v_{no}}{2}\frac{2C_{ss} + C_b}{2C_{ss} + C_b + C_f}\left(\frac{C_{ss}}{2C_b + C_{ss}}\right) \approx \frac{v_{no}}{2}\frac{1}{1 + 2C_b/C_{ss}} \ , \tag{3.79}$$

since $C_f \ll 2C_{ss} + C_b$. Since each channel receives noise from its two adjacent neighbors, this contribution must be counted twice. If we neglect the backplane capacitance, $v_{no1} = v_{no2} = v_{no}/2$ and the noise increases by a factor $\sqrt{1 + 0.5^2 + 0.5^2} = 1.22$. For a backplane capacitance $C_b = C_{ss}/10$, which is typical, the noise degrades by 16%. If the input impedances of the neighbor amplifiers Z_{i1} and Z_{i2} are sufficiently low, coupling to the next neighbors is negligible, so only the nearest neighbors contribute. Again, this underscores the importance of a low input impedance in optimizing noise.

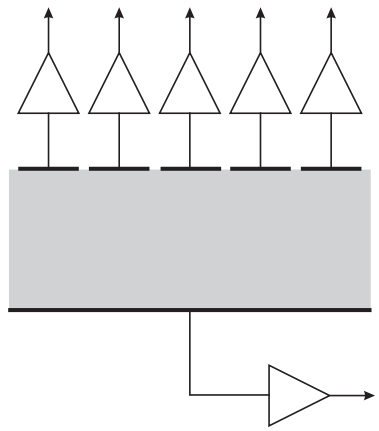

FIG. 3.18. Readout of the backside electrode of a strip or pixel detector provides a measurement of the total signal, regardless of the position.

3.11.2 *Backside readout*

Another situation where noise cross-coupling is important is the backside readout of a strip or pixel detector to provide a signal when any strip or pixel is struck. This is illustrated in Figure 3.18. The backside amplifier is in the signal path for all of the strip or pixel amplifiers, so its input impedance must be low to reduce signal cross-coupling to acceptable levels. The capacitance presented to the backside amplifier is C_d. To assess the cross-coupling of noise the model in Figure 3.16 can be applied, except that the coupling capacitance becomes C_d/n, where n is the number of strips or pixels. The additional noise injected from the backside amplifier (amplifier 1 in Figure 3.16) into a single electrode is

$$v_{no12} = \frac{i_{n1}}{\omega C_{f2}} = \frac{v_{no1}}{\dfrac{1}{\omega C_{f1}} + \dfrac{n}{\omega C_d}} \frac{1}{\omega C_{f2}} = \frac{v_{no1}}{\dfrac{C_{f2}}{C_{f1}} + n\dfrac{C_{f2}}{C_d}} \ . \tag{3.80}$$

Since the backside amplifier sees a larger capacitance than the strip or pixel amplifiers, its noise will be larger. Assume that both sides use the same pulse shaping, so we can ignore the factor F_S introduced in Section 3.10.1, as it will drop out. Then the backside amplifier's equivalent noise charge is $Q_{n1} = C_{f1} v_{no1}$ and the noise charge in the strip or pixel channel

$$Q_{n2} = C_{f2} \cdot v_{no12} = \frac{v_{no1}}{\dfrac{1}{C_{f1}} + \dfrac{n}{C_d}} = \frac{Q_{n1}}{1 + n\dfrac{C_{f1}}{C_d}} \ . \tag{3.81}$$

Thus, C_{f1} must be sufficiently large to reduce v_{no2} such that the contribution from the backside amplifier is minor compared to the noise Q_{n2} in the strip or

pixel channels. Since Q_{n1} will tend to be larger than Q_{n2} because of the larger capacitance, this drives feedback capacitor C_{f1} to larger values than usual.

Backside readout is often used to derive a timing signal. Since the pixel or strip amplifiers are in the current return path of the backside amplifier, their response time affects the backside signal. This places a constraint on their input time constants, which will tend to be more rigorous than for charge measurements alone.

3.12 Quantum noise limits in amplifiers

What is the lower limit to electronic noise? Can it be eliminated altogether, for example by using superconductors to eliminate thermal noise and avoiding devices that carry shot noise? The starting point is the uncertainty relationship

$$\Delta E \Delta t \geq \frac{\hbar}{2} \ . \tag{3.82}$$

Consider a narrow frequency band at frequency ω. The energy uncertainty can be given in terms of the uncertainty in the number of signal quanta

$$\Delta E = \hbar \omega \Delta N \tag{3.83}$$

and the time uncertainty in terms of phase

$$\Delta t = \frac{\Delta \varphi}{\omega} \ , \tag{3.84}$$

so that

$$\Delta \varphi \Delta N \geq \frac{1}{2} \ . \tag{3.85}$$

We assume that the distributions in number and phase are Gaussian, so that the equal sign applies.

Next, assume a noiseless amplifier with gain G, so that N_1 quanta at the input yield

$$N_2 = G N_1 \tag{3.86}$$

quanta at the output. Then the output must also obey the relationship

$$\Delta \varphi_2 \Delta N_2 = \frac{1}{2} \ . \tag{3.87}$$

However, since $\Delta N_2 = G \Delta N_1$ and since the output is shifted by a constant phase with respect to the input, $\Delta \varphi_2 = \Delta \varphi_1$,

$$\Delta \varphi_1 \Delta N_1 = \frac{1}{2G} \ , \tag{3.88}$$

which is smaller than allowed by the uncertainty principle.

This contradiction can only be avoided by assuming that the amplifier introduces noise per unit bandwidth of

$$\frac{dP_{no}}{d\omega} = (G-1)\hbar\omega \ , \tag{3.89}$$

which, referred to the input, is

$$\frac{dP_{ni}}{d\omega} = \left(1 - \frac{1}{G}\right)\hbar\omega \ . \tag{3.90}$$

If the noise from the following gain stages is to be small, the gain of the first stage must be large, and then the minimum noise of the amplifier is (Haus and Mullen 1962)

$$\frac{dP_{ni}}{d\omega} = \hbar\omega \ . \tag{3.91}$$

The quantum-limited spectral noise density is proportional to frequency. At the frequencies in the MHz range characteristic of semiconductor detector systems, the quantum noise limit is much below typical noise levels. However, at frequencies $> 10\,\mathrm{GHz}$ this is no longer the case.

This minimum noise limit applies only to phase-coherent systems. In systems where the phase information is lost, *e.g.* bolometers, this limit does not apply. At 100 GHz bolometers exhibit noise levels well below the amplifier quantum noise limit. For a detailed discussion see Caves (1982).

References

Armantrout, G.A. *et al.* (1972). Sensitivity problems in biological and environmental counting. *IEEE Trans. Nucl. Sci.* **NS-19/1** (1972) 107–116

Caves, C.M. (1982). Quantum limits on noise in linear amplifiers. *Phys. Rev.* **D26** (1982) 1817–1839

Haus, H.A. and Mullen, J.A. (1962). Quantum noise in linear amplifiers. *Phys. Rev.* **128** (1962) 2407–2413

Kraus, John D. (1986). *Radio Astronomy.* Cygnus-Quasar Books, Powell, ISBN 1-882484-00-2

Kraus, John D. (1988). *Antennas* (2nd edn). McGraw-Hill, New York, ISBN 0-07-035422-7, TK7871.6.K74

Motchenbacher, C.D. and Connelly, J.A. (1993). *Low-Noise Electronic System Design.* Wiley-Interscience, New York, ISBN 0-471-57742-1, TK7867.M692

Philippot, J.Cl. (1970). Automatic processing of diode spectrometry results. *IEEE Trans. Nucl. Sci.* **NS-17/3** (1970) 446–488

Radeka, V. (1974). Signal, noise and resolution in position-sensitive detectors. *IEEE Trans. Nucl. Sci.* **NS-21** (1974) 51–64

van der Ziel, A. (1986). *Noise in Solid State Devices and Circuits.* Wiley-Interscience, New York, ISBN 0-471-83234-0, TK7871.85.V34

4

SIGNAL PROCESSING

The raw detector signal must be processed to perform amplitude or time measurements. As optimizing either the amplitude or time resolution affects the pulse shape, signal processing is also called pulse shaping. Pulse shaping determines both the total noise and the peak signal amplitude at the output of the shaper. The analysis of pulse shapers involves three steps:

1. Evaluate the total noise at the shaper output.
2. Determine the pulse amplitude for a known input charge.
3. From the signal-to-noise ratio extrapolate to the input charge that yields a signal-to-noise ratio of unity. This is the equivalent noise charge.

4.1 Simple pulse shapers

A simple pulse shaper is shown in Figure 4.1. A high-pass filter sets the duration of the pulse by introducing a decay time constant τ_d. Next a low-pass filter increases the rise time to limit the noise bandwidth. The high-pass is often referred to as a "differentiator", since for short pulses it forms the derivative. Correspondingly, the low-pass is called an "integrator". Since the high-pass filter is implemented with a CR section and the low-pass with an RC, this shaper is referred to as a CR-RC shaper. Although pulse shapers are often more sophisticated and complicated, the CR-RC shaper contains the essential features of all pulse shapers, a lower frequency bound and an upper frequency bound.

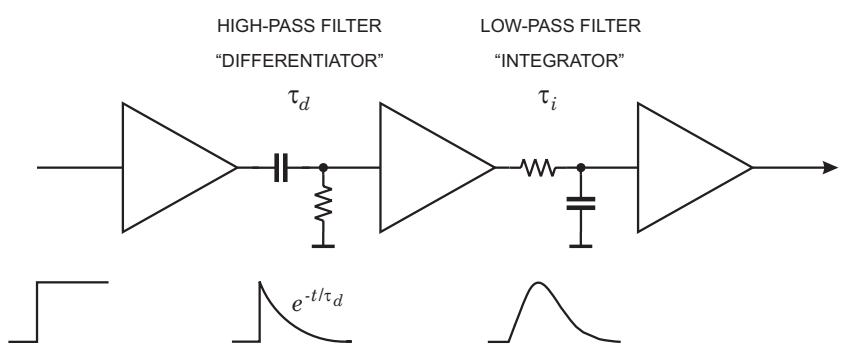

FIG. 4.1. A simple pulse shaper using a CR "differentiator" as a high-pass and an RC "integrator" as a low-pass filter.

The overall frequency response is the product of the individual frequency responses $G(f) = G_{int}(f) \cdot G_{diff}(f)$. Since in the Fourier transform a product $G_1(f) \cdot G_2(f)$ in the frequency domain is expressed in the time domain as the convolution

$$g(t) = g_1(t) * g_2(t) \equiv \int_{-\infty}^{\infty} g_1(\tau) g_2(t-\tau) d\tau \ , \tag{4.1}$$

the output pulse shape of the CR-RC shaper is the convolution of the input signal with the time responses of the individual stages

$$V_o(t) = V_i(t) * g_{int}(t) * g_{diff}(t) \ . \tag{4.2}$$

The resulting pulse shape for a unit step input

$$V_o(t) = \frac{\tau_d}{\tau_d - \tau_i} \left[e^{-t/\tau_d} - e^{-t/\tau_i} \right] \ , \tag{4.3}$$

where τ_d and τ_i are the differentiator and integrator time constants. With equal time constants $\tau_d = \tau_i$

$$V_o(t) = \left(\frac{t}{\tau} \right) e^{-t/\tau} \tag{4.4}$$

is a good representation. The output pulse assumes its maximum at the peaking time $T_P = \tau$. The noise performance of this simple shaper is only 36% worse than the optimum filter. It is also simple to evaluate, so it is useful for estimates.

4.1.1 *Effect of relative time constants*

Changing the time constants of the CR and RC sections in Figure 4.1 changes the noise bandwidth, so it will affect the noise level, but it also affects the signal amplitude. Figure 4.2 shows output pulse shapes for various combinations

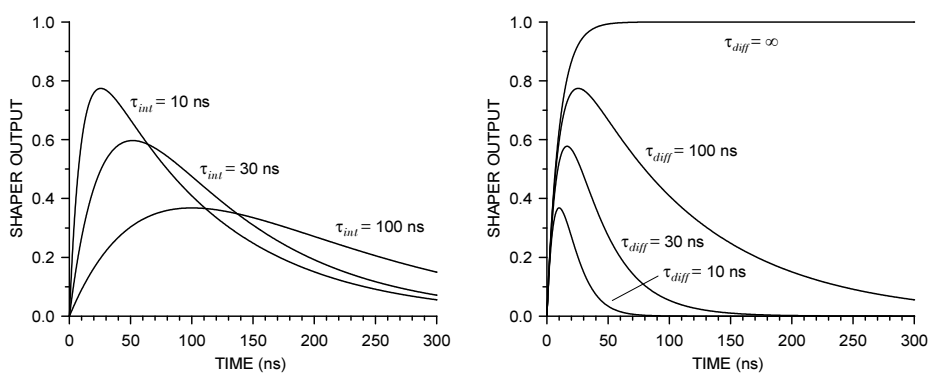

FIG. 4.2. Change in pulse amplitude *vs.* integrator time constant for a fixed differentiator time of 100 ns (left) and *vs.* differentiator time constant for a fixed integrator time constant of 10 ns (right).

of the CR and RC time constants. Figure 4.3 shows the corresponding changes in signal, noise and S/N. Although calculated for a specific set of detector and noise parameters, these curves represent general trends. In the case of a fixed differentiator time constant (left), when increasing the integrator time constant from 10 to 100 ns the noise decreases four-fold. However, this is partially compensated by the two-fold reduction in signal, so that the signal-to-noise ratio at 100 ns improves only by a factor two. For a fixed integrator time constant (right) the output noise level increases by 30% when the differentiator time constant increases from 10 to 100 ns, because reducing the lower cutoff frequency increases the noise bandwidth. In this case, however, the output pulse amplitude rises two-fold, so the signal-to-noise ratio still improves, despite the increase in noise.

Note that the need to limit the pulse width incurs a significant reduction in the output signal. Even at a relatively large differentiator time constant $\tau_d = 100\,\text{ns} = 10\tau_i$ the output signal is only 80% of the value for $\tau_d = \infty$, *i.e.* a system with no low-frequency roll-off. For a given pulse duration, *i.e.* differentiation time, the CR-RC shaper yields the optimum signal-to-noise ratio when the integrator and differentiator time constants are equal $\tau_d = \tau_i = \tau$. Then the peaking time, where the output pulse attains its maximum value, $T_P = \tau$.

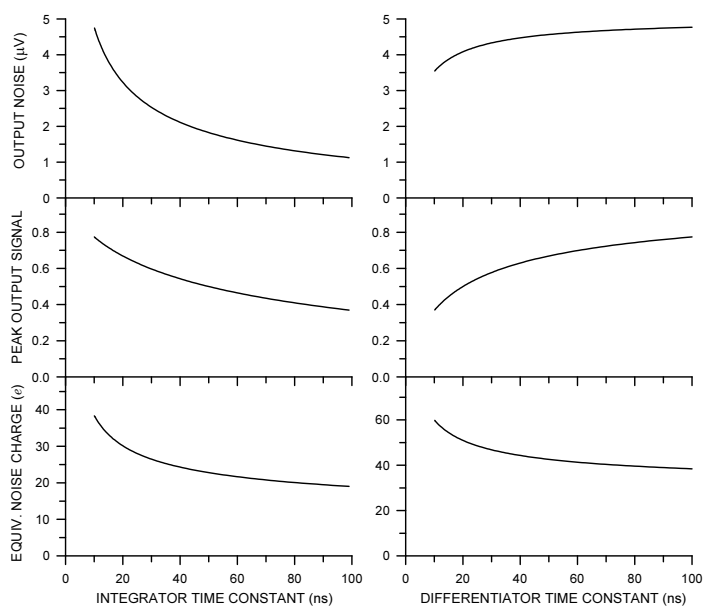

FIG. 4.3. Output noise (top), peak signal (middle), and equivalent noise charge (bottom) *vs.* integrator time constant for a fixed differentiator time of 100 ns (left) and *vs.* differentiator time constant for a fixed integrator time of 100 ns (right).

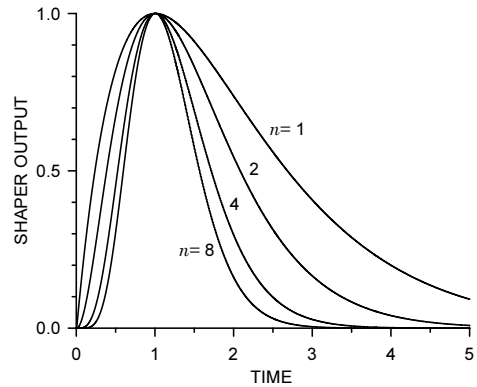

FIG. 4.4. Pulse shape *vs.* number of integrators in a *CR-nRC* shaper. The integration and differentiation time constants are scaled with the number of integrators ($\tau = \tau_{n=1}/n$) to maintain the peaking time.

After peaking the output of a simple *CR-RC* shaper returns to baseline rather slowly. The pulse can be made more symmetrical, allowing higher signal rates for the same peaking time. Very sophisticated circuits have been developed to improve this situation, but a conceptually simple way is to use multiple integrators. The resulting pulse shape

$$V_o(t) = \left(\frac{t}{\tau}\right)^n e^{-t/\tau} \tag{4.5}$$

is illustrated in Figure 4.4. Here the time constants are scaled with the number of integrators ($\tau = \tau_{n=1}/n$) to maintain the peaking time. Note that the peaking time is a key design parameter, as it dominates the noise bandwidth and must also accommodate the sensor response time. A detailed summary of multi-pole shapers and pulse shapes is given by Kowalski (1970).

Another consideration in the choice of time constants is the rise time of the input pulse applied to the shaper. Figure 4.1 shows a step input with zero rise time. This is convenient when characterizing the pulse shaper alone. In reality the rise time is increased by the collection time of the detector and the limited response time of the preamplifier. In many systems the input rise time is much smaller than the shaping time, so the step input is an acceptable approximation. However, when using short peaking times as is common in high-luminosity collider detectors, the sensor's collection time may be a substantial fraction of the shaper's peaking time. Furthermore, in the interest of reducing power consumption, the preamplifier bandwidth is often limited, so its rise time cannot be neglected. Then the preamplifier becomes part of the pulse shaper and the system must be analyzed as a whole.

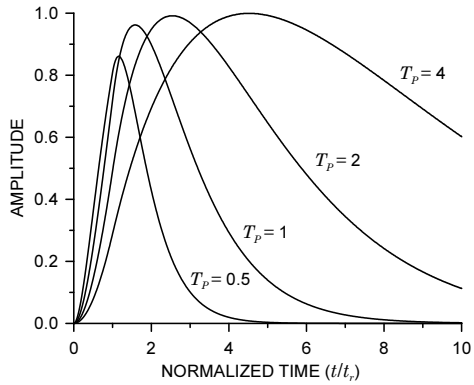

FIG. 4.5. Response of a *CR-RC* shaper with equal time constants to an input rise time t_r. All times are expressed in units of the rise time t_r. The shaper peaking times T_P range from 0.5 to $4t_r$.

Figure 4.5 shows the shaper output response to an input rise time t_r for *CR-CR* shapers with peaking times ranging from 0.5 to 4 times the rise time. The *CR* and *RC* time constants are equal. For a step input with an instantaneous rise the outputs would attain their peak amplitude of one at the nominal peaking time. The finite rise time delays the time of maximum signal for all shapers. This is most pronounced for the $T_P = 0.5t_r$ shaper, which with a step input peaks at $t = 0.5t_r$, but with the detector signal peaks at $t = 1.2t_r$. Furthermore, its peak amplitude is 14% smaller than for a step input. Since the noise level is independent of the input signal, this reduces the signal-to-noise ratio. The loss in pulse height is called "ballistic deficit". Although analytical techniques usually assume step inputs, analysis with realistic sensor pulses is straightforward using numerical techniques. Note that when the detector collection time is the dominant contributor to the input rise time, the peaking time of the output signal is not a measure of shaper noise performance, as the noise bandwidth is determined by the shaper's time constants, which are expressed in terms of the peaking time T_P with a step input.

4.2 Evaluation of equivalent noise charge

Analyzing the noise of a system utilizes three techniques. One can apply any one of these techniques alone, but fully understanding a "real" system requires all three. First is experiment – ultimately this describes how the system actually performs. Frequently, this is the only technique that physicists and circuit designers use. Second is numerical simulation. This is what integrated circuit designers do. When properly done, numerical simulations include all of the "non-ideal" contributions to system performance. They show how the system should behave and the comparison of simulations with measurements may indicate the need for

troubleshooting. Finally, the third technique is analytical simulation. Unlike the first two techniques, this provides an understanding of how individual components contribute and points the way for optimizations to be verified by numerical simulations and measurements.

4.2.1 *Experiment*

Inject an input signal with known charge using a pulse generator set to approximate the shape of the detector signal (to include possible ballistic deficit). Measure the pulse height spectrum using a multichannel pulse height analyzer. The peak centroid yields the signal magnitude and the peak width yields the noise (FWHM = $2.35Q_n$ rms)

If pulse-height digitization is not practical, one can measure total noise at the output of the pulse shaper using an rms voltmeter, a spectrum analyzer, or an oscilloscope. All of these techniques require some attention to proper instrumentation and measurement techniques.

1. Measure the noise level.

   (a) *Voltmeter:* The voltmeter must have adequate sensitivity and a sufficiently large bandwidth to include the full noise spectrum. Most high-frequency voltmeters measure peak amplitude, although the scale reads rms The rms reading is only correct for sinusoidal signals, where the peak-to-rms ratio is known. For a Gaussian distribution the peak reading depends on the averaging time, which is not always known. Since measuring the total noise of a spectral distribution requires integrating over the noise power, the correct instrument is a "true rms" voltmeter, which uses thermal sensors or electronically squares the input signal and displays the square root. Many true rms voltmeters are designed for low-frequency applications, so it is important to verify that the bandwidth is adequate; preferably the rated bandwidth extends well beyond the upper frequency cutoff of the system. It also doesn't hurt to read the operation manual. This is a broadband measurement, so it is prone to extraneous signals. Pulse shapers are often sensitive to frequencies commonly used for radio and television transmitters or industrial RF generators. Contamination by these signals will increase the measured value (see Chapter 9).

   (b) *Spectrum analyzer:* A spectrum analyzer shows the distribution of the noise *vs.* frequency and is a powerful diagnostic tool. Signals from radio or TV stations and RF generators appear as discrete peaks in the spectrum and can be eliminated from the digitized output. Again, it is important to verify that the analyzer measures true rms The spectrum analyzer provides discrete measurement values in N frequency bins Δf_n. Since the measured noise level depends on the resolution bandwidth, one must also verify that the indicated bandwidth is the

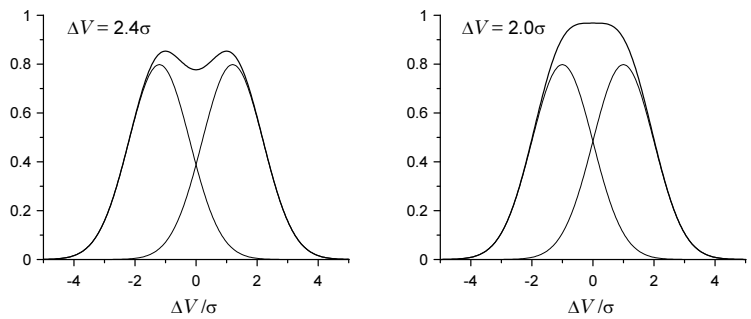

FIG. 4.6. The "dip" between two Gaussian distributions disappears when the spacing is reduced to 2σ. This can be used to measure noise with an oscilloscope.

noise bandwidth. Spectrum analyzers are available that indicate directly correct values of $\mathrm{V}/\sqrt{\mathrm{Hz}}$. The total noise is obtained by summing over the noise powers from all frequency bins

$$V_{no} = \sqrt{\sum_{n=0}^{N} [v_{no}^2(n) \cdot \Delta f]} \ .$$

(c) *Oscilloscope:* This is the most popular method and also the most prone to errors in the noise measurement. Since the noise distribution is Gaussian, its envelope is not well defined and the perceived boundaries depend on the intensity setting of the oscilloscope display. Thus, the "measured" value tends to lie between 2 and 3σ, where σ is equal to the rms noise. The accuracy and reproducibility of this measurement can be improved by exploiting the fact that two identical Gaussian distributions spaced by twice their rms value merge into a uniform distribution at the peak. To apply this, operate the oscilloscope in two-channel mode with a continuous trigger and the horizontal sweep rate set so the noise appears as a continuous band. Apply the amplifier output to both channels (AC coupled). Start with a large baseline offset; two bright bands of noise will be visible, separated by a dark band. Change the offset until the dark band just disappears, remove the signal, and measure the difference between the two baselines. As shown in Figure 4.6 the baseline difference is 2σ, *i.e.* twice the rms noise level.

2. With an oscilloscope measure the magnitude of the output signal V_{so} for a known input signal, either from a detector or from a pulse generator set up to approximate the detector signal.
3. Determine signal-to-noise ratio $S/N = V_{so}/V_{no}$ and scale to obtain the equivalent noise charge

$$Q_n = \frac{V_{no}}{V_{so}} Q_s \ .$$

4.2.2 *Numerical simulation (e.g. SPICE)*

This can be done with the full circuit including all extraneous components. The procedure is analogous to a measurement with a spectrum analyzer.

1. In a small-signal AC analysis determine the output noise *vs.* frequency and integrate by taking the discrete sum

$$V_{no} = \sqrt{\sum_{n=0}^{N} \left[v_{no}^2(n) \cdot \Delta f \right]} \ .$$

2. From a time-domain (pulse) analysis determine the magnitude of the output signal V_{so} for an input that approximates the detector signal.
3. Calculate the equivalent noise charge

$$Q_n = \frac{V_{no}}{V_{so}} Q_s \ .$$

The SPICE analysis is quite useful, since it can also tabulate the noise contributions of all components. It is quite common for many small contributions to add up to a substantial fraction of the total noise. Furthermore, "redesigning" the system is much quicker in software than in hardware.

4.2.3 *Analytical simulation*

Both measurements and numerical simulations provide the total noise and quantify specific noise contributions, but are not very helpful in understanding how individual noise sources contribute. Thus, analytical calculations are the third essential component of noise analysis. This will be demonstrated in more detail, so this is just a brief outline.

1. Identify individual noise sources and refer them to the input. For each source k determine the spectral distribution $v_{ni,k}^2(f)$. Some noise sources are physically present at the input, so this technique can be applied directly. Others are further "downstream"; here one evaluates the noise spectrum at the output and divides by the gain to refer it to the input.
2. Calculate the total noise at the shaper output ($G(f) =$ gain)

$$V_{no} = \sqrt{\int_0^\infty |G(f)|^2 \left(\sum_k v_{ni,k}^2(f) \right) df} \ \equiv \ \sqrt{\int_0^\infty |G(\omega)|^2 \left(\sum_k v_{ni,k}^2(\omega) \right) d\omega} \ .$$

3. Determine the peak signal output V_{so} for a known input charge Q_s and realistic detector pulse shape.

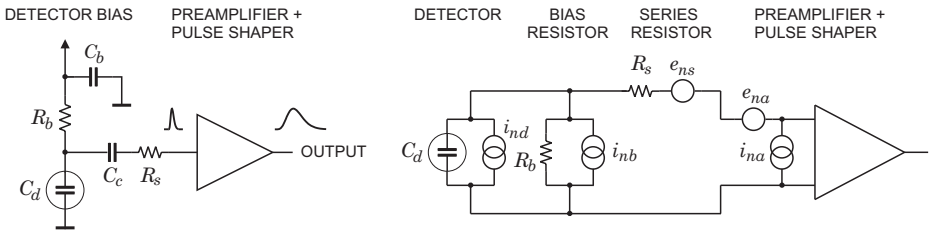

FIG. 4.7. A detector front-end circuit and its equivalent circuit for noise calculations.

4. Calculate the equivalent noise charge

$$Q_n = \frac{V_{no}}{V_{so}} Q_s \ .$$

4.3 Noise analysis of a detector and front-end amplifier

To determine how the pulse shaper affects the signal-to-noise ratio, consider the detector front-end in Figure 4.7. The detector is represented by the capacitance C_d and bias voltage is applied through the resistor R_b. The bypass capacitor C_b shunts any external interference coming through the bias supply line to ground. For high-frequency signals this capacitor appears as a low impedance, so for sensor signals the "far end" of the bias resistor is connected to ground and is parallel to the sensor. The coupling capacitor C_c blocks the sensor bias voltage from the amplifier input, which is why a capacitor serving this role is also called a "blocking capacitor". The series resistance R_s represents any resistance present in the connection from the sensor to the amplifier input. This includes the resistance of the sensor electrodes, the resistance of the connecting wires or traces, any resistance used to protect the amplifier against large voltage transients ("input protection"), and parasitic resistances in the input transistor.

The following analysis implicitly includes a constraint on the bias resistance. If the time constant $R_b C_d$ is small compared to the peaking time of the shaper T_P, the sensor will discharge through R_b and much of the signal will be lost. Thus, we have the condition $R_b C_d \gg T_P$, or $R_b \gg T_P/C_d$. The bias resistor must be sufficiently large to block the flow of signal charge, so that all of the signal is available for the amplifier. The role of the bias resistor is often misunderstood, with the interpretation that the signal current generated in the sensor flows through R_b and the resulting voltage drop is measured. In a well-designed system practically no signal current flows through R_b.

In the sensor the primary signal is a charge, but in the electronics the primary quantities are voltage and current. We will ultimately express the output noise in terms of a charge to allow a direct comparison with the primary signal deposition, but to accomplish this we need to analyze the contributions of the noise voltage and current sources and the bandwidth of the amplifier.

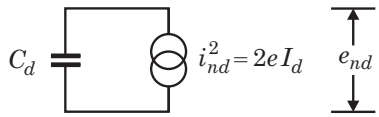

FIG. 4.8. Model showing how the sensor noise current translates into a noise voltage.

In this analysis we'll assume a voltage amplifier, so all noise contributions will be calculated as a noise voltage appearing at the amplifier input. We could also analyze the individual noise contributions in terms of current and obtain the same result as long as we're consistent. As shown in Chapter 3, the signal-to-noise ratio is unaffected by feedback, so the results apply to all configurations.

Steps in the analysis are:

1. Determine the frequency distribution of all noise voltages presented to the amplifier input from all individual noise sources.
2. Integrate over the frequency response of the shaper (for simplicity a CR-RC shaper) and determine the total noise voltage at the shaper output.
3. Determine the output signal for a known input signal charge. The equivalent noise charge (ENC) is the signal charge for $S/N = 1$.

The equivalent circuit for the noise analysis in Figure 4.7 includes both current and voltage noise sources. The "shot noise" i_{nd} of the sensor leakage current is represented by a current noise generator in parallel with the sensor capacitance. As discussed in Chapter 3, resistors can be modeled either as voltage or current generators. Generally, resistors shunting the input act as noise current sources and resistors in series with the input act as noise voltage sources.

This is why some in the detector community refer to current and voltage noise as "parallel" and "series" noise, although it is not clear why the physically meaningful terms "current" and "voltage" need to be replaced.

Since the bias resistor effectively shunts the input, as the capacitor C_b passes current fluctuations to ground, it acts as a current generator i_{nb} and its noise current has the same effect as the shot noise current from the detector. One can also model the shunt resistor as a noise voltage source and obtain the result that it acts as a current source. Choosing the appropriate model merely simplifies the calculation. Any other shunt resistances can be incorporated in the same way. Conversely, the series resistor R_s acts as a voltage generator. The electronic noise of the amplifier is described fully by the voltage and current sources e_{na} and i_{na}. Next, we discuss the individual noise contributions.

4.3.1 *Detector bias current*

All current sources shunting the input have infinite source resistance, the amplifier's input impedance is infinite, and current flow through R_b is negligible, so the noise current of the sensor flows through the detector capacitance as shown in Figure 4.8. The resulting voltage presented to the amplifier

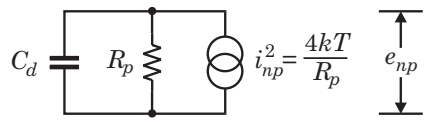

FIG. 4.9. Equivalent circuit for the noise contribution of resistors shunting the input.

$$e_{nd}^2 = i_{nd}^2 \frac{1}{(\omega C_d)^2} = 2eI_d \frac{1}{(\omega C_d)^2} \ , \tag{4.6}$$

so the noise voltage decreases with increasing frequency (shorter shaping time). The "white" noise spectrum of the sensor shot noise current has become "colored" (frequency-dependent) by flowing through the sensor capacitance. The same applies to any noise current injected into the input.

4.3.2 *Parallel resistance*

Any shunt resistance R_p acts as a noise current source. In Figure 4.7 the only shunt resistance is the bias resistor R_b. Additional shunt components in the circuit are the bias noise current source, which has infinite resistance by definition, and the sensor capacitance.

As shown in Figure 4.9 the noise current flows through the parallel combination of the resistance R_p and the sensor capacitance C_d. The resulting noise voltage applied to the amplifier input

$$\begin{aligned} e_{np}^2 &= \frac{4kT}{R_p} \left(\frac{R_p \cdot (-\mathbf{i}/\omega C_d)}{R_p + (-\mathbf{i}/\omega C_d)} \right)^2 \\ |e_{np}|^2 &= \frac{4kTR_p}{1 + (\omega R_p C_d)^2} \ . \end{aligned} \tag{4.7}$$

Integrating this distribution over all frequencies yields

$$v_n^2 = \int_0^\infty e_{np}^2(\omega) d\omega = \int_0^\infty \frac{4kTR_p}{1 + (\omega R_p C_d)^2} d\omega = \frac{kT}{C_d} \ , \tag{4.8}$$

which is independent of R_p. Commonly referred to as "kTC" noise (since the equivalent charge ${Q_n}^2 = {v_n}^2 {C_d}^2 = kTC_d$), this contribution is often erroneously interpreted as the "noise of the detector capacitance".

An ideal capacitor has no thermal noise; all noise originates in the resistor. So, why is the result independent of R_p? R_p determines the primary noise, but also the noise bandwidth of this subcircuit. As R_p increases, its thermal noise increases, but the noise bandwidth decreases, making the total noise independent of R_p.

This could lead one to believe that the noise contribution is independent of shaping time. However, if one integrates e_{np} over a bandwidth-limited system

$$v_n^2 = \int_0^\infty 4kTR_p \left| \frac{G(\mathbf{i}\omega)}{1 - \mathbf{i}\omega R_p C_d} \right|^2 d\omega \tag{4.9}$$

with $T_P \ll R_p C_d$ – as of interest here – the total noise decreases as R_p is made larger.

4.3.3 *Series resistance*

The noise voltage generator associated with the series resistance R_s is in series with the input, so it simply contributes

$$e_{nr}^2 = 4kTR_s \ . \tag{4.10}$$

4.3.4 *Amplifier input noise*

As discussed above, amplifiers have both white noise and excess ("$1/f$") noise components, so the equivalent input noise voltage

$$e_{na}^2 = e_{nw}^2 + \frac{A_f}{f} \ . \tag{4.11}$$

This noise voltage generator also adds in series with the other sources. The input noise current of the amplifier has the same effect as the detector bias current, so the analysis given in Section 4.3.1 can be applied.

4.3.5 *Cumulative input noise voltage*

The noise voltage generators are in series and simply add in quadrature, so the noise voltage spectral density at the amplifier input

$$\begin{aligned} e_{ni}^2(f) &= e_{nd}^2 + e_{np}^2 + e_{nr}^2 + e_{na}^2 = \\ &= \frac{2eI_d}{(\omega C_d)^2} + \frac{4kTR_p}{1 + (\omega R_p C_d)^2} + 4kTR_s + e_{na} + \frac{i_{na}^2}{(\omega C_d)^2} \ . \end{aligned} \tag{4.12}$$

Integrating over the cumulative noise spectrum at the amplifier output

$$V_{no}^2 = \int_0^\infty e_{no}^2(f) df = \int_0^\infty e_{ni}^2(f) \left| A_v \right|^2 df \tag{4.13}$$

and comparing to the output for a known input signal yields the signal-to-noise ratio. In this example the shaper is a simple CR-RC shaper, whose frequency

response of the is the product of the differentiator and integrator transfer functions

$$A_v = \frac{1}{1 - \dfrac{\mathbf{i}}{\omega\tau_d}} \cdot \frac{1}{1 + \mathbf{i}\omega\tau_i} \ , \tag{4.14}$$

so

$$|A|^2 = \frac{\tau_d^2}{(\tau_d + \tau_i)^2 + \left(\omega\tau_i\tau_d - \frac{1}{\omega}\right)^2} \ . \tag{4.15}$$

Integrating the noise spectrum at the amplifier output yields

$$\begin{aligned} V_{no}^2 = \frac{1}{4C_d^2}\left(\frac{4kT}{R_p} + 2eI_d + i_{na}^2\right)\frac{\tau_d^2}{\tau_d + \tau_i} + \\ + (4kTR_S + e_{na}^2)\frac{\tau_d}{\tau_i(\tau_d + \tau_i)} + A_f\frac{\tau_d^2}{\tau_d^2 + \tau_i^2}\log\left(\frac{\tau_d}{\tau_i}\right) \ . \end{aligned} \tag{4.16}$$

4.3.6 *Equivalent noise charge*

A signal charge Q_s yields a voltage at the amplifier input $V_s = Q_s/C_d$. Assume that the signal chain has no additional gain, so the signal voltage at the shaper input is $V_{si} = V_s$. Then the peak voltage at the shaper output (Gillespie 1953)

$$\begin{aligned} V_{so}(T_P) &\equiv A_{vs}V_s = \\ &= V_s\frac{\tau_d}{\tau_d - \tau_i} \cdot \left(\exp\left[-\frac{\tau_i}{\tau_d - \tau_i}\log\left(\frac{\tau_d}{\tau_i}\right)\right] - \exp\left[-\frac{\tau_d}{\tau_d - \tau_i}\log\left(\frac{\tau_d}{\tau_i}\right)\right]\right) \end{aligned} \tag{4.17}$$

and for $\tau_i = \tau_d$, $V_{so}(T_p) = \exp(-1)$. The signal-to-noise ratio

$$\frac{S}{N} = \frac{V_{so}(T_P)}{V_{no}} \ . \tag{4.18}$$

Expressed in terms of charge, $S/N = Q_s/Q_n$, so the equivalent noise charge

$$Q_n = Q_s\frac{V_{no}}{V_{so}} \ . \tag{4.19}$$

Analysis of this result shows that for a given differentiation time constant, minimum noise obtains when the differentiation and integration time constants are equal $\tau_i = \tau_d \equiv \tau$. Since τ_d sets the lower cutoff frequency, τ_i dominates the noise bandwidth. If $\tau_i \ll \tau_d$, the noise bandwidth is excessive. Reducing the upper cutoff frequency by increasing τ_i reduces the noise more than the signal until $\tau_i = \tau_d$ where further increases in τ_i reduce the signal more than the noise. This relationship between time constants applies only to a simple *CR-RC* shaper, but not to more sophisticated configurations. Even for a *CR-RC* shaper this criterion only applies when the differentiation time constant is the primary parameter, *i.e.* when the pulse width must be constrained. When the

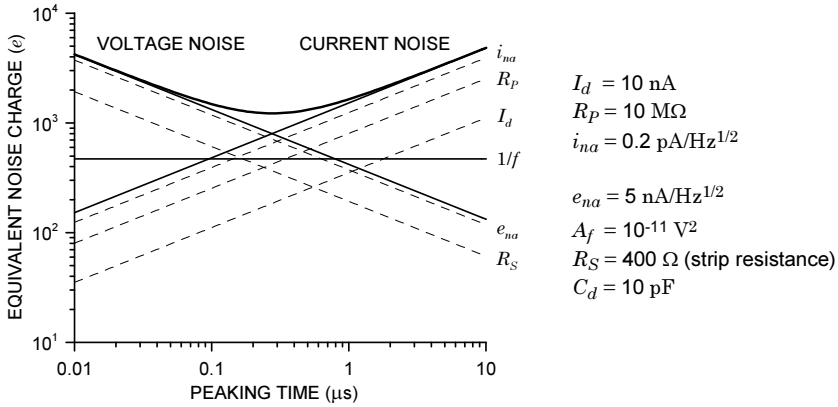

FIG. 4.10. Equivalent noise charge *vs.* *CR-RC* peaking time. The parameters contributing to the noise are shown to the right, grouped by current and voltage noise sources. With increasing shaping time the current noise contributions increase and the voltage white noise contributions decrease. The $1/f$ voltage noise portion is independent of shaping time. The cumulative noise assumes a minimum at the shaping time where the voltage and current noise contributions are equal.

rise time, *i.e.* the integration time constant, is the primary consideration, it is advantageous to make $\tau_d > \tau_i$, since the signal will increase more rapidly than the noise, as was shown above. One example is a system where the pulse height is sampled synchronously, *e.g.* in a colliding beam detector.

Using equal time constants $\tau_i = \tau_d \equiv \tau$, the equivalent noise charge using a *CR-RC* shaper

$$Q_n^2 = \left(\frac{\epsilon^2}{8}\right)\left[\left(2eI_d + \frac{4kT}{R_P} + i_{na}^2\right)\cdot\tau + \left(4kTR_S + e_{na}^2\right)\cdot\frac{C^2}{\tau} + 4A_fC^2\right] \quad (4.20)$$

The prefactor $\epsilon^2/8 = \exp(2)/8 = 0.924$ normalizes the noise to the signal gain. The first term combines all noise current sources and increases with shaping time. The second term combines all noise voltage sources and decreases with shaping time, but increases with sensor capacitance, due to the conversion of signal charge to voltage. The third term is the contribution of $1/f$ noise and, as a voltage source, also increases with sensor capacitance. The $1/f$ term is independent of shaping time, since for a $1/f$ spectrum the total noise depends on the ratio of upper to lower cutoff frequency, which depends only on shaper topology, but not on the shaping time.

Although Figure 4.7 shows only the detector capacitance shunting the input, the amplifier's input capacitance and any other stray capacitance have the same effect. Thus, the *total* input capacitance C is relevant for the equivalent noise charge.

Figure 4.10 shows a typical plot of equivalent noise charge *vs.* peaking time, with representative contributions from the individual noise sources. At short peaking times the voltage noise dominates, whereas at long peaking times the current noise takes over. Q_n assumes a minimum when the current and voltage noise contributions are equal. Increasing the voltage contribution, *e.g.* by increasing the sensor capacitance, shifts the voltage noise asymptote upwards and the optimum peaking time to longer values. Conversely, increased current noise shifts the optimum to smaller peaking times. The presence of $1/f$ noise flattens the minimum. The next section shows examples to illustrate this.

The contribution of noise current *vs.* shaping time is intuitive when viewed in the time domain. Since every shaper also acts as an integrator, one can view the total shot noise as the result of "counting electrons". Assume an ideal integrator that records all charge uniformly within a time T. The number of electron charges measured is $N_e = I_d T/e$. The associated noise is the fluctuation in the number of electron charges recorded $\sigma_n = \sqrt{N_e} \propto \sqrt{T}$.

Does this also apply to an AC-coupled system, where no DC component flows, so the integrated net charge is zero? Since shot noise is a fluctuation, the current undergoes both positive and negative excursions. Although the DC component is not passed through an AC coupled system, the excursions are. Since, on the average, each fluctuation requires a positive and a negative zero crossing, the process of "counting electrons" is actually the counting of zero crossings, which in a detailed analysis yields the same result.

For quick estimates one can use the following formula, which assumes an FET amplifier (negligible i_{na}).

$$Q_n^2 = 12 \left[\frac{e^2}{\mathrm{nA}\cdot\mathrm{ns}}\right] I_d \tau + 6\cdot 10^5 \left[\frac{e^2 \mathrm{k}\Omega}{\mathrm{ns}}\right] \frac{\tau}{R_p} + 3.6\cdot 10^4 \left[\frac{e^2 \mathrm{ns}}{(\mathrm{pF})^2 (\mathrm{nV})^2/\mathrm{Hz}}\right] e_n^2 \frac{C^2}{\tau} \ . \tag{4.21}$$

Here the equivalent noise charge is expressed in electrons. To convert to energy, $Q_n = 1\,e$ corresponds to 3.6 eV in Si and 2.9 eV in Ge.

4.4 Examples

4.4.1 *Photodiode readout*

The system described in this example was designed for medical imaging (positron emission tomography – PET). Images are formed by introducing a positron-emitting tracer into the body and recording the collinearly emitted annihilation gamma rays. The system utilizes both the direction of the emitted gamma and the time-of-flight (Choong *et al.* 2002). A detector module is shown in Figure 4.11. To obtain high efficiency for the 511 keV gamma radiation originating from electron-positron annihilation, the primary sensor is a scintillator coupled to a photomultiplier tube. To obtain the desired position resolution the cross section of each scintillator is about $3\times 3\,\mathrm{mm}^2$. The photomultiplier is used to provide the timing information. To reduce costs, an array of 64 BGO crystals is coupled to a one inch square photomultiplier tube with a single anode. An array of silicon

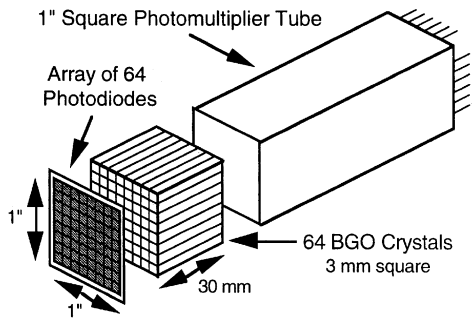

FIG. 4.11. An array of 64 BGO scintillators is read out by a photomultiplier (timimg) and a photodiode array (position).

photodiodes senses a portion of the scintillation light to provide the position information.

The signals are of order $1000\,e$, so low noise is essential. Achieving high quantum efficiency at the 480 nm emission wavelength of the BGO scintillators requires a thin dead layer on the photodiode to maximize quantum efficiency. A thin electrode has a high electrical resistance, which can be a significant noise source, so the photodiode and readout amplifier must be analyzed together. The photodiodes were especially designed for high quantum efficiency and low leakage current (Holland, Wang, and Moses 1997 and see Appendix A) and the readout amplifier is a custom designed IC. Figure 4.12 shows the quantum efficiency and the photodiode leakage ("dark") current and capacitance *vs.* bias voltage. As the depletion width increases with bias voltage the dark current increases, but the diode capacitance decreases. Since the current noise contribution increases with the square root of current, whereas the voltage noise contribution decreases

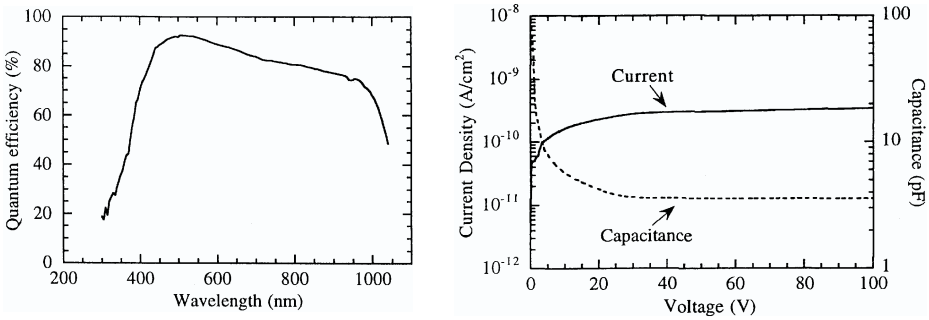

FIG. 4.12. Quantum efficiency of the photodiode array (left) and photodiode capacitance and leakage current *vs.* bias voltage (right). (Courtesy of S.E. Holland.)

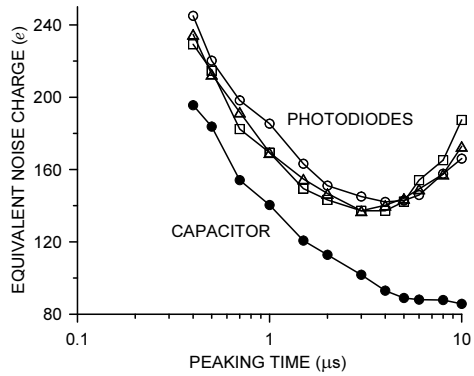

FIG. 4.13. Noise *vs.* peaking time with a purely capacitive input load (no shot noise) and with photodiodes.

linearly with capacitance, operating at minimum capacitance provides the minimum noise charge.

Figure 4.13 shows the noise *vs.* peaking time. The bottom curve shows the noise with a capacitor representing the photodiode. In the absence of diode leakage current the noise decreases as the peaking time is raised up to about 6 μs and then plateaus, an indication of $1/f$ noise. Connecting the photodiode introduces the current noise due to the leakage current, which increases with peaking time and yields a noise minimum at about 2 μs. Figure 4.14 shows an energy spectrum measured with the complete readout system.

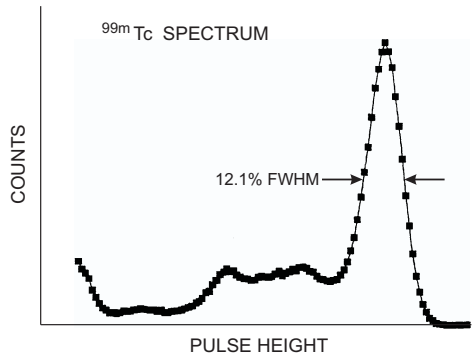

FIG. 4.14. Energy spectrum measured with a BGO scintillator and photodiode readout system.

4.4.2 *High-rate x-ray spectroscopy*

This example uses a silicon strip detector, but not in its usual role of providing position sensing, but in an x-ray spectroscopy system that achieves high rates by distributing the total rate over many parallel readout channels (Ludewigt *et al.* 1994). For example, a single readout channel with a shaping time of $1\,\mu$s limits the pulse rate to about $10^4\,\mathrm{s}^{-1}$. If a sensor is electrically segmented by subdividing its electrode into an array of strips or pixels, each electrode will capture only a fraction of the total rate. Thus, if a sensor electrode is segmented N-fold, the rate in a single electrode will be only $1/N$ of the total. This only applies to localized charge deposition, as is the case for gamma and x-rays of $< 30\,\mathrm{keV}$ in silicon, where the energy deposition is dominated by photoelectric absorption (Figure 1.20). Apart from reducing the rate per channel, segmentation also reduces the capacitance per electrode, so one can achieve the same noise level at a smaller shaping time, which further increases the rate capability. This detector was designed for use at synchrotron light sources, where the small beam spot allows the use of short strip electrodes of 2 mm length. The readout pitch is $100\,\mu$m. Figure 4.15 shows a cross section of the sensor (Ludewigt *et al.* 1996).

The sensor is read out by a custom-designed monolithic integrated circuit with 64 parallel readout channels (Krieger, Kipnis, and Ludewigt 1998). Each readout channel includes a charge-sensitive amplifier, a CR-RC^2 shaper with adjustable shaping time and gain, and an output buffer (Figure 4.16). The shaper outputs are fed to a bank of parallel analog-to-digital converters.

Figure 4.17 shows initial results of the noise *vs.* peaking time. Data are shown for the electronics with an open input and with purely capacitive loads of 0.38 and 0.75 pF. Here the lower three curves show the noise decreasing with increasing peaking time, but leveling off above 2 or $4\,\mu$s. The lower capacitance yields lower noise, as expected for the voltage noise contribution. At small peaking times the noise is dominated by the white voltage noise component, so it decreases

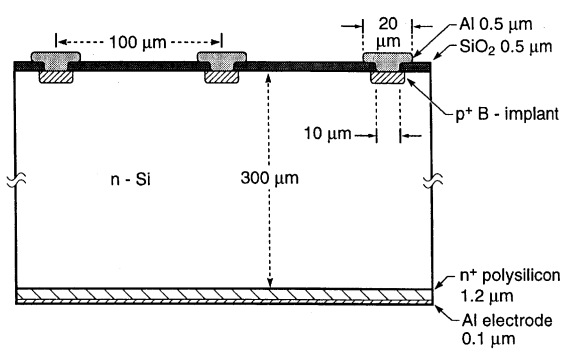

FIG. 4.15. Cross-section of the silicon strip sensor used for high-rate x-ray spectroscopy.

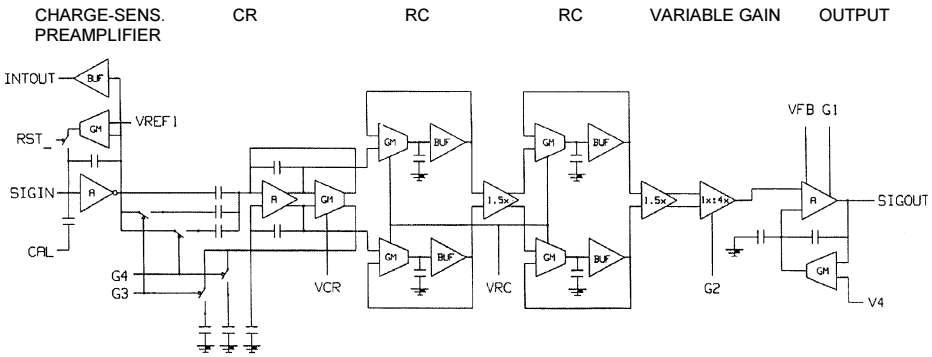

FIG. 4.16. Block diagram of a single channel of the high-rate x-ray detector readout IC. CR time constants are changed by switching capacitors and the integrator time constant is adjusted by changing the bandwidth of the "RC" gain stages.

with peaking time; at larger peaking times the noise is dominated by the $1/f$ component, whose contribution is independent of peaking time. Attaching the strip sensor introduces the shot noise from the reverse bias current, so the noise increases at larger peaking times (round solid dots). Since the bias current is strongly dependent on temperature, it can be reduced by cooling the sensor, as shown by the solid square symbols. Now the minimum noise is again dominated

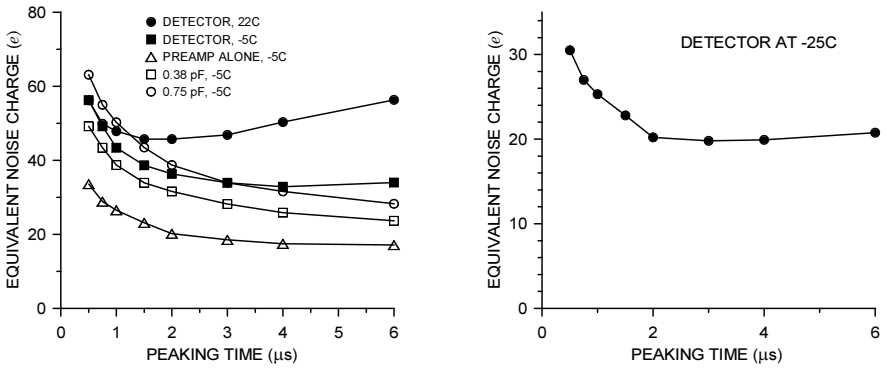

FIG. 4.17. Noise *vs.* peaking time for the high-rate x-ray detector, left the first prototype and right the optimized design. The open triangles show the electronic noise with an open input. The open squares and circles and show the noise for capacitive loads of 0.38 and 0.75 pF and the solid squares and circles show the noise with a sensor connected, both at room temperature and cooled to $-5\,^{\circ}$C. In the optimized high-rate x-ray detector system (right), at peaking times $> 2\,\mu$s the $1/f$ noise dominates, so the noise remains roughly constant.

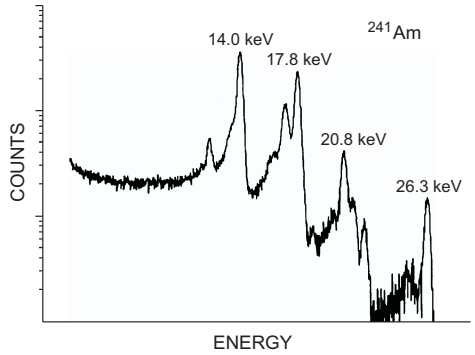

FIG. 4.18. Spectrum measured with the high-rate x-ray detector system.

by the $1/f$ noise of the electronics. Based on these results both the detector and the preamplifier were optimized to reduce both the white and $1/f$ contributions; the results are shown in the second panel of Figure 4.17 (Ludewigt *et al.* 1996). Now the current noise from the sensor is negligible because of cooling and the "flat" noise *vs.* peaking time indicates that $1/f$ noise dominates. Figure 4.18 shows an x-ray spectrum measured with the system. In this specific application, digital signal processing (discussed later) now produces comparable results with single channel readouts, but the principle of achieving high overall data rates through segmentation and parallel processing is the key to large-scale detectors at high-luminosity colliders and other applications. This will be discussed in Chapter 8.

4.5 Noise analysis in the time domain

The noise analysis of shapers is rather straightforward if the frequency response is known. On the other hand, since we are primarily interested in the pulse response, shapers are often designed directly in the time domain, so it seems more appropriate to analyze the noise performance in the time domain also.

Clearly, one can take the time response and Fourier transform it to the frequency domain, but this approach becomes problematic for time-variant shapers. The *CR-RC* shapers discussed up to now utilize filters whose time constants remain constant during the duration of the pulse, *i.e.* they are time-invariant. Many popular types of shapers utilize signal sampling or change the filter constants during the pulse to improve pulse characteristics, *i.e.* faster return to baseline or greater insensitivity to variations in detector pulse shape. These time-variant shapers cannot be analyzed in the manner described above. Various techniques are available, but some shapers can be analyzed only in the time domain. A commonly used time-variant filter is the correlated double-sampler (CDS).

The principle of a CDS shaper is shown in Figure 4.19. Input signals are superimposed on a slowly fluctuating baseline. To remove the baseline fluctua-

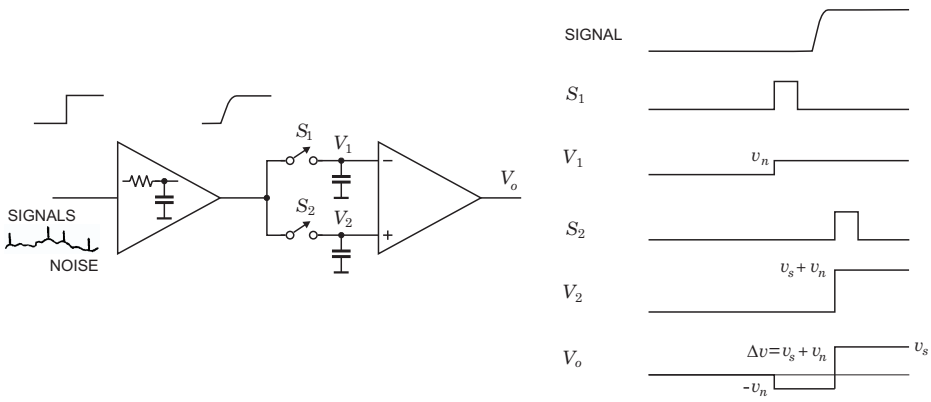

FIG. 4.19. Principle of a shaper using correlated double sampling.

tions the baseline is sampled prior to the signal by momentarily closing switch S_1 and the baseline level v_n is stored on the capacitor at the inverting input of the amplifier. Next, the signal plus baseline is sampled (S_2) and the signal plus baseline is stored on the capacitor at the non-inverting input. The amplifier forms the difference of the two inputs, so the net result removes the baseline and leaves the signal. The prefilter in the input amplifier is critical to limit the noise bandwidth of the system. Filtering after the sampler is useless, as noise fluctuations on time scales shorter than the sample time will not be removed. Here the sequence of filtering is critical, unlike a time-invariant linear filter where the sequence of filter functions can be interchanged. Correlated doubling sampling is widely used in monolithically integrated circuits, as many CMOS processes provide only capacitors and switches, but no resistors. A quantitative analysis of a CDS shaper will be presented in Section 4.5.4.

4.5.1 *Principles of noise analysis in the time domain*

The basis of noise analysis in the time domain is Parseval's theorem, which relates the amplitude response $A(f)$ to the time response $F(t)$:

$$\int_0^\infty |A(f)|^2 \, df = \int_{-\infty}^{\infty} [F(t)]^2 \, dt \ . \tag{4.22}$$

The left-hand side is essentially integration over the noise bandwidth. However, we'll use a more intuitive approach, first described in detail by F.S. Goulding (1972).

Noise is represented as a randomly recurring series of pulses. The magnitude of the noise is set by the rate of noise pulses. The pulse shapes are chosen to have a frequency spectrum corresponding to typical noise sources. For simplicity

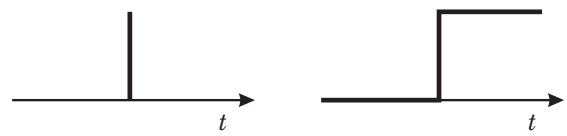

FIG. 4.20. A delta impulse (left) is infinitesimally short, but has unit area. Its frequency spectrum is uniform, *i.e.* white. A step impulse (right) increases to unit amplitude and stays at this level. Its amplitude (not power!) spectrum is proportional to $1/f$.

we'll just consider white noise sources. $1/f$ noise can be analyzed in the time domain, but it is more involved (Pullia 1998).

1. Voltage noise
   The frequency spectrum of thermal voltage noise sources at the input of the detector system is "white", *i.e.*
   $$A(f) = \text{const.} \tag{4.23}$$
   This is the spectrum of a δ impulse, shown in Figure 4.20.
2. Current noise
   The spectral density of the white shot noise current flowing through the capacitive reactance of the sensor is inversely proportional to frequency, *i.e.*
   $$A(f) \propto \frac{1}{f} \ . \tag{4.24}$$
   This is the spectrum of a step impulse, as shown in Figure 4.20. Intuitively, this is the result of a current pulse integrated on the sensor capacitance, which yields a fixed charge and a resulting voltage step. Note that here the amplitude falls with $1/f$, unlike $1/f$ noise, where the *power* falls off with $1/f$, so its amplitude has a $1/\sqrt{f}$ dependence.

The noise at the input of the amplifier is represented by a sequence of δ and step pulses whose rates determine the noise level. The shape of the primary noise pulses is modified by the pulse shaper; δ pulses become longer, step pulses are shortened. The shaper's response to step pulses is the same as for a signal. Delta pulses can be treated as two infinitesimally spaced step pulses of opposite sign, so the shaper output is the derivative of the step response. The noise level at a given measurement time T_m is determined by the cumulative effect (superposition) of all noise pulses occurring prior to T_m. Their individual contributions at $t = T_m$ are described by the shaper's "weighting function" $W(t)$ (Radeka 1972, 1974, Goulding and Landis 1982, Gatti and Manfredi 1986).

Consider a single noise pulse occurring in a short time interval prior to the measurement performed at time T_m. A typical current noise pulse at the shaper output is shown in Figure 4.21. The amplitude at $t = T_m$ is $a_n = W(T_m)$. If, on

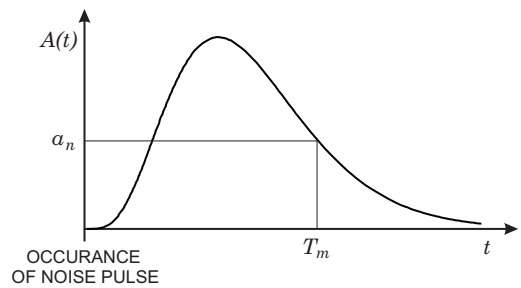

FIG. 4.21. A current noise pulse at the shaper output, occurring prior to the measurement performed at time T_m. The measured amplitude is a_n.

the average, $n_n dt$ noise pulses occur within a time interval dt, the fluctuation of their cumulative signal level at $t = T_m$ is proportional to $\sqrt{n_n dt}$. The magnitude of the baseline fluctuation is

$$\sigma_n^2(T) \propto n_n \left[W(T)\right]^2 dt \ . \tag{4.25}$$

For all noise pulses occurring prior to the measurement the cumulative fluctuation

$$\sigma_n^2 \propto n_n \int_0^\infty \left[W(t)\right]^2 \ dt \ , \tag{4.26}$$

where n_n determines the magnitude of the noise and the integral $\int_0^\infty \left[W(t)\right]^2 dt$ describes the noise characteristics of the shaper – the "noise index".

4.5.2 *The weighting function*

What is the weighting function $W(t)$? As noted above, a noise current pulse is represented by a step function. Thus, the corresponding noise pulse at the shaper output is the step response, as shown in Figure 4.1 for a *CR-RC* shaper, and the weighting function for current noise $W_i(t)$ is the shaper output as measured with an oscilloscope when a step is applied to the shaper input (or a delta current pulse is applied at the detector).

As noted above, voltage noise is represented by a delta impulse, which can be represented as the derivative of the step response. Thus, the weighting function for voltage noise

$$W_v(t) = \frac{d}{dt} W_i(t) \equiv W_i'(t) \ . \tag{4.27}$$

Figure 4.22 shows the weighting functions for two representative shaper responses, a $CR\text{-}RC^4$ and a trapezoidal shaper. The extended "flat top" of the trapezoidal shaper is advantageous when ballistic deficit is a concern.

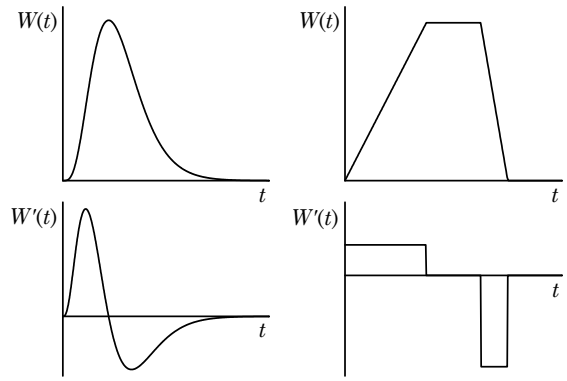

FIG. 4.22. Weighting functions for shapers with a CR-RC^4 (left) and "trapezoidal" (right) response.

Since the total noise fluctuation

$$\sigma_n^2 \propto n_n \int_0^\infty [W(t)]^2 \, dt$$

is determined by the integral over the weighting functions, the goal is to minimize the overall area by reducing the current noise contribution concurrently with minimizing the derivatives to reduce the voltage noise contribution. From these criteria we see that for a given pulse duration a symmetrical pulse with linear transitions provides the best voltage noise performance, as this minimizes the derivatives.

Quantitatively, the equivalent noise charge expressed in terms of the weighting function (Radeka 1974)

$$Q_n^2 = \frac{1}{2} i_n^2 \int_{-\infty}^{\infty} [W(t)]^2 dt + \frac{1}{2} C^2 e_n^2 \int_{-\infty}^{\infty} [W'(t)]^2 dt \ . \tag{4.28}$$

Since the integrals scale with the selected time scale, the weighting functions can be expressed in terms of a characteristic time. For a CR-RC^n shaper the peaking time is a good measure, but one could also choose the half-width of the pulse. The choice of the characteristic time is somewhat arbitrary, but as we will see below, the analysis often suggests the most meaningful choice. This leads to a general formulation of the equivalent noise charge

$$Q_n^2 = i_n^2 F_i T_S + e_n^2 F_v \frac{C^2}{T_S} + F_{vf} A_f C^2 \ , \tag{4.29}$$

where C is the sum of all capacitances shunting the input, the shape factors F_i, F_v, and F_{vf} depend on the shape of the pulse determined by the shaper for a

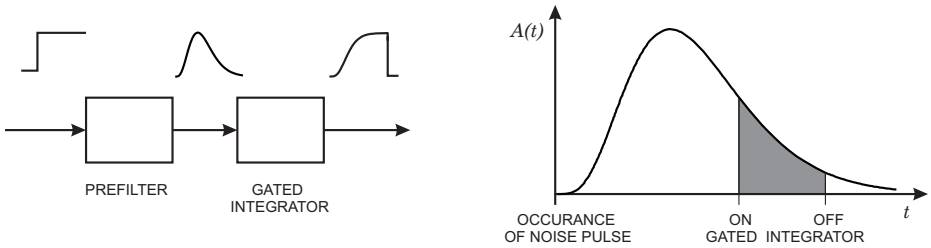

FIG. 4.23. In a gated integrator the signal is integrated over a selectable time. In this example the integrator is switched on prior to the signal pulse and switched off when the integrated signal has reached its maximum. Only the portion of a noise pulse within the integration time contributes to the output (shaded area).

step input, and T_S is the characteristic time, whose choice depends on the type of shaper. The shape factors F_i and F_v are easily calculated:

$$F_i = \frac{1}{2T_S} \int_{-\infty}^{\infty} [W(t)]^2 \, dt \quad \text{and} \quad F_v = \frac{T_S}{2} \int_{-\infty}^{\infty} \left[\frac{dW(t)}{dt} \right]^2 dt \; . \tag{4.30}$$

For time-invariant pulse shaping $W(t)$ is simply the system's impulse response (the output signal seen on an oscilloscope when a delta pulse is applied at the detector input) with the peak output signal normalized to unity. Viewing the result in both the time and frequency domain offers a simple interpretation of the current and voltage noise contributions. Increasing the duration of the shaper output pulse increases the integration time, so in the picture of "counting electrons" the noise increases with increasing shaping time. Conversely, increasing the derivative of $W(t)$ raises the upper cutoff frequency, so the noise bandwidth increases and the voltage noise contribution increases with decreasing shaping time.

4.5.3 *Time-variant shapers*

Time domain analysis simplifies the calculations for time-invariant shapers, since with a digitizing oscilloscope one can readily measure the step response and then numerically take the derivative. For these shapers the time domain approach is convenient, but not essential. However, as noted above, time-domain analysis is essential for time-variant filters. This is true except for a few exceptions, where a time-variant system can be analyzed by a time-invariant analogy. This will be shown later for correlated double sampling.

A simple example of a time-variant shaper is a gated integrator with prefilter, as illustrated in Figure 4.23. A prefilter limits the noise bandwidth and limits the pulse duration. The gated integrator integrates the prefiltered signal during a selectable time interval (the "gate"). In this example, the integrator is switched

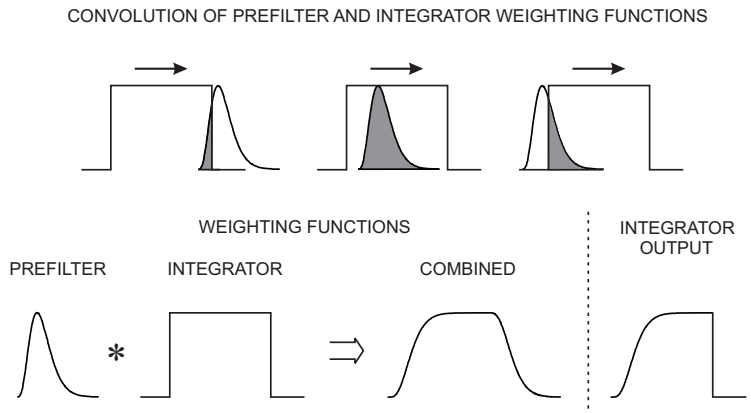

FIG. 4.24. Evolution of the weighting function for a time-invariant CR-RC^4 prefilter feeding a gated integrator. As shown above, the combined weighting function results from convolution of the prefilter and gated integrator weighting functions, *i.e.* the overlap integral as the gated integrator is "slid through" the prefilter response. Note that in a time-variant filter the weighting function and the output signal differ.

on prior to the signal pulse and switched off after a fixed time interval, selected to allow the output signal to reach its maximum.

Consider a noise pulse occurring prior to the "on time" of the integrator. The integrator captures only a portion of the noise signal, as shown in Figure 4.23. Since the time spacings between the begin of the noise pulses and the integration window are not correlated, the noise contribution is determined by integrating over all spacings. The weighting function is obtained by "sliding" the pulse through the integration window, *i.e.* by convolution. If W_1 is the weighting function of the *time-invariant* prefilter and W_2 is the weighting function of the *time-variant* stage, the weighting function

$$W(t) = \int_{-\infty}^{\infty} W_2(t') \cdot W_1(t - t')dt' \ . \tag{4.31}$$

Figure 4.24 shows the evolution of the weighting function for a time-invariant prefilter feeding a gated integrator. As for time-invariant shapers, the current noise contribution is determined by $W(t)$ and voltage noise contribution by the derivative $W'(t)$.

An important difference between time-variant and time-invariant pulse shapers is that in time-variant shapers the sequence of operations is critical. For this reason shapers with the same output pulse shape can exhibit different noise performance. For example, Goulding (1972) compares trapezoidal outputs implemented as time-invariant and time-variant shapers. The time-variant shaper has a current noise coefficient $F_i = 1.4$ and a voltage noise coefficient $F_v = 1.1$,

whereas the time-invariant shaper has $F_i = 0.8$ and $F_v = 2.2$. The time-variant trapezoid has more current noise and less voltage noise than the time-invariant version.

4.5.4 *Noise analysis of a correlated-double sample pulse shaper*

Correlated double sampling was already explained above. Now we apply time-domain noise analysis to this shaper, which is widely used in monolithic integrated circuit readouts. Referring to Figure 4.19, the signal is first sent through a prefilter, which is a simple RC low-pass filter with a time constant τ. Its step response is $1 - \exp(-t/\tau)$, where the peak amplitude is normalized to one. The weighting function is the convolution of the RC prefilter response with the two sequential samples of opposite sign. We assume that the amplitude is sampled during a time that is small compared to the time interval between samples T, so the sample pulse can be represented as a δ function. Thus, the weighting function is

$$\begin{aligned} t < 0: \;& W(t) = 0 \\ 0 \le t \le T: \;& W(t) = 1 - e^{-t/\tau} \\ t > T: \;& W(t) = \left(1 - e^{-t/\tau}\right) - \left(1 - e^{-(t-T)/\tau}\right) \\ & \quad = e^{-t/\tau}\left(e^{T/\tau} - 1\right) \; . \end{aligned}$$

4.5.4.1 *Current noise* The current (shot) noise contribution

$$Q_{ni}^2 = \frac{1}{2} i_n^2 \int_{-\infty}^{\infty} [W(t)]^2 dt \; . \tag{4.32}$$

The current noise index

$$\begin{aligned} N_i &= \int_{-\infty}^{\infty} [W(t)]^2 dt \\ N_i &= \int_0^T \left(1 - e^{-t/\tau}\right)^2 dt + \int_T^{\infty} e^{-2t/\tau}\left(e^{T/\tau} - 1\right)^2 dt \\ N_i &= T + \tau e^{-T/\tau} - \tau = T + \tau\left(e^{-T/\tau} - 1\right) \; . \end{aligned} \tag{4.33}$$

Increasing the sampling time T increases the noise, but since the expression in brackets is always negative, the effective integration time is less than T. To obtain the equivalent noise charge the noise must be normalized to the signal. The signal amplitude at the output of the prefilter at the time of the second sample

$$V_s/V_i = 1 - e^{-T/\tau} \; , \tag{4.34}$$

so that the equivalent noise charge due to the current noise becomes

$$Q_{ni}^2 = i_n^2 \tau \frac{1}{2(1-e^{-T/\tau})} \left(\frac{T/\tau}{1-e^{-T/\tau}} - 1 \right) \ . \tag{4.35}$$

For $T/\tau \gg 1$ the current noise Q_{ni} increases with $\sqrt{T}$.

Let's apply a reality check to this result. For pure shot noise the spectral noise density $i_n^2 = 2eI$. For $T = 0$ the noise cancels. In the limit where the sampling interval is much greater than the rise time of the prefilter, $T \gg \tau$,

$$Q_{ni}^2 \approx eI \cdot T \ ,$$

or expressed in electrons

$$Q_{ni}^2 \approx \frac{eI \cdot T}{e^2} = \frac{I \cdot T}{e} \tag{4.36}$$

$$Q_{ni} \approx \sqrt{N} \ , \tag{4.37}$$

where N is the number of electrons "counted" during the sampling interval T.

4.5.4.2 *Voltage noise* The voltage noise contribution

$$Q_{nv}^2 = C^2 e_n^2 \frac{1}{2} \int_{-\infty}^{\infty} [W'(t)]^2 dt \ . \tag{4.38}$$

The derivative of the weighting function

$$\begin{aligned} t < 0 : \ & W'(t) = 0 \\ 0 \le t \le T : \ & W'(t) = \frac{1}{\tau} e^{-t/\tau} \\ t > T : \ & W'(t) = \frac{1}{\tau} e^{-t/\tau} \left(1 - e^{T/\tau} \right) \ , \end{aligned}$$

and the voltage noise index

$$\begin{aligned} N_v &= \int_{-\infty}^{\infty} [W'(t)]^2 dt \\ N_v &= \int_0^T \left(\frac{1}{\tau} e^{-t/\tau} \right)^2 dt + \int_T^{\infty} \left(\frac{1}{\tau} e^{-t/\tau} \left(1 - e^{T/\tau} \right) \right)^2 dt \\ N_v &= \frac{1}{\tau} \left(1 - e^{-T/\tau} \right) \ . \end{aligned} \tag{4.39}$$

Normalizing to the signal amplitude at $t = T$, the equivalent noise charge due to voltage noise sources becomes

$$Q_{nv}^2 = \frac{C^2 e_n^2}{\tau} \frac{1}{2\left(1 - e^{-T/\tau}\right)} \ . \tag{4.40}$$

Note that the sampling time enters into the voltage noise contribution only through the amplitude normalization, whereas the current noise for $T/\tau \gg 1$ increases with $\sqrt{T}$.

We can also apply a reality check to this result. For $T = 0$ the noise index vanishes. In the limit $T \gg \tau$

$$Q_{nv}^2 = C^2 \cdot e_n^2 \cdot \frac{1}{2\tau} \ .$$

Compare this to the noise from an RC low-pass filter alone (*i.e.* the voltage noise at the output of the prefilter),

$$Q_n^2(RC) = C_i^2 \cdot e_n^2 \cdot \frac{1}{4\tau} \ .$$

From this we see that

$$\frac{Q_n(\mathrm{CDS})}{Q_n(RC)} = \sqrt{2} \ .$$

If the sample time is sufficiently large, the noise samples taken at the two sample times are uncorrelated, so the two samples simply add in quadrature and increase the prefilter's output noise by a factor $\sqrt{2}$.

4.5.4.3 *Total equivalent noise charge* The total equivalent noise charge

$$Q_n^2 = Q_{ni}^2 + Q_{nv}^2 = \frac{1}{2\left(1 - e^{-T/\tau}\right)} \left(i_n^2 \tau \left(\frac{T/\tau}{1 - e^{-T/\tau}} - 1 \right) + \frac{C^2 e_n^2}{\tau} \right) \ . \tag{4.41}$$

The noise charge depends on two shaper parameters, the prefilter time constant τ and the normalized sampling time T/τ. For any given value of T/τ, minimum noise obtains when the current and voltage noise terms are equal, which yields the optimum prefilter time constant

$$\tau^2 = \left(\frac{e_n^2 C^2}{i_n^2} \right) \frac{1 - e^{-T/\tau}}{\frac{T}{\tau} + e^{-T/\tau} - 1} \ . \tag{4.42}$$

Inserting the optimum time constant into eqn 4.41 gives the noise charge

$$Q_n^2 = \frac{i_n e_n C}{\left(1 - e^{-T/\tau}\right)} \sqrt{\frac{\frac{T}{\tau} - \left(1 - e^{-T/\tau}\right)}{1 - e^{-T/\tau}}} \ . \tag{4.43}$$

This yields minimum noise for $T/\tau = 1$ (1.0357 to be exact), but only for the optimum prefilter time constant

$$\tau^2 = 1.65 \left(\frac{e_n^2 C^2}{i_n^2} \right) \ . \tag{4.44}$$

For other values of τ the noise minimum occurs at different ratios T/τ.

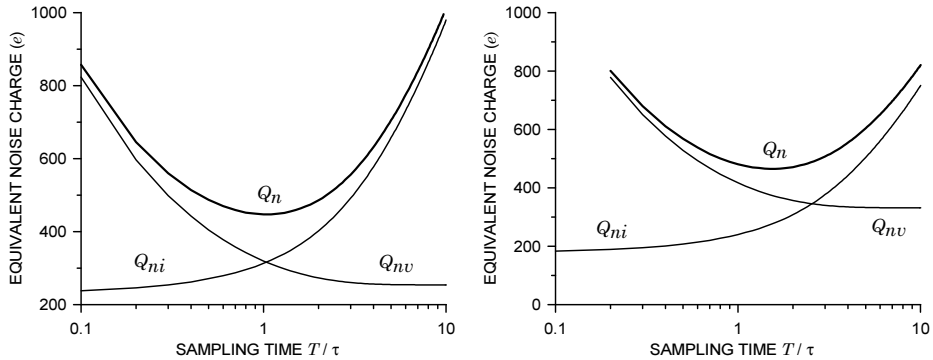

FIG. 4.25. Equivalent noise charge *vs.* sampling time for the optimum prefilter time constant of 1.7 μs (left) and a smaller time constant of 1.0 μs (right). The total input capacitance is 30 pF, the detector bias current 10 nA and the amplifier has an equivalent input noise of 2.5 nV/$\sqrt{\text{Hz}}$.

In the formalism developed for time-invariant shapers

$$Q_n^2 = i_n^2 T_S F_i + \frac{e_n^2 C^2}{T_S} F_v$$

we find that the characteristic time $T_S = \tau$ and

$$F_i = \frac{1}{2(1 - e^{-T/\tau})} \left(\frac{T/\tau}{1 - e^{-T/\tau}} - 1 \right) \tag{4.45}$$

$$F_v = \frac{1}{2\left(1 - e^{-T/\tau}\right)} \ . \tag{4.46}$$

Minimum noise obtains for the optimum values of τ and T/τ

$$Q_{n,min}^2 = 2 i_n e_n C \sqrt{F_i F_v} \ . \tag{4.47}$$

For $T/\tau = 1.0357$, $F_i = 0.47$ and $F_v = 0.78$. $\sqrt{F_i F_v} = 0.60$, whereas for a *CR-RC* shaper it is 0.92, so for equal values of τ the *CR-RC* shaper's noise will be 24% higher.

Figure 4.25 shows the noise *vs.* sampling time for a fixed prefilter time constant. The left panel is for the optimum shaping time, so the noise attains a shallow minimum at $T/\tau = 1$, as expected. In the second panel the prefilter time constant is half the optimum value. Now the minimum is at $T/\tau = 1.5$, but the noise is only 4% higher. At the optimum prefilter time constant the noise minimum occurs when the current and voltage noise contributions are equal, whereas for the non-optimum time constant this is not the case. At small sampling times the reduction in signal amplitude increases the equivalent noise charge. Figure

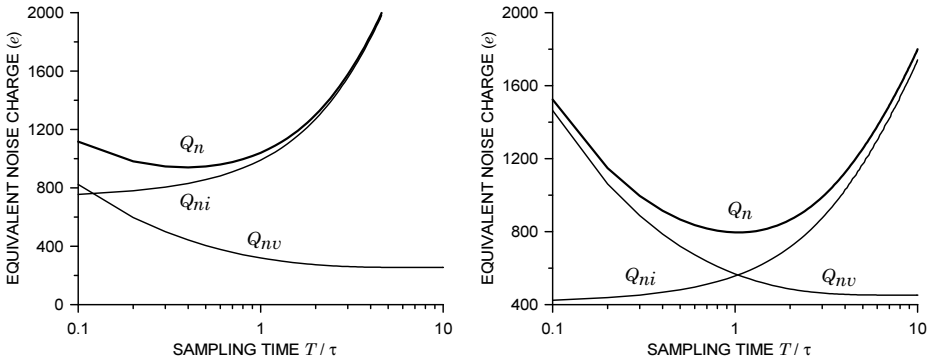

FIG. 4.26. Noise of a CDS shaper *vs.* sampling time at an elevated detector bias current of 100 nA. The left panel shows the noise for a prefilter time constant of 1.7 μs, the optimum for 10 nA bias current. The second panel is for $\tau = 540$ ns, the optimum for 100 nA bias current.

4.26 shows the effect of a ten times higher detector bias current, as could occur after radiation damage. The parameters are as above with the optimum prefilter time constant. The noise minimum is shifted to a smaller sampling time $T/\tau = 0.4$ ($T = 680$ ns) and the noise is approximately doubled to $940\,e$. Reducing the prefilter time constant to the optimum value of 540 ns for the higher noise current reduces the minimum noise to about $800\,e$ at $T/\tau = 1$ ($T = 540$ ns). At a bias current of 10 nA the reduced time constant of 540 ns provides a minimum noise of about $740\,e$ at $T/\tau = 0.5$ ($T = 270$ ns). The increased bias current carries an unavoidable noise penalty, but the choice of time constant together with an adjustment in sample time can reduce the change in noise over the course of radiation damage.

4.5.4.4 $1/f$ *noise* We have ignored the contribution of $1/f$ noise in the correlated double sampler, as evaluating the $1/f$ noise in the time domain is not straightforward (Pullia 1998). However, in this instance it can be calculated in the frequency domain (Kansy 1980, Lee *et al.* 2002) by analogy with a delay line pulse shaper (Knoll 1999). The correlated double sample has a frequency response

$$G_{CDS}(f) = 2\sin(\pi f T)\ , \tag{4.48}$$

which exhibits maxima at $fT = 1, 3, \ldots$ Thus, in the regime $fT \leq 1$ it acts as a high-pass filter. The low-pass response of the prefilter is necessary to attenuate the higher order peaks in the CDS response. The prefilter response

$$G_{LPF}(f) = \frac{1}{1 + \mathbf{i}\,(f/f_u)}\ , \tag{4.49}$$

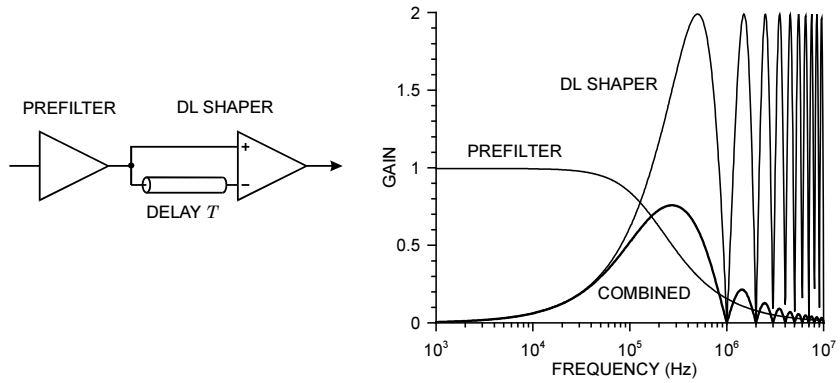

FIG. 4.27. Frequency response of an RC low-pass filter (LPF) with a cutoff frequency $1/2\pi\tau$ with $\tau = 1\mu$s, a delay line pulse shaper (DL) with delay time $T = 1\mu s$, and the combined response, plotted versus frequency. In reality the DL higher order gain oscillations extend to zero; here the calculation grid doesn't match perfectly.

where $f_u = 1/(2\pi\tau)$ is the upper cutoff frequency. Figure 4.27 shows the frequency response of the RC low-pass filter (integrator), the delay line shaper and the composite response $G(f)$. The delay line shaper's response peaks at $f = 1/2T$ and at odd multiples thereof. The low-pass response of the prefilter attenuates the higher order peaks. The frequency is normalized to the inverse delay time $1/T$. Figure 4.28 shows how the response curves change for non-optimum ratios of delay time to prefilter time constant $T/\tau = 0.5$ and 2. For $\tau = 0.5T$ the output noise increases and so does the pulse height, but not sufficiently to improve the

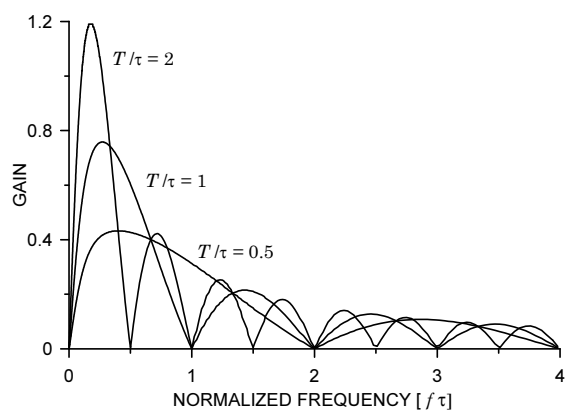

FIG. 4.28. Frequency response of a delay line shaper for $T/\tau = 0.5, 1$, and 2. The frequency is normalized to the inverse low-pass time constant $1/\tau$.

S/N with respect to $T/\tau = 1$. Conversely, $\tau = 2T$ reduces both the output noise and pulse height, again yielding inferior noise.

The total output noise for a $1/f$ input noise spectrum

$$v_{no}^2 = \int_0^\infty \frac{A_f}{f} \left|G(f)\right|^2 df \ . \tag{4.50}$$

The overall shaper response $G(f)$ was determined using SPICE and the integral evaluated numerically. For the optimum shaper constants $T/\tau = 1$ the total output noise voltage $v_{no}^2 = 1.04 A_f$. Normalizing to the output pulse height $V_{os}/V_i = 1 - e^{-T/\tau} = 0.645$ yields the shape factor $F_{vf} = 1.65$. Increasing the sampling time shifts the passband to lower frequencies and increases the integrated $1/f$ noise, yielding a shape factor $F_{vf} = 2.56$ for $T = 2\tau$. Reducing the sample time to $T = \tau/2$ also reduces F_{vf} to 1.04. Thus, although reducing the sample time in a CDS increases the cumulative contribution from shot noise and white voltage noise, it reduces the $1/f$ noise. Although not rigorously correct for a correlated double sampler, this does provide insight into the effect of the shaper constants and a reasonable estimate of $1/f$ noise.

4.6 Detector noise summary

Two basic noise mechanisms determine the equivalent noise charge, the input noise current spectral density i_n and input noise voltage e_n. For both time-invariant and time-variant shapers the equivalent noise charge

$$Q_n^2 = i_n^2 F_i T_S + e_n^2 F_v \frac{C^2}{T_S} + F_{vf} A_f C^2 \ , \tag{4.51}$$

where C is the sum of all capacitances shunting the input. F_i, F_v, and F_{vf} are determined by the frequency or time response of the shaper and T_S is a characteristic time, for example the peaking time of a CR-nRC shaped pulse or the prefilter time constant in a correlated sampler.

The shape factors F_i, F_v are easily calculated:

$$F_i = \frac{1}{2T_S} \int_{-\infty}^{\infty} \left[W(t)\right]^2 dt \ , \qquad F_v = \frac{T_S}{2} \int_{-\infty}^{\infty} \left[\frac{dW(t)}{dt}\right]^2 dt \ . \tag{4.52}$$

For time-invariant pulse shaping $W(t)$ is simply the system's impulse response (the output signal seen on an oscilloscope when a short current pulse is applied to the detector input) with the peak output signal normalized to unity. For a time-variant shaper the same equations apply, but the shape factors are determined differently.

Figure 4.29 illustrates the dependence of equivalent noise charge on basic noise parameters.

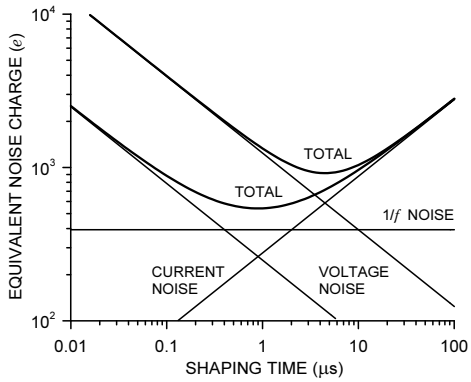

FIG. 4.29. Equivalent noise charge *vs.* shaping time. At small shaping times (large bandwidth) the ENC is dominated by voltage noise, whereas at long shaping times (large integration times) the current noise contributions dominate. The total noise assumes a minimum where the current and voltage contributions are equal. The "$1/f$" noise contribution is independent of shaping time and flattens the noise minimum. Increasing the voltage or current noise contribution shifts the noise minimum. Increased voltage noise is shown as an example.

- Current noise contribution increases with T_S.
- Voltage noise contribution from "white" noise sources decreases with increasing T_S.
- "$1/f$" voltage noise contribution is constant in T_S.
- Voltage noise contributions increase with the total capacitance shunting the input. This includes the sensor capacitance, the input capacitance of the input amplifying device and any other capacitance present at the input, for example wiring capacitance, connectors, *etc.*
- Minimum noise obtains at the shaping time where the current and voltage noise contributions are equal,

$$T_S = \frac{e_n}{i_n} C \sqrt{\frac{F_v}{F_i}} \ . \tag{4.53}$$

The minimum noise

$$Q_n^2 = 2 e_n i_n C \sqrt{F_i F_v} + F_{vf} A_f C^2 \ . \tag{4.54}$$

Typical values of F_i , F_v are shown in Table 4.1. Note that $F_i < F_v$ for higher order shapers. Shapers can be optimized to reduce current noise contribution relative to the voltage noise. This is a useful tool in mitigating radiation damage to semiconductor sensors, which leads to an increase in leakage current. Since the minimum noise obtained at the optimum shaping time is proportional to $\sqrt[4]{F_i F_v}$,

Table 4.1 Noise coefficients for various types of pulse shapers. The CAFE chip is a prototype IC designed for the ATLAS Semiconductor Tracker.

Shaper	F_i	F_v	$\sqrt[4]{F_i F_v}$
CR-RC Shaper	0.924	0.924	0.96
$CR-(RC)^4$ Shaper	0.45	1.02	0.82
$CR-(RC)^7$ Shaper	0.34	1.27	0.81
CAFE Chip	0.4	1.2	0.83
CDS	0.47	0.78	0.78
(opt. τ, $T/\tau = 1.04$)			

Table 4.1 also lists this quantity. As can be seen, the differences in optimum noise are not large, except for the simplest shaper, so other considerations such as rapid baseline recovery, sensitivity to detector leakage current, or simplicity are the deciding factors. A summary of noise contributions for a wide variety of shaper is given by Seller (1996).

A commonly used specification for the noise performance of a front-end system is the noise slope. The equivalent noise charge *vs.* capacitance ($C = C_d + C_a$)

$$Q_n = \sqrt{i_n^2 F_i T + (C_d + C_a)^2 e_n^2 F_v \frac{1}{T}} \ . \tag{4.55}$$

The derivative with respect to the sensor capacitance

$$\frac{dQ_n}{dC_d} = \frac{2C_d e_n^2 F_v \dfrac{1}{T_S}}{\sqrt{i_n^2 F_i T_S + (C_d + C_a)^2 e_n^2 F_v \dfrac{1}{T_S}}} \ . \tag{4.56}$$

If the current noise $i_n^2 F_i T_S$ is negligible,

$$\frac{dQ_n}{dC_d} \approx 2e_n \cdot \sqrt{\frac{F_v}{T_S}} \ . \tag{4.57}$$

The first factor is determined by the preamplifier and the second factor by the shaper. This is a useful specification, as it allows an estimate of the noise for different sensors. Frequently this is given as a specification for preamplifiers. However, this also requires knowledge of the shaper type and shaping time, as illustrated in Figure 4.30. Also, note that this parameterization is only valid if the input noise current is negligible, so it is not useful if the sensor shot noise contribution is significant or a if a bipolar transistor amplifier with significant base current noise is used. Nevertheless, this does not deter many practitioners from using this specification without regard to its validity.

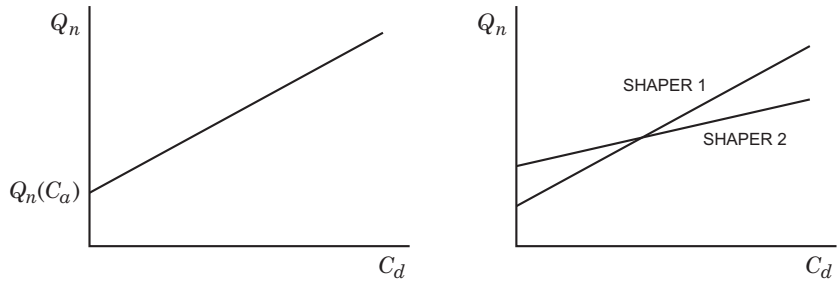

FIG. 4.30. In a system dominated by voltage noise the noise charge depends linearly on sensor capacitance. The intercept for zero sensor capacitance is determined by the additional capacitance C_a shunting the input. Note that the system noise depends on the shaper, as illustrated in the right-hand figure, which shows a second shaper with a longer shaping time, so it reduces the voltage noise at high C_d, but is more sensitive to current noise at low C_d.

When the noise slope is applicable, it can be used together with the zero intercept to determine the additional capacitance C_a in a system (amplifier input capacitance plus strays).

4.7 Threshold discriminator systems

Many systems detect merely the presence of a pulse. As this requires circuitry that senses whether a pulse exceeds a threshold, this is a crude amplitude measurement and is subject to the same considerations discussed in the previous sections. As illustrated in Chapter 1, noise affects not only the resolution of amplitude measurements, but also the determines the minimum detectable signal threshold.

Figure 4.31 shows a system that only records the presence of a signal if it exceeds a fixed threshold. Since this amplitude evaluation yields only a yes/no result, it is frequently called a binary readout.

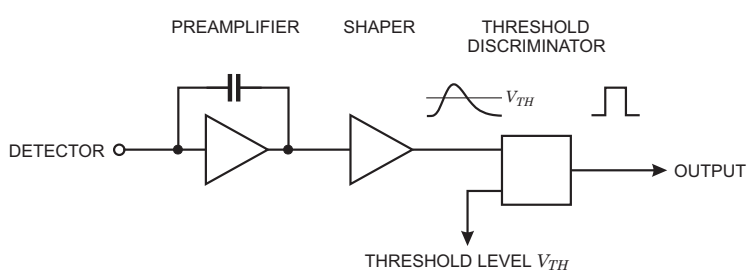

FIG. 4.31. A threshold discriminator (comparator) at the output of the shaper provides a digital output whenever the shaper output exceeds the threshold level V_{TH}.

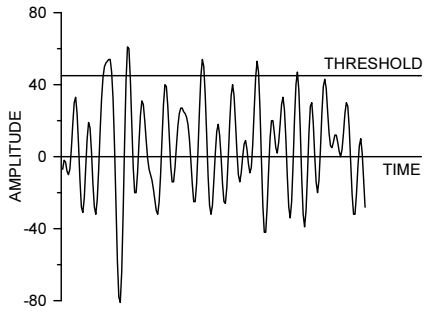

FIG. 4.32. Noise pulses will exceed an amplitude threshold with a rate dependent on the threshold setting.

How small a detector pulse can still be detected reliably? Consider the system at times when no detector signal is present. Noise will be superimposed on the baseline and some fraction of the noise pulses will cross the threshold, as illustrated in Figure 4.32. Since the amplitude distribution of the noise is Gaussian, some noise pulses will always cross the threshold regardless of the threshold setting, but the noise rate will vary with threshold. With the threshold level set to zero relative to the baseline, all of the signal pulses and all of the noise pulses will be recorded.

Assume that the desired signals are occurring at a certain rate. If the detection reliability is to be $> 99\%$, then the rate of noise hits must be less than

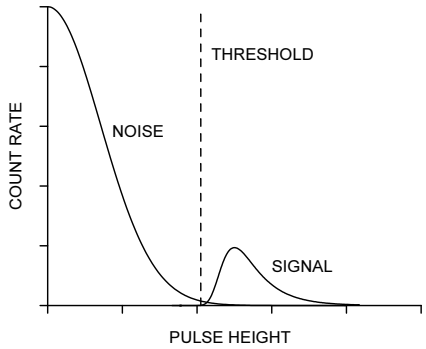

FIG. 4.33. In a binary system the threshold must be set low enough to capture most of the signal, but high enough to reduce the noise rate to an acceptable level. The noise rate is invariably much higher than the signal rate. For the sake of illustration the signal rate is shown much higher than typically acceptable. In this example the threshold setting is beginning to reject signal pulses, but a noticeable rate of noise pulses still exceeds threshold.

1% of the signal rate. The rate of noise hits can be reduced by increasing the threshold, but it cannot be set so high that > 1% of the signal pulses are lost. This is illustrated in Figure 4.33 for a Landau distribution, typical of minimum ionizing particles. The same considerations apply in a photon detector that relies on the detection of Compton scattered interactions.

4.7.1 *Noise rate*

At zero threshold the noise rate will be maximal and equal to f_{n0}. The value of f_{n0} is for noise pulses that cross the threshold with positive slope and will be determined later. If the distribution were dependent on amplitude alone, the integral over the Gaussian distribution (the error function) would determine the factor by which the noise rate f_{n0} is reduced,

$$\frac{f_n}{f_{n0}} = \frac{1}{Q_n\sqrt{2\pi}} \int_{Q_T}^{\infty} e^{-(Q/2Q_n)^2} dQ \ , \tag{4.58}$$

where Q is the signal charge, Q_n the equivalent noise charge, and Q_T the threshold level. However, since the pulse shaper broadens each noise impulse, the time dependence is equally important. For example, after a noise pulse has crossed the threshold, a subsequent pulse will not be recorded if it occurs before the trailing edge of the first pulse has dropped below threshold. Thus, we must consider the combined probability function for both the amplitude and time distributions.

The *combined probability function* for Gaussian time and amplitude distributions is illustrated in Figure 4.34. For illustration the widths of the noise and time distribution have been made equal, so the combined probability distribution is the circular contour plot in the upper right. The total noise rate is obtained by integrating over the combined probability density function in the regime that exceeds the threshold. This yields the expression for the noise rate as a function of threshold-to-noise ratio (Rice 1944):

$$f_n = f_{n0} \cdot e^{-Q_T^2/2Q_n^2} \ . \tag{4.59}$$

Although the signal and noise are expressed as charge, one can just as well use the corresponding voltage levels.

What is the noise rate at zero threshold f_{n0}? Since we are interested in the number of positive excursions exceeding the threshold, f_{n0} is half the frequency of zero-crossings. A detailed analysis of the time dependence (Rice 1944) shows that the frequency of zero crossings

$$f_0^2 = 4 \ \frac{\int_0^{\infty} f^2 A^2(f) df}{\int_0^{\infty} A^2(f) df} \ , \tag{4.60}$$

where $A(f)$ is the voltage or current gain *vs.* frequency. At the output of an ideal bandpass filter with lower and upper cutoff frequencies f_l and f_u the rate

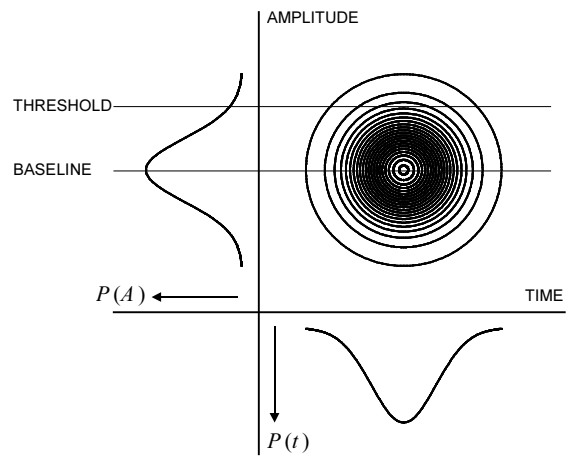

FIG. 4.34. The probability density function for the pulse rate exceeding a given threshold includes both the amplitude and time distributions. The relative rate is determined by integrating over the three-dimensional contour plot for amplitudes exceeding the threshold level.

$$f_0 = 2\sqrt{\frac{1}{3}\frac{f_u^3 - f_l^3}{f_u - f_l}} \ . \tag{4.61}$$

For a *CR-RC* filter with $\tau_i = \tau_d$ the ratio of cutoff frequencies of the noise bandwidth is $f_u/f_l = 4.5$, so to a good approximation one can neglect the lower cutoff frequency and treat the shaper as a low-pass filter, *i.e.* $f_l = 0$. Then

$$f_0 = \frac{2}{\sqrt{3}} f_u \ . \tag{4.62}$$

An ideal bandpass filter has infinitely steep slopes, so the upper cutoff frequency f_u must be replaced by the noise bandwidth. The noise bandwidth of an *RC* low-pass filter with time constant τ is $\Delta f_n = 1/4\tau$. Setting $f_u = \Delta f_n$ yields the frequency of zeros

$$f_0 = \frac{1}{2\sqrt{3}\tau} \tag{4.63}$$

and the frequency of noise hits *vs.* threshold

$$f_n = f_{n0} \cdot e^{-Q_T^2/2Q_n^2} = \frac{f_0}{2} \cdot e^{-Q_T^2/2Q_n^2} = \frac{1}{4\sqrt{3}\tau} \cdot e^{-Q_T^2/2Q_n^2} \ . \tag{4.64}$$

Since the approximation $f_l = 0$ makes the *CR-RC* filter equivalent to an amplifier with an upper cutoff frequency $f_u = 1/2\pi\tau$, the noise rate

$$f_n \approx f_u \cdot e^{-Q_T^2/2Q_n^2} \ . \tag{4.65}$$

The noise rate at zero threshold is approximately equal to the upper cutoff frequency of the system.

Thus, the required threshold-to-noise ratio for a given frequency of noise hits f_n is

$$\frac{Q_T}{Q_n} = \sqrt{-2\log(4\sqrt{3}f_n\tau)} \ . \tag{4.66}$$

Note that the threshold-to-noise ratio determines the product of noise rate and shaping time, *i.e.* for a given threshold-to-noise ratio the noise rate is higher at short shaping times. In other words, the noise rate for a given threshold-to-noise ratio is proportional to bandwidth and to obtain the same noise rate, a fast system requires a larger threshold-to-noise ratio than a slow system with the same noise level.

4.7.2 *Noise occupancy*

Frequently a threshold discriminator system is used in conjunction with other detectors that provide additional information, for example the time of a desired event. In a collider detector the time of beam crossings is known, so the output of the discriminator is sampled at specific times. The number of recorded noise hits then depends on

1. The sampling frequency (*e.g.* bunch crossing frequency) f_S.
2. The width of the sampling interval Δt, which is determined by the time resolution of the system.

The product $f_S\Delta t$ determines the fraction of time the system is open to recording noise hits, so the rate of recorded noise hits is $f_S\Delta t f_n$.

Rather than the rate, often it is more interesting to know the probability of finding a noise hit in a given interval, *i.e.* the occupancy of noise hits, which can be compared to the occupancy of signal hits in the same interval. This is the situation in a storage pipeline, commonly used in collider detectors. Here hits are time stamped and a specific time interval is read out after a certain delay time (*e.g.* trigger latency). Examples of such systems will be shown in Chapter 8.

The occupancy of noise hits in a time interval Δt is

$$P_n = \Delta t \cdot f_n = \frac{\Delta t}{4\sqrt{3}\tau} \cdot e^{-Q_T^2/2Q_n^2} \ , \tag{4.67}$$

i.e. the occupancy falls exponentially with the square of the threshold-to-noise ratio. The precise result depends on whether the pipeline is level or transition sensing. The above expression holds if only comparator transitions that occur within a time bucket are recorded. On the other hand, if the pipeline senses the presence of a level, the finite pulse width of the comparator output will increase the occupancy. For example, if the output pulse of the threshold comparator is equal to pipeline time interval Δt , the occupancy is doubled, as pulses occurring within Δt prior to the time slice will still "spill over" into the time of interest.

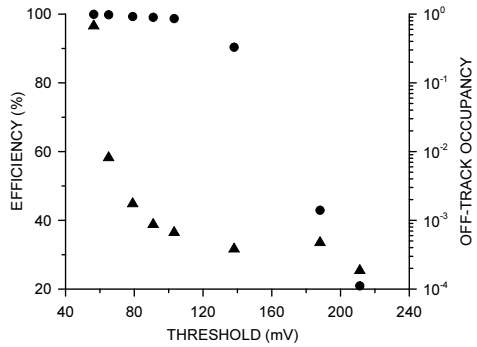

FIG. 4.35. Efficiency (circles) and occupancy (triangles) *vs.* threshold for a representative detector module. The signal is from minimum ionizing particles, so it has a Landau distribution. The data were taken in a tracking measurement, so the occupancy excludes reconstructed tracks, but still includes random hits, so the noise occupancy plateaus.

Figure 4.35 shows a representative plot of efficiency and occupancy *vs.* threshold. This plot demonstrates the desirable feature of a relatively broad threshold interval that yields both high efficiency and low noise occupancy.

4.7.3 *Measurement of noise in a threshold discriminator system*

The dependence of occupancy on threshold can be used to measure the noise level:

$$\log P_n = \log\left(\frac{\Delta t}{4\sqrt{3}\tau}\right) - \frac{1}{2}\left(\frac{Q_T}{Q_n}\right)^2 , \tag{4.68}$$

The slope of $\log P_n$ *vs.* Q_T^2 yields the noise level, *independently of the details of the shaper*, which affect only the offset. An example result is shown in Figure 4.36, taken in a test beam setup, but without beam.

Alternatively, the noise level can be determined from threshold or signal scans. The threshold is scanned while a fixed signal amplitude is applied to system, through the test input, for example. As the threshold level is scanned from low to high, initially all signal pulses will be recorded, but the rate will decrease as the threshold approaches the signal level. The transition is broadened by electronic noise, as illustrated in Figure 4.37. For Gaussian noise the distribution (often called "s-curve") is the error function, so the signal level sets the 50% point of the transition and the width of the transition is determined by the variance $\sigma_N = Q_n$. The threshold difference between the 16% and 84% levels is equal to $2\sigma_n = 2Q_n$.

An equivalent measurement can be performed by setting a fixed threshold and scanning the signal level. Noise levels extracted from occupancy scans and threshold or signal scans should yield the same result. However, frequently they

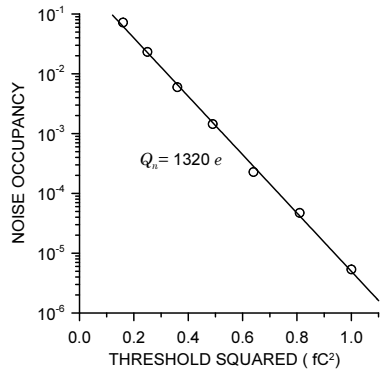

FIG. 4.36. Determination of electronic noise from a measurement of noise occupancy *vs.* threshold.

don't because of rate limitations in the circuitry, as at low thresholds the rate is dominated by noise hits. Scanning the signal level is less prone to these effects, as it begins at a very low rate and doesn't exceed the signal rate, provided the signal-to-noise ratio is sufficiently high.

4.8 Some other aspects of pulse shaping

4.8.1 *Baseline restoration*

Any series capacitor in a system prevents transmission of a DC component. A sequence of unipolar pulses has a DC component that depends on the duty factor, *i.e.* the event rate. As a result, the baseline shifts to make the overall transmitted charge equal zero, as shown in Figure 4.38.

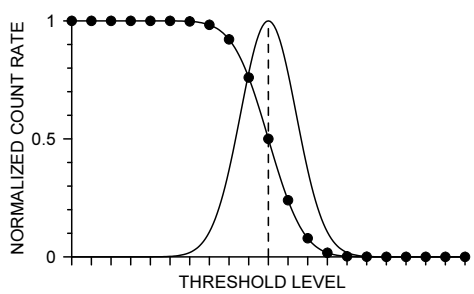

FIG. 4.37. Scanning the threshold level in a binary system at a fixed signal level yields the signal level and distribution in a binary system (curve with data points). Fitting the error function yields the noise level for Gaussian noise and – after differentiation – the noise distribution (middle curve).

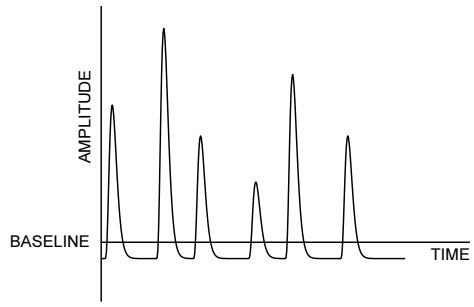

FIG. 4.38. In an AC coupled system a sequence of unipolar pulses will shift the baseline so that the net DC component is zero.

Random rates and random amplitudes lead to random fluctuations of the baseline shift, which is equivalent to an increase in noise. These shifts occur whenever the DC gain is not equal to the signal gain. Thus, baseline shifts can occur even in circuits that have a contiguous DC path from input to output, but tailor the frequency response through feedback networks.

If the signal rate is constant, the baseline shift is also constant, so the effect on resolution may be negligible, although one has to keep track of the rate when calibrating. At low rates (*i.e.* low occupancy) in collider experiments baseline shifts due to AC coupling are usually tolerable, but in very high rate systems baseline shifts are significant.

In non-accelerator measurements, x-ray spectroscopy for example, the effect of fluctuating random rates can be significant, especially in high-resolution systems operated at high photon rates. In this context one has to remember that a readout system that performs well in a high-energy physics environment may suffer severe degradation when exposed to fluctuating random rates.

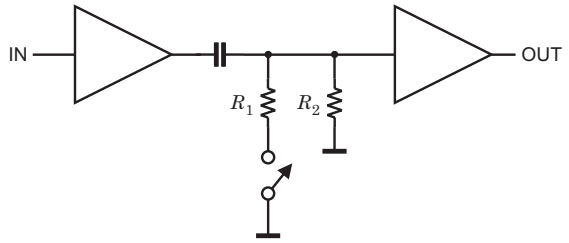

FIG. 4.39. Principle of a baseline restorer. In the absence of a signal the signal line is connected to ground through R1 to establish the baseline just prior to the arrival of a pulse. When a pulse arrives, the switch is opened and the input can follow the signal.

The baseline shift can be mitigated by a baseline restorer (BLR). The principle is illustrated in Figure 4.39. In the absence of a signal an electronic switch connects the signal line to ground to establish the baseline just prior to the arrival of a pulse. When a pulse arrives, the switch is opened and the input can follow the signal.

After the switch has been opened the baseline will adapt to the presence of the signal with a time constant determined by the coupling capacitor and R_2. This time constant must be much larger than the pulse width. After the signal the switch is closed again. Now R_1 sets the time constant with which the baseline returns to zero. The goal is to make this short.

Originally performed with diodes (passive restorer), baseline restoration circuits now tend to include active loops with adjustable thresholds to sense the presence of a signal (gated restorer). Asymmetric charge and discharge time constants improve performance at high count rates. An implementation of baseline restoration in monolithic integrated circuits for high-rate applications is described by Bevensee *et al.* (1996) and Dressnandt *et al.* (2001).

This is a form of time-variant filtering. Care must be exercised to reduce noise and switching artifacts introduced by the BLR. Good tail cancellation, explained in the next section, is crucial for proper baseline restoration.

4.8.2 *Tail (pole–zero) cancellation*

Pulse shapers are analyzed for a step input. In reality, the inputs have decay time constants, originating either in the sensor or the electronics. For example, the feedback capacitor in a charge sensitive preamplifier accumulates charge from multiple signals and ultimately the output voltage reaches the maximum allowed by the amplifier. Thus, the capacitor must be discharged. This is commonly done with a resistor. Now the output no longer a step, but decays exponentially, as shown in the left panel of Figure 4.40. This decay appears as a baseline undershoot following the signal, so a subsequent signal attains a reduced peak amplitude. Again, with random rates and varying amplitudes this degrades resolution.

The decay time constant is set to be large with respect to the shaping time, so the undershoot is small. However, the maximum time constant is set by the signal rate. In the limit, the average current $\langle i_S \rangle$ through the resistor may not cause a voltage drop $\langle i_S \rangle R_F$ that exceeds the output range of the amplifier. The rate capability is increased by reducing R_F, but this increases magnitude of the undershoot.

By adding a resistor R_{pz} to the subsequent pulse shaper's differentiator as shown in the right panel of Figure 4.40, the low frequency response is boosted to compensate for the decay of the signal applied to the shaper input. In circuit theory, the "pole" associated with the signal decay τ is compensated by a "zero" introduced by the time constant $R_{pz}C_{diff} = \tau$. If the decay is introduced by the discharge time constant of the preamplifier, then $R_F C_F = R_{pz} C_d$. In other applications, notably gaseous sensors, multiple decay components with different

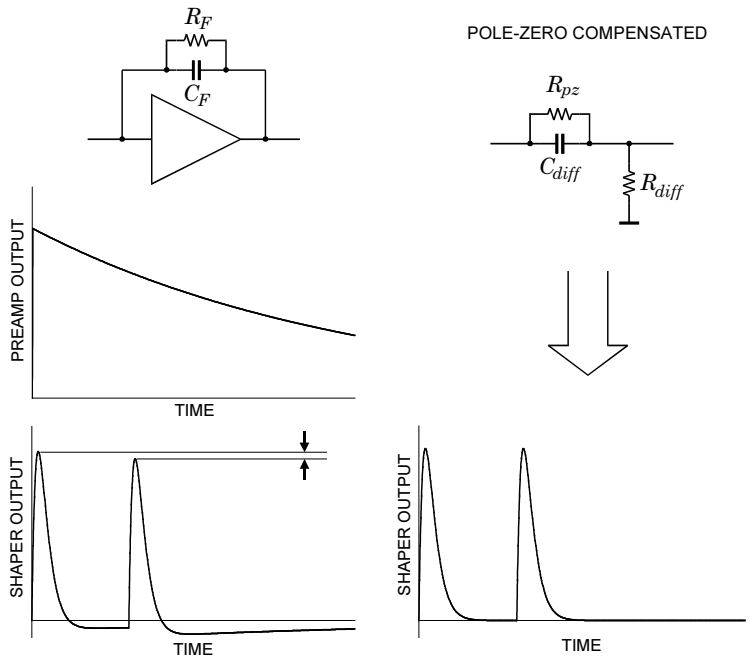

FIG. 4.40. A long decay time constant is transposed on the baseline and reduces the amplitude of the next pulse. Adding equalization circuitry to boost the frequency response in the appropriate range yields a constant baseline.

time constants are common. Multiple cancellation circuits can be introduced to deal with this. However, this only compensates for purely exponential decays.

An alternative to resistive discharge in the preamplifier is to use pulsed reset circuits (optical or transistor) that discharge the feedback capacitor when the output approaches the voltage limit. The discharge spike must be suppressed, but since the output of the preamplifier is a sequence of superimposed steps, pole–zero cancellation is not necessary.

4.8.3 *Bipolar vs. unipolar shaping*

As explained in the discussion on baseline restorers, a sequence of unipolar pulses passing through an AC coupled system leads to baseline shifts. This can be avoided by using pulse shapers with both a positive and negative lobe, so that the net charge of the shaped pulse is zero (although the peak amplitude is still proportional to signal charge). Generally, bipolar shapers can be constructed by adding a second differentiator to a unipolar shaper. Using equal time constants τ in a CR-CR-RC shaper yields the pulse shape

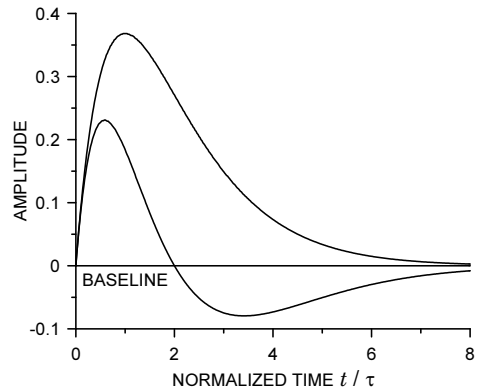

FIG. 4.41. Unipolar and bipolar pulses from CR-RC and CR^2-RC shapers. The time is normalized to the integrator and differentiator time constants τ.

$$V(t) = V_0 \left[\frac{t}{\tau} - \frac{1}{2} \left(\frac{t}{\tau} \right)^2 \right] e^{-t/\tau} \ , \tag{4.69}$$

which is plotted in Figure 4.41 together with the original CR-RC unipolar pulse shape for comparison. Since the second CR stage isn't an ideal differentiator, the zero crossing doesn't occur at the peak of the unipolar pulse. The reduction in peak amplitude visible in Figure 4.41 accounts for most of the 38% degradation in ENC relative to a CR-RC shaper, as the more rapid low-frequency roll-off hardly affects the noise bandwidth.

Bipolar shapers are usually frowned upon by purists, as the electronic noise is typically 25 – 50% worse than for the corresponding unipolar shaper. However, bipolar shaping eliminates rate dependent baseline shifts (as the DC component is zero) and pole–zero adjustment is less critical. Not all systems require optimum electronic noise, so operational robustness and user convenience may override. The most important feature of bipolar shapers may be the added suppression of low-frequency noise (see Chapter 9). In systems subject to external interference this can be crucial, so in practical systems "inferior" shapers often yield superior results.

4.9 Timing measurements

Pulse height measurements discussed up to now emphasize measurement of signal charge. Timing measurements seek to optimize the determination of the time of occurrence. Although, as in amplitude measurements, signal-to-noise ratio is important, the determining parameter is not signal-to-noise, but slope-to-noise ratio. This is illustrated in Figure 4.42, which shows the leading edge of a pulse fed into a threshold discriminator (comparator), a "leading edge trigger". The

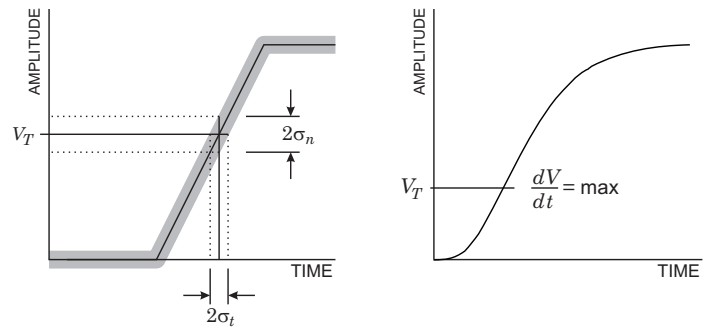

FIG. 4.42. Fluctuations in signal amplitude crossing a threshold translate into timing fluctuations (left). With realistic pulses the slope changes with amplitude, so minimum timing jitter occurs with the trigger level at the maximum slope.

instantaneous signal level is modulated by noise, where the variations are indicated by the shaded band. Because of these fluctuations, the time of threshold crossing fluctuates. By simple geometrical projection, the timing variance, or "jitter", is

$$\sigma_t = \frac{\sigma_n}{\frac{dV}{dt}\big|_{V_T}} \approx \frac{t_r}{S/N} \ . \tag{4.70}$$

Typically, the leading edge is not linear, so the optimum trigger level is the point of maximum slope, as shown in the second panel of Figure 4.42.

4.9.1 *Pulse shaping in timing systems*

Consider a system whose bandwidth is determined by a single RC filter. The time constant of the RC low-pass filter determines the rise time (and hence dV/dt) and the amplifier bandwidth (and hence the noise). The time dependence of the signal

$$V(t) = V_0(1 - e^{-t/\tau}) \ . \tag{4.71}$$

The rise time is commonly expressed as the interval between the points of 10% and 90% amplitude, so $t_r = 2.2\tau$. In terms of bandwidth

$$t_r = 2.2\tau = \frac{2.2}{2\pi f_u} = \frac{0.35}{f_u} \ . \tag{4.72}$$

For example, an oscilloscope with 100 MHz bandwidth has 3.5 ns rise time. In a cascade of amplifiers the individual rise times add in quadrature

$$t_r \approx \sqrt{t_{r1}^2 + t_{r2}^2 + \ldots + t_{rn}^2} \ . \tag{4.73}$$

These rules only apply to amplifiers with nRC-integrator responses, which is usually the case in amplifiers, as was shown in Chapter 2. However, digital

signal processing allows more complex response functions, so in these systems the validity of these relationships cannot be taken for granted. Digital oscilloscopes are a common example.

4.9.2 *Choice of rise time in a timing system*

Consider a detector pulse with peak amplitude V_0 and a rise time t_{rs} passing through an amplifier chain with a rise time t_{ra}. If the amplifier rise time is substantially greater than the signal rise time, so that the amplifier sets the overall rise time, the electronic noise

$$v_n \propto \sqrt{f_u} \propto \sqrt{\frac{1}{t_{ra}}} \tag{4.74}$$

and the signal slope

$$\frac{dV}{dt} \propto \frac{1}{t_{ra}} \propto f_u \ . \tag{4.75}$$

As the bandwidth f_u increases, the speed of the transition grows proportionally, whereas the electronic noise only increases with the square root of bandwidth. Thus, the gain in dV/dt outweighs increase in noise. If the amplifier is substantially faster than the signal rise time, further increases in amplifier speed increase the noise without substantially improving the overall rise time.

Quantitatively, the cumulative rise time at the amplifier output (discriminator output) is

$$t_r = \sqrt{t_{rs}^2 + t_{ra}^2} \ . \tag{4.76}$$

The electronic noise at the amplifier output is $v_{no}^2 = \int e_{ni}^2 df = e_{ni}^2 \Delta f_n$. For a single RC time constant the noise bandwidth

$$\Delta f_n = \frac{\pi}{2} f_u = \frac{1}{4\tau} = \frac{0.55}{t_{ra}} \ . \tag{4.77}$$

As the number of cascaded stages increases, the noise bandwidth approaches the signal bandwidth. In any case

$$\Delta f_n \propto \frac{1}{t_{ra}} \ . \tag{4.78}$$

The timing jitter

$$\sigma_t = \frac{V_{no}}{dV/dt} \approx \frac{V_{no}}{V_0/t_r} = \frac{1}{V_0} V_{no} t_r \propto \frac{1}{V_0} \frac{1}{\sqrt{t_{ra}}} \sqrt{t_{rs}^2 + t_{ra}^2} = \frac{\sqrt{t_{rs}}}{V_0} \sqrt{\frac{t_{rs}}{t_{ra}} + \frac{t_{ra}}{t_{rs}}} \ . \tag{4.79}$$

The second factor assumes a minimum when the rise time of the amplifier equals the collection time of the detector $t_{ra} = t_c$, as shown in Figure 4.43. However, the minimum is shallow, so approximate matching is adequate. The optimum timing resolution improves with decreasing signal rise time $\sigma_t \propto \sqrt{t_{rs}}$ and increasing signal amplitude V_0.

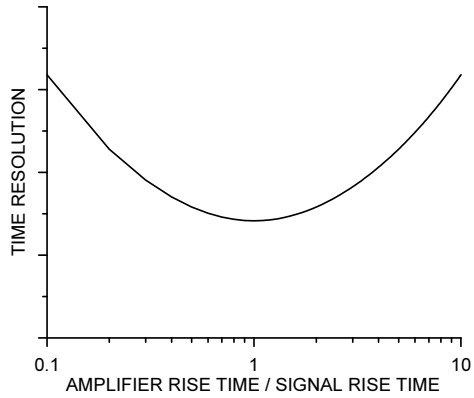

FIG. 4.43. The timing jitter assumes a minimum when the amplifier rise time matches the signal rise time.

The integration time should be chosen to match the rise time, but how should the differentiation time be chosen? As shown in Figure 4.2, the loss in signal can be appreciable even when the differentiation time constant is significantly greater than the integration time constant, *e.g.* $> 20\%$ for $\tau_{diff}/\tau_{int} = 10$. Since the time resolution improves directly with increasing peak signal amplitude, the differentiation time should be set as large as allowed by the required pulse rate.

4.9.3 *Time walk*

Up to now we have considered timing jitter, *i.e.* the timing variation for a fixed amplitude. In addition, the time at which the signal crosses a fixed threshold depends on pulse amplitude. As the amplitude varies, the timing signal shifts, so variations in signal amplitude will broaden the timing distribution. This phenomenon is called "time walk" and is illustrated in Figure 4.44.

As a result, the accuracy of timing measurement is limited by the combination of jitter (due to noise) and time walk (due to amplitude variations). If the rise time is known, "time walk" can be compensated in software event-by-event by measuring the pulse height and correcting the time measurement. This technique fails if both amplitude and rise time vary, as is common. Recall, that in semiconductor sensors the rise time combines both electron and hole components, so the slope has two components. In hardware, time walk can be reduced by setting the threshold to the lowest practical level, or by using amplitude compensating circuitry, which will be described in the following sections.

Before going into a detailed discussion of timing techniques we should bear in mind that many systems do not require the "perfect" timing system. In charged-particle tracking detectors all particles traverse the sensor, so the relative contribution of electrons and holes is always the same and rise-time compensation is not necessary. Minimum ionizing particles all deposit the same average energy

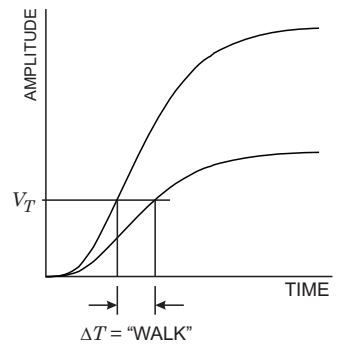

FIG. 4.44. The time at which a signal crosses a fixed threshold depends on the signal amplitude, leading to "time walk".

in the sensor, but the amplitude distribution is rather broad, so time walk is an issue. However, in x-ray detection dominated by "single point" photoelectric interactions, the hole (or electron) contribution can range from zero to 100%, so both the amplitude and the rise time can vary significantly. Currently, large scale x-ray detectors tend to measure amplitude alone, but experimenters are already compiling "wish lists" that include fast timing.

4.9.4 *Lowest practical threshold in leading edge triggering*

A single RC integrator has maximum slope at $t = 0$,

$$\frac{d}{dt}(1 - e^{-t/\tau}) = \frac{1}{\tau}e^{-t/\tau} \ . \tag{4.80}$$

However, the rise time of practically all fast timing systems is determined by multiple time constants. The effect of additional time constants can be visualized rather simply. For small t the slope at the output of a single RC integrator is approximately linear, so initially the pulse can be approximated as a ramp. The response of the following integrator to a ramp $V_i = \alpha t$ is

$$V_o = \alpha(t - \tau) + \alpha\tau e^{-t/\tau} \ . \tag{4.81}$$

Thus, the output is delayed by τ and curvature is introduced at small times, as shown is Figure 4.45. The output attains 90% of the input slope after $t = 2.3\tau$. Additional RC integrators introduce more curvature at the beginning of the pulse, as shown in the second panel of Figure 4.45. The delay for n integrators is $n\tau$. Since increased curvature at the beginning of the pulse limits the minimum threshold for good timing, one dominant time constant is best for timing measurements. This is unlike the situation in amplitude measurements, where multiple integrators are desirable to improve pulse symmetry and count rate performance.

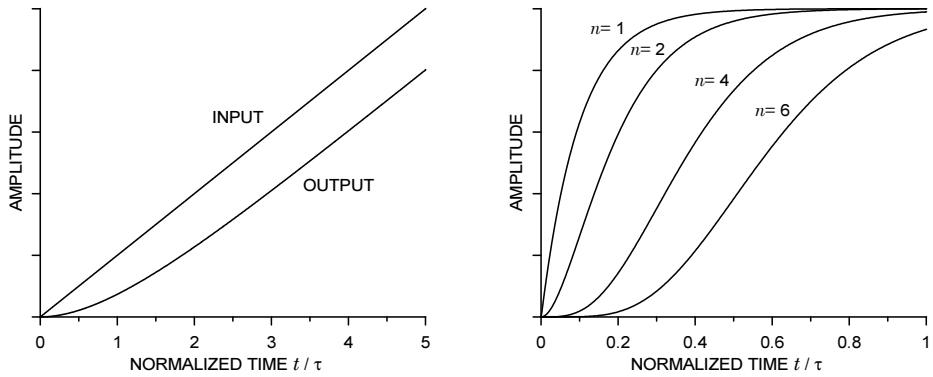

FIG. 4.45. A ramp applied to an RC integrator suffers a delay and the introduction of initial curvature (left). Additional integrators extend the curvature at small signal levels and further delay the signal. Curvature at low signal levels degrades the obtainable time resolution.

4.9.5 *Zero-crossing timing*

A conceptually simple technique to reduce time walk is to use the zero-crossing of a bipolar pulse, as shown in Figure 4.46. First consider a CR-RC shaper with equal time constants, so the output signal

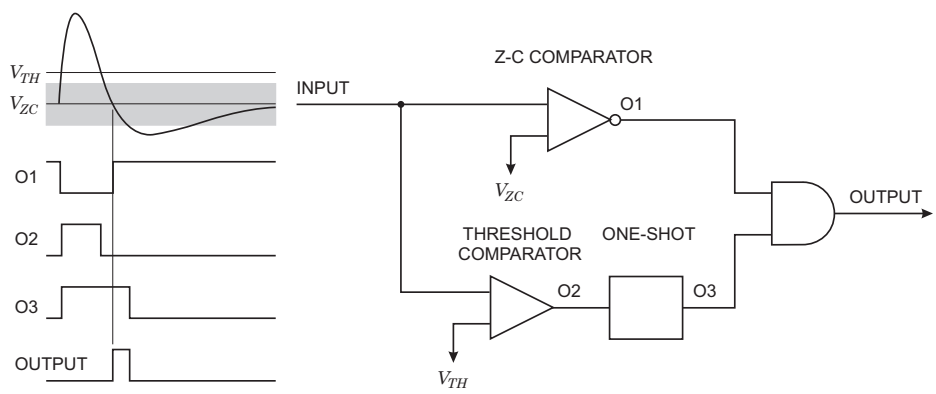

FIG. 4.46. Zero-crossing timing circuit. The zero-crossing timing comparator triggers continually on noise pulses. Only when the threshold comparator fires does the AND gate provide an output. The one-shot (monostable multivibrator) stretches the comparator output to ensure sufficient overlap with the Z-C comparator output. The threshold level is set well above the noise level, indicated as a gray band. Logic circuit symbols are explained in Chapter 5.

$$V_{uni}(T) = V_0 T e^{-T} \ . \tag{4.82}$$

The time is normalized to the time constant $T \equiv t/\tau$. To simplify the calculation the bipolar signal will be formed by an ideal differentiator. Then the bipolar output

$$V_{bip}(T) \equiv \frac{dV_{uni}(T)}{dt} = V_0 e^{-T} \left(1 - T\right) \ . \tag{4.83}$$

The zero crossing occurs at $T = 1$ or $t = \tau$, which is independent of pulse height.

This technique eliminates time walk, but at the expense of time jitter. The derivative of the bipolar pulse

$$\frac{dV_{bip}(T)}{dT} = -V_0 e^{-T} \left(2 - T\right) \ , \tag{4.84}$$

so the slope at the zero crossing point $T = 1$ is $dV_{bip}(\tau)/dt = -V_0/e = -0.37V_0$. Triggering on the leading edge of the unipolar signal, say at $V_T = 0.1V_0$ or $T = 0.11$, yields a slope (equation 4.83) at the trigger point $dV_{uni}(T)/dT = 0.8V_0$.

The noise bandwidth is essentially the same in both cases (the differentiator introduces a low-frequency cutoff, whereas the upper cutoff frequency remains about the same, so the noise bandwidth is only reduced slightly), but the bipolar signal has twice the time jitter because of the degraded slope.

The zero-crossing system requires an additional comparator to "arm" the zero-crossing signal, as shown in Figure 4.46. To eliminate walk the threshold of the zero-crossing comparator must be set to zero, but then it also triggers continually on noise. The threshold comparator fires on the leading edge and is set high enough to suppress triggers on noise. The timing signal is the AND of the threshold and timing comparator. Since the threshold comparator fires prior to the zero-crossing and only preselects signals, the time resolution is determined by the zero-crossing comparator.

The choice of unipolar *vs.* bipolar shaping depends on the range of signal amplitude, which determines whether time jitter or time walk is the dominant contribution to the overall time resolution. An additional consideration is the need for a second comparator in the zero-crossing system. However, this distinction dissolves when one attempts to use a unipolar pulse with a very low threshold. Then the rate of noise triggers may be excessive, but a second threshold comparator can provide amplitude discrimination, as in the zero-crossing system.

4.9.6 *Constant fraction timing*

Zero-crossing timing compensates for amplitude variations, but not for rise time. One technique that accomplishes both is constant fraction triggering. First we consider the amplitude compensation mode. The basic principle is to make the threshold track the signal. This can be achieved by deriving the trigger threshold from the signal, as shown in Figure 4.47. The trigger threshold is derived from the signal by passing it through an attenuator so that $V_T = fV_s$. In addition, the

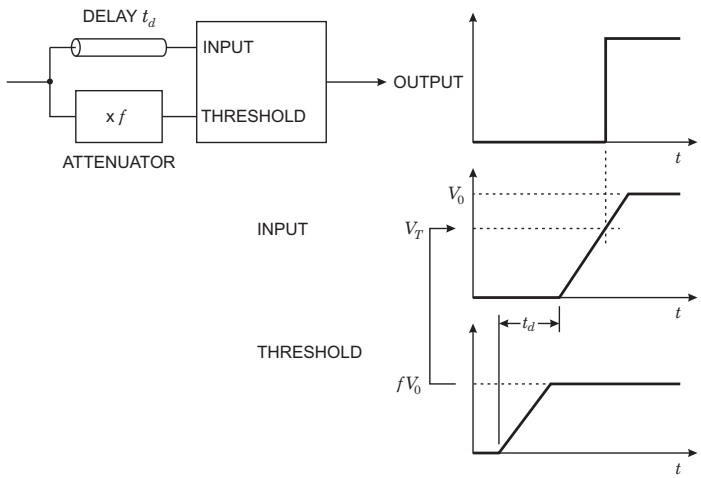

FIG. 4.47. Principle of a constant fraction timing circuit. The threshold V_T of a timing discriminator is derived from the pulse amplitude to compensate for time walk.

signal applied to the comparator input is delayed so that the transition occurs after the threshold signal has reached its maximum value $V_T = fV_0$. Delay lines can be implemented in integrated circuits as strip line spirals (Simpson *et al.* 1996), but require substantial area. As will be shown below, the circuit also functions well with delays less than the rise time, which also provide rise time compensation.

For simplicity assume a linear leading edge

$$\begin{aligned} V(t) &= \frac{t}{t_r} V_0 \qquad \text{for} \quad t \le t_r \\ V(t) &= V_0 \qquad\quad\;\; \text{for} \quad t > t_r \ , \end{aligned} \tag{4.85}$$

so the signal applied to the input is

$$V(t) = \frac{t - t_d}{t_r} V_0 \ . \tag{4.86}$$

When the input signal crosses the threshold level

$$fV_0 = \frac{t - t_d}{t_r} V_0 \tag{4.87}$$

and – provided that the delay time is greater than the rise time $t_d > t_r$ – the comparator fires at the time

$$t = ft_r + t_d \tag{4.88}$$

at a constant fraction of the rise time independent of peak amplitude. In reality, the signal pulse usually attains a cusp, so the condition for the delay is that is be approximately equal to the rise time.

If the delay t_d is reduced so that the pulse transitions at the signal and threshold inputs overlap, the threshold level

$$V_T = f\frac{t}{t_r}V_0 \tag{4.89}$$

and the comparator fires when

$$f\frac{t}{t_r}V_0 = \frac{t - t_d}{t_r}V_0 \ ,$$

at the time

$$t = \frac{t_d}{1-f} \qquad (t_d < (1-f)t_r) \tag{4.90}$$

independent of both amplitude and rise time (amplitude and rise-time compensation).

As shown in Chapter 2, the pulses in semiconductor detectors have two components of different slope, due to the different mobilities of electrons and holes. This does not allow full rise-time compensation. However, the circuit still compensates for amplitude and rise time variations if the pulses have a sufficiently large linear range that extrapolates to the same origin.

The condition for the delay must be met for the minimum rise time

$$t_d \leq (1-f)\ t_{r,min} \ . \tag{4.91}$$

In this mode the fractional threshold V_T/V_0 varies with rise time. For all amplitudes and rise times within the compensation range the comparator fires at the time

$$t_0 = \frac{t_d}{1-f} \ . \tag{4.92}$$

As noted above constant-fraction triggers are not commonly used in high-density ICs because the delay lines require substantial space. However, the basic principle can also be implemented with lumped circuit elements with some penalties in performance (Jackson *et al.* 1997).

4.9.7 *Fast timing – some results*

Figure 4.48 shows a time-walk measurement on an integrated circuit designed as a prototype for the ATLAS Semiconductor Tracker. The overall pulse shaping is equivalent to a CR-RC^4 shaper with a simple leading edge timing circuit. The change in time *vs.* amplitude maps directly to the rise time of the pulse. The time walk increases rapidly as the trigger threshold approaches the cusp of the signal. In this example the timing jitter is negligible at high signal levels, but becomes significant at low signal levels where the trigger level is near the cusp of the signal, where the time derivative is small.

Figure 4.49 shows results from a fast timing system using thin silicon sensors (Spieler 1982). This system was optimized for fast rise time, so the time jitter is in

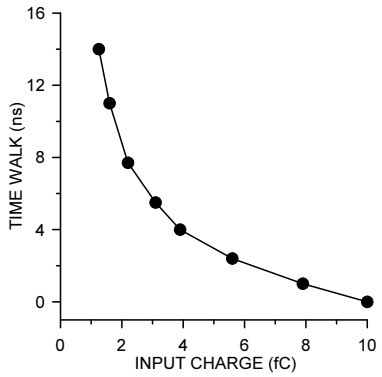

FIG. 4.48. Time walk of a strip detector system designed for an overall time resolution of 25 ns. The time jitter at 1.25 fC is 4 ns FWHM, so the total time distribution for 99% efficiency is contained within about 18 ns.

the range of picoseconds. As predicted, the time resolution improves with signal-to-noise ratio. For very large signals the time resolution plateaus. This can have two origins. In this measurement it was the inherent jitter of the time digitizer (see Chapter 5). Another limit is imposed by time jitter in the comparator. For small signal excursions around the trigger point, comparators can be considered as linear amplifiers, so they are also subject to timing jitter determined by the bandwidth (rise time) of the input stage and its electronic noise. A calculated curve using the measured rise time at the trigger input and the measured noise level is also shown. The experimental curve lies above the calculation because the timing discriminator limited the bandwidth and increased the rise time.

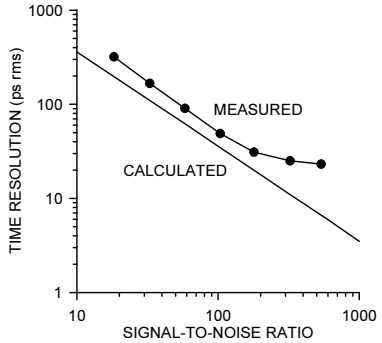

FIG. 4.49. Comparison between measured and predicted results for a timing system using thin silicon sensors. The rise time of 1.2 ns is the measured detector signal at the output of the amplifier chain.

References

Bevensee, B. *et al.* (1996). An amplifier-shaper-discriminator with baseline restoration for the ATLAS Transition Radiation Tracker. *IEEE Trans. Nucl. Sci.* **43/3** (1996) 1725–1731

Choong, W.-S. *et al.* (2002). A compact 16-module camera using 64-pixel CsI(Tl)/Si p-i-n photodiode imaging modules. *IEEE Trans. Nucl. Sci.* **NS-49/5** (2002) 2228–2235

Dressnandt, N. *et al.* (2001). Implementation of the ASDBLR straw tube readout ASIC in DMILL technology. *IEEE Trans. Nucl. Sci.* **48/4** (2001) 1239–1243

Gatti, E. and Manfredi, P.F. (1986). Processing the signals from solid-state detectors in elementary particle physics. *Riv. Nuovo Cimento* **9/1** (1986) 1–146

Gillespie, A.B. (1953). *Signal, Noise, and Resolution in Nuclear Counter Amplifiers*. Pergamon Press, New York

Goulding, F.S. (1972). Pulse shaping in low-noise nuclear amplifiers: a physical approach to noise analysis. *Nucl. Instr. Meth.* **100** (1972) 493–504

Goulding, F.S. and Landis, D.A. (1982). Signal processing for semiconductor detectors. *IEEE Trans. Nucl. Sci.* **NS-29/3** (1982) 1125–1141

Holland, S.E., Wang, N.W., and Moses, W.W. (1997). Development of low noise, back-side illuminated silicon photodiode arrays. *IEEE Trans. Nucl. Sci.* **NS-44/3** (1997) 443–447

Jackson, R.G. *et al.* (1997). Integrated constant fraction discriminator shaping techniques for the PHENIX lead-scintillator calorimeter. *IEEE Trans. Nucl. Sci.* **NS-44/3** (1997) 303–307

Kansy, R.J. (1980). Response of a correlated double sampling circuit to $1/f$ noise. *IEEE J. Solid-State Circuits* **SC-15/3** (1980) 373–375

Knoll, G.F. (2000). *Radiation Detection and Measurement* (3rd edn). Wiley, New York, ISBN 0-471-07338-5, QC 787.C6K56

Kowalski, E. (1970). *Nuclear Electronics*. Springer Verlag, New York, 1970

Krieger, B., Kipnis, I. and Ludewigt, B.A. (1998). XPS: a multi-channel preamplifier – shaper IC for x-ray spectroscopy. *IEEE Trans. Nucl. Sci.* **45/3** (1998) 732–734

Lee, T-H. *et al.* (2002). Analysis of 1/f noise in CMOS preamplifier with CDS circuit. *IEEE Trans. Nucl. Sci.* **NS-49/4** (2002) 1819–1823

Ludewigt, B. *et al.* (1994). A high rate, low noise, x-ray silicon strip detector system. *IEEE Trans. Nucl. Sci.* **NS-41/4** (1994) 1037–1041

Ludewigt, B. *et al.* (1996). Progress in multi-element silicon detectors for synchrotron XRF applications. *IEEE Trans. Nucl. Sci.* **NS-43/3** (1996) 1442–1445

Radeka, V. (1972). Trapezoidal filtering of signals from large germanium detectors at high rates. *Nucl. Instr. Meth.* **99** (1972) 525–539

Radeka, V. (1974). Signal, noise and resolution in position-sensitive detectors. *IEEE Trans. Nucl. Sci.* **NS-21** (1974) 51–64

Rice, S.O. (1944). Mathematical analysis of random noise. *Bell System Technical Journal* **23** (1944) 282–332 and **24** (1945) 46–156

Seller, P. (1996). Noise analysis in linear electronic circuits. *Nucl. Instr. Meth.* **A376** (1996) 229–241

Simpson, M.L. *et al.* (1996). A monolithic, constant-fraction discriminator using distributed R-C delay line shaping. *IEEE Trans. Nucl. Sci.* **NS-43/3** (1996) 1695–1699

Spieler, H. (1982). Fast timing methods for semiconductor detectors. *IEEE Trans. Nucl. Sci.* **NS-29/3** (1982) 1142–1158

5

ELEMENTS OF DIGITAL ELECTRONICS AND SIGNAL PROCESSING

5.1 Digital circuit elements

The basic difference between analog and digital signals is illustrated in Figure 5.1. Analog signals utilize continuously variable properties of the pulse to impart information, such as the pulse amplitude or pulse shape. Digital signals have constant amplitude, but the presence of the signal at specific times is evaluated, *i.e.* whether the signal is in one of two states, "low" or "high". However this still involves an analog process, as the presence of a signal is determined by the signal level exceeding a threshold.

Shannon (1949) described the transmission capacity of a digital link (bits per second):

$$C = B \log_2 \left(1 + \frac{S}{N}\right) \ , \tag{5.1}$$

where B is the bandwidth, S the signal (pulse amplitude), and N the noise. The noise enters, because near the switching threshold, digital elements are amplifiers. Although fundamentally limited by thermal noise as in analog circuits, the signal-to-noise ratio in digital systems is usually determined by cross-talk from other digital circuits. Thus, increasing the pulse amplitude will not improve the signal-to-noise ratio. Although digital systems are commonly described as a simple matter of "yes" or "no", real systems must also deal with "maybe".

5.1.1 *Logic elements*

Figure 5.2 illustrates several functions utilized in digital circuits ("logic" functions). An AND gate provides an output only when all inputs are high. An OR gives an output when any input is high. An eXclusive OR (XOR) responds when only one input is high. The same elements are commonly implemented with inverted outputs, then called NAND and NOR gates, for example. The D flip-flop

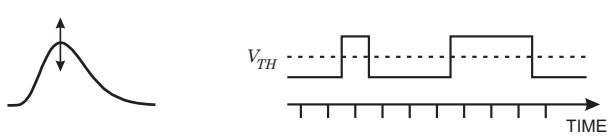

FIG. 5.1. Analog signals contain information in the form of amplitude (left). Digital signals have a fixed amplitude. Information is carried in the time structure of a pulse train (right).

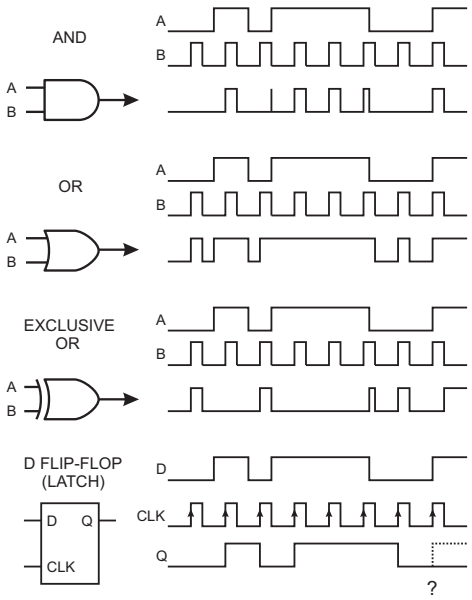

FIG. 5.2. Basic logic functions include gates (AND, OR, Exclusive OR) and flip-flops.

is a bistable memory circuit that records the presence of a signal at the data input D at the time of a signal transition at the clock input CLK. This device is commonly called a latch. Inverted inputs and outputs are denoted by small circles or by superimposed bars, *e.g.* $\overline{Q}$ is the inverted output of a flip-flop, as shown in Figure 5.3.

Logic circuits are fundamentally amplifiers, so they also suffer from bandwidth limitations. The pulse train of the AND gate in Figure 5.2 illustrates a common problem. The third pulse of input B is going low at the same time that input A is going high. Depending on the time overlap, this can yield a narrow output that may or may not be recognized by the following circuit. In an EX-OR this can occur when two pulses arrive nearly at the same time. The D flip-flop requires a minimum setup time for a level change at the D input to be recognized, so changes in the data level may not be recognized at the correct time. These marginal events may be extremely rare and perhaps go unnoticed. However, in complex systems the combination of "glitches" can make the system "hang up", necessitating a system reset. Data transmission protocols have been developed to detect such errors (parity checks, Hamming codes, etc.), so corrupted data can be rejected.

Some key aspects of logic systems can be understood by inspecting the circuit elements that are used to form logic functions. Figure 5.4 shows simple inverter circuits using MOS transistors. These devices will be described in the next chapter. At this point it is sufficient to know that in an NMOS transistor a conductive

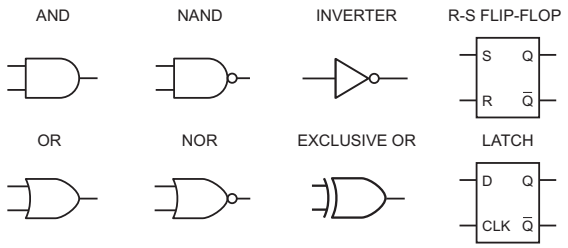

FIG. 5.3. Some common logic symbols. Inverted outputs are denoted by small circles or by a superimposed bar, as for the latch output $\overline{Q}$. Additional inputs can be added to gates as needed. An R-S flip-flop sets the Q output high in response to an S input. An R input resets the Q output to low.

channel is formed when the input electrode is biased positive with respect to the channel. The input, called the "Gate" (G), is capacitively coupled to the output channel connected between the "Drain" (D) and "Source" (S) electrodes. In the NMOS inverter applying a positive voltage to the gate makes the output channel conduct, so the output level is low. A PMOS transistor is the complementary device, where a conductive channel is formed when the gate is biased negative with respect to the source. Since the source is at positive potential, a low level at the inverter input yields a high level at the output. Regardless of the device and pulse polarity, the output pulse is always the inverse of the input. NMOS and PMOS inverters draw current when in their "active" state. Combining NMOS and PMOS transistors in a complementary MOS (CMOS) circuit allows zero current draw in both the high and low states with a substantial reduction in power consumption. A CMOS inverter is shown in Figure 5.5, which also shows how devices are combined to form a CMOS NAND gate. In the inverter the lower (NMOS) transistor is turned off when the input is low, but the upper (PMOS) transistor is turned on, so the output is connected to V_{DD}, taking the output

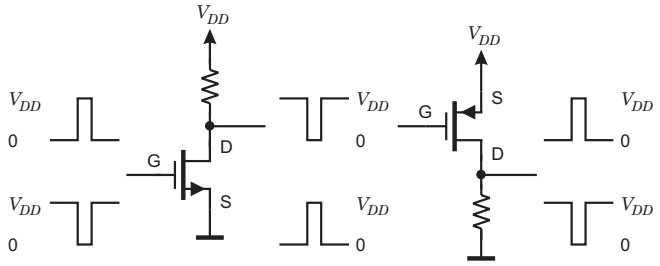

FIG. 5.4. In an NMOS inverter the transistor conducts when the input is high (left), whereas in a PMOS inverter the transistor conducts when the input is low (right). In both circuits the input pulse is inverted, whether the input swings high or low.

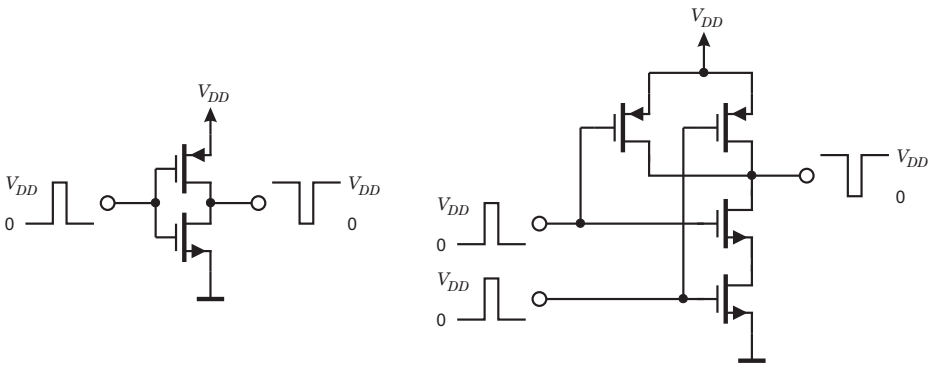

FIG. 5.5. A CMOS inverter (left) and NAND gate (right).

high. Since the current path from V_{DD} to ground is blocked by either the NMOS or PMOS device being off, the power dissipation is zero in both the high and low states. Current only flows during the level transition when both devices are on as the input level is at approximately $V_{DD}/2$. As a result, the power dissipation of CMOS logic is significantly less than in NMOS or PMOS circuitry. As will be discussed in the next chapter, the reduction in power only obtains in logic circuitry. CMOS analog amplifiers are not fundamentally more power efficient than NMOS or PMOS circuits, although CMOS allows more efficient circuit topologies.

5.1.2 *Propagation delays and power dissipation*

Logic elements always operate in conjunction with other circuits, as illustrated in Figure 5.6. The wiring resistance in conjunction with the total load capacitance increases the rise time of the logic pulse and as a result delays the time when

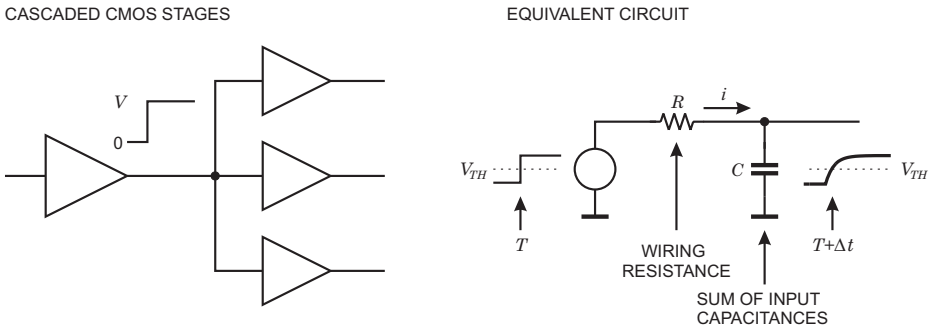

FIG. 5.6. The wiring resistance together with the distributed load capacitance delays the signal.

the transition crosses the logic threshold. The power dissipated in the wiring resistance R is

$$P = \int i^2(t) R dt \ . \tag{5.2}$$

The current flow during one transition

$$i(t) = \frac{V}{R} \exp\left(-\frac{t}{RC}\right) \ , \tag{5.3}$$

so the dissipated power per transition (either positive or negative)

$$P = \frac{V^2}{R} \int_0^\infty \exp\left(-\frac{t}{RC}\right) dt = CV^2 \ . \tag{5.4}$$

If transitions occur at a frequency f, the power dissipation

$$P = fCV^2 \ . \tag{5.5}$$

Thus, the power dissipation increases with clock frequency and the square of the logic swing.

Fast logic is time-critical. It relies on logic operations from multiple paths coming together at the right time. Valid results depend on maintaining minimum allowable overlaps (*e.g.* AND) and setup times (latches). Each logic circuit has a finite propagation delay, which depends on circuit loading, *i.e.* how many loads the circuit has to drive. In addition, as illustrated in Figure 5.6 the wiring resistance and capacitive loads introduces delay. This depends on the number of circuits connected to a wire or trace, the length of the trace and the dielectric constant of the substrate material. Relying on control of circuit and wiring delays to maintain timing requires great care, as it depends on circuit variations and temperature. In principle all of this can be simulated, but in complex systems there are too many combinations to test every one. A more robust solution is to use synchronous systems, where the timing of all transitions is determined by a master clock. Generally, this does not provide the utmost speed and requires some additional circuitry, but increases reliability. Nevertheless, clever designers frequently utilize asynchronous logic. Sometimes it succeeds ... and sometimes it doesn't.

5.1.3 *Logic arrays*

Commodity integrated circuits with basic logic blocks are readily available, *e.g.* with four NAND gates or two flip-flops in one package. These can be combined to form simple digital systems. However, complex logic systems are no longer designed using individual gates. Instead, logic functions are described in a high-level language (*e.g.* VHDL), synthesized using design libraries, and implemented as custom ICs – "ASICs" (application specific ICs) – or programmable logic arrays. In these implementations the digital circuitry no longer appears as an

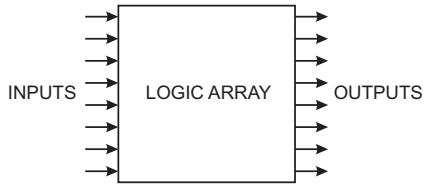

FIG. 5.7. Complex logic circuits are commonly implemented using logic arrays that as an integrated block provide the desired outputs in response to specific input combinations.

ensemble of inverters, gates, and flip-flops, but as an integrated logic block that provides specific outputs in response to various input combinations. This is illustrated in Figure 5.7. Field Programmable Gate or logic Arrays (FPGAs) are a common example. A representative FPGA has 512 pads usable for inputs and outputs, $\sim 10^6$ gates, and $\sim$ 100K of memory. Modern design tools also account for propagation delays, wiring lengths, loads, and temperature dependence. The design software also generates "test vectors" that can be used to test finished parts. Properly implemented, complex digital designs can succeed on the first pass, whether as ASICs or as logic or gate arrays.

5.2 Digitization of pulse height and time

For data storage and subsequent analysis the analog signal at the shaper output must be digitized. Important parameters for analog-to-digital converters (ADCs or A/Ds) used in detector systems are:

1. Resolution: The "granularity" of the digitized output.
2. Differential nonlinearity: How uniform are the digitization increments?
3. Integral nonlinearity: How much does the relationship of the digital output to the analog input deviate from strict proportionality?
4. Conversion time: How much time is required to convert an analog signal to a digital output?
5. Count-rate performance: How quickly can a new conversion commence after completion of a prior one without introducing deleterious artifacts?
6. Stability: Do the conversion parameters change with time?

Instrumentation ADCs used in industrial data acquisition and control systems share most of these requirements. However, detector systems place greater emphasis on differential nonlinearity and count-rate performance. The latter is important, as detector signals often occur randomly, in contrast to systems where signals are sampled at regular intervals. As in amplifiers, if the DC gain is not precisely equal to the high-frequency gain, the baseline will shift. Furthermore, following each pulse it takes some time for the baseline to return to its quiescent level. For periodic signals of roughly equal amplitude these baseline deviations

will be the same for each pulse, but for a random sequence of pulse with varying amplitudes, the instantaneous baseline level will be different for each pulse and affect the peak amplitude.

5.2.1 *ADC parameters*

5.2.1.1 *Digitizer resolution* Digitization incurs approximation, as a continuous signal distribution is transformed into a discrete set of values. To reduce the additional errors (noise) introduced by digitization, the discrete digital steps must correspond to a sufficiently small analog increment. For an accurate measurement, the resolution of the ADC must be significantly better than the noise level of the signal. Since pulse amplitudes varying within the digitization interval ΔV yield the same digitization result, the rms error

$$\sigma_v^2 = \int_{-\Delta V/2}^{\Delta V/2} \frac{v^2}{\Delta V}\, dv = \frac{\Delta V^2}{12} \tag{5.6}$$

or for an ADC with a full scale range V and n-bit resolution

$$\sigma_v^2 = \frac{2^{-2n} V^2}{12}\ . \tag{5.7}$$

This digitization noise must be smaller than the noise level of the analog input.

Another consideration is settling time. Given a single pole response $V(t) = V_0 \exp(-t/\tau)$, for a given precision $\Delta V/V_0$ the settling time $t = -\tau \log(\Delta V/V_0)$. To achieve a precision of 10^{-4} one must wait 9.2τ before acquiring the signal.

Apart from these considerations, the simplistic assumption is that the number of output bits n determines the digitizer resolution, $\Delta V = V/2^n$. For example, 13 bits yield $\Delta V/V = 1/8192 = 1.2 \cdot 10^{-4}$.

If we plot the probability *vs.* pulse amplitude that a pulse height corresponding to a specific output bin is actually converted to that address or bin, an ideal ADC would show the response illustrated in Figure 5.8. In reality, the channel profile is not rectangular as sketched above. Assigning analog amplitudes to digital bins involves a threshold comparator. As in every amplitude measurement,

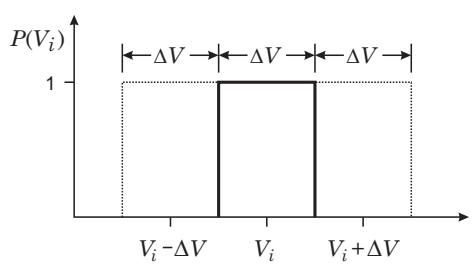

FIG. 5.8. Ideal ADC channel profiles.

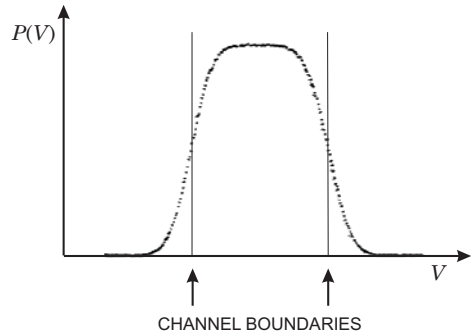

FIG. 5.9. Measured channel profile of a 13-bit ADC.

the accuracy of the threshold discrimination is subject to electronic noise. As a result, the edges of the channel profile will be "smeared" by electronic noise in the digitizer circuitry. Figure 5.9 shows the measured channel profile of a high-quality 13-bit ADC. In this example about 70% of the events within the channel boundaries are actually converted into the correct bin. The profiles of adjacent channels overlap, as shown in Figure 5.10.

These channel profiles were measured by scanning a precision pulser across a channel and recording the fraction of pulses converted into the proper digital bin. However, channel profile can be checked quickly by applying the output of a precision pulser to the ADC and carefully adjusting the output amplitude to the center of a digital bin. If the pulser output has very low noise, *i.e.* the amplitude jitter is much smaller than the voltage increment corresponding to one ADC channel, nearly all pulses will be converted to a single channel, with only a small fraction appearing in the neighbor channels. However, this is only true for well-designed ADCs. Figure 5.11 shows results from an ADC whose digital resolution is better than its analog resolution. In the 13-bit range the pulser

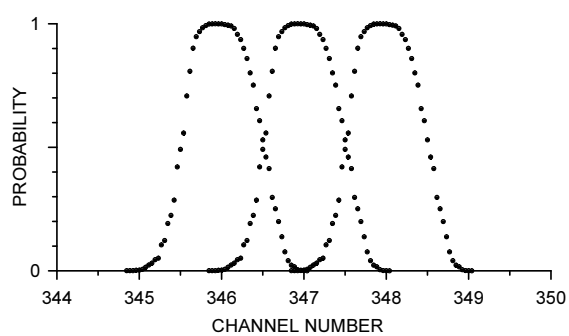

FIG. 5.10. The channel profiles of adjacent channel overlap.

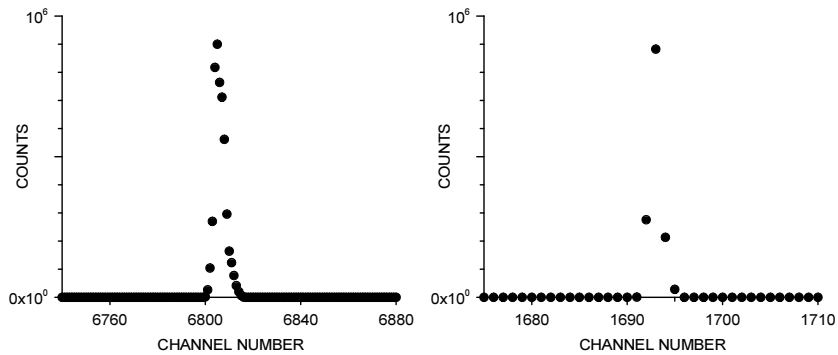

FIG. 5.11. Spectrum of a precision pulser centered within an ADC channel. The maximum number of counts per channel is about 10^6. In the 13-bit range (left) the signal is distributed over many channels. In the 11-bit range the spectrum is matches the digital resolution. Although this ADC can provide 13 bits of digital resolution, its analog resolution is only 10 – 11 bits.

signal is distributed over > 12 channels, whereas in the 11-bit range the digital resolution matches the analog resolution. Although this ADC can provide 13 bits of digital resolution, its analog resolution is only 10 – 11 bits, so the 12th and 13th bits are superfluous.

How much ADC resolution is required? If all counts of a peak fall in one bin, the resolution is ΔV. If the counts are distributed over several bins, peak fitting can yield substantially better resolution, depending on statistics. Figure 5.12 shows a signal with a constant width digitized with bin widths of $\Delta V = 2\sigma$, σ, 0.5σ, and 0.25σ. Fitting can determine the centroid position to a fraction of the bin width even with coarse digitization, if only one peak is present and the line shape is known. Five digitizing channels within a linewidth (FWHM) allow robust peak fitting and centroid finding, even for imperfectly known line shapes and overlapping peaks.

5.2.1.2 *Differential nonlinearity* Differential nonlinearity (DNL) is a measure of the uniformity of channel profiles over the range of the ADC. Depending on the nature of the distribution, either a peak or an rms specification may be appropriate:

$$\mathrm{DNL} = \max\left\{\frac{\Delta V(i)}{\langle \Delta V\rangle} - 1\right\} \quad \text{or} \quad \mathrm{DNL} = \mathrm{rms}\left\{\frac{\Delta V(i)}{\langle \Delta V\rangle} - 1\right\} \ , \tag{5.8}$$

where $\langle \Delta V\rangle$ is the average channel width and $\Delta V(i)$ is the width of an individual channel.

Differential nonlinearity of $< \pm 1\%$ max. is typical, but state-of-the-art ADCs can achieve 10^{-3} rms, *i.e.* the variation is comparable to the statistical fluctuation for 10^6 random counts. Instrumentation ADCs are often specified with an

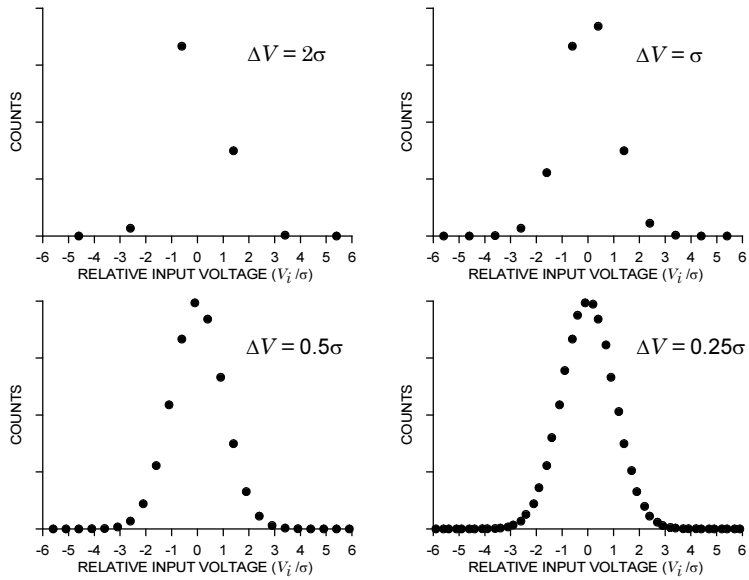

FIG. 5.12. Digitized spectra of a Gaussian peak whose width $\sigma = 1$. ADC resolution ΔV is increased by factors of two from 2σ (top left) to 0.25σ (bottom right).

accuracy of ± 0.5 LSB (least significant bit) or greater, so the differential nonlinearity may be 50% or more. If the differential nonlinearity exceeds ± 0.5 LSB, the conversion can be nonmonotonic. For certain analog values an increase in signal will lead to a decreased digitized result. Figure 5.13 shows some typical plots of differential nonlinearity, both with a suppressed zero, so that the DNL is visible. The signal spectrum was "white" and is a section of the Compton con-

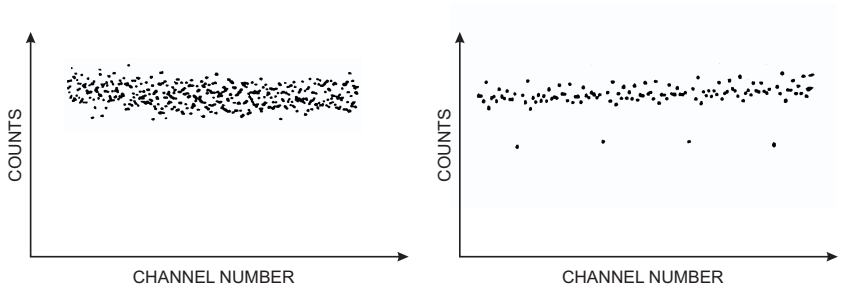

FIG. 5.13. Response of two ADC to a "white" spectrum, both vertical scales with suppressed zeros. The left-hand plot shows a random DNL distribution, whereas the right-hand plot (with 1/10 vertical scale) shows pronounced periodic structures.

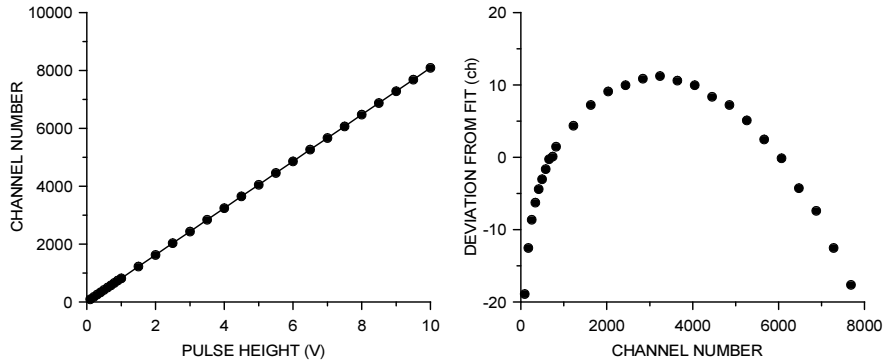

FIG. 5.14. Integral nonlinearity measurement on a 13-bit ADC. Left the digitized output is plotted *vs.* input amplitude. The right-hand plot shows the deviation of the digitized output from a straight-line fit.

tinuum from a plastic scintillator, so the spectrum is smooth. Sufficient counts were accumulated so that the statistical deviations were much smaller than the DNL. The left hand plot shows a random distribution of DNL, whereas the right hand plot shows periodic structures in the DNL. Poor ADC designs often show an odd-even effect, where the widths of alternating bins differ systematically.

5.2.1.3 *Integral nonlinearity* Integral nonlinearity measures the deviation from proportionality of the measured amplitude to the input signal level. Figure 5.14 shows the channel number *vs.* input amplitude and the deviation of the output from a straight-line fit.

The linearity of an ADC depends on the input pulse shape and duration, due to bandwidth limitations in the circuitry. The integral nonlinearity shown above

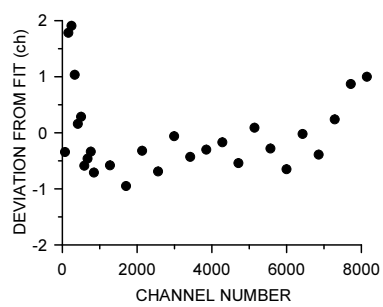

FIG. 5.15. Integral nonlinearity measured with a 3 μs wide pulse, instead of the 400 ns pulse width used in Figure 5.14.

was measured with a 400 ns wide input pulse. Increasing the pulse width to 3 μs improved the result significantly, as shown in Figure 5.15.

5.2.1.4 *Conversion time* During the digitization of a signal the system cannot accept a subsequent signal ("dead time"). The dead time results from several successive steps in the conversion process:

1. Signal acquisition time, which equals the time-to-peak plus settling time.
2. Conversion time, which can depend on pulse height.
3. Readout time to memory, which depends on speed of data transmission, buffer memory access and writing to mass storage.

In pulsed beam experiments dead time can be ignored if it is smaller than the pulse rate, so that conversion and data storage are complete before the next event. However, in continuous event streams, unless the event rate is very low, the measurement of yields or reaction cross-sections requires a measurement of dead time, *e.g.* with a reference pulser fed simultaneously into the spectrum. The total number of reference pulses issued during the measurement is counted and compared with the number of pulses recorded in the spectrum.

As will be seen below, the conversion time can depend on the pulse height. Does this mean that the efficiency is a function of pulse height? Usually not. If events in different parts of the spectrum are not correlated in time, *i.e.* random, they are all subject to the same average dead time (although this average will depend on the spectral distribution). However, be cautious when events are correlated. For example, in decay chains where the lifetime is less than the dead time, the daughter decay will be lost systematically.

5.2.1.5 *Count rate effects* Circuitry in ADCs is mostly analog, so as in amplifiers one often encounters internal baseline shifts with event rate or undershoots following a pulse. If signals occur at constant intervals, the effect of an under-

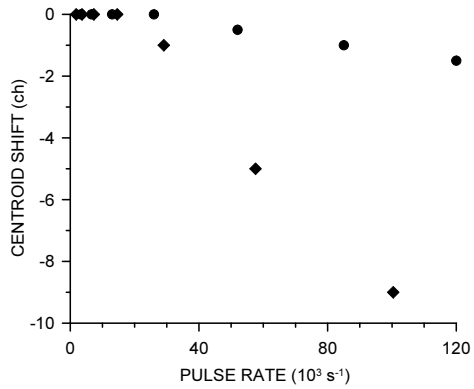

FIG. 5.16. Centroid shift *vs.* pulse rate of two 13-bit ADCs (8192 channels).

shoot will always be the same. However, in a random sequence of pulses, the effect will vary from pulse to pulse, which leads to spectral broadening.

Baseline shifts tend to manifest themselves as a systematic shift in centroid position with event rate. Measured results for two 13-bit ADCs subjected to a random rate are shown in Figure 5.16. At rates approaching $4 \cdot 10^4$ s^{-1} the centroid shift of the inferior unit is sufficiently large to cause significant resolution degradation.

5.2.1.6 *Stability* The conversion gain and baseline are subject to change with time and temperature. Stability *vs.* temperature is usually adequate with modern electronics in a laboratory environment, especially since temperature changes within an enclosure or integrated circuit are typically much smaller than ambient changes. However, in highly precise or long-term measurements one should monitor changes in gain and baseline of the overall system. A simple technique is to inject precision reference pulses to place a reference peak at both the low and high end of the spectrum. The difference between the two peaks yields the gain, and the position of either peak then determines the offset.

5.2.2 *Analog-to-digital conversion techniques*

Analog-to-digital converters suitable for the digitization of individual pulses tend to use variations of a few basic techniques. Here we just review some basic conversion principles to illustrate the strengths and weaknesses of different conversion techniques. Analog-to-digital converters are key components in many applications, so a wealth of literature can be found on the world wide web. Application

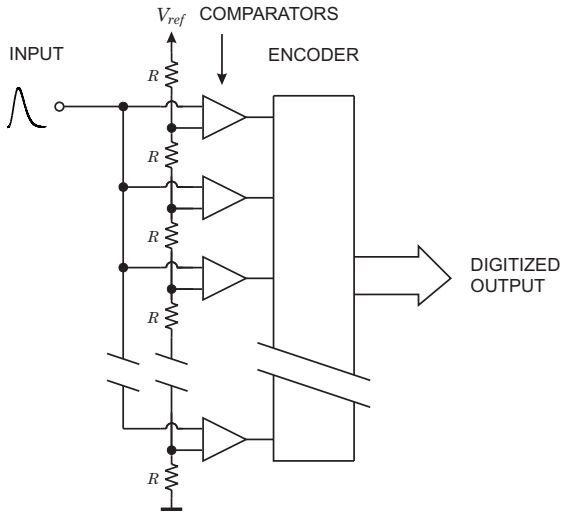

FIG. 5.17. Block diagram of a flash ADC.

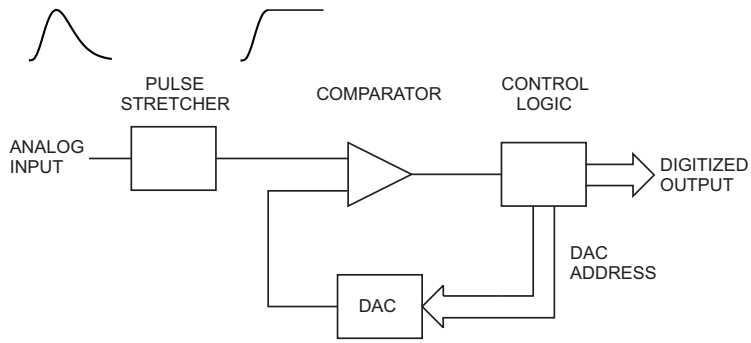

FIG. 5.18. Principle of a successive approximation ADC. The DAC is controlled to sequentially add levels proportional to $2^n, 2^{n-1}, \ldots 2^0$. The corresponding bit is set if the comparator output is high (DAC output < pulse height).

notes from major integrated circuit houses are a good source. Horowitz and Hill (1989) also discuss ADC techniques.

5.2.2.1 *Flash ADC* Conceptually, the simplest technique is flash conversion, illustrated in Figure 5.17. The signal is fed in parallel to a bank of threshold comparators. The individual threshold levels are set by a resistive divider. The comparator outputs are encoded such that the output of the highest level comparator that fires yields the correct bit pattern. The threshold levels can be set to provide a linear conversion characteristic where each bit corresponds to the same analog increment, or a nonlinear characteristic, to provide increments proportional to the absolute level, which provides constant relative resolution over the range.

The big advantage of this scheme is speed; conversion proceeds in one step and conversion times < 10 ns are readily achievable. The drawbacks are component count and power consumption, as one comparator is required per bin. For example, an eight-bit converter requires 256 comparators. The conversion is always monotonic and differential nonlinearity is determined by the matching of the resistors in the threshold divider. Only relative matching is required, so this topology is a good match for monolithic integrated circuits. Flash ADCs are available with conversion rates > 500 MS/s (megasamples per second) at eight-bit resolution. The power dissipation is about 5 W. A practical issue is the high input capacitance of many comparator inputs in parallel, so the driver must have sufficient current drive capability to charge up this capacitance commensurate with the fast conversion time. The required settling time increases the conversion time at high resolution, as $V = V_0(1 - e^{-t/\tau})$, so for the signal to approach its peak value to a precision of 10^{-3} requires a time of seven time constants τ.

5.2.2.2 *Successive approximation ADC* The most commonly used technique is the successive approximation ADC, shown in Figure 5.18. The input pulse is sent

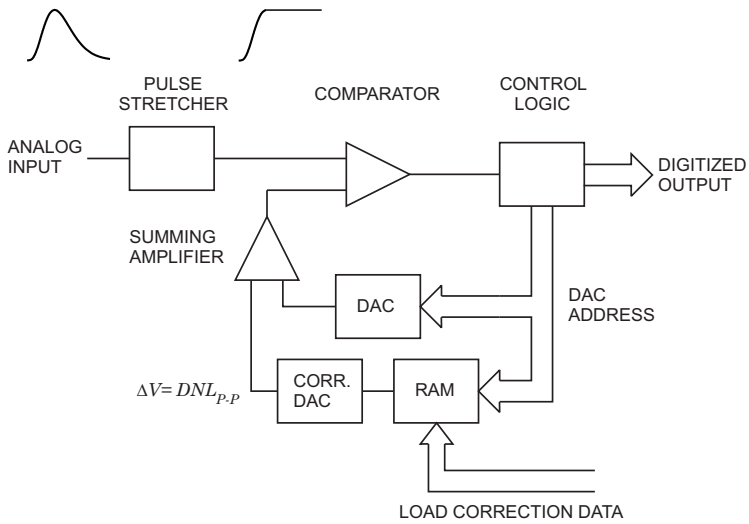

FIG. 5.19. A correction DAC can be used to improve differential nonlinearity.

to a pulse stretcher, which follows the signal until it reaches its cusp and then holds the peak value. The stretcher output feeds a comparator, whose reference is provided by a digital-to-analog converter (DAC). The DAC is cycled beginning with the most significant bits. The corresponding bit is set when the comparator fires, *i.e.* the DAC output becomes less than the pulse height. Then the DAC cycles through the less significant bits, always setting the corresponding bit when the comparator fires. Thus, n-bit resolution requires n steps and yields 2^n bins. This technique makes efficient use of circuitry and is fairly fast. High-resolution devices (16 – 20 bits) with conversion times of order μs are readily available. Currently, a 16-bit ADC with a conversion time of $1\,\mu$s (1 MS/s) requires about 100 mW.

A common limitation is differential nonlinearity, since the resistors that set the DAC levels must be extremely accurate. For DNL $< 1\%$ the resistor determining the 2^{12} level in a 13-bit ADC must be accurate to $< 2.4 \cdot 10^{-6}$. As a consequence, differential nonlinearity in high-resolution successive approximation converters is typically 10 – 20% and often exceeds the 0.5 LSB required to ensure monotonic response.

The differential nonlinearity can be corrected by various techniques. One is to average over many channel profiles for a given pulse amplitude, the "sliding scale" technique originated by Gatti (Cottini, Gatti, and Svelto 1963). Here an analog increment is added event-by-event and the digitized output is corrected accordingly. Thus, for a large number of events the conversion of a given pulse amplitude utilizes many states of the converter. For a random amplitude distribution this averages over many channel profiles and equalizes the differential

nonlinearity. When properly implemented this provides excellent differential nonlinearity with no significant degradation of the channel profile. However, flawed implementations are prone to step-like discontinuities in the DNL *vs.* amplitude.

Another technique is the "brute force" approach of using a correction DAC. The primary DAC output is adjusted by the output of a correction DAC to reduce differential nonlinearity. This is shown in Figure 5.19. Correction data are derived from a measurement of DNL. Corrections for each bit are loaded into the RAM, which acts as a lookup table. For each address of the main DAC the appropriate correction is applied to the correction DAC. The range of the correction DAC must exceed the peak-to-peak differential nonlinearity. If the correction DAC has N bits, the maximum DNL is reduced by $2^{-(N-1)}$ (if the deviations are symmetrical).

5.2.2.3 *Wilkinson ADC* The Wilkinson ADC (Wilkinson 1950) has traditionally been the mainstay of precision pulse digitization. The principle is shown in Figure 5.20. The peak signal amplitude V is acquired by a combined peak detector/pulse stretcher and transferred to a memory capacitor C. The output of the peak detector initiates the conversion process:

1. The memory capacitor is disconnected from the stretcher.

2. A current source is switched on to linearly discharge the capacitor with current I_R.

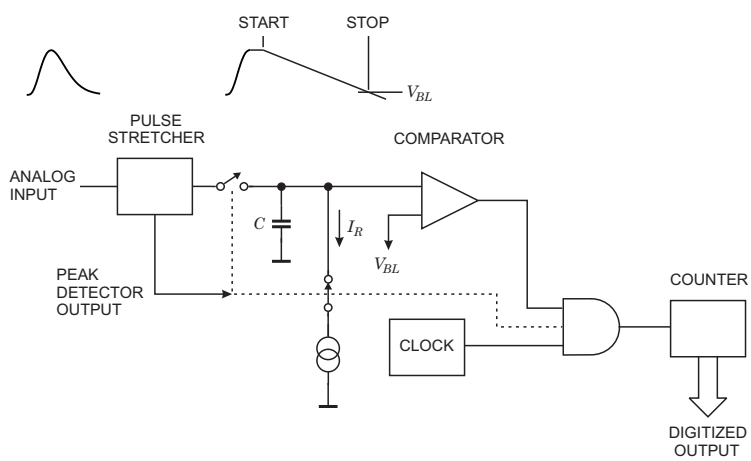

FIG. 5.20. Principle of a Wilkinson ADC. After the peak amplitude has been acquired, the output of the peak detector initiates the conversion process. The memory capacitor is discharged by a constant current while counting the clock pulses. When the capacitor is discharged to the baseline level V_{BL} the comparator output goes low and the conversion is complete.

3. Simultaneously with commencing the discharge a counter is enabled to determine the number of clock pulses until the voltage on the capacitor reaches the baseline level V_{BL}.

The time required to discharge the capacitor is a linear function of pulse height,

$$T_C = C \cdot \frac{V - V_{BL}}{I_R} \ , \tag{5.9}$$

so the counter content provides the digitized pulse height. The clock pulses are provided by a crystal oscillator, so the time between pulses is extremely uniform and this circuit inherently provides excellent differential linearity. The drawback is the relatively long conversion time T_C, which for a given resolution is proportional to the clock period T_{clk} and the pulse height, $T_C = n \times T_{clk}$ (n = channel number $\propto$ pulse height). For example, a clock frequency of 100 MHz provides a clock period $T_{clk} = 10$ ns and a maximum conversion time $T_C = 82\,\mu$s for 13 bits. Clock frequencies of 100 MHz are typical, but > 400 MHz have been implemented with excellent performance (DNL $< 10^{-3}$). This scheme makes efficient use of circuitry and allows low power dissipation. Wilkinson ADCs have been implemented in 128-channel readout ICs for silicon strip detectors (Garcia-Sciveres *et al.* 1999). Each ADC added only 100 μm to the length of a channel and a power of 300 μW per channel (see Chapter 8).

Many important details are not shown in Figure 5.20. For example, the beginning of the discharge must be synchronized with the clock. Switching the current source requires some time, which introduces nonlinearity for small signals. Cross-talk from the clock or counter to the analog circuitry can introduce correlations into the differential nonlinearity, as illustrated in Figure 5.13. It is tempting to utilize both the leading and trailing edge of the clock pulse to double the clock frequency and reduce conversion time. However, the duty cycle of the clock pulse must be constrained very accurately to 50% to avoid degradation of differential nonlinearity. This technique typically leads to odd–even structures in the DNL, so the least significant bit can become unusable. Simply suppressing this bit also reduces the conversion time two-fold, so "clock doubling" becomes self-defeating.

5.2.2.4 *Hybrid analog-to-digital converters* Conversion techniques can be combined to obtain high resolution and short conversion time. One example combines a flash ADC with a successive approximation or a Wilkinson (ramp run-down) converter. The fast flash ADC provides coarse conversion (*e.g.* 8 out of 13 bits) and the successive approximation or Wilkinson converter provides fine resolution. Since the second conversion range is small, the conversion time is significantly reduced. For example, a Wilkinson ADC covering 256 channels with a 100 MHz clock requires only 2.6 μs, which is comparable to a successive approximation ADC, but with superior differential nonlinearity.

Another approach is to use flash ADCs with sub-ranging. Not all applications require constant absolute resolution over the full range. Sometimes only relative

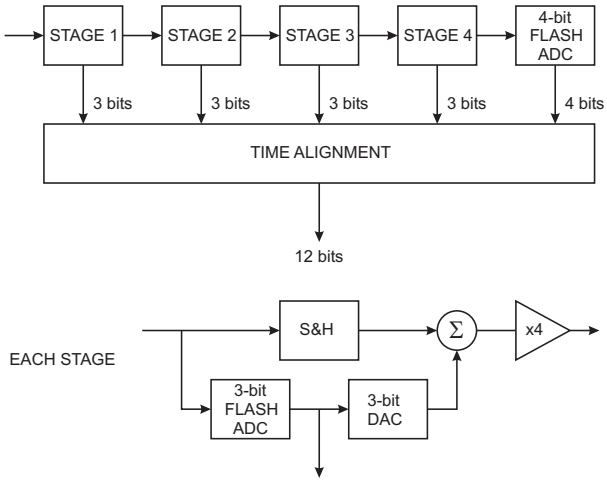

FIG. 5.21. A 12-bit pipelined ADC. The first four stages have a 3-bit output, but only 2-bit resolution, so the first four stages provide 8 bits. The last stage is a flash ADC that provides the final 4 bits.

resolution must be maintained, especially in systems with a very large dynamic range.

Sub-ranging utilizes a precision binary divider at the input to determine the coarse range and a fast flash ADC for fine digitization. One example is a fast digitizer that fits in phototube base and provides 17 to 18 bit dynamic range with 16 ns conversion time (Yarema *et al.* 1993, Zimmerman and Hoff 2004). The converter provides a digital floating point output (4 bit exponent, $8 + 1$ bit mantissa).

A popular architecture is the pipelined ADC, which consists of sequential conversion steps, as illustrated in Figure 5.21. The input to each stage is fed both to a sample and hold and a three-bit flash ADC. The sample and hold (S&H) maintains the signal level during conversion. The flash ADC quantizes its input to 3-bit accuracy. This output is fed to a DAC with 12-bit accuracy. The DAC's analog output is subtracted from the original signal and the difference signal is passed on to the next stage. The last 4 bits are resolved by a 4-bit flash ADC. As soon as a stage has passed its result to the next stage it can begin processing the next signal, so throughput is not determined by the total conversion time, but by the time per stage. Since outputs from individual stages appear sequentially, the outputs must be aligned in time to form the cumulative digitized output. Since the interstage gain is only four (rather than eight corresponding to 3 bits), each stage only contributes 2 bits of resolution. The extra bit is used for error correction. Commercially available pipelined ADCs provide 1 GS/s conversion rates with eight-bit resolution and a power dissipation of about 1.5 W. Note that

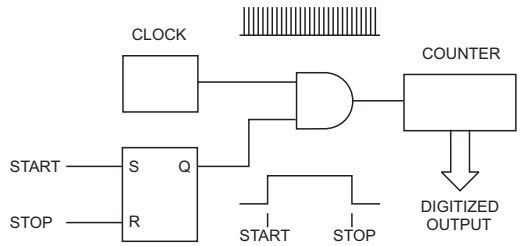

FIG. 5.22. The simplest form of time digitizer counts the number of clock pulses between the start and stop signals.

the effective resolution at the maximum sampling rate is less than the digital resolution.

Other techniques, the sigma-delta ADC being a notable example, measure incremental changes over the waveform. This architecture is popular in audio applications, so the frequencies are much lower than needed for the digitization of detector pulses.

5.3 Time-to-digital converters (TDCs)

Measurements of time intervals can utilize digital and analog techniques.

5.3.1 *Counter*

The combination of a clock generator with a counter is the simplest technique, shown in Figure 5.22. The clock pulses are counted between the start and stop signals, which yields a direct readout in real time. The limitation is the speed of the counter, which in current technology is limited to about 1 GHz, yielding a time resolution of 1 ns. Using the stop pulse to strobe the instantaneous counter status into a register provides multi-hit capability.

5.3.2 *Analog ramp*

Analog techniques are commonly used in high-resolution digitizers to provide resolution in the range of ps to ns. The principle is to convert a time interval into a voltage by charging a capacitor through switchable current source. The start pulse turns on the current source and the stop pulse turns it off. The resulting voltage on the capacitor C is $V = Q/C = I_T(t_{stop} - t_{start})/C$, which is digitized by an ADC. A convenient implementation switches the current source to a smaller discharge current I_R and uses a Wilkinson ADC for digitization, as illustrated in Figure 5.23. This technique provides high resolution, but at the expense of dead time and multi-hit capability.

5.3.3 *Digitizers with clock interpolation*

Integrated circuit technology makes it practical to implement clock interpolation to provide ps resolution together with multi-hit capability and no dead time

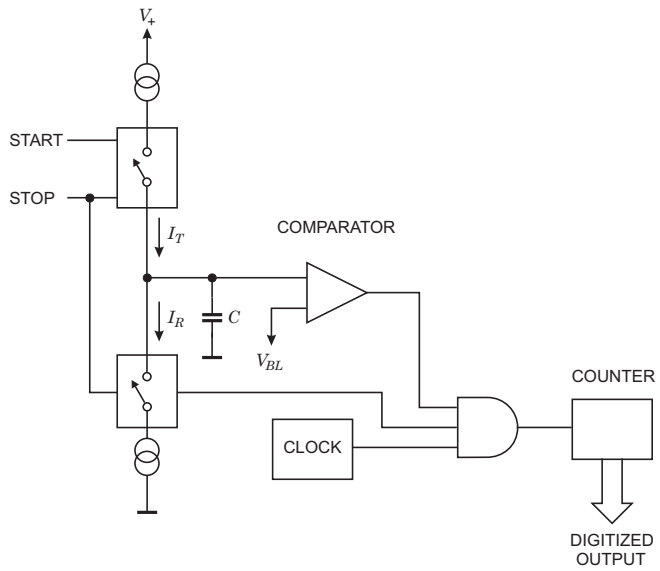

FIG. 5.23. Combining a time-to-amplitude converter with an ADC forms a time digitizer capable of ps resolution. The memory capacitor C is charged by the current I_T for the duration $T_{start} - T_{stop}$ and subsequently discharged by a Wilkinson ADC.

(Arai and Ohsugi 1986, 1989). A block diagram is shown in Figure 5.24. The clock period is interpolated by inverter delays (U1, U2, ...). The delay can be fine-tuned by adjusting the operating current of the inverters. This does not provide very tight control against temperature or voltage variations, so the delays are stabilized by a delay-locked loop referenced against the master clock, which is typically a very stable crystal oscillator, as indicated at the bottom of Figure 5.24 (Arai *et al.* 1998). The delay-locked loop ensures that the total delay of the interpolation chain is an integer multiple of the clock period. Devices with 250 ps resolution have been fabricated and tested for use in high-energy physics experiments, but the technique should be applicable to higher resolution digitizers.

5.4 Digital signal processing

Up to now we have utilized analog techniques for pulse shaping. However, filtering can also be applied in the digital domain. This is a topic worthy of a book in itself, so this will only be a brief introduction designed to provide some perspective relevant to large-scale detector systems. For a more detailed discussion of digital signal processing techniques see texts by Ifeachor and Jervis (1993), Oppenheimer and Schafer (1998), and others. For examples applied to detector pulse processing see Pullia *et al.* (2000) and Cardoso *et al.* (2004), which also give additional references.

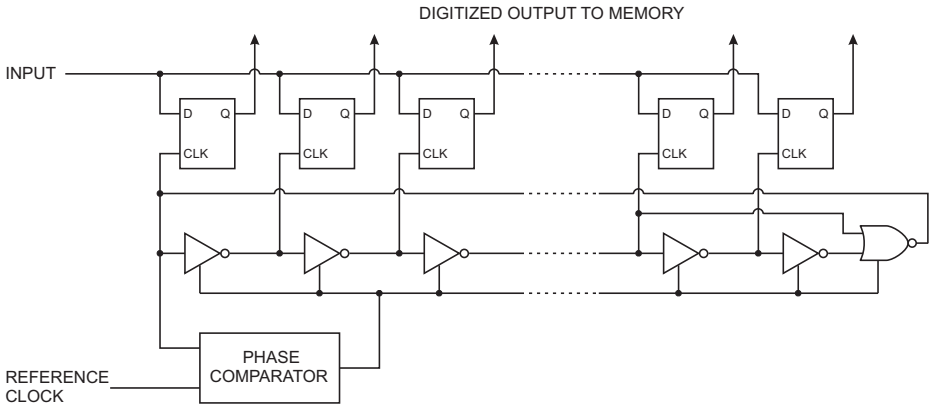

FIG. 5.24. Time digitizer using clock interpolation. The interpolator delays are controlled by a delay-locked loop referenced to the master clock oscillator (Arai *et al.* 1998).

First, the detector signal is sampled with a fast digitizer with sufficient resolution to reconstruct the pulse, as shown in Figure 5.25. Subsequently, a digital signal processor (DSP) applies the appropriate algorithms to filter the pulse and extract the pulse height (Figure 5.26). Digital signal processing allows great flexibility in implementing filtering functions. The software can be changed readily to adapt to a wide variety of operating conditions and it is possible to implement filters that are impractical or even impossible using analog circuitry. However, this comes at the expense of increased circuit complexity and increased demands on the ADC compared to analog shaping.

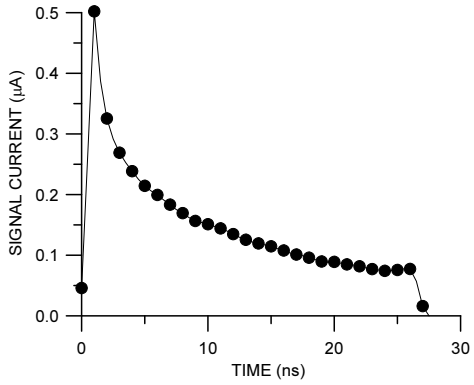

FIG. 5.25. Sampling a pulse to allow digital signal processing. The pulse shown is the current pulse from a strip detector.

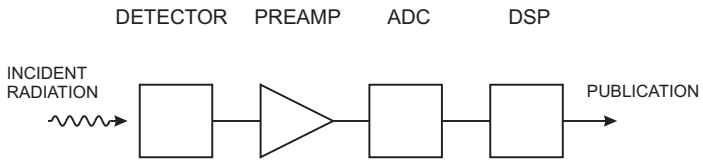

FIG. 5.26. Block diagram of a detector readout using digital signal processing.

Figure 5.27 illustrates how a filter function can be implemented using digital techniques. The amplitude of the input signal is multiplied at each discrete time step by a filter weighting function. The filter function can be calculated in real time by the DSP or it can be stored as values in a look-up table. This process could be applied to either a continuous or a digitized input signal. Subsequently the samples are integrated. Since the amplitudes add coherently, whereas the noise components add in quadrature, this yields a net improvement in signal-to-noise ratio. It is also rather straightforward to show that the optimum signal-to-noise ratio obtains when the weighting function has the same shape as the input signal. This is an example of a "matched filter". However this is only the optimum filter for retrieving the signal while retaining its shape. As we have seen, integrating the signal to extend its duration and then filtering decouples the choice of filter parameters from the original signal duration.

The simple scheme shown in Figure 5.27 requires that the time of the desired signals is known, so the weighting factors can be synchronized with the signal. This constraint is removed when the filtering is performed by convolution, so the DSP block in Figure 5.26 performs a sequence of multiplications and sums

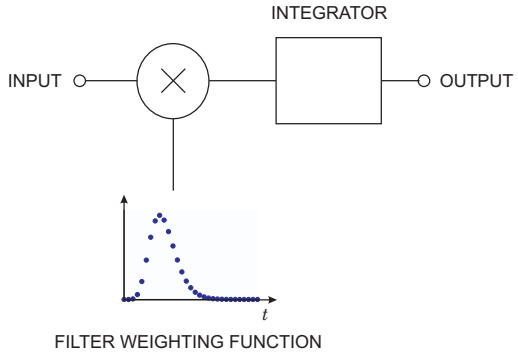

FIG. 5.27. In a simple digital filter the input signal is multiplied at each discrete time step by a filter weighting function.

$$S_o(n) = \sum_{k=0}^{N-1} W(k) \cdot S_i(n-k) \ , \tag{5.10}$$

where S_o and S_i are the output and input signals and W is the weighting function that yields the desired pulse shape. This is analogous to pulse shaping in analog systems (eqn 4.1). In digital signal processing this is referred to as a finite impulse response (FIR) filter, similar to an infinite impulse response (IIR) filter, which takes the sum to infinity. Specialized digital signal processors optimized to perform these functions are available, but FPGAs also allow very efficient implementations. Without special hardware, algorithms can be tested on a desktop computer using realistic detector pulses and noise spectra to assess artifacts in the output spectrum, for example using C^{++} functions (Embree and Danieli 1999).

The sample interval must be sufficiently small to capture the pulse structure. Figure 5.28 shows the same pulse as in Figure 5.25, but sampled at intervals of 4 ns instead of 1 ns. The sampling interval of 4 ns misses the initial peak.

This illustrates the Nyquist criterion. The ADC must digitize at greater than twice the rate of the highest frequency component in the signal. Apart from missing information on the fast components of the pulse, undersampling introduces spurious artifacts. With too low a sampling rate high frequency components will be "aliased" to lower frequencies, as shown in Figure 5.29.

To prevent aliasing, a low-pass filter must be introduced before the ADC. As a result, an additional analog block must be added to the signal processing chain (Figure 5.30). When an input frequency f_i is sampled at a rate f_s, the output frequencies can be reconstructed as $f_i \pm k f_s$, where k is any integer value. Thus, the input is aliased to both lower and higher frequencies and the prefilter ("anti-aliasing filter") is needed to exclude both possibilities. Every sampling process is subject to aliasing – *e.g.* also 2D or 3D image processing.

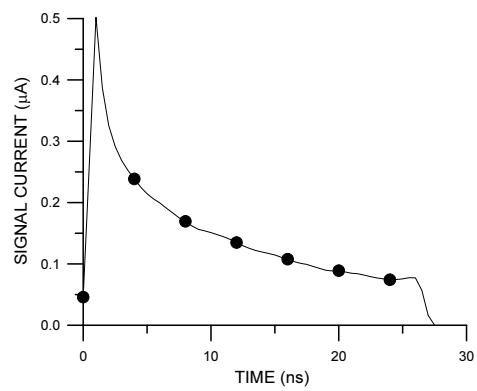

FIG. 5.28. Sampling at too low a rate does not preserve the full pulse structure.

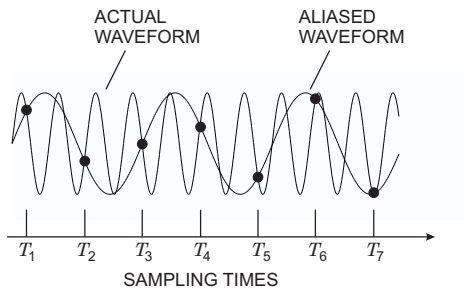

FIG. 5.29. Sampling at too low a rate "aliases" high-frequency components to lower frequencies.

The preamplifier is necessary to raise the level of the input noise sources such that the digitization noise of the ADC is negligible. As already noted in Section 5.2.1.1, the signal quantization inherent to the digitization process introduces quasi-random noise

$$\sigma_n = \frac{\Delta V}{\sqrt{12}} \ , \tag{5.11}$$

where ΔV is the signal increment corresponding to one bit. This quantization noise is increased by differential nonlinearity. When the Nyquist condition is fulfilled the noise is spread nearly uniformly and extends to 1/2 the sampling frequency f_S, so the spectral noise density

$$e_n = \frac{\sigma_n}{\sqrt{\Delta f_n}} = \frac{\Delta V}{\sqrt{12}} \cdot \frac{1}{\sqrt{f_S/2}} = \frac{\Delta V}{\sqrt{6 f_S}} \ . \tag{5.12}$$

Sampling at a higher frequency spreads the total noise over a larger frequency range, so oversampling can be used to increase the effective resolution.

From this we see that the front-end electronics and ADC must exhibit the same precision as in an analog system, *i.e.* the baseline and other pulse-to-pulse amplitude fluctuations must be less than order $Q_n/10$, *i.e.* typically 10^{-4} in high-resolution systems. For 10 V full scale at the ADC input in a high-resolution gamma-ray detector system, this corresponds to < 1 mV. In practice the effective

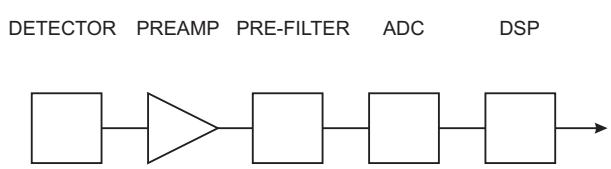

FIG. 5.30. A low-pass filter (prefilter) inserted in the ADC input prevents aliasing of high-frequency components into the desired frequency range.

resolution of ADCs suitable for these applications is commonly 2 bits worse than nominal, so this must be taken into account. At very high resolution the electronic noise of the ADC's input circuitry becomes the limit. For example, in a 24-bit ADC with a full-scale range of 10 V one bit corresponds to a voltage difference of 0.4 nV. The thermal noise of a 50 Ω resistor in 1 Hz bandwidth is more than twice as large. Furthermore, the dynamic range requirements for the ADC may be more severe than in an analog filtered system, as can be seen from the rather high peak-to-average ratio of the pulse in Figure 5.25. In any case, the ADC must provide high performance at short conversion times.

Today digital signal processing is technically feasible for some applications, *e.g.* detectors with moderate to long collection times (gamma and x-ray detectors), and systems are commercially available. Nevertheless, these systems tend to be complex and power-hungry.

In large-scale systems, however, the benefits are not so clear. Where intimate integration of sensors and electronics in a small volume is required, both circuit area and power dissipation are crucial considerations. Furthermore, these are special purpose systems. The electronics are specifically tailored to the sensor and application and do not need to be modified during the course of the experiments (the inevitable upgrades notwithstanding). Furthermore, simple analog filters usually provide results that are only slightly inferior to the optimized filters that a DSP system would allow.

The benefits of digital signal processing are:

1. Flexibility in implementing filter functions.
2. Filters are possible that are impractical in hardware.
3. Filter parameters can be changed simply.
4. Tail cancellation and pile-up rejection are easily incorporated.
5. Adaptive filtering can be used to compensate for pulse shape variations.

Where is digital signal processing appropriate? It provides clear benefits in systems that are highly optimized for resolution, high counting rates, and variable sensor pulse shapes.

Where is analog signal processing best (most efficient)? In systems that require fast time response the high power requirements of high-speed ADCs are prohibitive. Systems that are not sensitive to pulse shape can use fixed shaper constants and rather simple filters, which can be either continuous or sampled. For example, the APV25 chip described in Chapter 8 applies discrete sample processing using analog circuitry. Finally, in high density systems that require small circuit area and low power, analog filtering can efficiently transpose the relevant information to a frequency domain where digitization requirements are less demanding.

Given the dearth of good analog circuit designers and no prospects for improvement, it is often claimed that digital signal processing is a better match to available skills and avoids the need to understand the wide range of details that a sophisticated analog system requires. This argument is specious; both types

of systems require careful analog design. Nevertheless, progress in fast ADCs (precision, reduced power) will expand the range of DSP applications.

References

Arai, Y. and Ohsugi, T. (1986). An idea of deadtimeless readout system by using time memory cell. *Proceedings of the 1986 Summer Study on the Physics of the Superconducting Super Collider* pp. 455–457

Arai, Y. and Ohsugi, T. (1989). TMC – a CMOS time to digital converter VLSI. *IEEE Trans. Nucl. Sci.* **NS-36/1** (1989) 528–531

Arai, Y. *et al.* (1998). Time memory cell VLSI for the PHENIX drift chamber. *IEEE Trans. Nucl. Sci.* **NS-45/3** (1998) 735 – 739 and references therein.

Cardoso, J.M. *et al.* (2004). A high performance hardware reconfigurable hardware platform for digital pulse processing. *IEEE Trans. Nucl. Sci.* **NS-51/3** (2004) 921–925

Cottini, C., Gatti, E., and Svelto, V. (1963). A new method for analog-to-digital conversion. *Nucl. Instr. Meth.* **24** (1963) 241–242

Embree, P.M. and Danieli, D. (1999). *C^{++} Algorithms for Digital Signal Processing.* Prentice Hall PTR, Upper Saddle River. ISBN 0-13-179144-3, TK5102.9 E45

Garcia-Sciveres, M. *et al.* (1999). The SVX3D integrated circuit for deadtimeless silicon strip readout. *Nucl. Instr. Meth.* **A435** (1999) 58–64

Horowitz, P. and Hill, W. (1989). *The Art of Electronics.* Cambridge University Press, 2nd edition, ISBN 0-5213-7095-7

Ifeachor, E.C. and Jervis, B.W. (1993). *Digital Signal Processing – A Practical Approach.* Addison-Wesley, Wokingham. ISBN 0-201-54413-X, TK5102.I33

Oppenheimer, A.V. and Schafer, R.W. (1998). *Discrete-Time Signal Processing.* Prentice Hall, Upper Saddle River. ISBN 0-13-754920-2, TK5102.9.067

Pullia, A. *et al.* (2000). Quasi-optimum γ and x spectroscopy based on real-time digital techniques. *Nucl. Instr. Meth.* **A439** (2000) 378–384

Shannon, C.E. (1949). Communication in the presence of noise. *Proc. IRE* **37/1** (1949) 10–21, reprinted in *Proc. IEEE* **86/2** (1998) 447–457

Wilkinson, D.H. (1950). A stable ninety-nine channel pulse amplitude analyser for slow counting. *Proc. Cambridge Phil. Soc.* **46/3** (1950) 508–518

Yarema, R.J. *et al.* (1993). Wide range charge integrator and encoder ASIC for photomultiplier tubes. *IEEE Trans. Nucl. Sci.* **NS-40/4** (1993) 750–752

Zimmerman, T. and Hoff, J.R. (2004). The design of a charge-integrating modified floating-point ADC chip. *IEEE J. Solid-State Circuits* **SC-39/6** (2004) 895–905

6

TRANSISTORS AND AMPLIFIERS

As shown in Chapter 3 the electronic noise of a well-designed amplifier chain is determined by the first amplifying stage. Taking this a step further, in a well-designed amplifier stage, the noise is dominated by the input amplifying device, *e.g.* a bipolar transistor, a junction field effect transistor (JFET), or metal-oxide-semiconductor field effect transistor (MOSFET). At a basic level, one can consider individual amplifying devices as amplifiers, so the first stage must amplify its inherent noise to a level that overrides the noise of subsequent devices and other components. Understanding the noise of amplifying devices and its optimization requires some knowledge of device physics, so this chapter first describes the basic principles of bipolar and field effect transistors and how these devices are used in simple amplifiers. We then analyze the noise properties of FETs and bipolar transistors and turn to some illustrative amplifier designs that show how the various types of devices can be used effectively.

This chapter presents only an overview. The following books are recommended to those who wish to delve deeper. The texts by Shockley (1950) and Grove (1967) may appear dated, but they give excellent treatments of the basic physics. Sze (1981, 2002) provides a comprehensive overview of practically all semiconductor devices. Taur and Ning (1998) and Takeda, Yang, and Miura-Hamada (1995) give a more modern perspective relevant to modern integrated circuits. Tsividis (1987) covers MOS transistors in great detail. Baker, Li, and Boyce (1998) cover simulation and layout of integrated circuits. Wolf (1995, 2002) provides a detailed and quite up-to-date discussion of submicron integrated circuits. Clearly, this is not an exhaustive list and also reflects some personal proclivities (the primary sources often give the clearest explanations).

6.1 Bipolar transistors

Although the first patent awarded for a semiconductor amplifier described a field effect transistor, the first practical devices were bipolar transistors. The first bipolar transistors were point-contact transistors (Bardeen and Brattain 1948), which proved to be an evolutionary dead-end. All modern bipolar transistors are junction transistors (Shockley 1949), which fully exploit the presence of both majority and minority carriers. The principle is illustrated in the *npn* structure shown in Figure 6.1. The following explanation applies the signal to the base, which separates the input and output currents, but differs from many textbooks that apply the signal to the emitter. This has historical reasons, as the first transistors had poor current gain and frequency response, so applying the signal to the emitter was advantageous. This constraint was overcome decades ago and

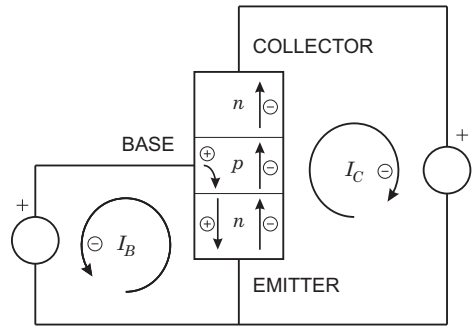

FIG. 6.1. Principle of a bipolar junction transistor.

commonly the signal is applied to the base. The different circuit topologies and their characteristics are described later.

The base and emitter form a diode, which is forward biased so that a base current I_B flows. The base current injects holes into the base–emitter junction. As in a simple diode, this gives rise to a corresponding electron current through the base–emitter junction.

If the potential applied to the collector is sufficiently positive so that the electrons passing from the emitter to the base are driven towards the collector, an external current I_C will flow in the collector circuit.

The ratio of collector to base current is equal to the ratio of electron to hole currents traversing the base–emitter junction. Assuming ideal diode behavior (as derived in Appendix E), the ratio of collector to base currents

$$\frac{I_C}{I_B} = \frac{I_{nBE}}{I_{pBE}} = \frac{D_n/N_A L_n}{D_p/N_D L_p} = \frac{N_D}{N_A}\frac{D_n L_p}{D_p L_n} \ . \tag{6.1}$$

If the ratio of doping concentrations in the emitter and base regions N_D/N_A is sufficiently large, the collector current will be greater than the base current. Thus, the device exhibits current gain. The gain can be increased further by narrowing the base width, as shown in Appendix G. Furthermore, we expect the collector current to saturate when the collector voltage becomes large enough to capture all of the minority carrier electrons injected into the base. With zero collector voltage the minority carriers in the base would simply recombine. Since the current inside the transistor includes both electrons and holes, the device is called a bipolar transistor. A quantitative description of the bipolar transistor is given in Appendix G.

Bipolar junction transistors allow much higher current gains than point-contact transistors, whose current gains were of order unity. However, since their output resistance was higher than the input resistance, point-contact transistors could still provide voltage gain and initially provided useful gain at higher

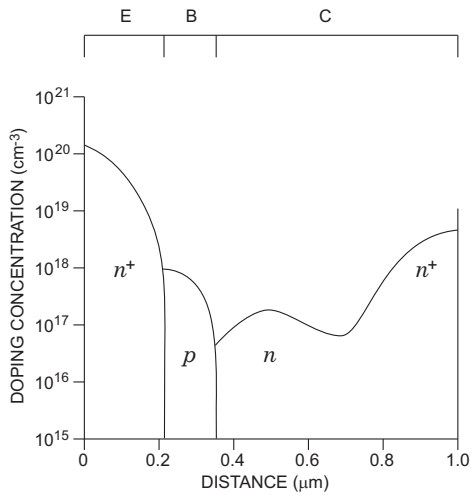

FIG. 6.2. Dimensions and doping profiles of the emitter (E), base (B), and collector (C) in a GHz bandwidth bipolar transistor.

frequencies than junction devices. Despite being obsolete for decades, the point-contact transistor still lingers on in some discussions of bipolar transistor properties. Unfortunately, this sometimes leads to erroneous conclusions.

Typical dimensions and doping levels of a modern high-frequency transistor (5 – 10 GHz bandwidth) are shown in Figure 6.2. Although this figure shows the transistor arranged horizontally, in the silicon wafer the structure extends from the surface into the silicon bulk, so the device is vertical. Furthermore, a practical device must be configured to avoid leakage paths at the device periphery that would bypass the device and also include isolation structures between adjacent devices. This is illustrated in Figure 6.3.

The base width, typically $0.2\,\mu$m or less in modern high-speed transistors, is determined by the difference in diffusion depths of the emitter and base regions, so it is not directly dependent on minimum feature size. Since the base width is much less than the recombination length L_n, it determines the charge profile, so in eqn 6.1 the recombination length is replaced by the base width. This further increases the current gain. Since the transit time through the base limits high-frequency performance, thin base regions are also necessary for high-speed devices. The thin base geometry and high doping levels make the base–emitter junction sensitive to large reverse voltages. Typically, base–emitter breakdown voltages for high-frequency transistors are but a few volts. A lightly doped layer between the base and collector (labelled "n-EPI" in Figure 6.3) increases the breakdown voltage of the collector.

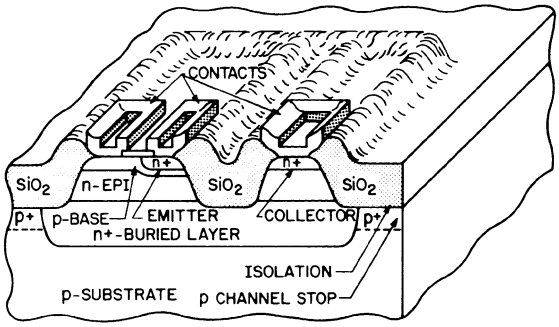

FIG. 6.3. Bipolar transistors are usually fabricated as vertical devices, with the emitter at the surface, the base one layer down, and the collector a relatively large tub in the bulk. The "n^+-buried layer" forms the collector with an intermediate lightly doped "n-epi" layer between the base and collector to increase the sustainable collector voltage. The "n^+-buried layer" extends laterally to the right to provide a collector contact at the surface. The p^+ channel stop at the right surrounds the transistor and isolates it from adjacent devices (From Sze 1981. ©John Wiley & Sons, reproduced with permission.)

Complementary devices can be formed by using either an *npn* or *pnp* sequence, as illustrated in Figure 6.4. The principle of operation is the same, except that the polarities of the applied voltages are reversed.

npn: positive collector–emitter and base–emitter voltages.

pnp: negative collector–emitter and base–emitter voltages.

The result of this simple analysis implies that for a given device the current gain should be independent of current. In reality this is not the case. Figure 6.5 shows the measured DC gain of a "general purpose" high-frequency transistor. The current gain peaks at about 8 mA and is roughly constant over only about an octave of current. At low currents the DC gain decreases due to recombination

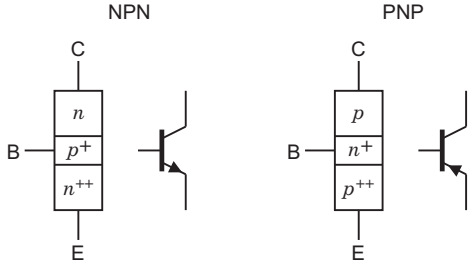

FIG. 6.4. Either *npn* (left) or *pnp* (right) transistors can be formed by juxtaposing the sequence of doping. The corresponding the circuit symbols are also shown.

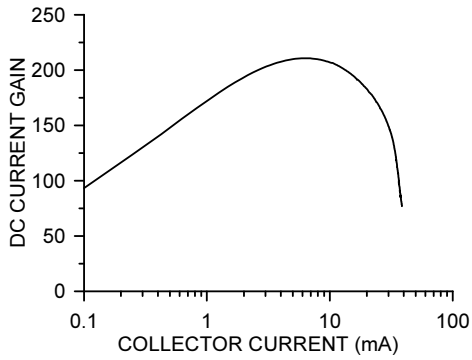

FIG. 6.5. Direct current gain *vs.* collector current of a "general purpose" high frequency transistor.

in the base–emitter depletion region. We'll return to this phenomenon in the next chapter when discussing radiation effects. At high currents the current gain drops because of resistive voltage losses, shifting of the high field region at the collector (increased base width), and loss of injection efficiency as the carrier density approaches the doping concentration, depending on the specific design.

For low-power applications the behavior at low currents is important. The "ideal" DC gain depends only on device and material constants, whereas the recombination depends on the local density of injected electrons and holes relative to the concentration of recombination centers. Thus, the relative degradation of DC gain due to recombination depends on the current density. Within the same fabrication process and at a given operating current, a large transistor will ex-

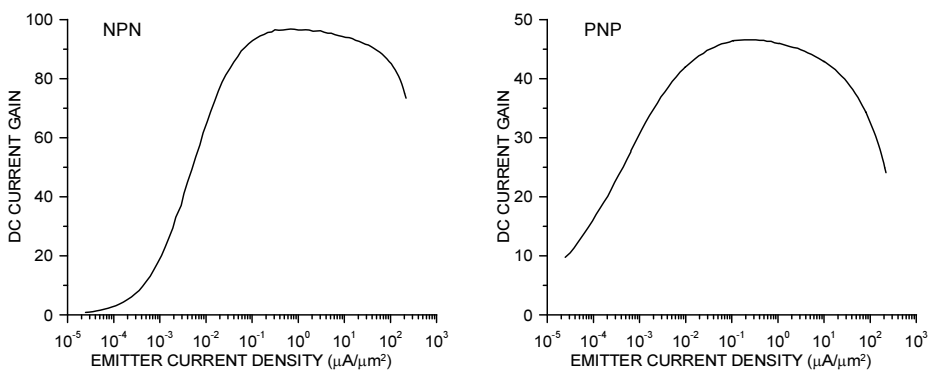

FIG. 6.6. DC gain *vs.* emitter current density of *npn* and *pnp* devices fabricated in a modern bipolar transistor integrated circuit process. The emitter area is 45 $(\mu\text{m})^2$, so to obtain the total emitter current the current density must be multiplied by 45.

hibit more recombination than a small transistor. Stated differently, for a given current, the large transistor will offer more recombination centers for the same number of carriers. As shown in Figure 6.6 modern devices exhibit DC gain that is quite uniform over orders of magnitude of emitter current and extending down to currents at the μA level.

The frequency response of the current gain is similar to that of the simple amplifier discussed in Chapter 2. At low frequencies the current gain is constant ($= \beta_{DC}$) and then drops off linearly with frequency. In the high frequency regime

$$\beta = -\mathbf{i}\frac{f_T}{f} \ , \tag{6.2}$$

where f_T is the frequency where the extrapolated current gain equals one. This is called the transit frequency. The gain–bandwidth product is constant, so the cutoff frequency, where the current gain rolls off by $1/\sqrt{2}$,

$$f_\beta = \frac{f_T}{\beta_{DC}} \ . \tag{6.3}$$

In a transistor with $f_T = 5\,\text{GHz}$ and a low-frequency current gain of 100 the current gain remains roughly constant up to 50 MHz and then falls off inversely proportional to frequency with a 90° phase shift.

6.1.1 *Bipolar transistors in amplifiers*

To form amplifiers transistors can be connected in three basic configurations, common-emitter, common-base, and common-collector, which will be described below. The three configurations have different properties, which complement each other. All three are frequently combined to obtain the desired characteristics of the overall amplifier.

The differential behavior, as for a small signal superimposed on the bias voltages, is the same for both *npn* and *pnp* devices, so the basic amplifier equations apply to both types. The availability of complementary transistors offers great flexibility in circuit design and also provides greater gain and bandwidth than pure *npn* or *pnp* designs. Figure 6.4 shows *npn* and *pnp* transistors together with their respective circuit symbols.

6.1.1.1 *Common-emitter amplifier* In a common-emitter amplifier the emitter is common to both the input and output. Figure 6.7 shows a common-emitter amplifier and its equivalent circuit.

The input signal is applied to the base, the output taken from the collector. The change in collector current ΔI_C in response to a base current ΔI_B is $\Delta I_C = \beta \Delta I_B$, so the change in output voltage $\Delta V_C = \Delta I_C R_L = \beta \Delta I_B R_L$, where $R_L = RR_C/R + R_C$ is the parallel combination of R_C and the external load R. Note that we must distinguish between the DC gain $\beta_{DC} = I_C/I_B$ and the differential current gain $\beta = dI_C/dI_B$, which applies to small signal variations superimposed on the DC bias.

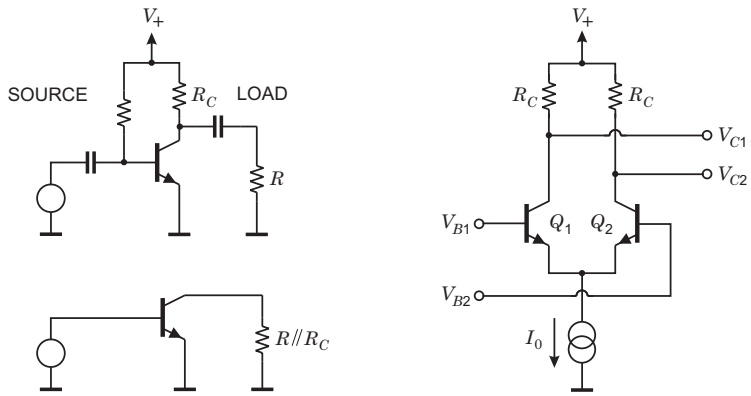

FIG. 6.7. Common-emitter amplifier with its equivalent circuit (left). Note that the effective load resistance is the parallel combination of R and R_C. The second panel shows a variant of the common-emitter amplifier configured as a differential pair.

Although the bipolar transistor is a current driven device, it is often convenient to consider its response to input voltage. Consider a transistor in the common-emitter configuration. The voltage gain

$$A_v = \frac{dV_{out}}{dV_{in}} = \frac{dI_C}{dV_{BE}} R_L = g_m R_L \ . \tag{6.4}$$

Since the dependence of base current on base–emitter voltage is given by the diode equation

$$I_B = I_R(e^{eV_{BE}/kT} - 1) \approx I_R e^{eV_{BE}/kT} \ , \tag{6.5}$$

the resulting collector current is

$$I_C = \beta_{DC} I_B = \beta_{DC} I_R e^{eV_{BE}/kT} \tag{6.6}$$

and the transconductance, *i.e.* the change in collector current *vs.* base–emitter voltage

$$g_m \equiv \frac{dI_C}{dV_{BE}} = \beta_{DC} I_R \frac{e}{kT} e^{eV_{BE}/kT} = \frac{e}{kT} I_C \ . \tag{6.7}$$

At a given temperature the transconductance depends only on collector current, so for any bipolar transistor – regardless of its internal design, whether "antique" or futuristic and regardless of the material, whether Si, Ge, SiGe or any other heterojunction – setting the collector current determines the transconductance. Since at room temperature $kT/e = 26\,\mathrm{mV}$,

$$g_m = \frac{I_C}{0.026\ [\mathrm{mV}]} \approx 40\left[\mathrm{V}^{-1}\right] I_C \ . \tag{6.8}$$

We can also interpret this result in the following manner. Because of the emitter diode's current–voltage characteristic, for small current excursions it behaves as a resistance

$$r_E = \frac{dV_{BE}}{dI_E} = \frac{kT}{eI_E} \ . \tag{6.9}$$

Since $I_E \approx I_C$, the transconductance $g_m \approx 1/r_E$.

Thus is not quite true, as the doped channel connecting the physical emitter to the external connection has some resistance, so the emitter resistance is the sum of the dynamic resistance r_E and the parasitic connection resistance. However, at low currents the parasitic resistance is usually negligible, so in small signal amplifiers the voltage gain is of a bipolar transistor is well-controlled by the emitter current and load resistance.

A variant of the common-emitter amplifier is the differential pair (sometimes called a "long tailed pair"), also shown in Figure 6.7. Both emitters are connected together to a current source I_0, so the total current is fixed. The quiescent current in each transistor is $I_0/2$, so the transconductance

$$g_{m1} = g_{m2} = \frac{e}{kT} \cdot \frac{I_0}{2} \ . \tag{6.10}$$

A small change in the input voltage to the first transistor $V_{B1} + \Delta V$ changes its collector current by

$$\Delta I_{C1} = \Delta V g_{m1} = \frac{\Delta V}{2} \cdot \frac{e}{kT} \cdot \frac{I_0}{2} \ . \tag{6.11}$$

Only half of the signal voltage is applied to each transistor, as the input divides equally between the two base–emitter junctions. Since the total current is fixed, the current of the second transistor Q_2 must decrease by the same amount, $\Delta I_{C2} = -\Delta I_{C1}$. Viewed in terms of emitter voltage, increasing V_{B1} pulls up the voltage at the common-emitter connection, so the base–emitter voltage of the second transistor is reduced and its current decreases. The collector voltages

$$\begin{aligned} V_{C1} &= V_+ - \left[\frac{I_0}{2} + \Delta V \cdot \frac{e}{kT} \cdot \frac{I_0}{4}\right] \cdot R_C \\ V_{C2} &= V_+ - \left[\frac{I_0}{2} - \Delta V \cdot \frac{e}{kT} \cdot \frac{I_0}{4}\right] \cdot R_C \ , \end{aligned} \tag{6.12}$$

so the change in output voltage

$$V_{C1} - V_{C2} = \Delta V \cdot \frac{e}{kT} \cdot \frac{I_0}{2} \ . \tag{6.13}$$

The differential voltage gain is half of a single transistor operating at the collector current I_0. However, the differential configuration has an important advantage, as the difference output $V_{C1} - V_{C2}$ depends only on the difference in input voltages $V_{B1} - V_{B2}$. The output voltage difference is independent of the

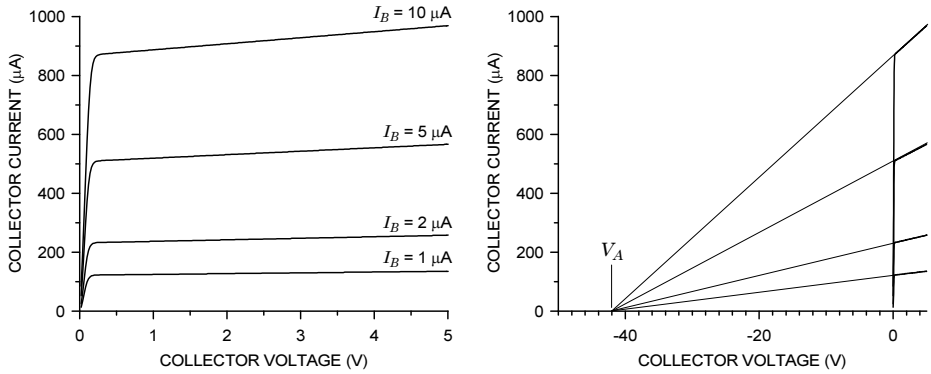

FIG. 6.8. Output characteristics of a bipolar transistor. Beyond the "knee" the extrapolated curves intersect at the same voltage, the "Early voltage" V_A.

supply voltage V_+, and also independent of of any "common mode" voltage at the input. Suppressing common mode components on the voltage supply line is important in reducing cross-talk between different channels in a large system, as the supply line impedance is never zero. As the differential pair operates at constant current, it does not impress voltage changes on the supply line, further reducing potential cross-talk. We'll return to this circuit topology in Section 6.4.

Since the maximum current draw of either transistor is limited to I_0, the circuit also acts as an amplitude limiter. The collector currents

$$I_{C1} = \frac{I_0}{1 + \exp\left(\dfrac{V_{B1} - V_{B2}}{kT/e}\right)} \quad \text{and} \quad I_{C2} = \frac{I_0}{1 + \exp\left(\dfrac{V_{B2} - V_{B1}}{kT/e}\right)}\,, \tag{6.14}$$

which yields a linear input voltage range of $\pm kT/e$. The maximum peak-to-peak differential output voltage swing is $I_0 R_C$, which corresponds to an input swing of about $\pm 4kT/e \approx \pm 100\,\text{mV}$, independent of the current I_0. The maximum input signal can be increased by inserting resistors R_E in the emitter connections of both transistors, which reduces the transconductance to $(r_E + R_E)^{-1}$.

A constraint on the obtainable voltage gain of an amplifier is imposed by the output characteristics of the transistor. These are shown schematically in Figure 6.8 for a transistor in a modern integrated circuit. At low collector voltages the field in the collector–base region is not sufficient to transport all injected carriers to the collector without recombination. At sufficiently high voltage the collector captures practically all available carriers, but the output current still increases gradually with voltage as the increased potential reduces the effective base width.

An interesting feature is that the extrapolated slopes in the saturation region intersect at the same voltage V_A for $I_C = 0$, the "Early voltage". This parameter sets the maximum voltage gain of the amplifier.

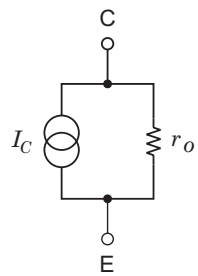

FIG. 6.9. Equivalent output circuit of a bipolar transistor. The slope of the output curve is equivalent to a resistance shunting the current source formed by the ideal transistor.

The finite slope of the output curves is equivalent to a current generator representing the ideal transistor shunted by a resistance, as shown in Figure 6.9. The shunt resistance

$$r_o = K\frac{V_A}{I_C} \ , \tag{6.15}$$

where V_A is the Early voltage and K is a device-specific constant of order 1, so usually it's neglected. Figure 6.8 shows an Early voltage of 42 V, so for a current of 1 mA the output resistance r_o is 42 kΩ.

In an amplifier the total load resistance is the parallel combination of the external load resistance and the output resistance r_o of the transistor. In the limit where the external load resistance is infinite, the load resistance is the output resistance of the transistor. Then the voltage gain is the maximum obtainable,

$$A_{v,\max} = \frac{dI_C}{dV_{BE}} r_o = g_m r_o \approx \frac{I_C}{kT/e}\frac{V_A}{I_C} = \frac{V_A}{kT/e} \ , \tag{6.16}$$

which at room temperature is about $40V_A$, so transistors with large Early voltages allow large voltage gains. To first order the maximum obtainable voltage gain is independent of current.

The effective load resistance is the parallel combination of the transistor's output resistance $r_o = KV_A/I_C$ and the external resistances R_C and R shown in Figure 6.7. In practice, the latter two resistances are usually selected to be much smaller than the transistor's output resistance, so the external resistances determine the total load resistance.

The bipolar transistor's input resistance

$$r_i = \frac{dV_{BE}}{dI_B} = \beta\ \frac{dV_{BE}}{dI_C} \ . \tag{6.17}$$

From the diode equation (eqn 6.5)

$$\frac{dV_{BE}}{dI_E} = \frac{kT}{eI_E} \ . \tag{6.18}$$

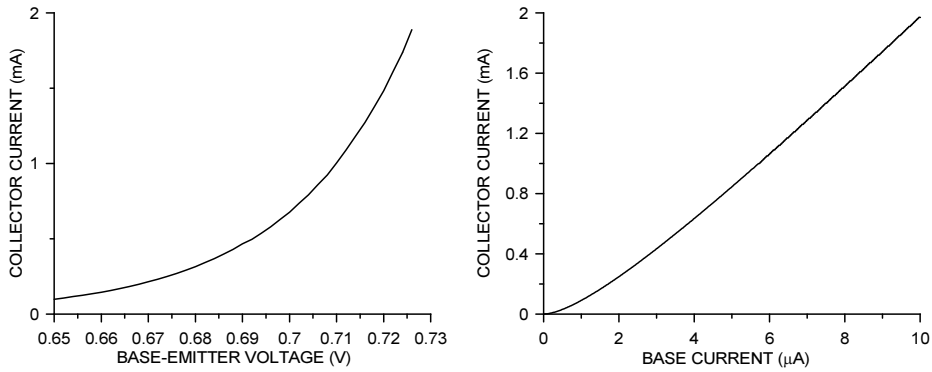

FIG. 6.10. Collector current *vs.* base–emitter voltage (left) and collector current *vs.* base current (right) in a typical bipolar transistor.

The emitter current $I_E = I_B + I_C = (I_C/\beta_{DC}) + I_C$. Since in a modern transistor $\beta_{DC} \gg 1$, $I_E \approx I_C$, so the input resistance

$$r_i \approx \beta \frac{dV_{BE}}{dI_E} = \beta \frac{kT}{eI_E} \approx \beta \frac{kT}{eI_C} \ . \tag{6.19}$$

The input resistance is proportional to the current gain and inversely proportional to the collector current. For $\beta = 100$ and $I_C = 1\,\text{mA}$, $r_i = 2600\,\Omega$.

We can also interpret this result in the following manner. As noted above, for small current excursions the emitter diode behaves as a resistance $r_E = kT/e$. Since $I_B \approx I_E/\beta$, the input resistance $r_i \approx \beta r_E$.

Although the bipolar transistor is often treated as a voltage-driven device, the exponential dependence of base current on input voltage means that the transconductance is very nonlinear. Figure 6.10 shows the collector current *vs.* base–emitter voltage and the collector current *vs.* base current in a typical bipolar transistor. With current drive the linearity is much better. In audio amplifiers, for example, nonlinearity causes distortion. Distortion can be limited by restricting the voltage swing, which to some degree is feasible because of the high transconductance. Distortion may also be reduced by negative feedback, but beginning with lower distortion has advantages in systems processing transients with large dynamic range.

6.1.1.2 *Common-base amplifier* In a common-base amplifier (Figure 6.11) the signal is applied to the emitter and the output taken from the collector. This configuration is used where a low input resistance is required.

$$r_i = \frac{dV_{EB}}{dI_E} \approx \frac{dV_{EB}}{dI_C} = r_E = \frac{1}{g_m} = \frac{kT}{e}\frac{1}{I_C} \tag{6.20}$$

Since at room temperature $kT/e = 26\,\text{mV}$, the input resistance $r_i = 0.026/I_C$, so $r_i = 26\,\Omega$ at $I_C = 1\,\text{mA}$. Put differently, the input resistance of a common-base

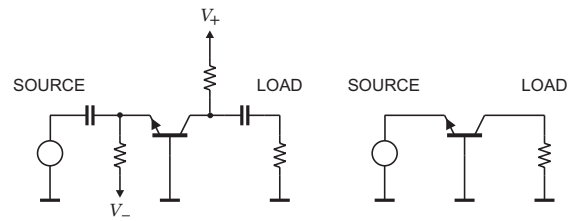

FIG. 6.11. Circuit diagram (left) and equivalent circuit (right) of a common-base amplifier.

stage equals the dynamic emitter resistance r_E and the input resistance is about $1/\beta$ times smaller than in the common-emitter configuration.

The presence of the source resistance in series with the emitter introduces negative feedback, as the voltage drop due to the emitter current is of opposite sign with respect to the input signal. This linearizes the voltage gain and also increases the output resistance, as expected for series feedback (see Appendix D).

6.1.1.3 *Common-collector amplifier* In the common-collector configuration (Figure 6.12) the signal is applied to the base and the output taken from the emitter. This circuit is commonly called an "emitter follower".

The load resistance R_L introduces local negative feedback. Since the emitter voltage follows the input voltage, the net base–emitter voltage applied to the transistor is reduced,

$$V_i = V_{BE} + I_E R_L \approx V_{BE} + \beta I_B R_L \ . \tag{6.21}$$

Since V_{BE} varies only logarithmically with I_B, it can be considered to be constant ($\approx 0.6\,\mathrm{V}$ for small signal transistors), so

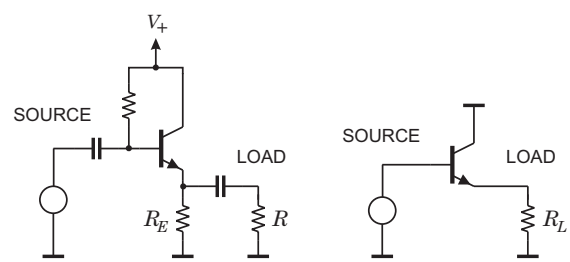

FIG. 6.12. Circuit diagram (left) and equivalent circuit (right) of a common-collector amplifier ("emitter follower"). As in Figure 6.7 the effective load resistance R_L is the parallel combination of R_E and R.

$$\frac{dV_i}{dI_i} = \frac{dV_i}{dI_B} \approx \beta R_L \ . \tag{6.22}$$

Thus, the input resistance depends on the load, $r_i \approx \beta R_L$.

Since $dV_{BE}/dI_B \approx$ const, the emitter voltage follows the input voltage, so the voltage gain cannot exceed one. The output resistance of the emitter follower

$$r_o = -\frac{dV_{out}}{dI_{out}} = -\frac{d(V_{in} - V_{BE})}{dI_E} \approx \frac{dV_{BE}}{dI_E} \approx \frac{1}{g_m} = r_E \ , \tag{6.23}$$

as $dV_{in}/dI_E = 0$, since the applied input voltage is independent of emitter current. At 1 mA current $r_o = 26\,\Omega$. Although the stage only has unity voltage gain, it does have current gain, so emitter followers are often used as output drivers. Furthermore, since in a common-emitter stage the input resistance $r_{i,CE} = \beta r_E$, whereas in the emitter follower it is $r_{i,CC} = \beta(r_E + R_L)$, the emitter follower allows a higher gain to be obtained from a preceding common-emitter or common-base stage. Section 6.4 shows how these stages can be combined to form a gain-block with high gain and bandwidth.

Although the emitter follower has unity voltage gain, certain common load conditions can lead to self-oscillation. Recall that in the high-frequency regime the current gain

$$\beta = -\mathbf{i}\frac{f_T}{f} \ .$$

If the emitter follower drives a purely capacitive load

$$X_C = -\mathbf{i}\frac{1}{2\pi f C} \ ,$$

the input impedance

$$Z_i = \beta X_C = \left(-\mathbf{i}\frac{f_T}{f}\right)\left(-\mathbf{i}\frac{1}{2\pi f C}\right) = -\frac{f_T}{2\pi f^2 C} \tag{6.24}$$

is real, but with a negative sign. A negative resistance corresponds to positive feedback and leads to self oscillation unless a sufficiently large dissipative element, *i.e.* a resistance, is connected in series. In practice, the external load is commonly shunted by an emitter resistor R_E, as shown in the left panel of Figure 6.12, so the load presented to the transistor is capacitive at frequencies $\gg (2\pi R_E C)^{-1}$. Since the negative component of the input resistance decreases linearly with frequency, reducing the output time constant $R_E C$ can stabilize the stage for a given series resistance at the input. The stage can also be stabilized by connecting the capacitive load through a series resistor, which is required by many high-speed operational amplifiers.

6.2 Field effect transistors

Field effect transistors (FETs) are the key elements in the high-density integrated circuitry that brought about the digital revolution. The GHz clocked CPUs, large

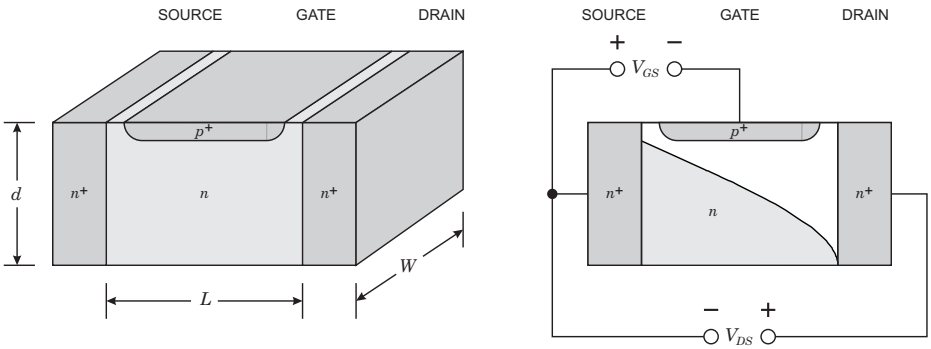

FIG. 6.13. Schematic illustration of a junction field effect transistor. A practical device would be embedded in a p-substrate.

memory ICs, and logic arrays would not be practical without FET technology. FETs are implemented as junction field effect transistors (JFETs) or metal-oxide-semiconductor field effect transistors (MOSFETs). MOSFETs provide by far the highest circuit density. The technologies are very different, but both types of FETs utilize a conductive channel whose resistance is controlled by an applied potential. However, as will be shown, the FET is more than just a voltage-controlled resistor. Additional effects are utilized to enhance its amplification. Key operating principles are common to both JFETs and MOSFETs, so the next section is also important for those who are interested only in MOSFETs.

Both JFETs and MOSFETs are conductivity modulated devices, utilizing only one type of charge carrier. Thus they are called unipolar devices, unlike bipolar transistors, for which both electrons and holes are crucial.

6.2.1 *Junction field effect transistors*

Historically, JFETs were the first practical FETs. They have been largely displaced by MOSFETs, but at low frequencies they provide superior noise performance. In JFETs a conducting channel is formed of n- or p-type semiconductor (GaAs, Ge or Si). Figure 6.13 shows an n-channel device. The n^+-drain, n^+-source, and intermediate n-type material form a conductive channel. The p^+-gate electrode forms a diode. In operation a positive voltage is connected to the drain and a negative voltage is connected to the gate, both relative to the source. When the gate diode is reverse-biased a depletion region forms and changes the profile of the conducting channel, as shown in the second panel of Figure 6.13. The basic structure is embedded in a p-substrate or p-well to confine the conducting channel. Next we consider what determines the longitudinal profile of the conducting channel.

First assume that the drain voltage is zero. Increasing the reverse gate potential will increase the depletion width, *i.e.* reduce the cross-section of the conducting channel, until the channel is completely depleted. The gate voltage where this

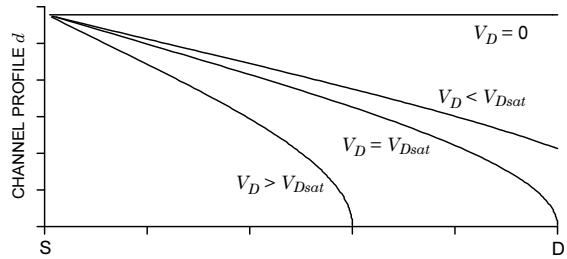

FIG. 6.14. Cross-section d of the conductive channel *vs.* distance from the source (S) towards the drain (D). The gate is at the upper edge of the plot and the conductive channel is the area below the curves. When the drain voltage equals the saturation voltage V_{Dsat}, the channel is fully depleted at the drain. With increasing drain voltage the pinch-off moves closer to the source ($V_D > V_{Dsat}$).

obtains is the "pinch-off voltage" V_P. Now set both the gate and drain voltages to zero. The channel will be partially depleted due to the "built-in" junction voltage. Next, apply a positive drain voltage. Since the drain is at a higher potential than the source, the effective depletion voltage increases in proximity to the drain, so the width of the depletion region will increase as it approaches the drain. If the sum of the gate and drain voltages is sufficient to fully deplete the channel, the device is said to be "pinched off". The drain voltage that achieves pinch off is called the saturation voltage V_{Dsat}, for reasons that will become apparent later. Increasing the drain voltage beyond this point moves the pinch-off point towards to the source. Figure 6.14 illustrates how the profile of the conductive channel changes with increasing drain voltage. The gate-to-channel potential along the channel length is determined locally by the voltage drop due to the current flowing through the resistive channel. As the channel cross-section decreases, the incremental voltage drop increases (as the current is constant), so the longitudinal drift field that determines the carrier velocity increases. The profiles shown in Figure 6.14 also include the effect of field-dependent mobility.

Pinching off the channel does not interrupt current flow. All thermally excited carriers have been removed from the depleted region, but there is a continuous potential drop from the source to the drain, so carriers originating from the resistive channel follow the potential drop to the drain. The gate voltage modulates the conductive portion of the channel, which "launches" carriers into the depleted region.

As already discussed in Chapter 2 the carrier mobility decreases at high fields. Thus, an increase in the electric field is not translated completely into an increase in velocity and at sufficiently high fields the velocity becomes independent of field. Since the velocity saturates at high fields, the current $I = N_C ev$ also saturates, since the number of carriers N_C remains constant. Thus, at high fields silicon acts as an incremental insulator ($dI/dV = 0$).

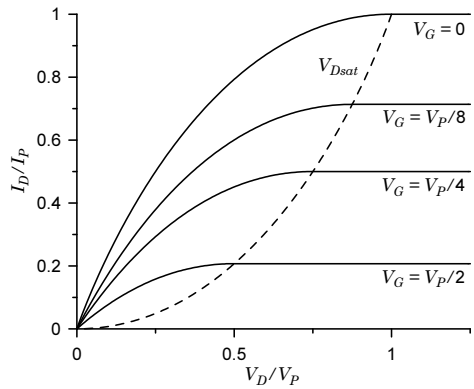

FIG. 6.15. Operating regimes of a field effect transistor.

As the drain voltage is increased beyond pinch-off, the additional voltage decreases the length of the resistive channel, but also increases the potential drop in the drain depletion region. As a result, the current increases only gradually with drain voltage. This is a result of the changing potential distribution along the channel in combination with the "incremental insulator" properties of the region between the conductive channel and the drain.

Although textbook equations for FET characteristics tend to use constant mobility, in reality velocity saturation is a key phenomenon in the mode where FETs are typically operated. At fields above about 10^5 V/cm practically all of the energy imparted by an increased field goes into phonon emission, but the mobility already decreases substantially at a field of 10^4 V/cm, corresponding to 1 V across 1 μm. Thus, the behavior of micron-scale devices is strongly affected by nonconstant mobility, leading to smaller transconductance than predicted by a constant mobility model. Sze (1981) discusses this in more detail. Nevertheless, the following discussion uses the constant mobility model, as it illustrates the key aspects.

6.2.1.1 *Current–voltage characteristics* In amplifiers the interesting operating regime is beyond pinch-off, *i.e.* drain voltages $> V_{Dsat}$. JFET output characteristics are illustrated in Figure 6.15, showing the linear region, where the channel behaves resistive, and the saturation region, where the device acts as a current source, *i.e.* exhibits a high output resistance. At very high drain voltages the device exhibits breakdown, whose onset is indicated by a rapid upturn in drain current.

1. "Linear Region"

At low drain and gate voltages, the resistive channel extends from the source to the drain. The current–voltage characteristic

$$I_D = I_P \left\{ \frac{3V_D}{V_P} - \frac{2}{V_P^{3/2}} \left[(V_D + V_{GS} + V_{bi})^{3/2} - (V_{GS} + V_{bi})^{3/2} \right] \right\} , \qquad (6.25)$$

where the pinch-off current

$$I_P = \frac{1}{6\varepsilon} \mu (eN)^2 d^3 \frac{W}{L} \qquad (6.26)$$

is determined by the carrier mobility μ, the doping level in the channel N, its depth d, and the channel's width W and length L. The pinch-off voltage

$$V_P = \frac{eNd^2}{2\varepsilon} . \qquad (6.27)$$

For a given gate-source voltage V_{GS} the drain current increases linearly for small drain voltages $V_D \ll V_{GS} + V_{bi}$. In this regime the FET can be used as a voltage controlled resistor, but with the caveat that the voltage swing due to any signal applied to the device must be small with respect to the gate voltage to avoid modulation of the resistance with signal level.

This constraint must also be observed when FETs are used as switches. When the channel is fully depleted, the switch is open with minimum drain–source capacitance, which is important for isolation at high frequencies. When the gate is biased below pinch off, the switch is closed, but it exhibits a finite resistance, which depends on the signal level. Minimum resistance obtains at $V_{GS} = -V_{bi}$, where the resistance still has a nonlinear component.

2. "Saturation Region"

When the gate and drain voltages are sufficiently high to pinch off the resistive channel, the drain current remains roughly constant with increasing drain voltage. This regime provides maximum voltage gain, so this is how FETs are commonly operated in amplifiers.

The drain saturation voltage V_{Dsat} increases as the gate voltage is changed from the static pinch off voltage V_P towards 0,

$$V_{Dsat} = V_P - V_{GS} - V_{bi} = \frac{eNd^2}{2\varepsilon} - V_{GS} - \frac{kT}{e} \log \left(\frac{NN_G}{n_i^2} \right) , \qquad (6.28)$$

where N_G is the doping level of the gate electrode. The corresponding drain saturation current

$$I_{Dsat} = I_P \left[1 - 3 \left(\frac{V_{GS} + V_{bi}}{V_P} \right) + 2 \left(\frac{V_{GS} + V_{bi}}{V_P} \right)^{3/2} \right] . \qquad (6.29)$$

These derivations are for a uniform doping distribution in the channel region (Shockley 1952, Sze 1981), which is not achieved in practical devices. However,

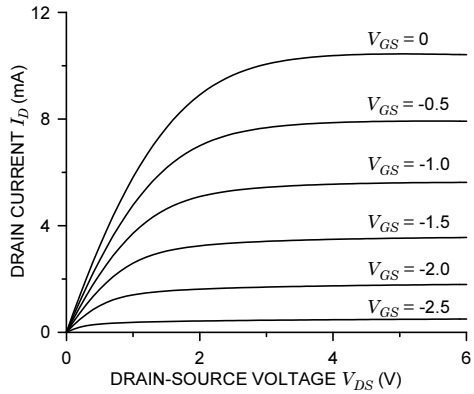

FIG. 6.16. Measured drain current *vs.* drain voltage for various gate voltages of a JFET.

for this and more realistic distributions the drain current in the saturation regime can be approximated closely as a parabolic function of gate voltage (Sevin 1965)

$$I_D = I_{DSS}\left(1 - \frac{V_{GS} + V_{bi}}{V_P}\right)^2 , \tag{6.30}$$

where I_{DSS} is the saturation drain current for $V_{GS} = 0$ (maximum current). Figure 6.16 shows the drain current *vs.* drain voltage for a range of gate voltages, measured on a commonly used JFET, a 2N4416. The "ohmic" region at low drain voltages is apparent. Saturation of the drain current at large drain voltages indicates a high output resistance.

From eqn 6.30 the transconductance

$$g_m = \left|\frac{dI_D}{dV_{GS}}\right| = \frac{2I_{DSS}}{V_P}\left(1 - \frac{V_{GS} + V_{bi}}{V_P}\right) , \tag{6.31}$$

which is maximum for $V_{GS} = 0$, *i.e.* maximum drain current. If $V_{bi} \ll V_P$,

$$g_m|_{V_G=0} \approx \frac{2I_{DSS}}{V_P} . \tag{6.32}$$

Combining eqns 6.30 and 6.31 shows that for a given device the transconductance depends primarily on drain current,

$$g_m = \frac{2\sqrt{I_{DSS}}}{V_P}\sqrt{I_D} . \tag{6.33}$$

The applied voltages only provide the boundary conditions to set up the required current. Thus, to maintain transconductance it is important to control

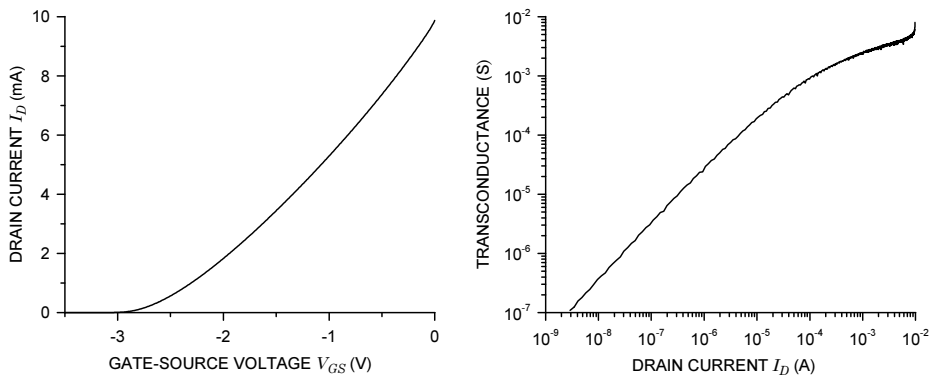

FIG. 6.17. Measured drain current *vs.* gate voltage (left) and transconductance *vs.* drain current of a JFET.

the current, rather than the gate voltage, which is more dependent on fabrication variations.

To see how device parameters affect the transconductance, we'll ignore the built-in voltage since it varies only weakly with doping (eqn 6.28). With this approximation

$$g_m|_{V_G=0} \approx \frac{2I_{DSS}}{V_P} \approx \frac{W}{L}\frac{\mu(Ne)^2 d^3}{3Ned^2} \propto \frac{W}{L}\mu N d \ . \tag{6.34}$$

Obviously, a high carrier mobility will increase the transconductance, since for a given carrier concentration this will increase the magnitude of the current.

1. The proportionality of transconductance to width W is trivial, since it is equivalent to merely connecting devices in parallel. Thus, the normalized transconductance g_m/W is useful in comparing technologies.
2. The transconductance increases with the number of carriers per unit length NWd and decreasing channel length L.
3. The transconductance increases with drain current

$$g_m = \left|\frac{dI_D}{dV_{GS}}\right| = \frac{2I_{DSS}}{V_P}\left(1 - \frac{V_{GS}+V_{bi}}{V_P}\right) = \frac{2\sqrt{I_{DSS}}}{V_P}\sqrt{I_D} \tag{6.35}$$

i.e. drain current is the primary operating parameter; the applied voltages are only the means to establish I_D.

All of these optimizations also increase the power dissipation. For low-power systems device optimization is more involved and will be discussed later.

Figure 6.17 shows measured JFET drain current *vs.* gate voltage and the transconductance *vs.* drain current, also for a 2N4416. At high drain currents the transconductance follows eqn 6.33 and increases with the square root of current. At low currents, however, the transconductance increases roughly linearly

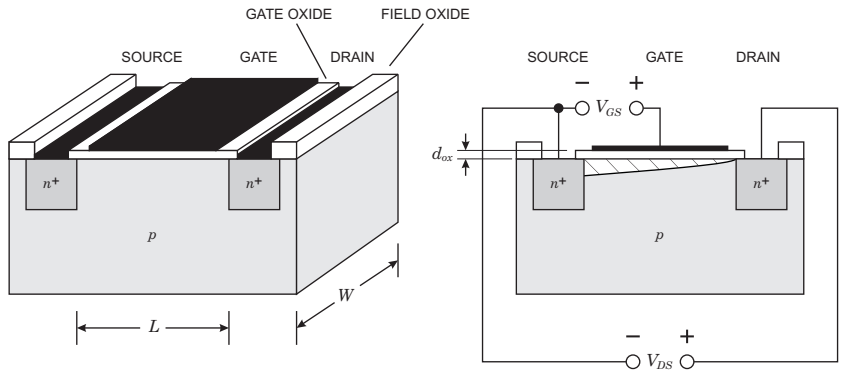

FIG. 6.18. Schematic structure of a MOSFET. A conducting channel (hatched region) is formed by applying a positive voltage to the capacitively coupled gate electrode. The field oxide establishes well-controlled surface conditions adjacent to the device.

with current. This behavior is not included in the simple abrupt junction model used to derive eqn 6.33. Although the physical mechanism is quite different, this behavior is similar to MOSFETs, as will be described in the next section.

The frequency response of an FET is limited by the charging time constant of the gate-to-channel capacitance C_{GC} together with the channel resistance, so the corner frequency

$$f_C = \frac{1}{2\pi} \cdot \frac{g_m}{C_{GC}} \approx \frac{1}{2\pi} \cdot \frac{g_m}{C_{GS}} \ . \tag{6.36}$$

The gate–channel capacitance is approximately equal to the gate–source capacitance. As this depends on the profile of the depletion region, it is correlated with the transconductance and in the saturation regime can be expressed as

$$C_{GC} = \frac{g_m^2}{2I_D} \frac{L^2}{\mu} \ , \tag{6.37}$$

which clearly shows the dependence on channel length L (Sevin 1965). In practice, a numerical simulation including velocity dependent mobility is necessary to predict actual device characteristics.

The FET still provides gain beyond f_C, but the input is no longer purely capacitive, as the channel resistance introduces a significant resistive component. The corner frequency f_C is typically in the GHz range, *i.e.* much higher than the typical frequencies of interest in semiconductor detector systems.

6.2.2 *Metal-oxide-semiconductor field effect transistors*

Unlike a JFET, where a conducting channel is formed by doping and its geometry is modulated by the applied voltages, the MOSFET changes the carrier

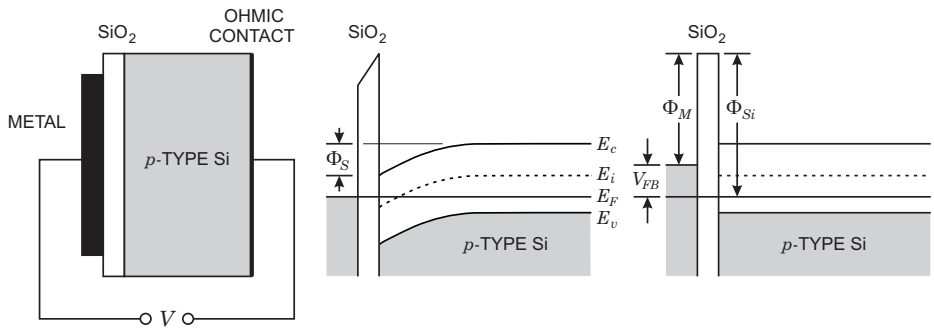

FIG. 6.19. A metal-oxide-semiconductor (MOS) capacitor is formed by a metallic electrode coupled to a semiconductor substrate through an insulator (left). The "dangling bonds" leave a net positive charge at the metal–Si interface, which bends the bands downward (middle). The ideal potential distribution is established by applying a voltage to straighten the bands, the "flat band voltage" V_{FB} (right).

concentration in the channel. A schematic drawing of a MOSFET is shown in Figure 6.18. The source and drain are n^+ regions in a p-substrate. The gate is capacitively coupled to the channel region through an insulating layer, typically SiO_2. Applying a positive voltage to the gate increases the electron concentration at the silicon surface beneath the gate. As in a JFET the combination of gate and drain voltages control the conductivity of the channel.

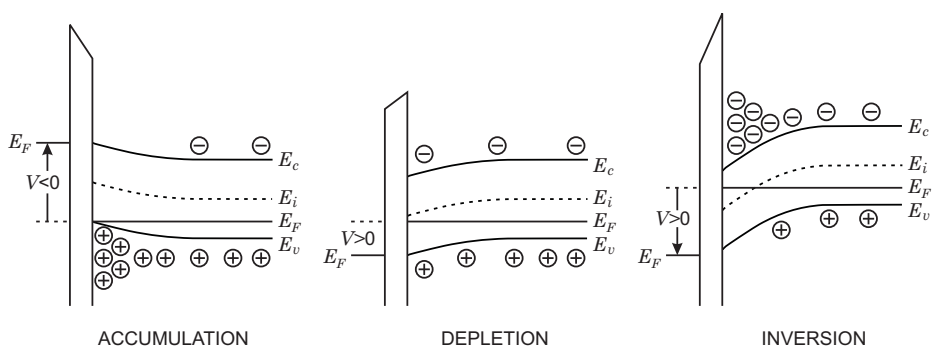

FIG. 6.20. The concentration of mobile carriers in the p substrate changes as the voltage applied to the metal electrode is varied. A negative potential bends the bands upward, so holes accumulate at the surface (left). A positive potential bends the bands downward, so the concentration of holes decreases (depletion, middle). Making the potential in more positive brings the Fermi level E_F above the intrinsic level E_i and electrons accumulate at the surface (inversion, right).

Formation of the conducting channel can be understood by analyzing a simple metal-oxide-semiconductor (MOS) capacitor, as illustrated in Figure 6.19 together with the potential levels and band structure. In equilibrium the chemical potential (the Fermi level) is constant throughout the system. This sets the levels of the metal and semiconductor relative to one another. The energy required to remove an electron from either the metal or the semiconductor is the work function, Φ_M and Φ_{Si}.

In its natural state, however, the band structure is not flat as just shown. The discontinuity in the crystal structure and charge trapped at the surface change the potential at the surface, so the bands bend as shown in middle panel of Figure 6.19.

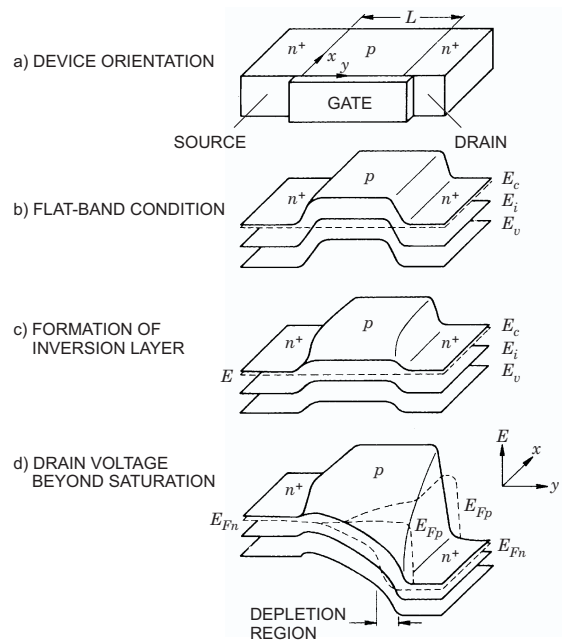

FIG. 6.21. Potential distribution along the channel and into the bulk of an n-channel MOSFET. The device orientation is shown at the top (a). In the second figure (b) the source and drain potentials are equal and the gate is adjusted to make the Fermi level constant throughout the device, so the bands are flat. Under these conditions there is no conducting channel between the source and drain. In the third figure (c) the gate voltage is made positive to invert the surface and form a conductive channel. In this mode the device is resistive. In the fourth figure (d) the drain voltage is made positive and adjusted beyond saturation. The Fermi levels for holes E_{Fp} and electrons E_{Fn} are also shown. Note the variation of E_{Fp} with depth (x coordinate). (From Sze 1981. ©John Wiley & Sons, reproduced with permission.)

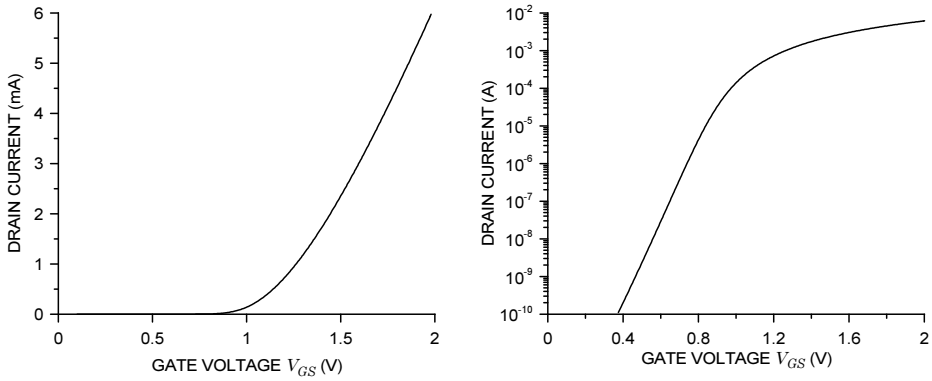

FIG. 6.22. Measured drain current *vs.* gate voltage plotted on linear (left) and logarithmic scales (right). The device is an n-channel MOSFET with a channel length of 0.8 μm and 100 μm width. The gate oxide is 20 nm thick. The threshold voltage V_T of this device is about 0.9 V.

As shown the surface potential Φ_S is positive, so the concentration of holes at the surface is reduced (depletion). Conversely, when $\Phi_S < 0$ the bands bend upwards, increasing the hole concentration at the surface (accumulation). When the surface potential is sufficiently positive the intrinsic level dips below the Fermi level, leading to an accumulation of electrons at the surface (inversion). This is illustrated in Figure 6.20 (band bending). In the absence of any special surface preparation the surface of silicon is n-type, *i.e.* p-type silicon inverts at the surface. For a comprehensive discussion of MOS physics see Nicollian and Brews (1982).

An n-channel MOSFET utilizes an n-channel in a p-substrate, so application of a positive potential to the gate forms the inversion layer needed for the channel. Figure 6.21 shows a three-dimensional representation of the potentials in the device for three characteristic operating conditions. As in the JFET, the combination of current flow in the channel and the applied potentials forms a depletion region that is greatest near the drain. At a sufficiently large drain potential the channel "pinches off".

Figure 6.22 shows the drain current *vs.* gate voltage of an n-channel MOSFET. Significant current flows when the gate voltage exceeds the "threshold voltage" V_T. This is a key parameter when the MOSFET is used as a switch. In this device the threshold voltage is about 0.9 V. However, when viewed on a logarithmic scale, it becomes apparent that the current doesn't cut off completely below threshold. At low gate voltages the charge density in the channel increases exponentially with surface potential, so the drain current increases exponentially with gate voltage. This regime is called "weak inversion". When the

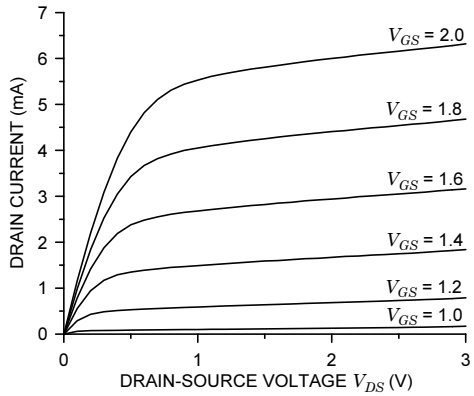

FIG. 6.23. Measured MOSFET output curves for gate voltages between 1 and 2 volts. The n-channel MOSFET has a channel length of 0.8 μm and 100 μm width.

gate is biased well above threshold the channel is in "strong inversion" and the current change is more gradual.

The output curves of a MOSFET resemble those of a JFET. Figure 6.23 shows the measured output curves for an n-channel MOSFET. At low drain voltages the drain current increases approximately linearly with drain voltage. In this regime – the "linear region" – the MOSFET acts like a resistor. At higher drain voltages – beyond the "saturation voltage" – the drain current tends to saturate, although it still increases gradually with drain voltage and exhibits a finite output resistance. The drain voltage required to attain saturation increases with operating current. Textbooks frequently give the saturation voltage as $V_{Dsat} = V_G - V_T$, where V_T is the threshold voltage, but this only holds for very small currents and in practice grossly underestimates the saturation voltage.

The saturation regime is most useful for amplifiers, as it maximizes both the transconductance and the output resistance. In saturation and strong inversion

$$I_D = \frac{W}{L}\frac{\mu C_{ox}}{2}(V_G - V_T)^2 \ , \tag{6.38}$$

where W is the width of the channel, L the length (measured from source to drain), μ the mobility of the carriers in the channel, and C_{ox} the gate capacitance per unit area ε_{ox}/d_{ox}. From this follows the transconductance

$$g_m = \frac{W}{L}C_{ox}\mu(V_G - V_T) = \frac{W}{L}\frac{\varepsilon_{ox}}{d_{ox}}\mu(V_G - V_T) = \sqrt{\frac{W}{L}\cdot\frac{\varepsilon_{ox}}{d_{ox}}\mu\cdot I_D} \ . \tag{6.39}$$

For a given device the transconductance depends primarily on the drain current I_D, as also shown for a JFET.

Figure 6.24 shows the transconductance *vs.* drain current for an NMOS transistor operated a sufficiently high drain voltage to ensure operation in the

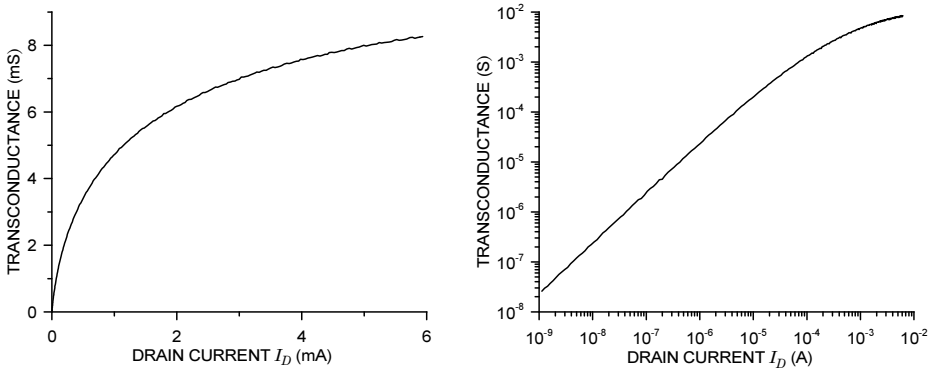

FIG. 6.24. Measured transconductance *vs.* drain current of an NMOS transistor with a channel length of 0.8 μm and width of 100 μm plotted on linear (left) and logarithmic scales (right).

saturation regime throughout the range of gate voltages used in the measurement. Plotted on a linear scale the parabolic dependence of transconductance on drain current is apparent. On a logarithmic scale we see that the device exhibits transconductance at currents well below threshold. In the subthreshold regime, where the channel is in weak inversion, the drain current increases exponentially with gate voltage and the transconductance increases linearly with current

$$g_m = \frac{I_D}{kT/e} \ , \tag{6.40}$$

as in a bipolar transistor.

In strong inversion, for a given width W and drain current I_D the transconductance is increased by decreasing the channel length L and the thickness of the gate oxide d_{ox}. In weak inversion, however, the transconductance is independent of device geometry and depends only on the drain current, similarly to a bipolar transistor. The transconductance is substantially lower, so this operating mode does not yield maximum gain for a given device. However, this is an important mode in low-power circuits and we'll consider it in more detail in Section 6.6.

To first order the input capacitance is formed primarily by the gate and channel with the gate oxide as dielectric, $C_{GS} = \varepsilon_0 \varepsilon_{\mathrm{SiO_2}} WL/d_{ox}$. This is increased by the fringing capacitance from the gate to the source and drain electrodes. In the saturation regime the gate–channel capacitance is modified by the longitudinal charge distribution in the channel.

6.2.3 *MOSFET types*

MOSFETs can be configured in various ways to provide great flexibility in circuit design. First, they can be implemented either as n-channel or p-channel devices.

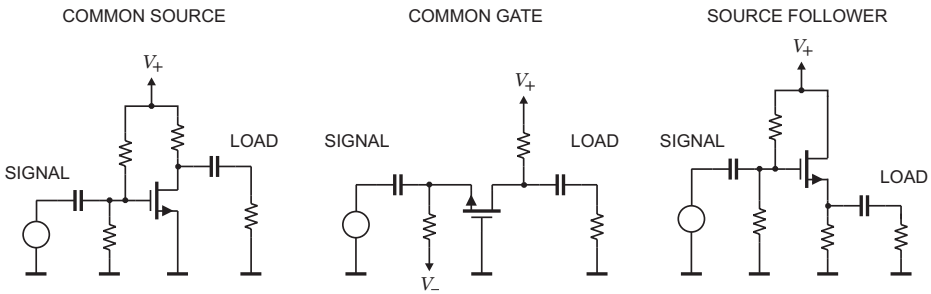

FIG. 6.25. Circuit topologies of FET amplifiers. Analogously to bipolar transistor amplifiers the devices can be operated in common-source (left), common-gate(middle) and common-drain (source follower, right).

The former uses highly doped n-regions to form the source and drain in a p-substrate. The latter is fully complementary and uses highly doped p-regions to form the source and drain in an n-substrate. The devices are identical, except that the polarities of the voltages are reversed and the carrier mobilities are different. Furthermore, the threshold voltage can be adjusted by a shallow surface layer with appropriate doping. Thus, at zero gate voltage devices can be designed to be normally on or normally off.

Complementary devices, *i.e.* n-channel and p-channel MOSFETs can be used together to form complementary MOS (CMOS) circuits, as was discussed in Chapter 5. NMOS devices must reside in a p-bulk, whereas PMOS devices must be in an n-bulk, so if the substrate is p-type, an n-well must be provided for the PMOS devices in a CMOS circuit (for a CMOS cross-section see Appendix A). Conversely, an n-substrate can be used with p-wells for the NMOS transistors.

6.2.4 *MOS Transistors in Amplifiers*

As shown for BJTs, three different circuit configurations are possible (Figure 6.25). The analysis follows the same lines as in Section 6.1.1 and the stages are used in similar ways. The most important difference to bipolar amplifiers is that the input resistance is high, as little or no gate current flows.

The voltage gain of the common-source stage

$$A_v = g_m R_L \ , \tag{6.41}$$

where R_L is the total load resistance at the collector, *i.e.* the parallel combination of collector resistor, transistor output resistance, and the input resistance of the following stage. However, neither the transconductance nor the transistor output resistance are as simply characterized as for bipolar transistor. The FET's transconductance depends on both current and device geometry, so it can vary widely from device to device. The same is true for the output resistance. Increasing the channel length tends to increase the output resistance. Generally,

for a given current the transconductance of a bipolar transistor will be substantially larger than in an FET. This is discussed in more detail later in Section 6.6.

The common-gate stage has properties similar to its bipolar transistor counterpart. The input resistance

$$r_i = \frac{1}{g_m} \tag{6.42}$$

and the output resistance is increased due to local negative feedback, as discussed in Appendix D.

Corresponding to the emitter follower, the common-drain stage or "source follower" has a voltage gain < 1 and its output resistance

$$r_o = \frac{1}{g_m} \tag{6.43}$$

is low, so its current drive capability is much higher than either the common-source or common-gate stages. Source followers are used to drive low-impedance or capacitive loads. Their input impedance is high, so they are useful in isolating the output of a common-source amplifier, which needs a high impedance load to provide voltage gain, from a low impedance cable or other low-impedance stage.

6.3 Noise in transistors

6.3.1 *Noise in field effect transistors*

The primary noise sources in field effect transistors are

- thermal noise in the channel and
- gate current.

The analysis of these noise sources is greatly simplified at frequencies where the transit times in the device are much smaller than the period $1/f$. In the circuits of primary interest here the relevant frequencies are much smaller than the intrinsic cutoff frequencies of the device (MHz *vs.* GHz), so the change in phase of a signal traversing the device is negligible and for all practical purposes fluctuations are coupled instantaneously to the device electrodes.

Thermal velocity fluctuations of the charge carriers in the channel superimpose a noise current on the quiescent output current. The spectral density of the noise current at the drain of a JFET or MOSFET is

$$i_{nD}^2 = \frac{N_{tot}e}{L^2}\mu_0 4kT_e \ . \tag{6.44}$$

The current fluctuations depend on the number of charge carriers in the channel N_{tot} and their thermal velocity, which in turn depends on their temperature T_e and low field mobility μ_0. Finally, the induced current scales with $1/L$ because of Ramo's theorem.

To make practical use of the above expression it is necessary to express it in terms of directly measurable device parameters. Since the transconductance in the saturation region

$$g_m \propto \frac{W}{L}\mu N e d = \mu \frac{N_{tot} e}{L^2} \ , \tag{6.45}$$

one can express the noise current as

$$i_{nD}^2 = \gamma_n g_m 4kT_0 \ , \tag{6.46}$$

where $T_0 = 300\,\mathrm{K}$ and γ_n is a semi-empirical constant that depends on the carrier concentration in the channel and the device geometry. It also accounts for the fact that the electron temperature can be higher than room temperature because phonon emission at high fields introduces velocity fluctuations that can exceed thermal velocity.

The output noise current is referred to the input by dividing by the transconductance, yielding an equivalent input noise voltage spectral density

$$e_n^2 = \frac{i_{nD}^2}{g_m^2} = \gamma_n \frac{4kT_0}{g_m} \ . \tag{6.47}$$

The noise coefficient γ_n is usually given as 2/3, but is typically in the range 0.5 – 1. In this expression the temperature dependence is implicit in g_m. When the device is cooled the electron temperature doesn't decrease correspondingly. However, the equivalent input noise voltage does improve because the mobility, and thus g_m, increases.

In a JFET the gate noise current has two components. The first is the shot noise associated with the reverse bias current of the gate–channel diode. The gate current I_G carries shot noise, so its spectral noise density is $i_n^2 = 2eI_G$. The second contribution to the input noise current originates from the thermal noise of the channel, which couples through the gate–channel capacitance to the gate electrode. In a JFET this coupling capacitance is about 2/3 of the gate–source capacitance (van der Ziel 1963, Johnson 1966). Thus, the total input noise current

$$i_{nG}^2 = 2eI_G + e_n^2 \omega^2 \frac{2}{3} C_{GS}^2 \ . \tag{6.48}$$

Figure 6.26 shows the resulting noise model.

When an impedance Z is connected between the gate and the source, the gate noise current will flow through this impedance and generate a voltage at the gate

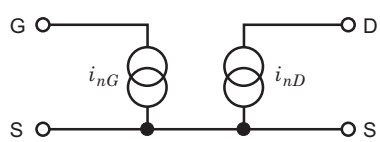

FIG. 6.26. Noise sources in FETs.

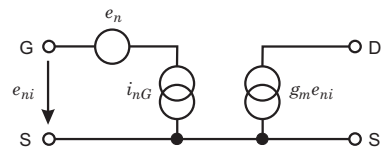

FIG. 6.27. Noise model for JFETs and MOSFETs.

$e_{nG} = Zi_{nG}$, leading to an additional noise current at the output $g_m e_{nG}$. Then the total noise current at the output becomes

$$i_{no}^2 = i_{nD}^2 + (g_m Z i_{nG})^2 \ . \tag{6.49}$$

To allow a direct comparison with the input signal this cumulative noise will be referred back to the input to yield the equivalent input noise voltage

$$e_{ni}^2 = \frac{i_{no}^2}{g_m^2} = \frac{i_{nD}^2}{g_m^2} + Z i_{nG}^2 \equiv e_n^2 + Z i_{nG}^2 \ . \tag{6.50}$$

Figure 6.27 shows the corresponding model. Note that the total input noise voltage is not physically present at the input. This is only true of the noise current component.

In MOSFETs the DC component of the gate current is very low (at oxide thicknesses $< 10\,\mathrm{nm}$ it is dominated by electron tunneling through the oxide). Here the capacitive coupling from the channel to the gate is the dominant noise current source. To a good approximation (Shoji 1966)

$$i_{nGc}^2 \approx \frac{1}{2} kT\omega^2 \frac{C_{iG}^2}{g_{msat}} \ , \tag{6.51}$$

i.e. the input noise current increases with frequency. C_{iG} is the portion of the input capacitance that couples to the channel, so it excludes the fringing capacitance to the drain and source, for example. In both JFETs and MOSFETs the capacitively coupled noise current is correlated with the equivalent input noise voltage, so when this contribution becomes significant the cross-correlation term (van der Ziel 1963) must be included as described in Chapter 3.

Figure 6.27 applies to both JFETs and MOSFETs. The total input noise current $i_n^2 = i_{nG}^2 + i_{nGc}^2$. The noise current flows through the source impedance, producing an input noise voltage, which together with the equivalent input noise voltage of the transistor e_n yields the total input noise e_{ni}. This in turn translates to the output through the transconductance g_m to yield a noise current at the output $g_m e_{ni}$.

The input noise current and voltage translate into the equivalent noise charge

$$Q_n^2 = i_n^2 F_i T_S + e_n^2 C_i^2 \frac{F_v}{T_S} \ . \tag{6.52}$$

For a representative JFET $g_m = 0.02\,\mathrm{S}$, $C_i = 10\,\mathrm{pF}$, and $I_G < 150\,\mathrm{pA}$. If we set $F_i = F_v = 1$,

$$Q_n^2 = 1.9 \cdot 10^9 \left[e^2/\mathrm{s}\right] T_S + \frac{3.25 \cdot 10^{-3} \left[e^2\,\mathrm{s}\right]}{T_S} \ . \tag{6.53}$$

As the shaping time T_S decreases, the current noise contribution decreases and the voltage noise contribution increases. For $T_S = 1\,\mu\mathrm{s}$ the current contribution is $44\,e$ and the voltage contribution $57\,e$, so the two contributions are roughly equal.

6.3.1.1 *Optimization of device geometry* For a given device technology and normalized operating current I_D/W both the transconductance and the input capacitance are proportional to device width W, so that the ratio

$$\frac{g_m}{C_i} = \mathrm{const} \ . \tag{6.54}$$

Then the signal-to-noise ratio can be written as

$$\begin{aligned} \left(\frac{S}{N}\right)^2 &= \frac{(Q_s/C)^2}{e_n^2} = \frac{Q_s^2}{(C_d + C_i)^2} \frac{g_m}{4kT_0\Delta f} \\ \left(\frac{S}{N}\right)^2 &= \frac{Q_s^2}{\Delta f} \frac{1}{4kT_0} \left(\frac{g_m}{C_i}\right) \frac{1}{C_i \left(1 + \dfrac{C_d}{C_i}\right)^2} \ . \end{aligned} \tag{6.55}$$

From this we see that S/N is maximized when $C_i = C_d$ (capacitive matching).

When $C_i \ll C_d$, the detector capacitance dominates, so the effect of increased transistor capacitance is negligible. As the device width is increased the transconductance increases and the equivalent noise voltage decreases, so S/N improves. When $C_i > C_d$ the equivalent input noise voltage decreases as the device width is increased, but only with $1/\sqrt{W} \propto 1/\sqrt{C_i}$, so the increase in capacitance overrides, decreasing S/N. Note that capacitive matching relies on the linkage between device width and transconductance. Merely increasing the device capacitance without a corresponding decrease in noise brings no benefit. Adapting the device to the sensor can be accomplished by connecting transistors in parallel or by selecting the device width, as described in Section 6.6.

6.3.1.2 *Minimum obtainable noise charge* Device scaling can be used to determine the minimum obtainable noise charge for a given device technology. The transconductance of an FET increases with drain current as shown in Figure 6.24. However, noise only decreases up to a certain current. The reason is that the noise from parasitic source and gate resistances becomes significant. Furthermore, since the equivalent noise charge depends on the total capacitance at the input node, we must consider not just the gate–channel capacitance, but also any other parasitic capacitances associated with the device structure.

Assume that a transistor of width W assumes its minimum noise at a current I_D with an associated transconductance g_m. Since the parasitic gate and source resistances are both inversely proportional to device width, the optimum current density I_D/W will be the same for all widths of transistors using the same technology (and device length). Thus, to obtain minimum noise one can tailor the FET to a given detector by scaling the device width and keeping the current density I_D/W constant.

Within this framework one can characterize the device technology by the normalized transconductance and input capacitance

$$g'_m = \frac{g_m}{W} \quad \text{and} \quad C'_i = \frac{C_i}{W} \tag{6.56}$$

and use these quantities to scale to any other device width. Since the equivalent input noise voltage

$$e_n^2 \propto \frac{1}{g_m} \ , \tag{6.57}$$

the normalized input noise voltage is

$$e'_n = e_n \sqrt{W} \ . \tag{6.58}$$

Using these quantities the equivalent noise charge can be written as

$$Q_n^2 = 4kT_0 \frac{\gamma_n}{W g'_m} \frac{F_v}{T_S} (C_d + C_s + W C'_i)^2 \ , \tag{6.59}$$

where C_s is any stray capacitance present at the input in addition to the detector capacitance C_d and the FET capacitance WC'_i. For $WC'_i = C_d + C_s$ the noise attains its minimum value

$$Q_{n,min} = \sqrt{\frac{16kT_0}{\kappa_n} \frac{F_v}{T_S} (C_d + C_s)} \ , \tag{6.60}$$

where

$$\kappa_n \equiv \frac{g_m}{\gamma_n C_i} \tag{6.61}$$

is a figure of merit for the noise performance of the FET. Note that the device input capacitance WC'_i includes the gate–channel capacitance and the fringing capacitance from the gate to the source and drain.

As an example, an n-channel CMOS transistor with $1.2\,\mu$m channel length at a current density $I_D/W = 0.3\,\mathrm{A/m}$ has a ratio of transconductance to input capacitance ratio $g_m/C_i = 3 \cdot 10^9\,\mathrm{s}^{-1}$. Assume $\gamma_n = 1$. Using a *CR-RC* shaper with a 20 ns shaping time and an external capacitance $C_d + C_s = 7.5\,\mathrm{pF}$, the minimum noise $Q_{n,min} = 88\,\mathrm{aC} = 546\,e$, achieved at a device width $W = 5\,\mathrm{mm}$, and a drain current of 1.5 mA.

The obtainable noise improves with the inverse square root of the shaping time, up to the point where $1/f$ noise becomes significant. For example, at $T_S = 1\,\mu$s, the minimum noise $Q_{n,min} = 12\,\mathrm{aC} = 77\,e$, although in practice additional noise contributions will increase the obtainable noise beyond this value.

6.3.2 *Low-frequency excess noise ("1/f noise")*

The preceding discussion has neglected $1/f$ noise, which adds a constant contribution independent of shaping time

$$Q_{nf}^2 \propto A_f C^2 \ . \tag{6.62}$$

Although excess low-frequency noise is determined primarily by the concentration of unwanted impurities and other defects, their effect in a specific technology is also affected by device size. For some forms of $1/f$ noise

$$A_f = \frac{K_f}{WLC_G^2} \ , \tag{6.63}$$

where C_G is the gate–channel capacitance per unit area, and K_f is an empirical constant that is device and process dependent. Typical values of the noise constant for various device types are

p-MOSFET	$K_f \approx 2 \cdot 10^{-32}\,\mathrm{C^2/cm^2}$
n-MOSFET	$K_f \approx 5 \cdot 10^{-31}\,\mathrm{C^2/cm^2}$
JFET	$K_f \approx 10^{-33}\,\mathrm{C^2/cm^2}$

Specific implementations can improve on these values. One should note that this model is not universally applicable, since excess noise usually does not exhibit a pure $1/f$ dependence; especially in PMOS devices one often finds several slopes. In practice, one must test the applicability of this parameterization by comparing it with data before applying it to scaled amplifiers. Nevertheless, as a general rule, devices with larger gate area $W \cdot L$ tend to exhibit better "$1/f$" noise characteristics.

The frequency where the "$1/f$" noise intersects the white noise, the "corner frequency" $f_c = A_f/e_n^2$, is often used as a measure of "$1/f$" noise. However, this can be misleading, as for the same level of "$1/f$" noise a higher white noise level will reduce the corner frequency. FETs typically exhibit noise corner frequencies in the kHz range. Specially designed JFETs can do better, whereas in small MOSFETs the noise corner can be in the range of hundreds of kHz.

6.3.3 *Noise in bipolar transistors*

In bipolar transistors the shot noise from the base current is important. As in FETs the transit times in modern bipolar transistors are much smaller than shaping times of interest in our applications, so we can neglect frequency dependencies and the noise model is greatly simplified. The basic noise model shown in Figure 6.28 is the same as used for FETs, but the magnitude of the input noise current is much greater, as the base current will be $1 - 100\,\mu$A rather than $< 100\,$pA as in a JFET or fA in a MOSFET.

The base current noise is the shot noise associated with the component of the emitter current provided by the base.

$$i_{nB}^2 = 2eI_B \ . \tag{6.64}$$

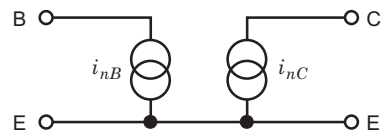

FIG. 6.28. Noise model of a bipolar transistor showing the primary noise sources.

The noise current in the collector is the shot noise originating in the base–emitter junction associated with the collector component of the emitter current.

$$i_{nC}^2 = 2eI_C \ . \tag{6.65}$$

Following the same argument as in the analysis of the FET, the output noise current is equivalent to an equivalent noise voltage

$$e_n^2 = \frac{i_{nC}^2}{g_m^2} = \frac{2eI_C}{(eI_E/kT)^2} \approx \frac{2(kT)^2}{eI_C} \ . \tag{6.66}$$

The noise equivalent circuit is shown in Figure 6.29, where i_n is the base shot noise current i_{nB}.

The equivalent noise charge (where T_S is the shaping time)

$$Q_n^2 = i_n^2 F_i T_S + e_n^2 C^2 \frac{F_v}{T_S} = 2eI_B F_i T_S + \frac{2(kT)^2}{eI_C} C^2 \frac{F_v}{T_S} \ . \tag{6.67}$$

Since $I_B = I_C/\beta_{DC}$,

$$Q_n^2 = 2e\frac{I_C}{\beta_{DC}} F_i T_S + \frac{2(kT)^2}{eI_C} C^2 \frac{F_v}{T_S} \ . \tag{6.68}$$

The current noise term increases with I_C, whereas the second (voltage) noise term decreases with I_C. As a result the equivalent noise charge attains a minimum

$$Q_{n,min}^2 = 4kT \frac{C}{\sqrt{\beta_{DC}}} \sqrt{F_i F_v} \tag{6.69}$$

at a collector current

$$I_C = \frac{kT}{e} C \sqrt{\beta_{DC}} \sqrt{\frac{F_v}{F_i}} \frac{1}{T_S} \ . \tag{6.70}$$

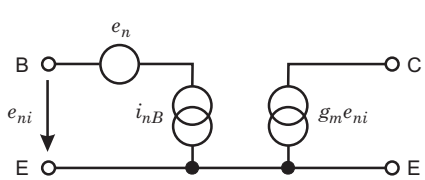

FIG. 6.29. Noise model of a bipolar transistor.

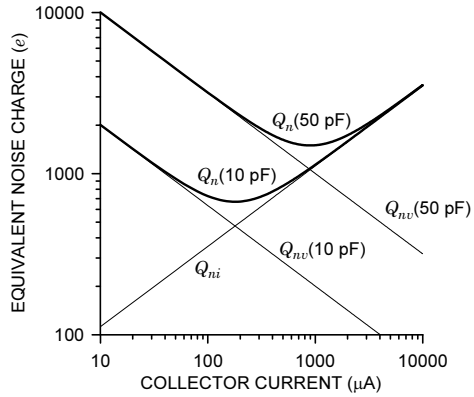

FIG. 6.30. Equivalent noise charge Q_n of a bipolar transistor *vs.* collector current, showing the current noise Q_{ni} and the voltage noise contributions Q_{nv} for 10 and 50 pF ($\tau = 25$ ns). When the total capacitance increases from 10 pF (bottom) to 50 pF (top), the minimum noise increases and the optimum collector current shifts upwards to increase the transconductance.

Figure 6.30 shows the calculated noise for sensor capacitances of 10 and 50 pF using a shaper with $F_i = 0.4$ and $F_v = 1.2$. At collector currents below optimum the collector shot noise dominates, because of reduced transconductance, whereas at high currents the base current shot noise takes over. Increasing the capacitance at the input shifts the collector current noise curve (equivalent input noise voltage contribution) upwards, so the minimum noise increases and shifts to a higher current. When the capacitance is increased to 50 pF, the minimum noise increases by a factor $\sqrt{5}$ and the optimum collector current is shifted five times higher.

Figure 6.31 shows the effect of changing the shaping time. Reducing the shaping time from 100 ns to 10 ns does not change the minimum noise, but shifts the optimum collector current higher. For a given shaper, the minimum obtainable noise is determined only by the total capacitance at the input and the DC gain of the transistor, not by the shaping time. The shaping time only determines the current at which this minimum noise is obtained.

The following relationships provide a simple estimate of obtainable BJT noise. For a *CR-RC* shaper the minimum noise

$$Q_{n,min} = 772 \left[\frac{e}{\sqrt{\text{pF}}}\right] \cdot \frac{\sqrt{C}}{\sqrt[4]{\beta_{DC}}} \tag{6.71}$$

obtained at a collector current

$$I_c = 26 \left[\frac{\mu\text{A} \cdot \text{ ns}}{\text{pF}}\right] \cdot \frac{C}{\tau}\sqrt{\beta_{DC}} \tag{6.72}$$

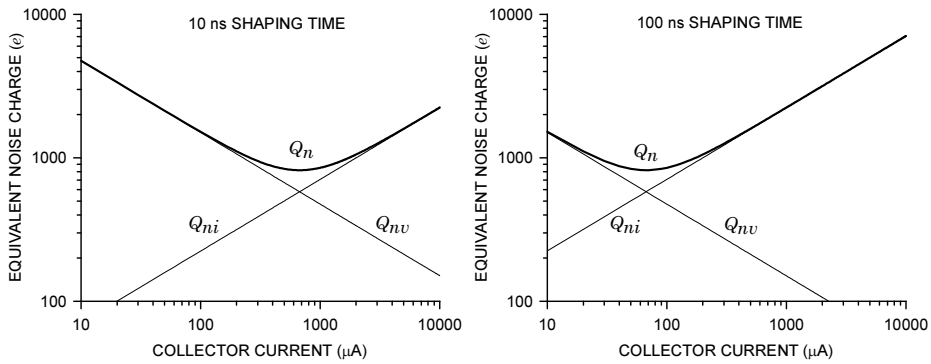

FIG. 6.31. When the shaping time is changed from 10 ns (left) to 100 ns (right) the minimum noise achievable with a bipolar transistor remains the same, but it obtains at a lower collector current.

Since typically $\beta_{DC} \approx 100$, these expressions allow a quick and simple estimate of the noise obtainable with a bipolar transistor. Setting $\beta_{DC} = 100$, the minimum noise

$$Q_{n,min} \approx 250 \left[\frac{e}{\sqrt{\mathrm{pF}}} \right] \cdot \sqrt{C} \ . \tag{6.73}$$

In shapers other than the simple *CR-RC* configuration both the minimum obtainable noise and the optimum current will be modified slightly as shown in eqns 6.69 and 6.70.

Low-frequency noise in bipolar transistors tends to be lower than in FETs, with typical noise corners in the 1 – 10 Hz range.

6.3.4 *Comparison between bipolar and field effect transistors*

The noise characteristics of bipolar transistors differ from field effect transistors in four important aspects:

1. The equivalent input noise current cannot be neglected, due to base current flow.
2. The total noise does not decrease monotonically with increasing device current.
3. The minimum obtainable noise does not depend on the shaping time.
4. The input capacitance is usually negligible.

The last statement requires some explanation. The input capacitance of a bipolar transistor is dominated by two components:

1. The geometrical junction capacitance, or transition capacitance C_{TE}.
2. The diffusion capacitance C_{DE}.

The transition capacitance in small devices is typically about 0.5 pF. The diffusion capacitance depends on the current flow I_E through the base–emitter junction and on the base width W, which sets the diffusion profile (Cooke 1971).

$$C_{DE} = \frac{\partial q_B}{\partial V_{BE}} = \frac{eI_E}{kT}\left(\frac{W}{2D_B}\right) \equiv \frac{eI_E}{kT} \cdot \frac{1}{\omega_{Ti}} \ . \tag{6.74}$$

D_B is the diffusion constant in the base and ω_{Ti} is a frequency that characterizes carrier transport in the base. ω_{Ti} is roughly equal to the frequency f_T where the current gain of the transistor is unity.

Inserting some typical values, $I_E = 100\mu$A and $\omega_{Ti} = 10$ GHz, yields $C_{DE} = 0.4$ pF. The transistor input capacitance $C_{TE} + C_{DE} = 0.9$ pF, whereas FETs providing similar noise values at comparable currents have input capacitances in the range 5 – 10 pF.

Except for low-capacitance detectors, the current dependent part of the BJT input capacitance is negligible, so it will be neglected in the following discussion. For practical purposes the amplifier input capacitance can be considered constant at 1 – 1.5 pF.

This leads to another important conclusion. Since the primary noise parameters do not depend on device size and there is no significant linkage between noise parameters and input capacitance, *capacitive matching does not apply to bipolar transistors.*

Indeed, capacitive matching is a misguided concept for bipolar transistors. Consider two transistors with the same DC gain but different input capacitances. Since the minimum obtainable noise

$$Q^2_{n,min} = 4kT\frac{C}{\sqrt{\beta_{DC}}}\sqrt{F_i F_v} \ , \tag{6.75}$$

increasing the transistor input capacitance merely increases the total input capacitance C and the obtainable noise.

Since the base current noise increases with shaping time, bipolar transistors are only advantageous at short shaping times. With current technologies FETs are best at shaping times greater than 50 – 100 ns, but decreasing feature size of MOSFETs will improve their performance.

6.3.5 *Noise optimization – capacitive matching revisited*

"Capacitive matching" is often presented as a universal criterion for noise optimization with capacitive sources. The results derived for bipolar transistors already show that capacitive matching does not apply in all amplifiers. This discussion is supposed to clarify where capacitive matching is useful and where it isn't.

Consider the an array of n amplifiers with both current and voltage noise whose inputs are connected in parallel, so each sees the same signal, and whose outputs are summed (Figure 6.32). As in previous derivations of the equivalent

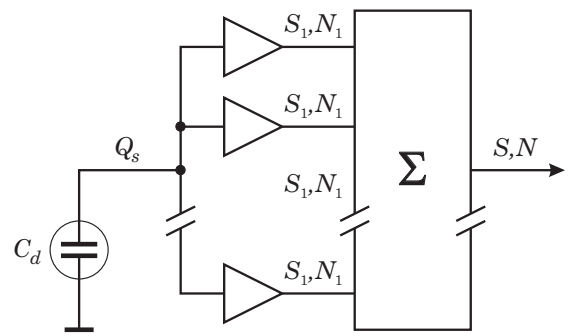

FIG. 6.32. An array of amplifiers whose inputs are connected in parallel and outputs are summed.

noise charge, we assume voltage-sensitive amplifiers with purely capacitive inputs. Furthermore, to simplify the analysis, the amplifiers do not utilize feedback.

Of course, in considering the current and voltage noise contributions of this amplifier array, one can follow a formal argument based on the noise charge

$$Q_n^2 = i_n^2 F_i T_S + e_n^2 C^2 \frac{F_v}{T_S} \ . \tag{6.76}$$

Since the current noise contribution does not depend on capacitance, matching the amplifier input capacitance to the detector capacitance does not affect this term. On the other hand, since the voltage contribution does depend on capacitance, a correlation between e_n and C can yield an optimization condition. To provide more insight into this argument, we'll consider the signal and noise contributions separately.

6.3.5.1 *Current noise* For the noise currents originating in the individual amplifiers, the common connection to the signal source is a summing node, so if i_{n1} is the equivalent noise current of a single amplifier, for n amplifiers the total input noise current flowing through the signal source impedance is

$$i_n = \sqrt{n} \cdot i_{n1} \ . \tag{6.77}$$

The flow of this current through the input impedance Z_i formed by the parallel connection of the detector capacitance and amplifier capacitances gives rise to a noise voltage

$$v_n = i_n Z_i \ . \tag{6.78}$$

This voltage is applied in parallel to all amplifier inputs, so at the output of an individual amplifier (assuming a gain A) the noise level is

$$N_1(n) = A v_n = A\sqrt{n} i_{n1} Z_i \ . \tag{6.79}$$

At the output of the summing circuit, the cumulative noise from all amplifier outputs is simply n times larger, as all amplifiers see the same input.

$$N(n) = nN_1(n) = An^{3/2} i_{n1} Z_i \ . \tag{6.80}$$

The magnitude of the signal applied to all amplifiers is the same, since the amplifiers respond to voltage and all inputs are connected in parallel. For a signal current i_s the input voltage is

$$v_s = i_s Z_i \ . \tag{6.81}$$

In the summed output the signals add coherently, so that

$$S(n) = nAi_s Z_i \tag{6.82}$$

and the signal-to-noise ratio

$$\frac{S(n)}{N(n)} = \frac{nAi_s Z_i}{An^{3/2} i_{n1} Z_i} = \frac{i_s}{\sqrt{n} \cdot i_{n1}} = \frac{1}{\sqrt{n}} \frac{S(1)}{N(1)} \tag{6.83}$$

is degraded by a factor $1/\sqrt{n}$ with respect to the single amplifier.

Varying the amplifier input capacitance is irrelevant. As the total input capacitance increases, the noise voltage developed at the input decreases with Z_i, but so does the signal voltage $v_s = Q_s/C$, so the signal-to-noise ratio is unaffected.

6.3.5.2 *Voltage noise* The voltage noise contribution differs from the current noise in an important aspect. Voltage noise is not additive at the input.

This statement can be justified with two arguments, the first more physical and the second more formal.

1. Voltage noise tends to originate within a device (*e.g.* thermal noise of an FET channel or collector shot noise in a BJT) and appears as a noise current at the output, which is mathematically transformed to the input. This noise voltage is not physically present at the input and is not affected by any connections or components in the input circuit.
2. The noise voltage sources that represent all voltage noise contributions of a given amplifier are in series with the individual inputs. Since the input impedance of the amplifier is postulated to be much higher than the source impedance (amplifier input capacitance $\ll$ detector capacitance), the source impedance is by definition negligible in comparison, so the noise voltage associated with a given amplifier only develops across the input of that amplifier.

Assume that each amplifier has an input referred noise v_{n1} and an input capacitance C_{i1}. Then the input signal voltage

$$v_s = \frac{Q_s}{C} \ , \tag{6.84}$$

where C is the total input capacitance including the detector,

$$C = C_d + nC_{i1} \ . \tag{6.85}$$

The signal at each amplifier output is

$$S_1 = Av_s = A\frac{Q_s}{C_d + nC_{i1}} \ . \tag{6.86}$$

The noise at each amplifier output is

$$N_1 = Av_{n1} \ . \tag{6.87}$$

After summing the n outputs the signal-to-noise ratio

$$\frac{S(n)}{N(n)} = \frac{nS_1}{\sqrt{n}N_1} = \sqrt{n}\frac{S_1}{N_1} = \sqrt{n}\frac{A\dfrac{Q_s}{C_d + nC_{i1}}}{Av_{n1}} = \frac{Q_s}{v_{n1}/\sqrt{n}} \cdot \frac{1}{C_d + nC_{i1}} \ , \tag{6.88}$$

which assumes a maximum when $C_d = nC_{i1} = C_i$. Under this "capacitive matching" condition $\sum C_{i1} = C_d$ the signal-to-noise ratio

$$\frac{S}{N} = \frac{Q_s}{v_{n1}}\sqrt{\frac{C_d}{C_{i1}}}\frac{1}{C_d + nC_{i1}} = \frac{Q_s}{v_{n1}}\sqrt{\frac{C_d}{C_{i1}}}\frac{1}{2C_d} \tag{6.89}$$

or

$$\frac{S}{N} = \frac{1}{2}\frac{Q_s}{v_{n1}\sqrt{C_{i1}}} \cdot \frac{1}{\sqrt{C_d}} \ . \tag{6.90}$$

Since v_{n1} and C_{i1} are properties of the individual amplifier, *i.e.* constants, the signal-to-noise ratio decreases with the square root of detector capacitance.

This relationship only holds if

1. the noise of the input amplifier/device decreases with increasing input capacitance;
2. the input capacitance is scaled with the detector capacitance ("capacitive matching").

The first point is critical; if the noise voltage of a device and its input capacitance are not correlated, capacitive matching is deleterious. A specific example is a MOSFET operated in weak inversion. Its transconductance depends only on current, independent of geometry. If power consumption is to be kept constant, increasing the size of the device at the same operating current will not increase the transconductance, but it will increase the input capacitance and, as a result, the equivalent noise charge. Later in Section 6.6 we'll see examples of noise optimization in weak and moderate inversion.

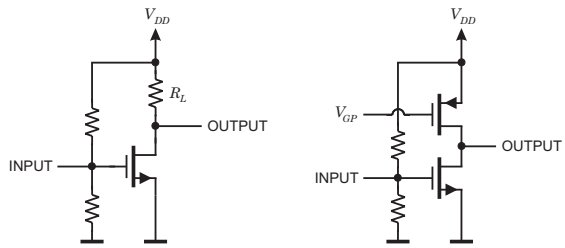

FIG. 6.33. A simple MOSFET amplifier with a resistive load (left) and with a PMOS transistor as an active load (right).

6.4 Composite amplifiers

Up to now, we have only considered amplifiers with a single transistor, as shown in the first panel of Figure 6.33. We'll use this as a starting point to demonstrate why more complex amplifiers are needed. The gain of the simple amplifier

$$A_v = g_m(R_L||r_o) = g_m \frac{R_L r_o}{R_L + r_o} \ , \tag{6.91}$$

where R_L is the load resistor and r_o the output resistance of the transistor. Let's assume that the transistor is to operate at a drain voltage of $V_D = 3\,\mathrm{V}$ with a drain current $I_D = 1\,\mathrm{mA}$ and that the supply voltage is $V_{DD} = 6\mathrm{V}$. Then $R_L = (V_{DD} - V_D)/I_D = 3\,\mathrm{k}\Omega$. How does this compare to the output resistance of the transistor?

Figure 6.34 shows representative output curves of NMOS and PMOS transistors. The transition into the saturation regime of the PMOS device is much "softer" than of the NMOS transistor, as expected from the lower mobility of

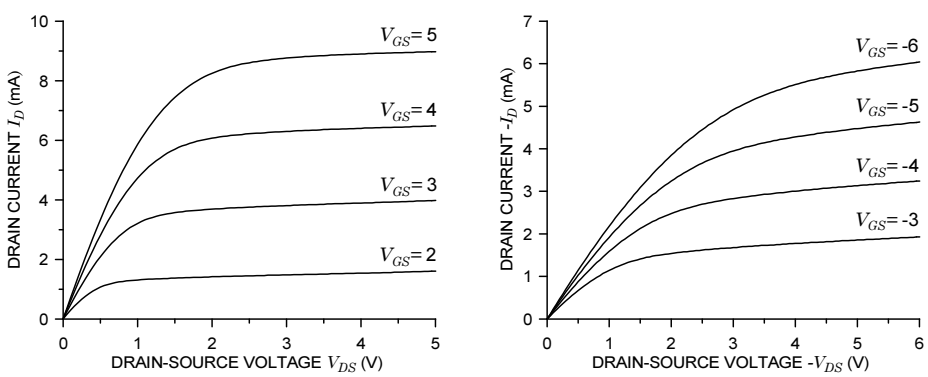

FIG. 6.34. Output curves for NMOS (left) and PMOS (right) transistors, both with $W = 25\,\mu\mathrm{m}$ and $L = 0.8\,\mu\mathrm{m}$.

holes, so they require higher voltages to attain velocity saturation. The smaller transconductance of the PMOS device is also clearly visible from the smaller spacing of the V_{GS} curves. However in both devices the output resistance at 1 mA is about 10 kΩ, so the voltage gain is limited by the load resistor. A higher voltage gain can be attained by using an active load, as shown in the right-hand part of Figure 6.33. Now the effective load resistance is the parallel combination of the NMOS and PMOS output resistance in parallel, *i.e.* 5 kΩ. Can this be improved further? Increasing the channel length raises the output resistance, as shown in Figure 6.35, which also shows how the output resistance falls with increasing drain current. Increasing the device width reduces the output resistance proportionally, $r_o \propto 1/W$, as this is equivalent to connecting devices in parallel, so the normalized output resistance $r_o W$ is plotted *vs.* the normalized drain current I_D/W to yield curves that can be scaled to any device width and current in the 0.8 μm process used for these devices. Increasing the channel length also reduces the transconductance, so this is most effectively applied in the load transistor, rather than the input device.

Figure 6.36 shows how the differential amplifier topology discussed in Section 6.1.1.1 can be applied in a MOSFET amplifier. M1–M2 and M3–M4 comprise the differential gain stages. M22+M23 and M14+M15 are the current sources

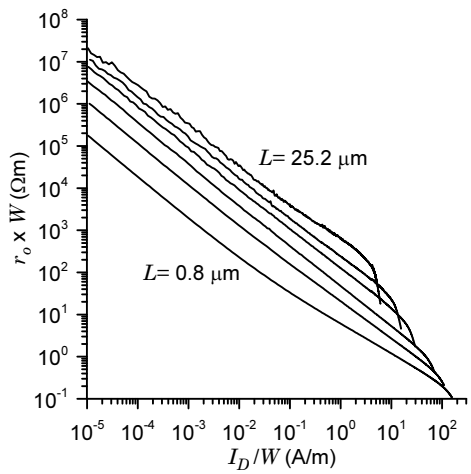

FIG. 6.35. Normalized output resistance $r_o \times W$ of PMOS transistors *vs.* normalized drain current I_D/W for channel lengths of 0.8, 1.2, 2.0, 5.2, 10, and 25.2 μm. The output resistance is obtained by dividing the normalized output resistance by the device width. The drain voltage $V_{DS} = 3$ V, so the downward kink in the curves indicates where the increasing gate voltage brings the output out of the saturation regime. The "ripples" in the curves are artifacts of the measurement system that are exaggerated by the required numerical differentiation.

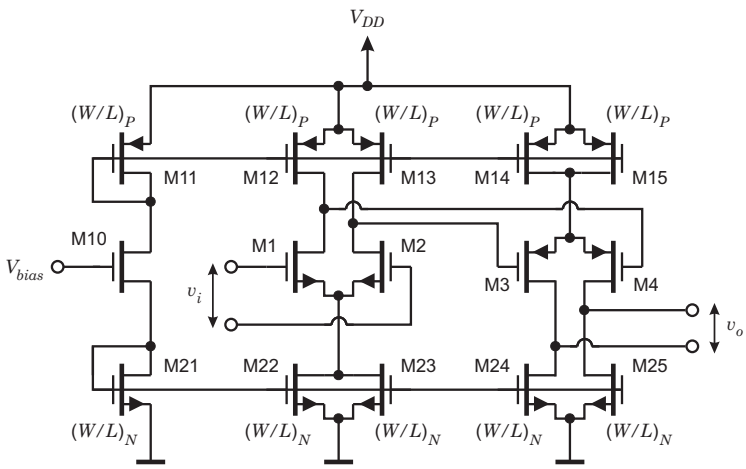

FIG. 6.36. A differential amplifier cascade implemented in CMOS technology. The input signal is applied to M1–M2 and the output taken from M3–M4.

that set the standing current for the NMOS and PMOS differential pairs. M12, M13 and M24, M25 are the respective loads. M10, M11, and M21 establish the bias voltages for the current sources and loads.

The bias circuit demonstrates current mirroring, a powerful technique in matching or scaling the operating currents of different stages. In the first gain stage the total currents through M12+M13 and M22+M23 must be equal. The chain M11–M10–M21 is designed so that the current flowing through it is $I_0/2$, which sets the appropriate gate–source voltages at M11 and M21. The gate–source voltage of M21 is transferred to M22 and M23, all of which have the same geometry, so the gate voltage will establish the same current flow in each transistor, *i.e.* the current is "mirrored" from M21 to M22 and M23. Correspondingly, the gate–source voltage of M11 is transferred to the load transistors M12 and M13. The current source uses paralleled transistors M22 and M23 to establish the total current I_0, whereas the current flowing through the load transistors M12 and M13 is $I_0/2$. The second stage devices M14–M15 and M24–M25 are biased correspondingly. In principle, the current sources M14+M15 and M22+M23 could each be one transistor with twice the width, but since edge effects can be important, it is safest to use multiple transistors with the same geometry. The bias transistor M10 establishes the appropriate voltage drop between M11 and M21 to match the drain–source voltage of the gain stages.

As described in Section 6.1.1.1 this circuit is insensitive to changes in the supply voltage V_{DD} and also to common mode components at the amplifier inputs ("common mode rejection"). The cascade of NMOS and PMOS stages also allows the quiescent level at the output to match the input.

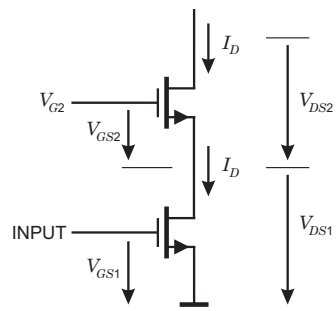

FIG. 6.37. Cascode amplifier.

As described above, the output resistance of a MOSFET can be raised by increasing the channel length. However, this also requires a higher gate–source voltage, which increases the saturation voltage, so this techniques is often limited by the available supply voltage. Furthermore, increasing the channel length reduces the achievable transconductance. The output resistance of the input device can be increased while maintaining the transconductance by applying the configuration shown in Figure 6.37. The lower "input" transistor operates in common-source and provides transconductance. The upper "cascode" transistor operates with a fixed gate voltage, so for the signal drain current of the input transistor it operates in common-gate and presents a low input resistance to the lower transistor. The current gain of the cascode transistor is unity, so the transconductance of the combination equals that of the input transistor. The output resistance of the cascode transistor is increased because the output resistance of the input transistor is in series with the source. If the output voltage increases, the current draw of the cascode transistor increases. However, the increased current gives an increased voltage drop across the output resistance of the input transistor, so the gate–source voltage of the cascode transistor decreases, which counteracts the increase in current. Thus, the output resistance increases by virtue of this local negative feedback (see Appendix D).

Mathematically, this is straightforward to derive without any knowledge of electronics. We consider the effect of changing the output voltage with zero input signal. The drain voltage at the output of the cascode

$$V_{DS} = V_{DS1}(V_{GS1}, I_D) + V_{DS2}(V_{GS2}, I_D) \ . \tag{6.92}$$

The total differential

$$dV_{DS} = dV_{GS1}\frac{dV_{DS1}}{dV_{GS1}} + dI_D\frac{dV_{DS1}}{dI_D} + dV_{GS2}\frac{dV_{DS2}}{dV_{GS2}} + dI_D\frac{dV_{DS2}}{dI_D} \ . \tag{6.93}$$

Since $dV_{GS1} \equiv 0$ and $dV_{DS}/dI_D = r_o$,

$$dV_{DS} = dI_D r_{o1} + r_{o1} dI_D \frac{r_{o2} dI_D}{dV_{GS2}} + r_{o2} dI_D \tag{6.94}$$

and

$$\frac{dV_{DS}}{dI_D} = r_{o1} + r_{o1}\frac{r_{o2}dI_D}{dV_{GS2}} + r_{o2} \ . \tag{6.95}$$

Using $dI_D/dV_{GS} = g_m$, the output resistance

$$r_o = r_{o1} + r_{o2} + r_{o1}r_{o2}g_{m2} \ . \tag{6.96}$$

The cascode increases the output resistance by an additional term, which is the output resistance of the input stage multiplied by the voltage gain of the cascode stage. Although we've used the nomenclature for MOSFETs, the same analysis applies to other devices.

The transistor pair forming the cascode acts as a single device with the transconductance of the input device, but with a much larger output resistance. The same technique can also be applied to the load, if lengthening the channel is not adequate to achieve the desired gain.

With these techniques we can increase the amplifier gain. The second key ingredient is bandwidth. As already shown in Chapter 2, any capacitance C_o present at the output causes the voltage gain to roll off at frequencies above $\omega_u = 1/R_LC_o$, where R_L is the effective load resistance of the amplifier, which for active loads is the parallel combination of the output resistances of the gain and load devices. We increased the voltage gain, by increasing R_L, but this also reduced the bandwidth. In this simple circuit the product of voltage gain and bandwidth

$$A_v\omega_u = g_mR_L\frac{1}{R_LC_o} = \frac{g_m}{C_o} \ . \tag{6.97}$$

The gain–bandwidth product is independent of load resistance and equal to the frequency where the amplifier gain is unity, the "unity gain frequency" ω_0. Thus, at any frequency well above the cutoff frequency, where the gain drops linearly with frequency, the product of voltage gain and frequency is equal to the gain–bandwidth product ω_0. Negative feedback applied around an amplifier (see Appendix D), maintains the gain–bandwidth product, so one can use a gain stage with a large gain–bandwidth product and by setting the gain achieve the desired bandwidth. The excess gain ("reserve gain") at low frequencies is useful in stabilizing the DC baseline and also increases linearity. From this we learn that desirable features in an amplifying device or gain stage are high transconductance and low output capacitance.

The capacitance at the output node C_o depends on circuit topology and basic characteristics of the IC technology used. In our applications the circuit bandwidth is determined less by the inherent device speed, than by the device capacitance and stray capacitance to the substrate. Thus, reducing the capacitive load at the high impedance node is crucial in maximizing the bandwidth. In a MOSFET the relevant capacitances are the input capacitance of the following stage and the capacitance of the drain implants to the bulk. The capacitance of the connecting traces also adds, but with short trace lengths this is negligible. The capacitance of the drain implants usually dominates, so reducing the device

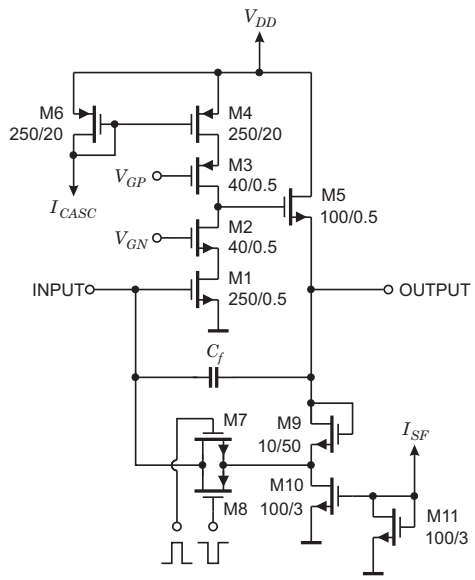

FIG. 6.38. Circuit diagram of a charge-sensitive amplifier using cascode circuitry. Transistor sizes are indicated as W/L, *e.g.* M1 has $W = 250\,\mu$m and $L = 0.5\,\mu$m.

width is beneficial. Another technique is the use of an annular device geometry, where the drain is a small island in the center.

On the other hand, increasing transconductance to improve noise and speed ultimately leads to increased device width W, which increases the output capacitance. The cascode allows use of a wide input device coupled with a narrow cascode device, so the cascode provides both increased low-frequency gain and gain–bandwidth product. This combination of features makes the cascode one of the key circuits in detector front-ends.

Figure 6.38 illustrates how these techniques can be applied to a charge-sensitive amplifier. Unlike commercial operational amplifiers, the input stage is not formed by a differential pair of transistors, *i.e.* with complementary noninverting and inverting inputs. In a differential pair the noise current of one device modulates the current of the other, so the combined effect of both devices increases the equivalent input noise voltage by $\sqrt{2}$. Thus, for the same noise level as a single transistor, each transistor in the differential pair requires at least twice the current, so the pair more than quadruples the current draw. Instead the input stage is "single ended". This comes at the expense of "power supply rejection ratio", *i.e.* the sensitivity of the circuit to spurious signals introduced through the power supply V_{DD} or ground. Indeed, single-ended circuitry typically shows gain for signals on the power supply lines. In high density detector

systems the need to reduce power tends to prevail over the robustness of fully differential circuitry.

Cascodes are used both in the gain and load portions. M1 and M2 comprise the transconductance stage and M3 and M4 the load. M1 is sized to provide the required transconductance for the desired noise. M2 and M3 have a substantially smaller width to reduce the capacitance at the high impedance node. The transconductance of M4 is not important, but it should have a large output resistance (see eqn 6.96). To avoid additional capacitive loading by the subsequent stage, the output is fed through a source follower M5. The high input impedance of the source follower also maintains the voltage gain independent of load. The source follower also provides current drive capability.

The detector is capacitively fed back through C_f, but DC feedback must also be provided, both to set the DC operating point and to discharge the feedback capacitor. This is achieved by a CMOS pair M7 and M8 in parallel with the feedback capacitor. When DC biased the transistors act as resistors to continuously discharge the feedback capacitor and provide DC feedback to maintain the DC level at the amplifier output. This can also be accomplished by only one transistor. The second transistor is important in the second operating mode, which uses pulsed feedback. Here the reset transistors are driven by a short pulse, which discharges the feedback capacitor and also charges the amplifier input to the proper potential. A transistor pair is used to cancel the charge injection through the gate–channel capacitance. The NMOS and PMOS transistors are driven with pulses of opposite polarity. The device geometries of M7 and M8 are small to reduce the gate–channel capacitance, but sized individually to match their gate capacitances and cancel the injected charge.

Transistor M9 shifts the output level to set the voltage at the juncture of M2 and M3 to approximately $V_{DD}/2$. When the feedback switch is closed this voltage is $V_{GS}(\mathrm{M1}) + V_{DS}(\mathrm{M9}) + V_{GS}(\mathrm{M5})$. Voltages V_{GN} and V_{GP} set the drain voltages of M1 and M4.

Another detail is the current mirror M6 that sets the operating current. Since both the noise and speed of the amplifier are set by the transconductance of the input device, and the transconductance is set primarily by current, controlling the current is crucial. The current mirror allows the cascode current to be set by an external current, which is not subject to device parameter variations on the chip. This is especially useful in circuits subject to radiation damage, which will be discussed in the next chapter. Setting the current may be as simple as connecting a resistor from the I_{CASC} port to ground. Since the voltage drop across the resistor is typically several times than the gate–source voltage of the current mirror, the dependence of I_{CASC} to transistor threshold voltage variations is reduced.

To first order the gain–bandwidth product f_0 and the feedback network set the upper cutoff frequency $f_u = f_0/(1 + C_d/C_f)$ and thus the rise time constant $\tau = 1/(2\pi f_u)$. However, as discussed briefly in Chapter 2 and treated in Appendix D, in a feedback amplifier the phase shift is also critical. At frequencies

well below the cutoff frequency the phase shift is zero (or in an inverting amplifier 180°). At the cutoff frequency the complex load incurs an additional 45° and at frequencies well above the cutoff frequencies the load is imaginary (*i.e.* purely capacitive) and the additional phase shift is constant at 90°. If these were the only phase shifts, the feedback loop would be unconditionally stable. However, additional time constants are introduced by the additional devices that affect the signal path.

This amplifier topology can provide a gain–bandwidth product of about 1 GHz with < 1 mW power dissipation. The dominant pole set is by the transconductance of M1 and the total capacitance at the cascode output. However, additional time constants are introduced at the juncture of M1 and M2, which sets the minimum width of M2, and in the source follower. In the design shown in Figure 6.38 the second pole is at several hundred MHz, so the reset path must include some attenuation (provided by M9) to prevent self-oscillation during reset, as the reset switch establishes unity gain.

This illustrates why even a "single stage" amplifier can be limited to closed loop bandwidths substantially smaller than the unity gain frequency. It also shows that if feedback is to be applied around a cascade of amplifiers, the cutoff frequencies of the individual amplifiers cannot be the same. The overall frequency response of the cascade must be dominated by one amplifier and the cutoff frequencies of the other stages must be well above the frequency where the loop gain is unity. Kipnis, Spieler, and Collins (1994) describe another variant of a cascode preamplifier using bipolar transistors showing the gain and phase response.

The frequency break points are usually called "poles", as feedback systems are often analyzed using Laplace transforms, where the time constants appear as poles in complex frequency space. An ideal amplifier with just one time constant has a "single pole response". Many commercially available operational amplifiers are made unconditionally stable for all gains by deliberately reducing the gain–bandwidth product at the first cutoff frequency such that all higher poles are well above the unity gain frequency. However, the freedom in design comes at the expense of bandwidth.

The loop gain in normal operation of the circuit in Figure 6.38 is

$$A_{vL} = A_{v0} \frac{C_f}{C_d + C_f} \ , \tag{6.98}$$

where C_d is the detector capacitance and A_{v0} is the open loop gain of the amplifier. In Figure 6.38 this is the voltage gain measured between the input and output in the absence of feedback. When the reset switch M7–M8 is closed, the impedance in the feedback path is much smaller, so the loop gain moves to an unstable regime. To compensate for this the output signal is connected to the reset switch through the voltage divider M19–M10. This scheme allows a higher bandwidth (faster response time) for signal acquisition than an amplifier designed for unconditional stability under all feedback conditions.

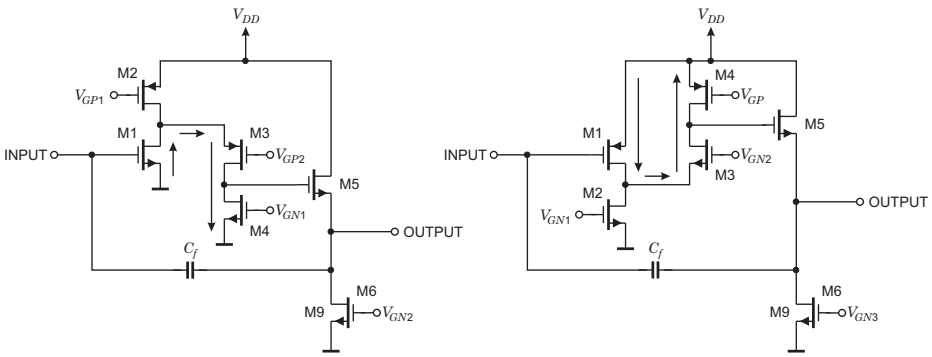

FIG. 6.39. The folded cascode implemented with an NMOS input (left) and a PMOS input transistor (right). The arrows show the signal current path of the input transistor M1. The bulk of M1's standing current is provided by M2.

A popular variant of the cascode amplifier is the "folded" cascode shown in Figure 6.39. This circuit avoids two conflicts inherent to the linear cascode. As the input transistor determines the equivalent noise, it must operate at a relatively high current. For a high output resistance, however, low current is advantageous. The folded cascode addresses these requirements by separating the DC and signal current paths. Since M3 presents a low input resistance, the output resistance of M2 is sufficiently high without an additional cascode transistor, so that practically all of M1's signal current flows into the low input resistance presented by M3. Since only a fraction of M1's standing current flows though the load transistor M4, the desired high load resistance can be achieved without a cascode load, further reducing the total voltage requirements. The folded cascode can be implemented with either NMOS or PMOS inputs. The latter provides lower $1/f$ noise, which in some applications outweighs the somewhat inferior transconductance of the PMOS input transistor. Some commonly used implementations of the folded cascode are quite prone to spurious pickup and cross-talk. This will be discussed in Chapter 9.

Although CMOS was used in these examples, similar circuits can be implemented with bipolar transistors or combinations of bipolar and MOS devices (using bi-CMOS integrated circuit processes, for example). Indeed, the basic considerations and circuit derivations apply to any transconductance device, be it a bipolar transistor, JFET, MOSFET or vacuum tube. Bipolar transistors provide higher transconductance and generally allow lower power dissipation in amplifiers, whereas CMOS provides added flexibility in the choice of device geometry. CMOS also provides higher circuit density and in mixed analog-digital systems the advantages of a full CMOS implementation on a single chip usually outweigh transconductance considerations.

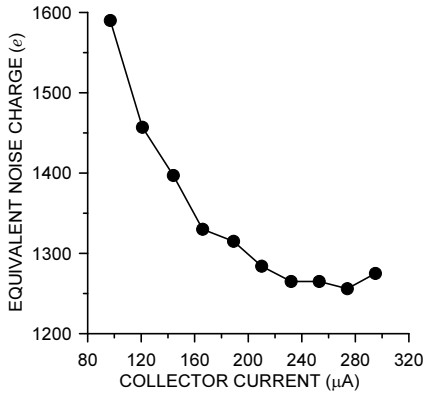

FIG. 6.40. Measured noise *vs.* collector current of a strip detector module using a bipolar transistor front-end and a 12 cm long strip sensor. The peaking time at the output of the pulse shaper is about 25 ns.

This discussion was not meant to be an exhaustive treatment of amplifier design, but is intended to provide a glimpse of the rationale behind some common circuits. Many other circuit techniques and more detailed analyses are needed to implement a full system and a vast body of literature is available to delve more deeply into the subject.

6.5 Overall noise of a detector module

Figure 6.40 shows the measured noise of a complete detector module *vs.* collector current of the bipolar transistor input stage. The noise decreases with collector current and reaches a minimum at about 250 μA, which was the design value. The circuit was not designed to operate at higher currents than 300 μA, so the noise could not be measured beyond the minimum. The simulated noise levels are 1460 e at 150 μA and 1230 e at 300 μA, in good agreement with the measurements.

Although the first amplifying device should dominate the system's electronic noise, many other contributions add up to increase the overall noise. The individual contributions may be small, but there are many, so that in practice the overall noise level is 10 – 20% higher than for the input device alone. Table 6.1 shows a representative breakdown for a strip detector system with a bipolar transistor front-end.

The input transistor contributes only 50 – 60% of the total noise, with the collector and base current contributing about 80% of that and the base resistance the remainder. Substantial contributions come from the neighbor amplifiers and strips, as was discussed in Chapter 3. The thermal noise of the strip resistance can be reduced by connecting the amplifier at the midpoints of the strips, rather than at the ends. Contributions from other components in the preamplifier circuit

(the load transistor with its emitter resistor and the cascode transistor) add small but noticeable contributions.

6.6 Optimization for low power

Optimizing the readout electronics in large vertex or tracking detector systems is not optimization of one characteristic such as noise alone, but finding an optimum compromise between noise, speed, and power consumption.

For both BJTs and FETs, the minimum obtainable noise under optimum scaling increases with the square root of detector capacitance, although the physical origins for this behavior are quite different in the two types of devices. However, in low-power systems the minimum obtainable noise values obtained from the equations for both FETs and BJTs should be viewed as limits, not necessarily as desirable goals, since they are less efficient than other operating points.

First, consider two input transistors, which provide the same overall noise with a given detector, but differ in input capacitance. Since the sum of detector

Table 6.1 Percent contributions to the total noise power in a detector module, simulated for 12 cm long strips and a bipolar transistor front-end (Kipnis 1996). The strip resistance is 15 Ω/cm and the post-irradiation fluence is 10^{14} p/cm. The analysis was performed for amplifiers connected either at the ends of the strips or at the midpoint of the strip electrodes to reduce the effect of the strip resistance. The individual contributions listed account for about 95% of the total noise.

Noise source	Center-tap pre-irradiation	End-tap pre-irradiation	Center-tap post-irradiation	End-tap post-irradiation
Total noise (e)	1370	1510	1475	1555
Input transistor	61.8	48.7	60.5	53.1
Neighbor amplifier	13.5	9.8	6.7	5.4
Strip resistance (center strip)	7.0	21.8	3.7	12.7
Strip resistance (neighbor)	2.3	7.1	1.1	3.8
Feedback resistor	6.2	5.3	6.2	5.6
Leakage current (center strip)	–	–	12.2	11.0
Load transistor	1.1	1.0	1.2	1.1
Emitter resistor of load transistor	2.5	1.9	1.7	1.4
Cascode transistor	–	–	1.5	1.3

and input capacitance determines the voltage noise contribution, the transistor with the higher input capacitance must have a lower equivalent noise voltage e_n, *i.e.* operate at higher current. In general, low capacitance input transistors are preferable, and systems where the total capacitance at the input is dominated by the detector capacitance are more efficient than systems that are capacitively matched. Capacitive matching should be viewed as a limit, not as a virtue.

6.6.1 *Optimum operating current*

Both the equivalent input noise voltage

$$e_n^2 \approx \frac{4kT}{g_m}$$

and the gain–bandwidth product

$$f_0 = \frac{g_m}{2\pi C_o}$$

depend on the transconductance g_m of the input transistor. The capacitance C_o at the node where voltage gain obtains invariably limits the obtainable circuit rise time, rather than the inherent speed of the transistor. From this we see that increasing transconductance improves both noise and speed. The transconductance depends primarily on device current. In a bipolar transistor

$$g_m = \frac{I_C}{kT/e} \ . \tag{6.99}$$

In a MOSFET in strong inversion

$$g_m = \sqrt{\frac{W}{L} \cdot \frac{\varepsilon_{ox}}{d_{ox}} \mu \cdot I_D} \ , \tag{6.100}$$

so for a given device width W, reducing the channel length L or gate oxide thickness d_{ox} should increase the transconductance. The choice of bulk material determines the carrier mobility μ and the gate oxide's dielectric constant ε_{ox}. However, this simple scaling rule only applies in strong inversion, whereas MOSFETs in large detector arrays are best operated in weak or moderate inversion. In weak inversion, the dependence of transconductance on current is the same as for a bipolar transistor, so it depends only on current and not on device geometry. The moderate inversion regime is the transition from weak inversion (low current) to strong inversion (high current) and its dependence is more complicated.

Since transconductance sets both the noise and speed, power efficiency improves when we increase the ratio of transconductance to drain current g_m/I_D. In a bipolar transistor the transconductance is proportional to collector current (eqn 6.99), so g_m/I_C is constant

$$\frac{g_m}{I_C} = \frac{e}{kT} \tag{6.101}$$

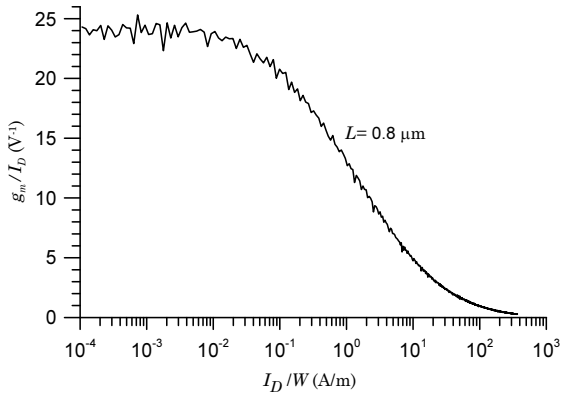

FIG. 6.41. Normalized transconductance g_m/I_D *vs.* normalized drain current I_D/W measured on an NMOS transistor with $0.8\,\mu$m channel length. The transconductance is determined by differencing the raw measured I_D *vs.* V_{GS} data, so the irregularities in the curves are due to the differential nonlinearity of the digitizer in the measurement system.

and equal to $(26\,\mathrm{mV})^{-1} \approx 40\,\mathrm{V}^{-1}$. In an FET the dependence of transconductance on drain current is more complicated. Increasing the device width W at constant current density is equivalent to connecting multiple devices operating at the same current in parallel, so to yield a universal curve Figure 6.41 shows the normalized transconductance g_m/I_D *vs.* normalized drain current I_D/W. This curve applies to all transistors using the same technology and channel length.

At low currents the MOSFET starts out with constant g_m/I_D. This is the weak inversion regime. Theoretically, $g_m/I_D = e/kT = 38.4\,\mathrm{V}^{-1}$, but in this device it is $26\,\mathrm{V}^{-1}$. This is typical and is due to charge states at the Si–SiO_2 interface. One then sees a rapid decrease in the regime $0.1 < I_D/W < 10\,\mathrm{A/m}$ (moderate inversion) and finally a gradual decrease at $I_D/W > 10\,\mathrm{A/m}$ (strong inversion). Note that although g_m/I_D is decreasing with current, the transconductance itself is increasing, as was shown in Figure 6.24.

The strong inversion regime is most commonly used, especially when minimum noise is required, since it yields the highest transconductance. Note, however, that the abscissa of Figure 6.41 is logarithmic, and that the high transconductance in strong inversion comes at the expense of substantial current. Furthermore, increasing V_{GS} to increase current also increases the output saturation voltage, so the required drain–source voltage also increases. In systems where both speed and noise must be obtained at low power, for example HEP tracking detectors, the moderate inversion regime is advantageous, as it still provides 20 to 50% of the transconductance at roughly 1/10 the power.

Reducing the channel length improves power efficiency, as shown in Figure 6.42. These data were measured on devices with channel lengths ranging from 0.8

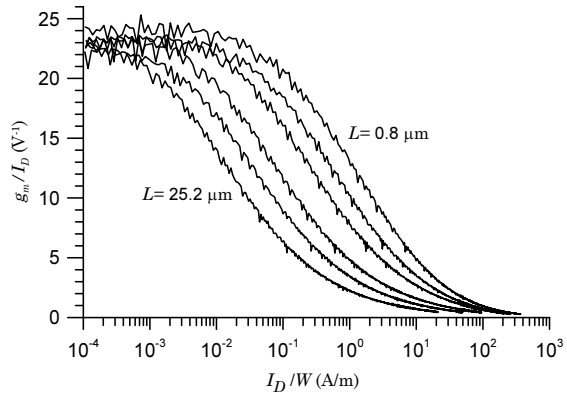

FIG. 6.42. Normalized transconductance g_m/I_D *vs.* normalized drain current I_D/W for channel lengths of 0.8, 1.2, 2.0, 5.2, 10.0, and 25.2 μm. All devices were fabricated on the same die in a 0.8 μm CMOS process. The irregularities in the curves are due to the differential nonlinearity of the digitizer in the measurement system.

to 25.2 μm, all on the same chip and fabricated in a 0.8 μm process. For example, the 0.8 μm channel length device shows $g_m/I_D = 24$ at $I_D/W = 10^{-3}$ and $g_m/I_D \approx 1$ at $I_D/W = 100$. The transition from weak to strong inversion shifts to higher currents as the channel length is reduced. At $I_D/W = 0.1$ the 0.8 μm long device yields $g_m/I_D = 21$, whereas 25 μm long devices yield $g_m/I_D = 6$. Thus, reducing the channel length allows more efficient circuitry, although not as predicted by the strong inversion formula.

The best power efficiency obtains at the highest normalized transconductance g_m/I_D that will provide the desired noise level. Uniquely associated with this value of g_m/I_D is a current density $(I_D/W)_{g_m/I_D}$, which for a given technology depends on the channel length. While keeping the current density constant, one can adjust the width to change the transconductance. As the width is changed the drain current $I_D = W \cdot (I_D/W)_{g_m/I_D}$ changes proportionally. This value of drain current sets the transconductance $g_m(W) = W \cdot (I_D/W)_{g_m/I_D}(g_m/I_D)_{selected}$. Thus, both the drain current and the transconductance scale proportionally to width, as does the FET's input capacitance. As the width is increased the equivalent noise charge decreases until the input capacitance equals the sensor capacitance. With further increases in width the increase in capacitance outweighs the decrease in noise voltage, so the noise charge increases. If the minimum noise is too high, one chooses a lower value of g_m/I_D, which will achieve a given transconductance at a smaller device width, so capacitive matching will occur at a higher transconductance. Thus the minimum noise will be lower, albeit at the expense of power dissipation. This procedure is illustrated in Figure 6.43.

For example, assume that the desired noise level is 1000 e. A normalized transconductance $g_m/I_D = 24$ (weak inversion) allows a minimum noise of 1400 e

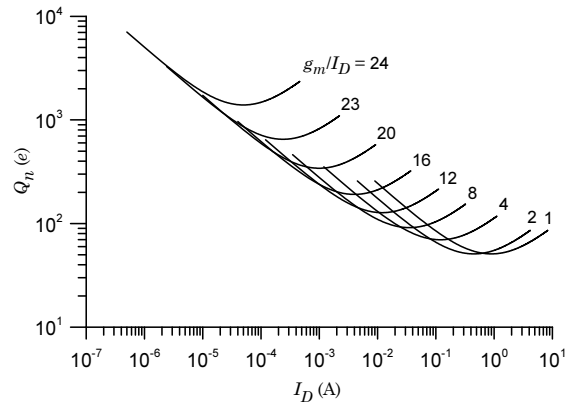

FIG. 6.43. Equivalent noise charge *vs.* drain current for various values of normalized transconductance g_m/I_D (corresponding to unique current densities I_D/W). The calculation assumes a detector capacitance of 10 pF and a transistor input capacitance of 1 fF per μm width. In the low-current regime the asymptote for all curves follows the relationship $I_D \propto 1/Q_n^2$.

at a drain current of 50 μA. Increasing the current density so that $g_m/I_D = 20$ shifts the operating mode towards moderate inversion and yields a minimum noise of 340 e at a drain current of 1 mA. However, following the $g_m/I_D = 20$ curve to smaller drain currents (device widths) provides the desired 1000 e noise level at a drain current of 30 μA, less than the 50 μA needed for 1400 e noise at $g_m/I_D = 24$. Going to much smaller values of g_m/I_D yields the desired 1000 e noise at higher currents. The choice of g_m/I_D is not very critical; $g_m/I_D = 22$ or 23 gives practically the same result.

This illustrates that capacitive matching is not a good criterion for systems where low power is important. Near capacitive matching the device width (and hence the current) can be reduced significantly without a substantial increase in noise. For example, at $g_m/I_D = 24$ allowing a 10% increase in noise reduces the device current to 40% of the current at capacitive matching. For currents well below the noise minimum all curves follow the relationship $I_D \propto 1/Q_n^2$, so for constant supply voltage the required power increases with the inverse square of the required noise charge, which depends on the signal magnitude provided by the sensor.

When scaling the device width at constant current density, the equivalent input noise voltage

$$e_n^2 \propto \frac{1}{g_m} \propto \frac{1}{I_D} \ . \tag{6.102}$$

Since the equivalent noise charge

$$Q_n^2 \propto e_n^2 C^2 \propto \frac{C^2}{I_D} \ , \tag{6.103}$$

when operating well below capacitive matching, the required power for a given noise level increases with the square of sensor capacitance,

$$P_D \propto I_D \propto C^2 \ . \tag{6.104}$$

A similar result obtains for bipolar transistors. The most efficient operating regime with respect to power is at a current well below the current needed for minimum noise

$$I_C = \frac{kT}{e} C_{tot} \sqrt{\beta_{DC}} \sqrt{\frac{F_v}{F_i}} \frac{1}{T_S} \ . \tag{6.105}$$

In this regime the noise is dominated by voltage noise, so

$$Q_n^2 \approx \frac{2(kT)^2}{eI_C} C^2 \frac{F_v}{T_S} \tag{6.106}$$

and for a given noise level C^2/I_C is constant, so the required power

$$P \propto I_C \propto C^2 \ . \tag{6.107}$$

In this form of optimum scaling the required power in the input device scales with the square of capacitance at the input. The required power also increases with the square of the desired signal-to-noise ratio. Even when the required noise level is close to the minimum noise, one can operate well below the optimum current, as the noise minimum is rather shallow.

These scaling rules represent optimum scaling, where the device input capacitance is negligible, *i.e.* well below capacitive matching. For comparison, from Figure 6.43 a 2000-fold increase in current reduces the minimum noise at capacitive matching from 1400 to $50\,e$, whereas optimum scaling requires that the current increase only 800-fold. The difference is due to the penalty incurred by the device capacitance.

As illustrated above, the optimization is iterative and usually involves several components, but basically one starts with a circuit topology and chooses a normalized drain current I_D/W that provides an adequate gain–bandwidth product. For example, in Figure 6.38 this involves setting the ratio of device widths in the cascode. If the capacitive load at the output node is dominated by device widths, then scaling all widths while scaling the current simultaneously to maintain the current density I_D/W, will maintain the amplifier's gain–bandwidth product g_m/C_o.

6.6.2 *Technology improvements*

To what extent do improvements in device technology improve amplifier performance? Can we simply rely on Moore's Law to meet future needs? There is

a widespread tendency to expect the miracles of modern technology to make possible what is impossible today.

The preceding section underscores the importance of transconductance and its relationship to power dissipation. Low electronic noise levels require sufficient transconductance coupled with acceptable input capacitance. In addition, large detector arrays require that these parameters obtain at low power. These requirements militate against many novel technologies that appear to simplify fabrication and reduce cost. Examples are amorphous silicon transistors or thin film transistors deposited by inkjet printing. All of these devices suffer from low mobilities and, hence, low transconductance. Nanotechnology offers the potential of very small devices, and the notion of "self-assembly" will appeal to anyone who has constructed a complex detector, but nanotransistors will require nanosensors to reduce capacitance to match the small transconductance. Nanometer thin sensors in trackers yield correspondingly small signals, which require lower noise and drive up front-end power. These novel devices will make inroads as switching devices, but their applicability to low-noise analog circuits is dubious. For the applications considered here crystalline devices appear to offer the most realistic prospects for technological improvements. The basic scaling rules discussed above still apply, but we seek improvements in the normalized transconductance g_m/I.

The transconductance of MOSFETs in weak inversion depends on current alone, so the parameter that can be improved is input capacitance. If smaller feature sizes push the weak inversion regime to higher current densities, a narrower device will provide the required transconductance at a lower input capacitance. However, a thinner gate oxide increases the capacitance per unit area, so the reduction in width must be balanced against the normalized input capacitance $C_i/W \propto L/d_{ox}$.

For MOSFETs in strong inversion

$$\frac{g_m}{\sqrt{I_D}} = \sqrt{\frac{W}{L} \cdot \frac{\varepsilon_{ox}}{d_{ox}} \mu} \ , \tag{6.108}$$

so high-density processes with shorter channel lengths L and thinner gate oxides will provide higher transconductance at a given current. Furthermore, the moderate inversion regime occurs at more favorable currents, as shown in Figure 6.42. However, it is not clear that moving to yet smaller channel lengths than shown in Figure 6.42 will provide similar benefits in the moderate inversion regime that is most interesting for low-noise amplifiers. Figure 6.44 compares measured data for NMOS devices with $L = 0.8$ and $0.3\,\mu$m channel lengths, fabricated in 0.8 and $0.25\,\mu$m CMOS processes, respectively. In the higher density process the transition from weak to moderate inversion occurs in the same current range as in the $0.8\,\mu$m MOSFET and the normalized transconductance in weak inversion is distinctly lower.

Why does the $0.3\,\mu$m channel length not show the expected improvement? Scaling to smaller feature size involves more than lateral scaling, *i.e.* resolution

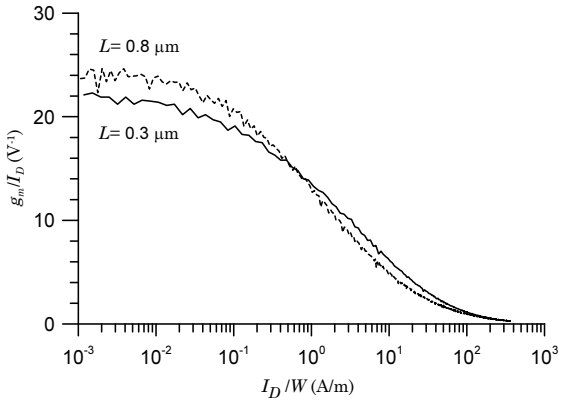

FIG. 6.44. Comparison of normalized transconductance g_m/I_D *vs.* normalized drain current I_D/W for NMOS devices with $L = 0.8$ and $0.3\,\mu$m channel lengths, fabricated in 0.8 and $0.25\,\mu$m CMOS processes, respectively.

in lithography. The vertical dimensions, *i.e.* the depth of the source and drain implants must also be reduced to avoid spreading the channel into the bulk, which reduces transconductance. The gate oxide must also be thinned. All this reduces the maximum operating voltage. In digital circuitry this implies smaller logic swings, so threshold control and noise immunity are concerns. In analog circuitry the dynamic range is reduced, as the maximum signal level is reduced while the electronic noise levels remain essentially the same. In some fabrication processes this is addressed by providing two choices of oxide thickness to allow "low-voltage" and "high-voltage" devices. Clearly, this comes at the expense of process complexity.

In modern $0.3\,\mu$m devices the output characteristics are still quite good, as shown in Figure 6.45. This is unlike older implementations where short channel devices had low output resistances reminiscent of triode vacuum tubes. The benefits of velocity saturation are apparent, as small drain voltages across the short channel length produce high longitudinal fields.

Silicon integrated circuit technology is very flexible and still offers many new possibilities. Reduced feature size does provide substantial benefits in digital circuitry, both in circuit density and power. For analog applications the optimization is more complex and requires concurrent process improvements beyond mere feature size. Although the limits of scaling to ever smaller devices are a genuine concern, it is not clear what the limits will be.

Bipolar transistors provide the highest transconductance per unit current, in practice outperforming MOSFETs even in the weak inversion regime. In bipolar transistors this ratio is set by basic physics, so it is unaffected by improved process technology, increased device speed, or the use of heterojunction devices, *e.g.* SiGe devices. Furthermore, bipolar transistors tend to have substantially

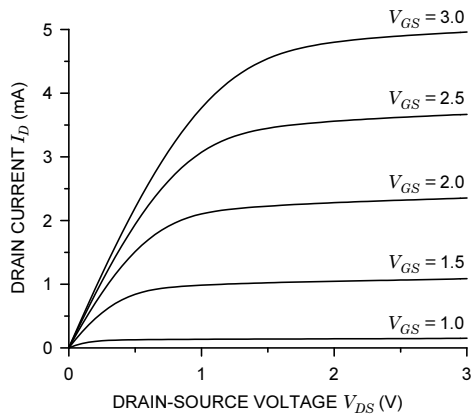

FIG. 6.45. Output curves of an NMOS transistor with $W = 9.45\,\mu$m and $L = 0.3\,\mu$m.

lower input capacitance for comparable noise levels, further reducing power requirements. Although transconductance per unit current in bipolar transistors is independent of technology, high-density processes tend to improve contamination control, which improves the DC gain at low currents and thus extends the usable operating range to lower currents. Furthermore, faster devices tend to reduce parasitic base and emitter resistances. In the past, bipolar processes have suffered from low circuit density, but the cellular telephone market has promoted mixed technology BiCMOS processes, which combine high frequency SiGe bipolar transistors with high-density CMOS. This is a very attractive option.

6.7 Power dissipation of an active pixel array *vs.* strip readout

To close this chapter we'll apply some of the principles developed above to a simple feasibility test. Early in the development of custom integrated circuits for vertex detectors a commonly raised objection to random-access pixel arrays was that the power dissipation would be prohibitive. If a strip readout required 2 mW per strip on an 80 μm pitch, *i.e.* 250 mW per cm width, could it be practical to read out 15 000 pixels per cm^2? We can use the scaling rules developed above to obtain a rough estimate of the power dissipation in the front-end.

Assume that a strip detector covers a certain area with n strips and a pixel detector covers the same area with $n \times n$ pixels. The strip width and pixel size are small compared to the sensor thickness, so the capacitance is dominated by the strip-to-strip or pixel-to-pixel fringing capacitance. Then the capacitance is proportional to the periphery (pitch p and length l). Thus, the strip capacitance is proportional to $2(p + l) \approx 2l$ and the pixel capacitance proportional to $4p$. Thus,

$$C_{pixel} \approx \frac{2}{n} C_{strip} \ . \tag{6.109}$$

As discussed above, under optimal scaling the power dissipation of the readout amplifier for a given noise level is proportional to the square of capacitance $P \propto C^2$, so the power per pixel

$$P_{pixel} \approx \frac{4}{n^2} P_{strip} \ . \tag{6.110}$$

Since there are n times as many pixels as strips, the power in all pixels' front-ends

$$P_{pixel,tot} \approx \frac{4}{n} P_{strip} \ . \tag{6.111}$$

From this we see that increasing the number of readout channels can reduce the total power dissipation.

The circuitry per cell does not consist of the amplifier alone, so a fixed power P_0 per cell must be added, bringing up the total power by $n^2 P_0$, so these savings are only realized in special cases. Nevertheless, this estimate shows that power dissipation is not a fundamental hurdle in the implementation of highly segmented arrays. In practice, random addressable pixel arrays for high-luminosity colliders have been implemented with overall power densities (total power/detector area) comparable to strip detector systems.

References

Bardeen, J. and Brattain, W.H. (1948). The transistor, a semiconductor triode. *Phys. Rev.* **74** (1948) 230–231

Baker, R.J., Li, H.W. and Boyce, D.E. (1998). *CMOS Circuit Design, Layout, and Simulation.* IEEE Press, New York. ISBN 0-7803-3416-7, TK7871.99. M44B35 1997

Cooke, H.F. (1971). Microwave transistors: theory and design. *Proc. IEEE* **59/8** (1971) 1163–1181

Grove, A.S. (1967). *Physics and Technology of Semiconductor Devices.* Wiley, New York. ISBN 0-471-32998-3

Johnson, H. (1966). Noise in field effect transistors, Chapter 6 in *Field Effect Transistors.* T. Wallmark and H. Johnson (editors), Prentice Hall, Englewood Cliffs

Kipnis, I., Spieler, H., and Collins, T. (1994). An analog front-end bipolar-transistor integrated circuit for the SDC silicon tracker. *IEEE Trans. Nucl. Sci.* **NS-41/4** (1994) 1095–1103

Kipnis, I. (1996). *Noise Analysis due to Strip Resistance in the ATLAS SCT Silicon Strip Module.* LBNL-3907, available at http://www-atlas.lbl.gov/strips/doc/reports.html

Nicollian, E.H. and Brews, J.R. (1982). *MOS (Metal Oxide Semiconductor) Physics and Technology.* Wiley, New York. ISBN 0-471-08500-6, TK7871.99. M44N52

Sevin, L.J. (1965). *Field-Effect Transistors.* McGraw-Hill, New York. TK7872. T73 S45

Shockley, W. (1949). The theory of $p-n$ junctions in semiconductors and $p-n$ junction transistors, *Bell System Tech. Journal* **28** (1949) 435–489

Shockley, W. (1950). *Electrons and Holes in Semiconductors*. van Nostrand, Princeton

Shockley, W. (1952). A unipolar field-effect transistor, *Proc. IRE* **40** (1952) 1365–1376

Shoji, M. (1966). Analysis of high-frequency thermal noise of enhancement mode M.O.S. field effect transistors, IEEE Trans. Electron Devices **ED-13/6** (1966) 520–524

Sze, S.M. (1981). *Physics of Semiconductor Devices* (2nd edn). Wiley, New York. ISBN 0-471-05661-8, TK7871.85.S988 1981

Sze, S.M. (2002). *Semiconductor Devices – Physics and Technology*. Wiley, New York. ISBN 0-471-33372-7, TK7871.85.S9883 2001

Takeda, E., Yang, C.Y. and Miura-Hamada, A. (1995). *Hot-Carrier Effects in MOS Devices*. Academic Press, San Diego. ISBN 0-12-682240-9, TK7871.99. M44T35 1995

Taur, Y. and Ning, T.H. (1998). *Fundamentals of Modern VLSI Devices*. Cambridge University Press, Cambridge. ISBN 0-521-55056-6, TK7871.99.M44T38 1998

Tsividis, Y.P. (1987). *Operation and Modeling of the MOS Transistor*. McGraw-Hill, New York. ISBN 0-07-065381-X, TK7871.99.M44T77 1987

Wolf, S. (1995). *Silicon Processing for the VLSI Era, Volume 3 – The Submicron MOSFET*. Lattice Press, Sunset Beach, ISBN 0-961672-5-3

Wolf, S. (2002). *Silicon Processing for the VLSI Era, Volume 4 – Deep-Submicron Process Technology*. Lattice Press, Sunset Beach, ISBN 0-9616721-7-X

van der Ziel, A. (1963). Gate noise in field effect transistors at moderately high frequencies. *Proc. IRE* **51** (1963) 461–467

7

RADIATION EFFECTS

Radiation-resistant electronics have been integral to the aerospace, nuclear reactor, and weapons communities for many years, but only rather recently have they become important for particle accelerators and accelerator-based experiments. The SSC made the design of radiation-resistant detectors and electronic readout systems a key design consideration for high-energy physics experimentalists. The energy frontier has now shifted to the LHC, which requires even higher luminosities to achieve its physics goals. Even at existing machines, for example the Tevatron at FNAL and the B factories at KEK and SLAC, radiation-hard electronics are required in the innermost vertex detector and tracking systems. On the accelerator side, higher beam currents and the increased sophistication of monitoring and diagnostic systems are bringing the need for radiation-resistant electronics to the forefront of designers' concerns.

Although one can argue that vacuum tubes are extremely radiation-hard, the complexity of today's electronics systems restricts our focus to semiconductor devices. For all practical purposes this leaves us with silicon and gallium-arsenide devices. For a variety of reasons silicon transistors and integrated circuits comprise the bulk of radiation-hard electronics. Although initial results on GaAs detectors after neutron irradiation appeared promising, charged particles produced inferior results. In designing SSC and LHC detectors, no compelling justification was found for GaAs electronics in any radiation-sensitive application. GaAs was often cited as "known to be radiation-hard", because of the radiation resistance of GaAs FETs. Indeed, they are superior to Si MOSFETs, but Si JFETs offer comparable radiation resistance, as discussed below in Section 7.3.2. Indeed, in most areas silicon technology provides critical performance advantages. For these reasons, despite the fascinating physics of compound semiconductors, this chapter will emphasize silicon technology.

For many years an active scientific community has studied radiation effects in semiconductor electronics, producing a wealth of data and a detailed understanding of many phenomena. Although access to some of these results and techniques is restricted, most of the data and papers are in the public domain and readily accessible. Much has been published on basic damage mechanisms and on device properties for specific applications. However, when attempting to apply this information to an area outside the traditional purview of the radiation effects community, key pieces of information needed to link basic damage mechanisms to usable design guidelines were often missing. This was very clear in the development of detectors for the SSC and LHC, where both the application of detectors with deep depletion regions and novel circuit designs combining

low noise, high speed, and low power pushed developments into uncharted territory. Radiation damage studies on detectors had focussed on high resolution gamma-ray or charged particle spectroscopy in nuclear physics with very different requirements (Kraner 1982). Some early experiments indicated that silicon detectors could function in high energy applications at fluences of order $10^{14}\,\mathrm{cm}^{-2}$ (Borgeaud *et al.* 1983, Kondo *et al.* 1985, Weilhammer 1985), but it took over a decade, many measurements, and much detailed analysis to understand the phenomena and ultimately learn how to modify the material to extend its lifetime. Some data were not sufficiently appreciated, for example early reports of "type inversion" (Kuznetsov *et al.* 1975). Meanwhile, system functionality at fluences of order $10^{15}\,\mathrm{cm}^{-2}$ and ionization doses of 100 Mrad has been demonstrated and now work continues to extend limits by another order of magnitude.

This is a very complicated field and developing a general road map is not easy. Nevertheless, one can apply a few fundamental considerations to understanding the effects of radiation on various device types in specific circuit topologies and narrow the range of options that must be studied in detail. That is the thrust of this discussion. This cannot be an exhaustive treatment and the reader should consult the literature for more detailed coverage.

Holmes-Siedle and Adams (2002) provide a modern overview. The books by Messenger and Ash (1986), Van Lint *et al.* (1980) and Srour *et al.* (1984) are good references for general principles. Ma and Dressendorfer (1989) wrote what is still the definitive work on radiation effects in MOS devices. Oldham (1999) adds important updates. Most papers on radiation effects in semiconductor devices are presented at the IEEE Nuclear and Space Radiation Effects Conference and published in the annual conference issue (usually December) of the IEEE Transactions on Nuclear Science. Additional papers, primarily from the high energy physics community, are presented at the IEEE Nuclear Science Symposium and published in the conference issue of the IEEE Transactions on Nuclear Science. Other conferences on detector instrumentation tend to publish their proceedings in Nuclear Instruments and Methods.

7.1 Radiation damage mechanisms

First the basic phenomena will be outlined, before entering into a more detailed discussion. Semiconductor devices are affected by two basic radiation damage mechanisms:

- Displacement damage: Incident radiation displaces silicon atoms from their lattice sites. The resulting defects alter the electrical characteristics of the crystal.
- Ionization damage: Energy absorbed by ionization in insulating layers, usually SiO_2, liberates charge carriers, which diffuse or drift to other locations where they are trapped. This leads to unintended concentrations of charge and, as a consequence, parasitic fields.

Both mechanisms are important in detectors, transistors and integrated circuits. Some devices are more sensitive to ionization effects, some are dominated by displacement damage. Hardly a system is immune to either one phenomena and most are sensitive to both.

Ionization effects depend primarily on the absorbed energy, independent of the type of radiation. At typical incident energies ionization is the dominant absorption mechanism, so ionization damage is proportional to energy absorption per unit volume (dose), usually expressed in rad or gray ($1\,\mathrm{rad} = 100\,\mathrm{erg/g}$, $1\,\mathrm{Gy} = 1\,\mathrm{J/kg} = 100\,\mathrm{rad}$). Since the charge liberated by a given dose depends on the absorber material, the ionizing dose must be referred to a specific absorber, for example $1\,\mathrm{rad(Si)}$, $1\,\mathrm{rad(SiO_2)}$, $1\,\mathrm{rad(GaAs)}$, or in SI units $1\,\mathrm{Gy(Si)}$, *etc.*

Displacement damage depends on the nonionizing energy loss, *i.e.* energy and momentum transfer to lattice atoms, which depends on the mass and energy of the incident quanta. A simple measure as for ionizing radiation is not possible, so that displacement damage must be specified for a specific particle type and energy.

In general, radiation effects must be measured for both damage mechanisms, although one may choose to combine both, for example by using protons, if one has sufficient understanding to unravel the effects of the two mechanisms by electrical measurements. Even nonionizing particles can deposit some ionization dose via recoils, but this contribution tends to be very small: $2 \cdot 10^{-13}\,\mathrm{rad(Si)}$ per $1\,\mathrm{MeV\,neutron/cm^2}$, for example (Messenger and Ash 1986).

To set the scale, consider a tracking detector operating at the LHC with a luminosity of $10^{34}\,\mathrm{cm^{-2}s^{-1}}$. In the innermost volume of a tracker the particle flux from collisions is $n' \approx 2 \cdot 10^9/r_\perp^2\,\mathrm{cm^{-2}s^{-1}}$, increasing roughly twofold in the outer layers due to interactions and loopers. At $r_\perp = 30\,\mathrm{cm}$ the particle fluence after one year of operation ($10^7\,\mathrm{s}$) is about $2 \cdot 10^{13}\,\mathrm{cm^{-2}}$. A fluence of $3 \cdot 10^{13}\,\mathrm{cm^{-2}}$ of minimum ionizing particles corresponds to an ionization dose of $1\,\mathrm{Mrad}$, obtained after 1.5 years of operation. Albedo neutrons from a calorimeter could add a yearly fluence of $10^{12} - 10^{13}\,\mathrm{cm^{-2}}$.

7.1.1 *Displacement damage*

An incident particle or photon capable of imparting an energy of about $25\,\mathrm{eV}$ to a silicon atom can dislodge it from its lattice site. Displacement damage creates defect clusters. For example, a $1\,\mathrm{MeV}$ neutron transfers about $60 - 70\,\mathrm{keV}$ to the Si recoil atom, which in turn displaces roughly 1000 additional atoms in a region of about $0.1\,\mu\mathrm{m}$ size. Displacement damage is linked to nonionizing energy loss (NIEL, Burke 1986), which is not proportional to the total energy absorbed, but depends on the particle type and energy. Nonionizing energy loss for a variety of particles has been calculated over a large energy range. Figure 7.1 shows the displacement damage *vs.* energy for neutrons, protons, pions, and electrons, plotted relative to $1\,\mathrm{MeV}$ neutrons. Although not all data are verified quantitatively, these curves can be used to estimate relative effects.

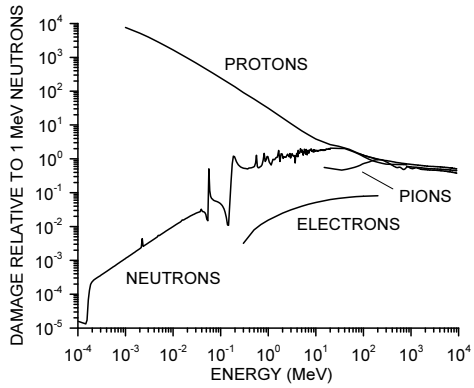

FIG. 7.1. Displacement damage *vs.* energy for neutrons, protons, pions, and electrons, plotted relative to 1 MeV neutrons. The data were compiled by Lindström (2000), based on Griffin *et al.* (1993), Konobeyev (1992), Huhtinen and Aarnio (1993), and Summers *et al.* (1993).

X-rays do not cause direct displacement damage, since momentum conservation sets a threshold energy of 250 keV for photons. ^{60}Co gamma-rays, commonly used in radiation testing, cause displacement damage primarily through Compton electrons and are about three orders of magnitude less damaging per photon than a 1 MeV neutron (Srour *et al.* 1979). Table 7.1 gives a rough comparison of displacement damage for several particles and energies.

Although nonionizing energy loss is an intuitively convenient measure, the details of defect formation also play a key role. For 1 MeV neutrons the initial vacancy distribution is highly clustered, whereas for 10 MeV protons the distribution is quite uniformly distributed and 24 GeV protons form a mixture of clustered and uniformly distributed damage sites (Huhtinen 2002). In GaAs this leads to significant differences between neutron and proton damage (Rogalla *et al.* 1997; Chilingarov, Meyer, and Sloan 1997a). Following the NIEL model initial irradiations of GaAs devices were performed exclusively with neutrons and supported the notion that GaAs is more radiation resistant than Si. However, with proton irradiation – more representative of a hadron collider environment –

Table 7.1 Rough comparison of displacement damage for several particles and energies.

Particle	proton	proton	neutron	electron	electron
Energy	1 GeV	50 MeV	1 MeV	1 MeV	1GeV
Relative damage	1	2	2	0.01	0.1

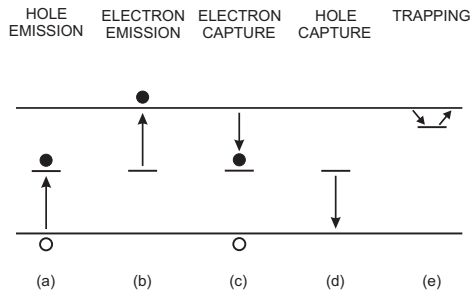

FIG. 7.2. Emission and capture processes through intermediate states. The arrows show the direction of electron transitions.

the contrary proved to be true. In a microscopic analysis Huhtinen (2002) argues that there is no obvious reason why NIEL scaling should be valid, so the prudent approach is to measure the effects of radiation with particles and energies representative of the actual radiation environment. Messenger *et al.* (2004) discuss limits to the application of the NIEL concept.

Displacement damage manifests itself in three important ways:

- Formation of mid-gap states, which facilitate the transition of electrons from the valence to the conduction band. In depletion regions this leads to a generation current, *i.e.* an increase in the current of reverse-biased *pn*-junctions. In forward biased junctions or nondepleted regions mid-gap states facilitate recombination, *i.e.* charge loss.
- States close to the band edges facilitate trapping, where charge is captured and released after some time.
- A change in doping characteristics (effective donor or acceptor density).

The role of mid-gap states is illustrated in Figure 7.2 (for a quantitative treatment see Appendix F). Because interband transitions in Si require momentum transfer ("indirect bandgap"), direct transitions between the conduction and valence bands are extremely improbable (unlike GaAs, for example). The introduction of intermediate states in the forbidden gap provides "stepping stones" for emission and capture processes. The individual steps, emission of holes or electrons and capture of electrons or holes, are illustrated in Figure 7.2. As shown in Figure 7.2 (a) the process of hole emission from a defect can also be viewed as promoting an electron from the valence band to the defect level. In a second step (b) this electron can proceed to the conduction band and contribute to current flow – generation current. Conversely, a defect state can capture an electron from the conduction band (c), which in turn can capture a hole (d). This "recombination" process reduces current flowing in the conduction band.

Since the transition probabilities are exponential functions of the energy differences, all processes that involve transitions between both bands require mid-gap states to proceed at an appreciable rate. Given a distribution of states these

processes will "seek out" the mid-gap states. Since the distribution of states is not necessarily symmetric, one cannot simply calculate recombination lifetimes from generation currents and vice versa (as is possible for a single mid-gap state, commonly assumed in textbooks).

Whether generation or recombination dominates depends on the relative concentration of carriers and empty defect states. In a depletion region the conduction band is underpopulated, so generation prevails. In a forward-biased junction carriers flood the conduction band, so recombination dominates. Figure 7.2 (e) also shows a third phenomenon; defect levels close to a band edge will capture charge and release it after some time, a process called "trapping".

In a radiation detector or photodiode system the increased reverse-bias current increases the electronic shot noise. The decrease in carrier lifetime due to trapping incurs a loss of signal as carriers recombine while traversing the depletion region. Defect states can act as donors, acceptors, or be electrically neutral. The predominant charge states formed in Si are acceptor-like, which in sufficient concentration affect the net space charge in the active region. The space charge determines the voltage required for full charge collection. The same phenomena occur in transistors, but are less pronounced, depending on device type and structure. Displacement damage effects will be discussed in more detail in Section 7.2.

7.1.2 *Ionization damage*

As in the detector bulk, electron–hole pairs are created in the oxide. The ionization energy $E_i = 18\,\mathrm{eV}$. The electrons are quite mobile and move to the most positive electrode. Holes move by a rather complex and slow hopping mechanism, which promotes the probability of trapping in the oxide volume with an associated fixed positive charge. Holes that make it to the oxide–silicon interface can be captured by interface traps, originating from the lattice mismatch at the oxide–silicon interface or impurities. This is illustrated in Figure 7.3, which shows a schematic cross-section of an n-channel MOSFET. As discussed in Chapter 6 a positive voltage applied to the gate electrode "inverts" the adjacent surface of the p-silicon bulk and forms a conductive channel between the n^+ doped source and drain electrodes. Holes freed in the oxide by radiation accumulate at the oxide–silicon interface. The positive charge build-up at the silicon interface requires that the gate voltage be adjusted to more negative values to maintain the negative charge in the channel.

Trapped oxide charge can also be mobile, so that the charge distribution generally depends on time, and more specifically, how the electric field in the oxide changes with time. The charge state of a trap depends on the local quasi-Fermi level (see Appendix E), so the concentration of trapped charge will vary with changes in the applied voltage and state-specific relaxation times. As charge states also anneal, ionization effects depend not only on the dose, but also on the dose rate. Figure 7.3 also shows a thick field oxide, which serves to control the silicon surface charge adjacent to the FET and prevent the formation of parasitic

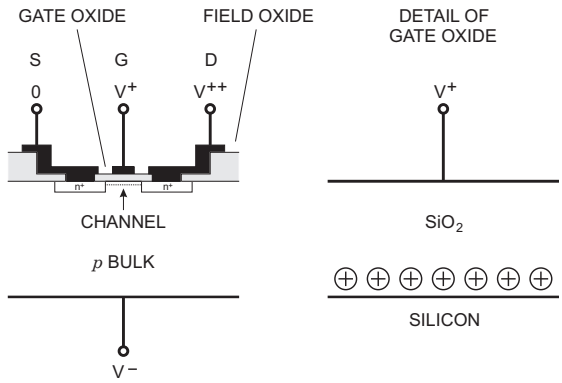

FIG. 7.3. Schematic cross-section of an n-channel MOSFET (left). A detail of the gate oxide shows the trapped holes at the oxide–silicon interface (right).

channels to adjacent devices. The same positive charge build-up as in the gate oxide also occurs here, indeed it can be exacerbated because the field oxide is quite thick. We'll return to these phenomena in Sections 7.3.1 and 7.3.3. For a detailed discussion, see the texts by Ma and Dressendorfer (1989) and Oldham (1999).

In summary, ionization effects are determined by

- interface trapped charge,
- oxide trapped charge,
- the mobility of trapped charge, and
- the time and voltage dependence of charge states.

Although the primary radiation damage depends only on the absorbed ionizing energy, the resulting effects of this dose depend on the rate of irradiation, the applied voltages and their time variation, the temperature, and the time variation of the radiation itself. Ionization damage manifests itself most clearly in MOS field effect transistors, so it will be discussed in more detail in that section.

7.2 Radiation damage in diodes

Diode structures are basic components of more complex devices, for example bipolar transistors, junction FETs and integrated circuits. Since the characteristics of diode depletion regions depend primarily on bulk properties, displacement damage is the key damage mechanism. Reverse-biased diodes with large depletion depths are used as radiation detectors and photodiodes. Because of their large depletion depths, typically hundreds of microns, detector diodes are very sensitive to bulk damage and extensive work by the high-energy physics community has produced many insights into bulk radiation effects. Affected are the detector leakage current, the doping characteristics, and charge collection.

A theoretical analysis from first principles is quite complex, due to the many phenomena involved. Take doping changes as an example. A key mechanism is that defects are mobile. Silicon interstitials are quite active and displace either P donors or B acceptors from substitutional sites and render them electrically inactive. These interstitial dopants together with oxygen, commonly present in the lattice as an impurity, react in very different ways with vacancies to form complexes with a variety of electronic characteristics (see Tsveybak *et al.* 1992, Huhtinen 2002 and references therein). Fortuitously, although a multitude of competing effects can be invoked to predict and interpret experimental results, the data can be described by rather simple parameterizations.

7.2.0.1 *Reverse bias current* The increase in reverse bias current (leakage current) is linked to the creation of mid-gap states, as discussed in Appendix F. Experimental data are consistent with a uniform distribution of active defects in the detector volume. The diode bias current after irradiation

$$I_d = I_0 + \alpha \cdot \Phi \cdot Ad \ , \tag{7.1}$$

where I_0 is the bias current before irradiation, α is a damage coefficient dependent on particle type and fluence, Φ is the particle fluence, and the product of detector area and thickness Ad is the detector volume. For 650 MeV protons $\alpha \approx 3 \cdot 10^{-17}$ A/cm (Barberis *et al.* 1993, Chilingarov *et al.* 1995) and for 1 MeV neutrons (characteristic of the albedo emanating from a calorimeter) $\alpha = 4 \cdot 10^{-17}$ A/cm (Moll 1999). The parameterization used in eqn. 7.1 is quite general, as it merely assumes a spatially uniform formation of electrically active defects in the detector volume, without depending on the details of energy levels or states. Unlike a high-quality diode, where reverse current saturates well below the depletion voltage, in a radiation-damaged diode the reverse bias current increases roughly with the square root of voltage, due to the uniform distribution of radiation damage sites.

After initial defect formation the leakage current decreases with time. Several processes appear to contribute to the annealing process, as a sum of exponentials with various time constants gives a good fit to the data (Wunstorf 1992, Chilingarov *et al.* 1995). The ROSE collaboration (Lindström 2001) adopted a prescription in which the leakage current is measured after 80 min. annealing at 60 °C. This procedure yields the same damage coefficient for a wide range of irradiations. The results apply to both n- and p-type material, whether float zone, Czochralski, or epitaxially grown, and over a wide range of resistivities.

The coefficients given above apply to operation at 20 °C. The reverse bias current is strongly dependent on temperature. Even after rather low fluences the generation current dominates (see Appendix F), so the reverse bias current

$$I_R(T) \propto T^2 e^{-E/2kT} \ . \tag{7.2}$$

For radiation damaged samples an activation energy $E = 1.2$ eV has provided a good fit (Barberis *et al.* 1993, Gill *et al.* 1992, Ohsugi *et al.* 1988), whereas

unirradiated samples usually exhibit $E = 1.12\,\mathrm{eV}$, the gap energy. The ratio of currents at two temperatures T_1 and T_2 is

$$\frac{I_R(T_2)}{I_R(T_1)} = \left(\frac{T_2}{T_1}\right)^2 \exp\left[-\frac{E}{2k}\left(\frac{T_1 - T_2}{T_1 T_2}\right)\right] \ . \tag{7.3}$$

In practice, the variation of leakage current with temperature is very reproducible from device to device, even after substantial doping changes due to radiation damage. The leakage current can be used for dosimetry and diodes are offered commercially specifically for this purpose.

7.2.0.2 *Doping characteristics* The effect of displacement damage on doping characteristics has been investigated extensively in the course of detector studies for the LHC and is still the subject of ongoing study. Measurements on a variety of strip detectors and photodiodes by groups in the U.S., Japan and Europe have shown that the effective doping of n-type silicon initially decreases, appears intrinsic (*i.e.* undoped), and then turns p-like ("type inversion"), with the effective doping (or more accurately, space charge) increasing with fluence.

One contribution is donor removal (Pitzl *et al.* 1992, Giubellino *et al.* 1992), but this by itself would only reduce the n-type nature of the material. "Type inversion" is consistent with the notion that acceptor sites are formed by the irradiation. Before irradiation the depletion region shows a space charge from the dopant host atoms, so in n-type material the space charge is positive. The applied bias voltage is required not only to deplete the detector, but also to collective mobile charge against the field set up by the space charge. Displacement damage forms acceptor-like states, which are populated by electrons from the bulk through thermal excitation, so they form a negative space charge. Initially, the positive space charge decreases as new acceptor states neutralize original donor states. At some fluence the two balance, so the space charge is zero, and beyond this fluence the acceptor states dominate yielding a net negative space charge. Note that "type inversion" is not associated with the creation of mobile holes, so although the material appears p-like, it is not the same as conventional p-doped material (Li 1994). As the change in space charge results from multiple sequential processes, its evolution is temperature dependent, as will be described below.

Operationally, the change in space charge appears as a change in doping level, so the net space charge is commonly referred to as an effective doping level N_{eff}. The detector functions as before and no change in bias polarity is needed, but to transport charge through the full detector thickness d the voltage must be raised proportionally to the increase in space charge

$$V = \frac{e}{2\varepsilon}\,|N_{eff}|\,d^2 \ . \tag{7.4}$$

In analogy to conventional diode operation this is often referred to as the "depletion voltage", although the device is devoid of mobile charge even at smaller

voltages. Since the radiation-induced space charge results from thermal excitation, it can be removed by cooling to cryogenic temperatures (Palmieri *et al.* 1998). This will be discussed later. Note that even after strong type inversion radiation damage is extremely dilute, *e.g.* a space charge concentration of $10^{15}\,\mathrm{cm}^{-3}$ compared to the concentration of Si lattice atoms of $5 \cdot 10^{22}$.

Although the basic phenomena were identified in the early 1990s (Pitzl *et al.* 1992, Lemeilleur *et al.* 1992, Ziock *et al.* 1994) it took roughly a decade to develop a comprehensive parameter set and microscopic interpretations of the phenomena (for summary reports see Lindström, Moll, and Fretwurst 1999 and Lindström 2001). A major breakthrough was achieved by the successful application of "defect engineering". Since the key aspect is not the formation of defects, but the process that makes them electrically active, various techniques to impede the formation of electrically active sites were considered early on. Oxygen captures vacancies and carbon captures interstitials, so both could divert defect evolution to a less deleterious path. Early tests yielded inconclusive results (Li *et al.* 1992). Meanwhile, oxygen has been shown to be very effective, but unfortunately it was not recognized initially because tests were conducted with neutrons, whereas its benefits are most pronounced under charged particle irradiation. Oxygen concentrations in the range $10^{17} - 10^{18}\,\mathrm{cm}^{-3}$ are necessary. In float-zone silicon this is obtained by prolonged growth of a thermal oxide on the silicon surface. Concurrently, oxygen diffuses into the silicon bulk and builds up the required concentration. Czochralski-grown silicon inherently has a high oxygen content, but it is not suitable for high-quality detectors because of a high concentration of dislocations and other defects. However, after substantial radiation damage this becomes less important and use of Czochralski-grown silicon in radiation detectors is being investigated. This is an attractive option, as Czochralski material is commonly used in integrated circuits (with a thin high-quality epitaxially grown surface layer that accommodates the circuitry). However, float zone material is also needed for high-voltage switching MOSFETs and silicon controlled rectifiers (SCRs), so a large commercial market is pushing float-zone material towards larger diameters and 15 cm diameter wafers are readily available.

7.2.1 *Contributions to N_{eff}*

Overall, four components contribute to the change in space charge: donor removal, build-up of stable charge, beneficial annealing, and anti-annealing. The first two depend only on fluence and show no temperature-dependent evolution, so they are called "stable damage". The others depend on defect dynamics and show a significant temperature dependence. The parameters adopted below are from Lindström (2001) and Moll (1999).

1. Stable damage

The change in space charge after exposure to a fluence Φ is

$$\Delta N_C = N_{C0}\left(1 - e^{-c\Phi}\right) + g_C\Phi \ . \tag{7.5}$$

The first term describes donor removal. $N_{C0} = \eta N_d$ is the concentration of removable donors, where $\eta \approx 0.7$ for standard float-zone silicon. In oxygen doped silicon $\eta = 0.45$ for neutron irradiation and 1.0 for protons. The parameter $c = 1 - 3 \cdot 10^{-13}\,\mathrm{cm}^2$. The second term describes acceptor formation, where the parameter $g_C = 1.5 \cdot 10^{-2}\,\mathrm{cm}^{-1}$ for neutrons and $g_C = 1.9 \cdot 10^{-2}\,\mathrm{cm}^{-1}$ for protons. In oxygen-doped material $g_C = 2.0 \cdot 10^{-2}\,\mathrm{cm}^{-1}$ for neutrons and $g_C = 5.3 \cdot 10^{-3}\,\mathrm{cm}^{-1}$ for protons.

2. Beneficial annealing

This term describes a recovery from the change in space charge,

$$\Delta N_a(t) = g_a \Phi e^{-t/\tau_a} \ . \tag{7.6}$$

The prefactor $g_a = 1.8 \cdot 10^{-2}\,\mathrm{cm}^{-1}$ and the dominant activation energy $E_a = (1.09 \pm 0.09)\,\mathrm{eV}$, so the time constant is given by

$$\frac{1}{\tau_a} = k_{0a} e^{-E_a/kT} \ . \tag{7.7}$$

In standard silicon the parameter $k_{0a} = 2.8 \cdot 10^{13}\,\mathrm{s}^{-1}$, so at a temperature of 20 °C the time constant $\tau_a \approx 55\,\mathrm{h}$. In oxygenated silicon $k_{0a} = 2.2 \cdot 10^{13}\,\mathrm{s}^{-1}$ and at a temperature of 20 °C the time constant $\tau_a \approx 70\,\mathrm{h}$.

3. Anti-annealing

The concentration and time evolution of acceptor-like sites

$$\Delta N_Y = g_Y \Phi \left(1 - \frac{1}{1 + t/\tau_Y}\right) \ , \tag{7.8}$$

where $g_Y = 5.2 \cdot 10^{-2}\,\mathrm{cm}^{-1}$ for neutrons and $g_Y = 6.6 \cdot 10^{-2}\,\mathrm{cm}^{-1}$ for protons in standard silicon. The time constant is given by

$$\frac{1}{\tau_Y} = k_{0Y} e^{-E_Y/kT} \ , \tag{7.9}$$

with an activation energy $E_Y = (1.31 \pm 0.04)\,\mathrm{eV}$ and a prefactor $k_{0Y} = 8.0 \cdot 10^{14}\,\mathrm{s}^{-1}$, so at a temperature of 20 °C the time constant $\tau_a \approx 480\,\mathrm{h}$. At short annealing times the time dependence is well described as a first order process, indicative that initially a first order process dominates and then other processes become effective.

In oxygenated silicon the parameters improve, especially for protons. Neutrons show $g_Y = 4.8 \cdot 10^{-2}\,\mathrm{cm}^{-1}$ and protons $g_Y = 2.3 \cdot 10^{-2}\,\mathrm{cm}^{-1}$. The time constant at 20 °C is 800 d for neutrons ($k_{0Y} = 4.8 \cdot 10^{14}\,\mathrm{s}^{-1}$) and 950 d for protons ($k_{0Y} = 4.04 \cdot 10^{14}\,\mathrm{s}^{-1}$).

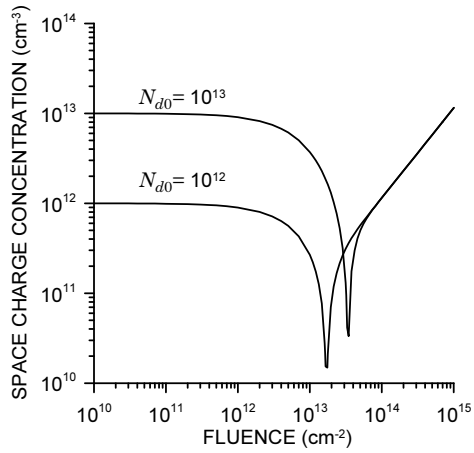

FIG. 7.4. Calculated space charge *vs.* high-energy proton fluence for silicon with initial donor concentrations N_{d0} of 10^{12} and $10^{13}\,\mathrm{cm}^{-3}$. With an infinitely fine calculation grid both distributions would dip to zero.

As a cumulative effect of these processes the change in space charge

$$\Delta N_{eff}(\Phi) = \Delta N_C(\Phi) + \Delta N_a(\Phi, T, t) + \Delta N_Y(\Phi, T, t) \tag{7.10}$$

and the effective doping concentration N_{eff} of n-type starting material

$$N_{eff} = N_d - \Delta N_{eff}\ , \tag{7.11}$$

where a positive or negative sign of N_{eff} denotes whether the effective doping is n- or p-like. Type inversion from n- to p-like silicon occurs at a fluence of about $10^{13}\,\mathrm{cm}^{-2}$, as is shown in Figure 7.4 for two levels of initial doping density. Figure 7.4 shows only the stable component, as would be obtained at low temperature.

Very high resistivity silicon ($\rho > 10\,\mathrm{k}\Omega\,\mathrm{cm}$ or $N_d < 4 \cdot 10^{11}\,\mathrm{cm}^{-3}$) is often highly compensated, $N_{eff} = N_d - N_a$ with $N_d \sim N_a \gg N_{eff}$, so that minute changes to either donors or acceptors can alter the net doping concentration significantly. Then the above equations must be modified accordingly. Moderate resistivity n-type material ($\rho \approx 1-5\,\mathrm{k}\Omega\,\mathrm{cm}$) used in large area tracking detectors is usually dominated by donors.

Figure 7.5 shows the evolution of the beneficial annealing and anti-annealing *vs.* time at a temperature $20\,^\circ\mathrm{C}$ and $-5\,^\circ\mathrm{C}$ after a fluence burst of $10^{14}\,\mathrm{cm}^{-2}$ protons. At room temperature beneficial annealing reduces the change in space charge at times $< 10^6\,\mathrm{s}$ ($\sim 12\,\mathrm{d}$). Then anti-annealing dominates, increasing N_{eff} to about $6 \cdot 10^{12}\,\mathrm{cm}^{-3}$ over the course of a year. If the starting donor concentration is $1 \cdot 10^{12}\,\mathrm{cm}^{-3}$ this requires a bias voltage of 360 V. Operating the detector at $-5\,^\circ\mathrm{C}$ delays anti-annealing.

In reality the increase in fluence and the annealing proceed concurrently, so Figure 7.5 doesn't apply directly, but it illustrates that operating the detector at low temperature and only allowing warm-up during annual maintenance periods is critical. ATLAS chose an operating temperature of $-7\,^{\circ}$C. The desire to reduce leakage current could drive the operating temperature even lower, but then the desirable effect of beneficial annealing is suppressed. Figure 7.6 shows predictions for the ATLAS pixel detector. The accelerated anti-annealing during the maintenance periods is clearly visible.

Operation can be extended by utilizing detector configurations that do not require that charge traverse the whole detector. As shown in Chapter 2 the induced signal current in highly segmented detectors peaks near the electrode. Since charge motion near the opposite electrode doesn't contribute much to the total integrated charge signal, a useful signal can still be obtained for carriers that only partially traverse the detector. In practice this is obtained by implementing the detector with n^+ electrodes in an n-substrate. After inversion this behaves like n^+ electrodes in a p-substrate. When operated at less than the "depletion" voltage electrons are still collected at the n-electrodes, but not from the full detector thickness, so the signal is reduced. The detector ceases to be usable when the signal-to-noise ratio becomes too small, but this is a "soft" failure mode (Unno *et al.* 1996, 1997), so the limits shown in Figure 7.6 can be extended. Small electrode areas reduce both the reverse bias current and the capacitance, so reduced noise extends the operation of pixel detectors with respect to strip devices. The ATLAS pixel system has noise levels $\sim 200\,e$ (Grosse-Knetter 2004), whereas the strip systems operate at $\sim 1500\,e$ (Turala 2001).

7.2.2 *Trapping*

As discussed in Appendix F and in Chapter 2, when trapping centers are present a change in carrier concentration from equilibrium decays with a lifetime

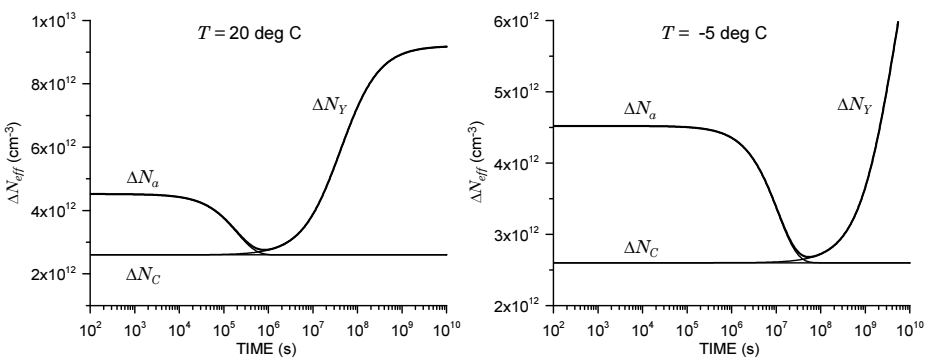

FIG. 7.5. Evolution of beneficial annealing ΔN_a and anti-annealing ΔN_Y *vs.* time at $20\,^{\circ}$C and $-5\,^{\circ}$C after a proton fluence burst of $10^{14}\,\mathrm{cm}^{-2}$.

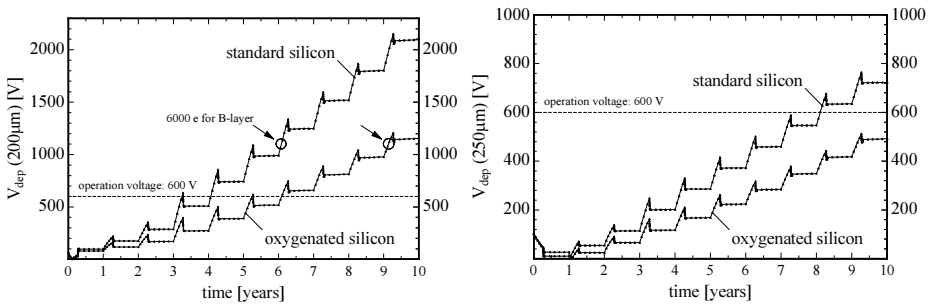

FIG. 7.6. Voltage required for full charge collection *vs.* time simulated for the ATLAS pixel detector. In the innermost layer (left) the total fluence over ten years is $3 \cdot 10^{15}$ and in the second layer (right) at $r = 10\,\mathrm{cm}$ it is $3 \cdot 10^{15}\,\mathrm{cm}^{-2}$, dominated by charged pions. The detectors are designed for a maximum operating voltage of 600 V. Oxygenated silicon extends the lifetime of the inner layer to 6 years and the 10 cm layer to > 10 years. Simulations by R. Wunstorf (RD48 1999).

$$\tau = \frac{1}{v_{th}\sigma N_t} \ . \tag{7.12}$$

The underlying assumption is that the thermal velocity v_{th} is large compared to the drift velocity, so that during charge collection the integrated path length of the superimposed random motion is proportional to the thermal velocity. In high quality diodes the lifetime is $10 - 30\,\mathrm{ms}$. Displacement damage adds traps of a different nature, so

$$\tau = \frac{1}{v_{th}\left(\sigma_0 N_{t0} + \sigma_r N_{tr}\right)} \ , \tag{7.13}$$

where σ_0 and N_{t0} characterize the prevailing traps in the original diode and σ_r and N_{tr} describe the traps formed during irradiation. Since the concentration of traps is proportional to fluence, $N_{tr} \propto \Phi$, the lifetime can be described by

$$\frac{1}{\tau} = \frac{1}{\tau_i} + \frac{\Phi}{K} \ , \tag{7.14}$$

where τ_i is the initial lifetime (Grove 1967, Messenger and Ash 1986). After very little fluence the damage term prevails, so

$$\tau \approx \frac{K}{\Phi} \ . \tag{7.15}$$

For a point deposition of charge, the net signal charge is proportional to $\exp(-t_c/\tau)$, as discussed in Chapter 2, so reducing the collection time t_c mitigates the effect of trapping. Since either the operating voltage is increased or

depletion widths are reduced at damage levels where charge trapping is appreciable, fields tend to be higher and collection times decrease automatically with radiation damage, provided the detector can sustain the higher fields. The mobilities of electrons and holes appear to be unaffected by heavy radiation damage (Brodbeck *et al.* 2002).

Kramberger *et al.* (2002) irradiated a diverse set of samples with initial resistivities ranging from 1 to 15 kΩ cm using both standard and oxygenated silicon with neutrons, pions, and protons to fluences up to $2 \cdot 10^{14}\,\mathrm{cm}^2$ and obtain for

$$\begin{aligned}
\text{neutrons}: &\quad K_e = (2.44 \pm 0.06) \cdot 10^6\,\mathrm{s/cm}^2 \qquad K_h = (1.67 \pm 0.06) \cdot 10^6\,\mathrm{s/cm}^2,\\
\text{pions}: &\quad K_e = (1.75 \pm 0.06) \cdot 10^6\,\mathrm{s/cm}^2 \qquad K_h = (1.30 \pm 0.03) \cdot 10^6\,\mathrm{s/cm}^2,\\
\text{protons}: &\quad K_e = (1.79 \pm 0.06) \cdot 10^6\,\mathrm{s/cm}^2 \qquad K_h = (1.30 \pm 0.03) \cdot 10^6\,\mathrm{s/cm}^2.
\end{aligned}$$

The errors are statistical and don't include a 10% uncertainty in the measured fluence. Krasel *et al.* (2004) irradiated oxygenated detectors with 24 GeV/c protons to neutron-equivalent fluences ranging from $2 \cdot 10^{13}$ to $9 \cdot 10^{14}\,\mathrm{cm}^{-2}$ and found

$$\begin{aligned}
K_e &= (1.95 \pm 0.06) \cdot 10^6\,\mathrm{s/cm}^2\\
K_h &= (1.98 \pm 0.07) \cdot 10^6\,\mathrm{s/cm}^2
\end{aligned}$$

for electrons and holes, respectively. In test beam measurements of four prototype ATLAS pixel modules exposed to a neutron-equivalent fluence of $1.1 \cdot 10^{15}\,\mathrm{cm}^{-2}$ Troncon *et al.* find an average lifetime of $(4.1 \pm 0.3 \pm 0.6)$ ns (Troncon 2004).

Note that although the trapping constants K_e and K_h are similar, since holes move about three times slower than electrons, their effective drift length is three times smaller than for electrons. Thus, detector structures that emphasize the electron signal provide a higher charge collection efficiency.

Assuming $K \approx 2 \cdot 10^6$, the carrier lifetime at a fluence of 10^{14} is 20 ns. In a 300 μm thick detector operated at 300 V the average field is 10^4 V/cm, so the drift velocity $v = 7 \cdot 10^6$ cm/s. The collection time is 4 ns, so little charge is lost to trapping. At a fluence of 10^{15}, however, the lifetime drops to about 2 ns. Operating a 250 μm thick detector at 600 V yields an average field of $2.4 \cdot 10^4$ V/cm. Because of nonconstant mobility the drift velocity only increases to $v = 9 \cdot 10^6$ cm/s, practically saturation velocity, so the estimated drift time is 3 ns and a substantial charge loss is expected. These are only estimates to illustrate the problem. Although the potential distribution appears to be as expected for a partially depleted detector with the maximum field at the n-contact dropping linearly into the bulk, an additional high field region appears near the p-contact (Beattie *et al.* 1998, Wunstorf 1992, Krasel *et al.* 2004, Castaldini *et al.* 2002, Eremin, Verbitskaya, and Li 2002). More detailed simulations indicate 50 – 60% charge collection efficiency for tracks traversing the detector after a fluence of $10^{15}\,\mathrm{cm}^{-2}$, falling off to about 25% after a fluence of $5 \cdot 10^{15}$ (Krasel *et al.* 2004). Test beam measurements of complete ATLAS pixel modules using n-side

readout of "n-on-n" pixels (see Section 7.5) in oxygenated silicon achieved nearly full charge collection efficiency in 250 μm thick sensors operated at 700 V after exposure to a neutron-equivalent fluence of $1.1 \cdot 10^{15}\,\mathrm{cm}^{-2}$ (Troncon 2004).

7.2.3 *Ionization effects*

The basic detector is insensitive to ionization effects. In the bulk, ionizing radiation creates electrons and holes that are swept from the sensitive volume; charge can flow freely through the external circuitry to restore equilibrium. A potential problem lies in the peripheral structures, the oxide layers that are essential to controlling leakage paths at the edge of the diode and to preserving interelectrode isolation in segmented detectors.

The positive space charge due to hole trapping in the oxide and at the interface (see Figure 7.3) attracts electrons in the silicon bulk to the interface. These accumulation layers can exhibit high local electron densities and form conducting channels, for example between the detector electrodes. This is especially critical at the "ohmic" electrodes in double-sided detectors, where the absence of pn-junctions makes operation reliant on full depletion of the silicon surface (see Appendix A). Even without radiation, the silicon surface tends to be n-type, so the ohmic side of n-type detectors is inherently more difficult to control (Barberis *et al.* 1994, Wheadon *et al.* 1994). Before irradiation the interstrip resistance tends to be very high, so small increases in surface charge have a big effect. However, to prevent signal leakage from one strip to another, the requirement is only that the interstrip impedance remain large with respect to the input impedance of the amplifier, which in strip detector systems is typically 100 – 1000 ohms.

Some detectors include integrated coupling capacitors and biasing networks. Biasing structures such as punch-through resistors and MOSFET structures are subject to ionization damage (Azzi *et al.* 1996). Although these devices can remain functional, substantial changes in voltage drop have been reported for punch-through and accumulation layer devices, whereas measurements on polysilicon resistors irradiated to 4 Mrad (65 MeV protons) show no effect (Kubota *et al.* 1991).

7.3 Radiation damage in transistors and integrated circuits

In principle, the same phenomena discussed for detectors also occur in transistors, except that the geometries of transistors are much smaller (depletion widths $< 1\,\mu$m) and the typical doping levels are higher ($> 10^{15}\,\mathrm{cm}^{-3}$).

7.3.1 *Bipolar transistors*

The most important damage mechanism in bipolar transistors is the degradation of DC gain at low currents. The damage mechanism is the same that causes increased leakage current in detectors, formation of mid-gap states by displacement damage. The difference is that the base–emitter junction is forward biased, so the high carrier concentration in the conduction band tips the balance from

generation to recombination (see Figure 7.2 and Appendix F). The fractional carrier loss depends on the relative concentrations of injected carriers and defects. Consequently, the reduction of DC gain due to radiation damage depends on current density. For a given collector current a small device (small emitter area) will suffer less degradation in DC gain than a large one.

Since the probability of recombination depends on the transit time through the junction region, reduced base width will also improve the radiation resistance. Base width is strongly linked with device speed, so that the reduction in DC gain β_{DC} from its pre-irradiation value β_0 scales inversely with a transistor's unity gain frequency f_T (Messenger and Ash 1986 and see Appendix G),

$$\frac{1}{\beta} - \frac{1}{\beta_0} = K\Phi \propto \frac{\Phi}{f_T} \ . \tag{7.16}$$

Since IC technology is driven primarily by device speed, mainstream market forces indirectly improve the radiation resistance of bipolar transistor processes. Mid-gap states also limit the low current performance before irradiation. Over the past decade, evolutionary improvements in contamination control and process technology have also yielded substantially better low-current performance. Measurements on bipolar transistors from several vendors have shown that processes not specifically designed for radiation resistance are indeed quite usable in severe radiation environments, even at low currents. (Cartiglia *et al.* 1992, Kipnis, Spieler, and Collins 1994, Spencer *et al.* 1995).

Changes in doping levels due to radiation have little effect in bipolar transistors. Typical doping levels in the base and emitter are $N_B = 10^{18}$ and $N_E = 10^{20}\,\mathrm{cm}^{-3}$. In the collector depletion region doping levels are smaller, typically 10^{16}, rising to 10^{18} or 10^{19} at the collector contact. At these levels the change in doping level due to displacement damage ($\Delta N_A \approx 10^{12}\,\mathrm{cm}^{-3}$ at $\Phi = 10^{14}\,\mathrm{cm}^{-2}$) is negligible, although local device temperatures may be sufficiently high that anti-annealing leads to noticeable effects.

Figure 7.7 shows measured DC gain for *npn* and *pnp* bipolar transistors irradiated to a fluence of $1.2 \cdot 10^{14}\,\mathrm{cm}^{-2}$ with 800 MeV protons (Kipnis, Spieler, and Collins 1994). These devices exhibited $f_T = 10\,\mathrm{GHz}$ for the *npn* and 4.5 GHz for the *pnp* transistors. In the CAFE chip, a prototype design for the ATLAS silicon tracker, the *npn* input device was operated at a current density of about $2\,\mu\mathrm{A}/(\mu\mathrm{m})^2$, where the post-radiation current gain decreased to about 60% of its initial value. Although a smaller transistor would deteriorate less, the thermal noise contribution of the parasitic base resistance would be excessive, so a compromise is necessary. No measurable changes in transconductance were measured, as expected. The output resistance of these devices decreased by $< 10\%$ after irradiation.

The data shown above are for a junction-isolated IC process. Modern high-density bipolar transistor processes utilize oxide isolation, which is subject to charge build-up as discussed in Section 7.1.2. Figure 7.8 illustrates the effect.

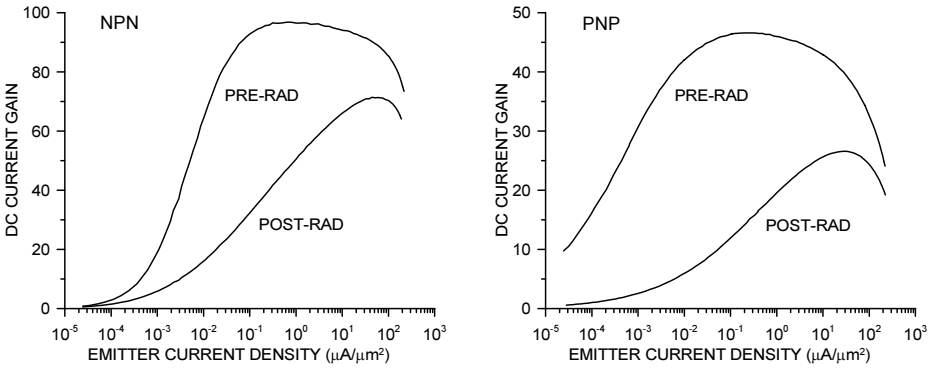

FIG. 7.7. DC gain of *npn* and *pnp* transistors before and after irradiation to a fluence of $1.2 \cdot 10^{14}\,\mathrm{cm}^{-2}$ (800 MeV protons).

The build-up of positive charge in the oxide adjacent to the base–emitter junction attracts electrons, which provide a leakage path that degrades the current gain (Pease 1983). The same phenomenon provides a path between adjacent devices. The process of positive charge build-up in the oxide is similar to that in the gate oxide of MOSFETs, but with the important difference that the oxide is practically field-free. This has a significant effect on hole capture and recombination. It has been found that trapped charge provides a deterrent to further charging (Witczak *et al.* 1998, Graves *et al.* 1998). However, at low production rates the initially formed charge can decay and allow further charge build-up. The result is that the degradation of current gain depends on the dose rate. At high dose rates the process is space-charge limited, but at low dose rates it can

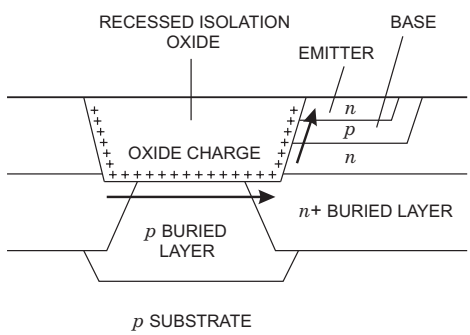

FIG. 7.8. Positive charge accumulation in the isolation oxide adjacent to an *npn* transistor leads to parasitic current paths (arrows). The horizontal arrow represents leakage between adjacent devices and the arrow towards the upper right degrades the current gain.

develop and lead to enhanced low dose-rate sensitivity (ELDRS). Experiments on MOS capacitors indicate that this is a universal effect in thick oxides irradiated at zero or very low fields (Fleetwood *et al.* 1994). This phenomenon has also afflicted bipolar logic ICs that were considered to be extremely radiation resistant. Pease (2003) gives a comprehensive survey of the history and recent results. The results are strongly process dependent and a reliable scaling procedure that relates accelerated testing to normal operating conditions has not been established.

This is a severe problem, as exposures to the 5 or 10 Mrad level needed to assess devices for the LHC are conducted on a time scale of days or weeks ($\sim$ 10 rad/s), whereas 5 Mrad in 10 years corresponds to 0.05 rad/s, assuming the canonical 10^7 s exposure time per year. The data shown in Figure 7.7 are for junction-isolated devices fabricated in an "obsolete" integrated circuit process. Measurements on devices in a standard commercial oxide isolated process that yielded good results at high dose rates proved to be marginal at low dose rates, whereas devices using a radiation-hard oxide isolated process have provided acceptable results (Dabrowski *et al.* 2000).

Noise degradation has been measured on individual transistors and complete preamplifier circuits. The results are consistent with the measured degradation in DC gain and no change in transconductance or parasitic resistances, as expected. As derived in Chapter 6 the optimum noise of a bipolar transistor

$$Q^2_{n,\min} = 4kT\frac{C}{\sqrt{\beta_{DC}}}\sqrt{F_i F_v}\ . \tag{7.17}$$

If the optimum operating current

$$I_C = \frac{kT}{e}C\sqrt{\beta_{DC}}\sqrt{\frac{F_v}{F_i}}\frac{1}{T_S} \tag{7.18}$$

is adjusted as radiation damage progresses, the noise of the input device will degrade with the square root of current gain. Many contributions enter into the overall noise, as shown in Table 6.1, so the overall noise degradation may be smaller than predicted by eqn 7.17. Figure 7.9 shows the measured spectral output noise density of a monolithically integrated preamplifier before and after irradiation to a fluence of $1.2{\cdot}10^{14}\,\mathrm{cm}^{-2}$ with 800 eV protons (Kipnis, Spieler, and Collins 1994). The gain increased by a few percent after irradiation, so the input noise increase is somewhat smaller than shown. In these and other measurements the noise after irradiation is explained by the increase in base current shot noise.

7.3.2 *Junction field effect transistors (JFETs)*

JFETs (silicon or GaAs) can be quite insensitive to both ionization and displacement effects, as they are majority carrier devices. In this context, the important feature is that device characteristics are determined essentially by the geometry and doping level of the channel. Typical doping levels are $10^{15} - 10^{18}\,\mathrm{cm}^{-3}$, so

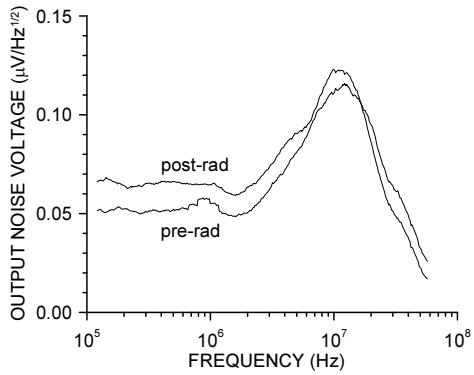

FIG. 7.9. Noise of a bipolar transistor preamplifier before and after irradiation to a fluence of $1.2 \cdot 10^{14}\,\mathrm{cm}^{-2}$ with 800 MeV protons.

the effect of radiation-induced donor removal or acceptor states is small. Silicon JFETs exhibit very good radiation resistance. Measurements on both standard commercial devices and custom designed integrated circuits have shown minimal changes in gain at fluences $> 10^{14}$ neutrons/cm^2 and ionization doses up to 100 Mrad (Citterio *et al.* 1992, 1995; Radeka *et al.* 1993). Low frequency noise ($f < 100$ kHz) may increase by an order of magnitude, but at high frequencies very little change in noise is observed. Measurements of Si JFETs at 90 K also show excellent radiation characteristics (Citterio *et al.* 1995).

In some applications, analog storage circuitry for example, gate leakage current is important. Generation current in the gate depletion region due to displacement damage can affect the gate current strongly. Measurements on commercial JFETs irradiated by high-energy electrons to 100 Mrad ($\Phi \approx 10^{15}\,\mathrm{cm}^{-2}$) show the gate reverse current increasing 100-fold from an initial value of 70 pA (Stephen 1985). Here one should choose the smallest geometry device commensurate with other requirements.

At this point it is worth noting that the superior radiation resistance claimed for GaAs ICs has more to do with the use of JFETs or MESFETs (a Schottky barrier JFET) than the properties of the semiconductor. These devices are more radiation resistant than silicon MOSFETs, but suffer from a much lower circuit density.

7.3.3 *Metal-oxide-silicon field effect transistors (MOSFETs)*

Within the FET family, MOSFETs present the most pronounced ionization effects, as the key to their operation lies in the oxide that couples the gate to the channel. As described above and illustrated in Figure 7.3, positive charge build-up due to hole trapping in the oxide and at the interface shifts the gate voltage required for a given operating point to more negative values. The processes lead-

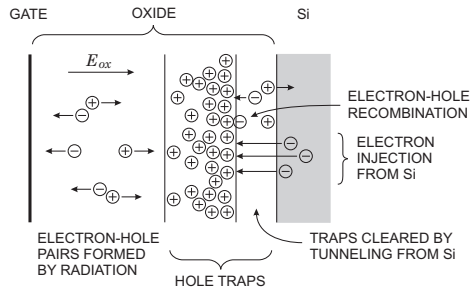

FIG. 7.10. Schematic representation of hole accumulation and removal processes in SiO_2. (Adapted from Boesch *et al.* 1986. ©IEEE, reproduced with permission.)

ing to hole accumulation and removal in an NMOS device are summarized in Figure 7.10.

Electron–hole pairs are formed by incident radiation. The electrons move rapidly under the influence of the applied field towards the positively biased gate electrode. Holes move much more slowly towards the SiO_2–Si interface. Some fraction of the holes is trapped in the oxide. The holes that reach the Si interface recombine with electrons from the bulk. The critical process is the trapping of holes, which leads to charge build-up and, as a consequence, a shift in the gate voltage required for a given current flow. In well-controlled oxides the hole traps extend $5 - 10$ nm from the oxide–Si interface into the oxide. The number of holes available for trapping increases linearly with oxide thickness and the transit time. As a result the trapping probability also increases roughly linearly with thickness and the overall threshold shift for a given dose is approximately proportional to the square of the oxide thickness.

Within a region extending $2.5 - 5$ nm into the oxide, the probability of electrons tunneling from the silicon bulk into the oxide is sufficiently high to remove the trapped hole charge by recombination. This is believed to be the dominant process in the long term annealing of the observed voltage shifts. The concentration of trapped holes too far from the bulk for significant electron tunneling is reduced in part by recombination with primary electrons drifting through the oxide and by hot electrons injected from the channel.

Since the trapped charge modifies the field distribution in the oxide, which in turn affects the carrier motion and trap population, the resulting shift in gate voltage with dose is nonlinear. Furthermore, since trapped holes can be released by thermal excitation, and the migration of trapping sites from the oxide–Si interface into the bulk is also thermally driven, the change in required gate voltage becomes a complex function of temperature and dose rate.

The gate voltage shift is typically expressed in terms of threshold voltage V_T, which roughly marks the onset of appreciable current flow. This shift affects the operating points in analog circuitry and switching times in digital circuitry.

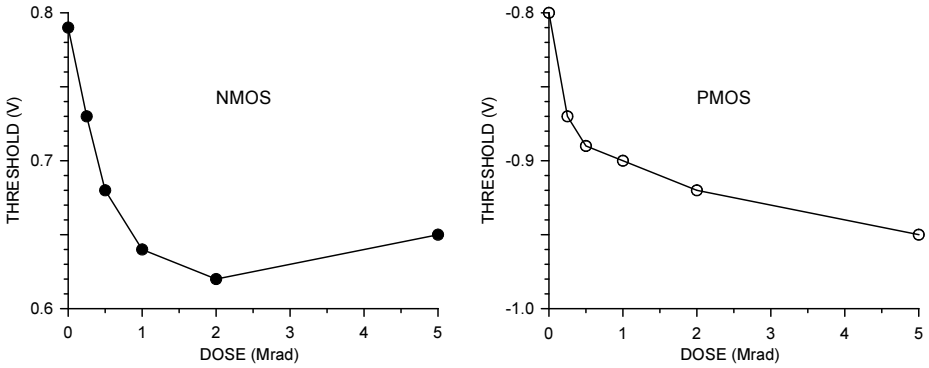

FIG. 7.11. Threshold voltage shifts for radiation-hardened NMOS and PMOS transistors *vs.* ^{60}Co radiation dose in Mrad(Si).

Reducing the thickness of the gate oxide d_{ox} greatly improves the radiation resistance; gate voltage shifts scaling with d_{ox}^2 to d_{ox}^3 for a given dose have been observed for oxide thicknesses > 20 nm (Ma and Dressendorfer 1989).

Typical threshold shifts for a 1.2 μm radiation-hardened CMOS IC process with a 20 nm thick gate oxide are shown in Figure 7.11 (Dabrowski *et al.* 1991). After exposure to 5 Mrad(Si) of ^{60}Co irradiation, NMOS thresholds shift by 200 mV and PMOS levels change by 150 mV. For both NMOS and PMOS devices the threshold voltage shifts to more negative values as expected from positive charge build-up in the oxide. The slight upturn above 2 Mrad in the NMOS curve is typical and reflects the build-up of interface states (Ma and Dressendorfer 1989). About 70% of the threshold shifts occur during the first 250 krad, also a typical phenomenon. Measurements to 125 Mrad on a similar process show a total threshold shift of 400 mV for NMOS and 100 mV for PMOS with little increase beyond 10 Mrad (Seller *et al.* 1995). Conventional CMOS ($> 0.5\,\mu$m feature size) typically shows similar threshold shifts at 20 – 40 krad.

As noted above, it was always assumed that shrinking feature size accompanied by thinner gate oxides would improve the radiation resistance of standard MOS processes. For thinner oxides, however, the radiation-induced threshold shifts are dramatically reduced, since the electron tunneling rate is now sufficiently large to neutralize the trapped holes by recombination. The effect of electron tunneling is shown in Figure 7.12. These thin oxides are integral to "deep submicron" processes. Measurements on standard commercial 0.25 μm devices show negligible threshold shifts and show good noise performance measured to 100 Mrad (Campbell *et al.* 1999, Snoeys *et al.* 2000). Complex integrated circuits have remained operational beyond 100 Mrad (Einsweiler 2004). Commercial "deep submicron" processes have rendered "radiation-hard" CMOS largely obsolete in the detector community.

The transconductance of MOSFETs in strong inversion is hardly affected by radiation, provided the drain current is maintained. Clearly, if the operating point is set by voltages, threshold shifts will affect the device current, and as a consequence, the transconductance. Controlling the device current via current mirrors, as shown in Chapter 6, circumvents this problem. The weak inversion slope, however, is strongly affected by surface charge build-up. This is clearly demonstrated in Figure 7.13 (Dabrowski *et al.* 1991).

NMOS devices are more sensitive than PMOS devices. For example, to operate a 1.2 μm NMOS transistor in moderate inversion one might choose a normalized drain current $I_D/W = 0.3\,\mathrm{A/m}$, yielding $I_D = 3\,\mathrm{mA}$ for a 1 mm wide transistor. The normalized transconductance $g_m/I_D = 15.4\,\mathrm{V}^{-1}$ or $g_m = 4.6\,\mathrm{mS}$ before irradiation. After exposure to 5 Mrad $g_m/I_D = 11.8\,\mathrm{V}^{-1}$ or $g_m = 3.5\,\mathrm{mS}$. Typically, the NMOS devices suffered a 20 – 30% degradation, whereas the PMOS devices were quite insensitive to radiation, with only a few percent decrease in transconductance at 5 Mrad. About half of the observed change at 5 Mrad occurred before attaining a dose of 1 Mrad.

Extensive noise measurements have been performed at the University of Pennsylvania (Tedja *et al.* 1992) and by a UCSC/LBNL group (Dabrowski *et al.* 1991). In the latter, spectral noise density was measured over a frequency range of 10 kHz to 10 MHz before and after ^{60}Co irradiation to a dose of 5 Mrad(Si). The noise was measured at three representative drain current densities I_D/W. Again, these data can be scaled to any device width, where the noise scales with $W^{-1/2}$. White noise was evaluated at high frequencies and is characterized by the noise coefficient $\gamma_n = {e_n}^2 \cdot g_m/4kT$ to assess the inherent noise properties

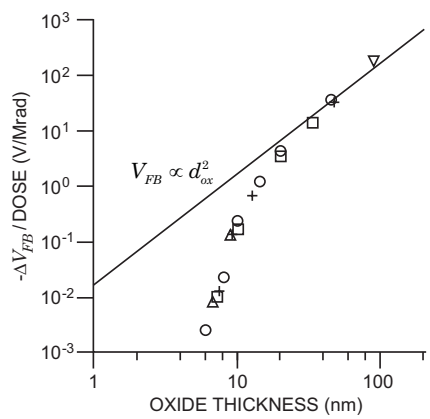

FIG. 7.12. Change in flat-band voltage V_{FB} *vs.* gate-oxide thickness d_{ox} for various devices irradiated at 80 K with the same oxide field $E_{ox} = 2 \cdot 10^6\,\mathrm{V/cm}$. A line showing the dependence $V_{FB} \propto d_{ox}^2$ is shown for comparison. (Saks *et al.* 1984. ©IEEE, reproduced with permission.)

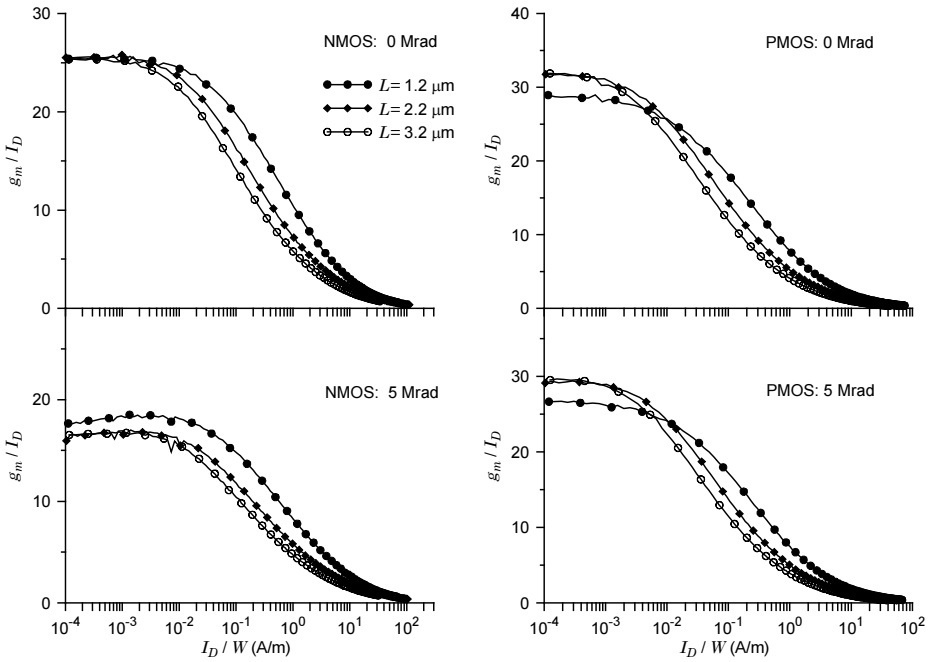

FIG. 7.13. Normalized transconductance g_m/I_d *vs.* normalized drain current I_d/W for NMOS and PMOS transistors with channel lengths of 1.2, 2.2 and 3.2 μm before and after ^{60}Co irradiation to 5 Mrad(Si).

independent of transconductance. Results for various device geometries and current densities are shown in Table 7.2. For these measurements the substrate was biased at the source potential. Although the observed degradation is quite small in some cases, typically it is quite substantial and would need to be compensated for by a considerably higher operating current. The difference between the NMOS and PMOS results is striking. The NMOS devices show a much greater degradation. The PMOS devices also exhibit substantially less low-frequency noise. The low-frequency noise spectral density of the NMOS devices can be described by $e_n^2 = A_f/q + B$, where q ranges from 0.8 to 1.0 and is constant for all currents for the same geometry. The changes in q after a dose of 5 Mrad are of order 0.1. The noise coefficient A_f is about $1.0 - 1.5 \cdot 10^{-30}\,\mathrm{V}^{-2}$ pre-rad and 5 to $10 \cdot 10^{-30}\,\mathrm{V}^{-2}$ post-rad. Before irradiation, A_f scales well with inverse gate area, but no clear pattern is observed after irradiation. The low frequency noise behavior of the PMOS devices is more complex and cannot be parameterized as simple $1/f$ noise, but the devices exhibit substantially better noise than the NMOS transistors.

Due to the presence of mobile trapped charge, threshold behavior can become quite difficult to predict when the gate voltage changes appreciably with

varying duty cycles, as in logic circuitry. Detectors and analog circuitry are simpler by comparison, since the voltage levels are either static or change with a fixed period, as in analog pipelines, for example. In general, when performing ionization damage tests devices must be operated at typical operating voltages and digital circuitry must be clocked at frequencies and patterns approximating typical operation.

Generally speaking, both bulk and SOI (silicon on insulator) CMOS are subject to the effects described above. SOI is often cited as a specifically radiation-hard technology because of its resistance to transient radiation effects, primarily latchup due to photocurrents developed at high intensity bursts of radiation ($> 10^6 - 10^7$ rad/s) typical of nuclear detonations (latchup is explained in Appendix A). Although SOI can provide superior device speed because of reduced stray capacitance, this technology is not inherently more resistant to radiation in our applications. If anything, the additional oxide interfaces tend to complicate matters and at this time most radiation-resistant CMOS processes are on bulk silicon.

Table 7.2 Noise coefficients $\gamma_n = e_n^2 \cdot g_m/4kT$ for NMOS and PMOS transistors of various widths and lengths, operated at current densities $Id/W = 0.03, 0.10$ and 0.3 A/m, before and after ^{60}Co irradiation to 5 Mrad(Si).

Type	NMOS	PMOS	NMOS	PMOS	NMOS	PMOS	NMOS	PMOS
Width (μm)	75	75	1332	1332	888	888	1332	1332
Length (μm)	1.2	1.2	1.2	1.2	2.2	2.2	3.2	3.2
$I_d/W = 0.03$								
0 Mrad			0.81	0.61	0.64	0.59	0.66	0.5
5 Mrad			2.17	0.84	1.00	0.58	1.50	0.69
$I_d/W = 0.1$								
0 Mrad	1.10	0.70	1.20	1.10	0.80	0.80	0.80	0.66
5 Mrad	3.80	1.10	3.40	1.60	1.30	0.90	1.70	0.70
$I_d/W = 0.3$								
0 Mrad	1.60	1.30	2.00	1.70	1.10	1.00	1.10	0.77
5 Mrad	5.00	2.90	4.80	2.70	1.60	1.40	1.20	0.81

7.3.4 *Radiation effects in integrated circuit structures*

The preceding discussion has emphasized the properties of individual devices. In integrated circuits many devices are placed close together. As mentioned above, the silicon surface is naturally n-type, so isolation structures are required to preclude unwanted cross-coupling between devices. Two basic techniques are used:

- junction isolation, where reverse-biased pn-junctions provide both ohmic and capacitive isolation.
- oxide isolation, where oxide layers with carefully controlled interface properties deplete the adjacent silicon of mobile charge.

Detailed information on these processes can be found in texts on IC technology, for example in the comprehensive series by Wolf (1990, 1995, 2002).

Junction isolation is very robust, but requires substantial additional space. Oxide isolation allows higher packing densities and is used by most high-density IC processes. All CMOS processes utilize some form of oxide isolation, whereas bipolar transistor processes can be found with both junction and oxide isolation. Under irradiation the oxide layers used for isolation (field oxide) suffer from the same phenomena described for the gate oxide of MOSFETs (see field oxide in Figure 7.3). Since isolation oxides are thicker than gate oxides, more electron–hole pairs are formed by incident radiation. Furthermore, the fields in the isolation oxide tend to be much lower, so charge trapping in the oxide will be exacerbated. Developing radiation-hard isolation oxides (field oxides) was a major challenge in the development of high-density radiation-hard CMOS and remains one of the few "secret" process ingredients (for a basic discussion see Ma and Dressendorfer 1989). In commercial "deep submicron" processes NMOS leakage is the main problem. This is eliminated by using enclosed geometry transistors and p^+ guard rings. A small drain is surrounded by the gate, which is in turn enclosed by the source (Snoeys *et al.* 2000).

Inherently radiation-hard devices, notably JFETs and bipolar transistors, are often implemented in nonhardened oxide-isolated processes. Here radiation effects in the isolation structures can severely affect the radiation resistance of the devices. Clues to the importance of such parasitic ionization effects can be gleaned from a comparison of neutron and photon irradiations. As discussed in the context of low dose rate effects in bipolar transistor ICs, the suitability of these processes must be determined case-by-case.

IC processes also use special device structures to facilitate the integration of different device types. A prime example is the lateral *pnp* transistor, a structure more compatible with a standard CMOS process than "classic" vertical bipolar transistors. In a lateral transistor the emitter, base and collector are arranged along the surface of the silicon with large-area exposure to oxide interfaces. Unlike vertical bipolar transistors, lateral devices are very susceptible to ionizing radiation, as surface leakage causes severe degradation of DC gain. Lateral *pnp*

transistors can be used as current sources or high impedance loads, if the biasing circuitry is designed to accommodate substantial increases in base currents.

Digital circuitry is susceptible to single event upset. In modern devices the charge required to change the state of a switching device is so small that it can be introduced by local charge deposition from single particles. It is more probable for heavily ionizing particles (alphas, for example), but minimum ionizing particles can produce Si recoils that have a significantly higher energy loss dE/dx. This leads to "soft errors" that can be detected by appropriate error codes. CMOS circuitry includes parasitic bipolar transistors that can form a *p-n-p-n* thyristor structure (Appendix A). This structure can latch up in a nonrecoverable state. The probability of single event processes can be minimized by appropriate process and circuit design. For a survey see Holmes-Siedle and Adams (2002). As an example, see the investigation of single event upset in optical links for the ATLAS silicon strip detector by Dowell *et al.* (2002).

7.4 Dosimetry

Dosimetry is fairly straightforward in irradiations with charged particles, but is often flawed in ^{60}Co irradiations used in assessing ionization effects. The absorption length of the ~ 1 MeV photons emitted in ^{60}Co decays is ~ 5 cm (Figure 1.20), much greater than the depth of the devices to be tested. As MOSFETs and oxide layers are at the surface, very few photons are absorbed in the device structures to be tested. However, photons scattered from the surroundings (concrete walls or packaging materials, for example) excite x-rays with a high absorption probability (dose enhancement, see Ma and Dressendorfer 1989, Srour *et al.* 1984). As a consequence, the dose in ^{60}Co irradiations may be substantially higher or lower than estimated from source activity alone. Reproducible circumstances are established by ensuring "charge transfer equilibrium". As Compton scattering dominates, the deposited dose is due to the energy deposited by the Compton electrons. By enclosing the sample with material that ensures that the flux of Compton electrons entering the sample equals the flux exiting the sample, the dose is uniform and well-established throughout the sample. Ideally, the absorber would be silicon, but this is quite cumbersome, so higher density materials such as lead are used. However, the x-rays from the non-silicon absorber can lead to dose enhancement, so to suppress them an additional layer of aluminum is included, adjacent to the sample. This creates a "silicon-like" environment, as the Al and Si x-rays are of similar energy. A suitable equilibrium shield consists 1.5 mm of lead followed by 0.7 mm of aluminum (Ma and Dressendorfer 1989). X-ray generators are convenient for use in radiation testing, but the strong dependence of absorption on depth for the typical energy spectra make dosimetry very difficult. Nevertheless, they can be quite effective in comparative measurements or process monitoring.

7.5 Mitigation techniques

Although within a given technology little can be done to reduce radiation damage to an individual device, many techniques can be applied to reduce the effects of radiation damage to an overall system. The goal of radiation-hard design is not so much to obtain a system whose characteristics do not change under irradiation, rather than to maintain the required performance characteristics over the lifetime of the system. The former approach tends to utilize mediocre to poor technologies that remain so over the course of operation. The latter starts out with superior characteristics, which gradually deteriorate under irradiation. Depending on the specific system, these designs may "die" gradually, although at some fluence or dose a specific circuit, typically digital, may cease to function at all.

7.5.1 *Detectors*

Increased detector leakage current has several undesirable consequences.

1. The integrated current over typical signal processing times can greatly exceed the signal.
2. Shot noise increases.
3. The power dissipated in the detectors increases (bias current times voltage).

Since the leakage current decreases exponentially with temperature, cooling is the simplest technique to reduce diode leakage current. For example, reducing the detector temperature from room temperature to 0 °C reduces the bias current to about 1/6 of its original value.

Detector power dissipation is a concern in large-area silicon detectors for the LHC, where the power dissipation in the detector diode itself can be of order $1 - 10\,\mathrm{mW/cm^2}$. Since the leakage current is an exponential function of temperature, local heating will increase the leakage current, which will increase the local heating, and so on, ultimately taking the device into thermal runaway. To avoid this potentially catastrophic failure mode, the cooling system must be designed to provide sufficient cooling of the detector, a challenging (but doable) task in a system that is to have zero mass. An implementation and results are shown in the next chapter, Section 8.6.5.

Reducing the integration time reduces both baseline changes due to integrated detector current and shot noise. Clearly, this is limited by the duration of the signal to be measured. To some degree, circuitry can be designed to accommodate large baseline shifts due to detector current, but at the expense of power. AC coupled detectors eliminate this problem. In instrumentation systems that require DC coupling, correlated double sampling techniques can be used to sample the baseline before the signal occurs and then subtract from the signal measurement.

One of the most powerful measures against detector leakage current is segmentation. For a given damage level, the detector leakage current per signal channel can be reduced by segmentation. If a diode with a leakage current of $10\,\mu$A is subdivided into 100 subelectrodes each with its own signal processing

channel, the bias current in each channel will be 100 nA and shot noise reduced by a factor of 10. This is why large area silicon tracking detectors can survive in the LHC environment. Fortuitously, increased segmentation is also required to deal with the high event rate. Pixel detectors with small electrode areas offer great advantages in this regard.

The most severe restriction on radiation resistance is imposed by type inversion in the sensor, when the net space charge becomes so large that the detector will no longer sustain the required voltage for full charge collection. This is especially critical for position-sensing detectors with electrodes on both sides (double-sided detectors), for which full collection is essential (see Section 2.5.3).

One can circumvent the type-inversion limit by using back-to-back single-sided detectors. The initial configuration uses n-type segmented strip electrodes on n bulk, with a contiguous p electrode on the backside. Initially, the pn-junction is at the backside. This does require full depletion in initial operation, but this is no problem for the nonirradiated device and becomes easier to maintain as increasing fluence moves the bulk towards type inversion. After type inversion the charge collection region extends from the n electrodes. Since most of the signal charge is induced when the carriers are near the strip or pixel electrodes, this provides good efficiency even when operating at less than "depletion" voltage (Unno *et al.* 1996, 1997).

Since highly damaged detectors are largely devoid of mobile charge, they appear approximately ohmic. This means that reverse bias is not essential to obtain low acceptable bias currents and the detectors also function under forward bias. Chilingarov and Sloan (1997b) demonstrated good charge collection efficiency with forward bias voltages much smaller than the reverse bias required for the same signal. Although the bias current is larger than for reverse bias, this results in less power dissipation in the detector and also simplifies the detector design. However, the smaller fields also lead to longer collection times, so the effect of trapping is exacerbated. To some degree this is ameliorated by the fact that electron traps are more readily filled by the larger standing current (Beattie *et al.* 2000).

A innovative approach to avoiding large operating voltages while reducing collection times is the "3D detector" (Parker, Kenney, and Segal 1997). Rather than forming the diode between opposite faces of a wafer, the electrodes are alternating columns of n^+ and p^+ material that are are normal to the surface of the p bulk. Columns have been implemented with 300 μm depth and a diameter about 5% of the depth (Kenney *et al.* 1999). In one set of test devices the pitch within a row is 200 μm and adjacent n- and p-rows are offset by 100 μm so that the distance that must be depleted is $\sim 140\,\mu$m. By appropriate choice of geometry the distance between n- and p-columns can be much smaller than the wafer thickness, so that both the required voltage for full collection and the collection times are reduced. For example, a spacing of 50 μm betwen n- and p-columns would yield full charge collection at order of magnitude smaller voltage than conventional devices. Devices have been irradiated to fluences $> 5 \cdot 10^{14}\,\mathrm{cm}^{-2}$

of 24 GeV/c pions with good results. The same technology can be applied as trenches at the edges of the detector to provide "active" edges to avoid the dead area at the edges of conventional devices. This is very useful in tiling detectors to form large area arrays for tracking or x-ray imaging (Kenney *et al.* 2001). A similar principle has been applied to planar devices by placing alternating n^+ and p^+ electrodes on opposite faces of a double-sided sided detector (Li *et al.* 2002).

Another path that is being pursued is operation at cryogenic temperatures. Since the build-up of space charge results from the population of acceptor-like states through thermal excitation, heavily damaged detectors can be resuscitated by cooling to liquid nitrogen temperature (Palmieri *et al.* 1998). Charge trapping is not suppressed, so roughly half of the charge is recovered when operating a 300 μm thick detector at 250 V after irradiation to a neutron fluence of $2.2 \cdot 10^{15}\,\mathrm{cm}^{-2}$. Cryogenic operation of a readout IC fabricated in 0.25 μm CMOS technology has also been demonstrated (Anelli *et al.* 2003). Silicon detectors operating at 130 K are in use as high-intensity radiation monitors (Palmieri *et al.* 2003). For an overview of cryogenic silicon detector technology see Abreu *et al.* (2003).

One way to avoid the limits of silicon is to use different materials. Diamond has been shown to be quite radiation resistant, with a 15% degradation in signal-to-noise after exposure to a 24 GeV proton fluence of $2.2 \cdot 10^{15}\,\mathrm{cm}^{-2}$ (Adam *et al.* 2003). Diamond is an insulator, so bias currents are very low without resorting to *pn*-junctions and their inherent doping effects. However, the large bandgap also reduces the charge yield. Minimum ionizing particles on average produce 36 electron–hole pairs per micron, about half the charge obtained with silicon. This is partially mitigated by a smaller dielectric constant, relative to silicon ($\varepsilon \approx 5.5$ *vs.* 12.9). Apart from cost, a major limitation is the obtainable drift length, which depends greatly on the growth techniques. In polycrystalline material drift lengths of $\sim 200\,\mu$m have been achieved (Adam *et al.* 2003). Diamond pixel sensors have been operated successfully with readout ICs designed for ATLAS silicon pixel system (Keil *et al.* 2003). Single crystal synthetic diamond is now available and has been applied in beam monitoring (Edwards *et al.* 2004).

4H-SiC is available as single crystals and has also shown interesting results. The radiation resistance of GaN is also being investigated. Materials for very high luminosity colliders are under investigation by the RD50 collaboration. An interim status report by Moll (2003) gives an overview.

7.5.2 *Electronics*

The design of the electronic systems is governed by changes in transistor parameters under irradiation, but circuit design and – at a higher level – architecture are at least as important. Amplifiers are sensitive to changes in gain, bandwidth, and noise, so that effects on transconductance and noise parameters are important. Comparators used for threshold determination and timing rely critically on threshold shifts. Analog storage cells and switched capacitor systems tend to

be sensitive to leakage currents. Digital circuitry is affected by threshold shifts that affect propagation delays and by device transconductance, which determines switching speed.

Shorter shaping times improve tolerance to leakage currents. In high rate systems, fast response time is needed anyway, so experimental desires and engineering considerations interfere constructively. Since the system must be designed to tolerate a substantial shot noise current, utilization of bipolar junction transistors becomes very attractive, since the base shot noise becomes a minor contribution (in contrast to systems that emphasize noise minimization, as in x-ray spectrometry or liquid argon calorimetry).

In amplifiers bipolar transistor circuitry offers power advantages over CMOS. In logic circuitry, especially at low overall switching rates, CMOS is advantageous both because of power consumption and circuit density. For example, the on-detector silicon tracker front-end designed for the ATLAS experiment at the LHC uses bipolar transistor technology for the amplifier–pulse shaper–comparator and radiation-hard CMOS for a clock-driven digital pipeline buffer and data readout (Dabrowski *et al.* 2000). In amplifiers, bipolar transistors offer higher bandwidth for a given power and superior device matching, which is a prime consideration in highly segmented systems with a correspondingly large number of channels. Threshold shifts in bipolar transistors are quite small with excellent matching between devices. JFETs yield excellent noise performance in applications where power consumption and circuit density are not prime considerations. Even when a CMOS front-end is chosen, because of the use of a switched capacitor analog memory, or the desire to combine the analog and digital circuitry on the same chip, amplifiers can be made quite radiation resistant, since the circuitry can be made to adjust for shifts in threshold voltage.

This principle was illustrated in Figure 6.38, where the current of the input cascode is controlled by an external current via a current mirror. The gate voltages of the cascode transistors can be chosen to accommodate threshold shifts. This approach does not maintain the DC output level, so baseline shifts must be corrected for by correlated double sampling or rendered irrelevant by AC coupling. The operating voltage and the gate voltages of the cascode transistors must be chosen to accommodate the threshold shifts, so overall power dissipation will be higher than needed without radiation damage. Techniques of this type can provide radiation-resistant amplifiers with radiation-soft transistors.

In general, the use of fully differential circuitry and current mirrors yields circuitry whose operating point relies primarily on relative device matching. (Spieler 1994, Kipnis, Spieler, and Collins 1994, Spencer *et al.* 1995) Changes in threshold voltages or current gain in adjacent devices tend to track after radiation damage, so the circuit will maintain its operating point. As noted in Chapter 6 differential input stages are usually ruled out by power considerations, but differential circuitry can be employed efficiently farther downstream. Circuitry should also be designed to minimize single-point failure modes. Failure of common bias net-

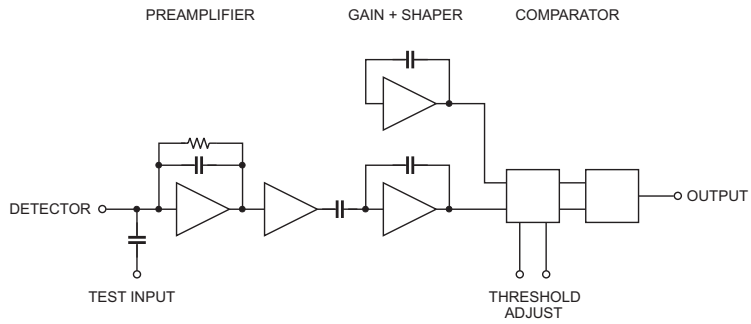

FIG. 7.14. Block diagram of a binary readout channel.

works will cause all associated circuitry to fail. Local biasing with highly parallel architectures reduces these problems.

Figure 7.14 shows a block diagram to illustrate the approach taken in a representative design (Kipnis, Spieler, and Collins 1994). This system only records the presence of a detector signal, so each channel includes an amplifier, pulse shaper and threshold comparator. The input transistor is biased through a current mirror, as just described. The gain stages must provide sufficient gain so that the threshold voltage at the comparator is sufficiently high to provide good channel-to-channel and chip-to-chip uniformity. Since the first two stages are single-ended to reduce power consumption, substantial circuit complexity would be necessary to maintain DC stability, so AC coupling is introduced at the input of the third stage. From here on the circuitry is differential. The third stage is still single-ended, but it is replicated as a dummy amplifier to bias the second input of the differential comparator. The dummy amplifier is included in each individual channel to obtain optimal parameter tracking under radiation damage, and also to maintain a parallel architecture and reduce single-point failure modes. The threshold level is applied differentially to exploit device tracking during irradiation. In modern designs this is digitally controlled by a digital-to-analog converter (DAC). The calibration circuitry provides a means to monitor the gain. Similar principles have been applied to bipolar transistor and CMOS circuitry. Maintaining threshold matching under irradiation is a challenge, so some implementations employ trim DACs on each channel to correct the master threshold level channel-by-channel (Dabrowski *et al.* 2000, Blanquart *et al.* 2002, and examples in Chapter 8).

CMOS logic circuitry does not offer the flexibility of self-adjusting circuitry. Since the threshold shifts of n- and p-MOSFETs are not complementary, circuit switching thresholds change. At very high damage levels the device transconductance also suffers due to build-up of interface charge and increased scattering of charge carriers in the channel. Both effects change propagation delays, which can lead to race conditions (mismatches in propagation delays of streams whose

results are combined) that cause circuit failure. These problems require careful design. On the other hand, digital circuitry can be used to control operating points and enhance radiation resistance by adjusting for changes in analog circuitry. Examples are shown in Chapter 8.

7.5.3 *Summary*

Judicious evaluation of the radiation fields coupled with a stringent analysis of application requirements have yielded electronic systems that perform well to ionizing doses of 100 Mrad and particle fluences of to $10^{15}\,\mathrm{cm}^{-2}$. Some examples will be discussed in Chapter 8. Developing radiation-resistant systems does require great attention to detail and substantially more testing effort than conventional designs, but the effort is necessary to support the need for ever increasing luminosity. For many applications we are limited less by technology than by ingenuity.

References

Abreu, M. *et al.* (2003). Recent progress in low-temperature silicon detectors. *Nucl. Phys. B (Proc. Suppl)* **125** (2003) 169–174

Adam, W. *et al.* (2003). The development of diamond tracking detectors for the LHC. *Nucl. Instr. and Meth.* **A514** 79–86

Anelli, G. *et al.* (2003). A high-speed low-noise transimpedance amplifier in a 0.25 μm CMOS technology. *Nucl. Instr. and Meth.* **A512** (2003) 117–128

Azzi, P. *et al.* (1996). Radiation damage experience at CDF with SVX'. *Nucl. Instr. and Meth.* **A383** (1996) 155–158

Barberis, E. *et al.* (1993). Temperature effects on radiation damage to silicon detectors. *Nucl. Instr. and Meth.* **A326** (1993) 373–380

Barberis, E. *et al.* (1994). Capacitances in silicon microstrip detectors. *Nucl. Instr. and Meth.* **A342** (1994) 90–95

Beattie, L.J. *et al.* (1998) The electric field in irradiated silicon detectors. *Nucl. Instr. and Meth.* **A418** (1998) 314–321

Beattie, L.J. *et al.* (2000). Forward-bias operation of Si detectors: a way to work in high-radiation environment. *Nucl. Instr. and Meth.* **A439** (2000) 293–302

Blanquart, L. *et al.* (2002). Analog front-end cell designed in a commercial 0.25 μm process for the ATLAS pixel detector at LHC. *IEEE Trans.Nucl.Sci.* **49** (2002) 1778–1782

Boesch, H.E. *et al.* (1986). Saturation of threshold voltage shift in MOSFETs at high total dose. *IEEE Trans. Nucl. Sci.* **NS-33/6** (1986) 1191–1197

Borgeaud, P. *et al.* (1983). The effect of radiation on the energy resolution of ion-implanted silicon detectors. *Nucl. Instr. and Meth.* **211** (1983) 363–367

Brodbeck, T.J. *et al.* (2002). Carrier mobilities in irradiated silicon. *Nucl. Instr. and Meth.* **A477** (2002) 287–292

Burke E.A. (1986). Energy Dependence of proton-induced displacement damage in silicon. *IEEE Trans. Nucl. Sci.* **NS-33/6** (1986) 1276–1281

Campbell, M. *et al.* (1999). A pixel readout chip for 10–30 Mrad in standard 0.25μm CMOS. *IEEE Trans. Nucl. Sci.* **NS-46/3** (1999) 156–160

Cartiglia, N. *et al.* (1992). Radiation hardness measurements on bipolar test structures and an amplifier-comparator Circuit. *Conference Record of the IEEE Nuclear Science Symposium, Oct. 25–31, 1992, Orlando, Florida.* Vol. 2, pp. 819–821. ISBN 0-7803-0883-2

Castaldini, A. *et al.* (2002). Electric field distribution in irradiated silicon detectors. *Nucl. Instr. and Meth.* **A476** (2002) 550–555

A. Chilingarov *et al.* (1995). Radiation effects and operational projections for silicon in the ATLAS inner detector. *Nucl. Instr. and Meth.* **A360** (1995) 432

A. Chilingarov, Meyer, J.S. and Sloan, T. (1997a). Radiation damage due to NIEL in GaAs particle detectors. *Nucl. Instr. and Meth.* **A395** (1997) 35–44

A. Chilingarov, A. and Sloan, T. (1997b). Operation of heavily irradiated silicon detectors under forward bias. *Nucl. Instrum. and Meth.* **A399** (1997) 35–37

Citterio, M. *et al.* (1992). A Study of low noise JFETs exposed to large doses of gamma rays and neutrons. *Conference Record of the IEEE Nuclear Science Symposium, Oct. 25–31, 1992, Orlando, Florida.* Vol. 2, pp. 794–796. ISBN 0-7803-0883-2

Citterio, M. *et al.* (1995). Radiation effects at cryogenic temperatures in Si-JFET, GaAs MESFET, and MOSFET devices. *IEEE Trans. Nucl. Sci.* **NS-42/6** (1995) 2266–2270

Dabrowski, W. *et al.* (1991). Noise measurements on radiation-hardened CMOS transistors. *Conference Record of the 1991 IEEE Nuclear Science Symposium and Medical Imaging Conference, Nov. 2–9, 1991, Santa Fe, New Mexico.* Vol. 3, pp. 1536–1540. IEEE catalog no. 91CH3100-5

Dabrowski, W. *et al.* (2000). Design and performance of the ABCD chip for the binary readout of silicon strip detectors in the ATLAS semiconductor tracker. *IEEE Trans.Nucl.Sci.* **NS-47** (2000) 1843–1850

Dowell, J.D. *et al.* (2002). Single event upset studies with the optical links of the ATLAS semiconductor tracker. *Nucl.Instrum. and Meth.* **A481** (2002) 575–584

Edwards, A.J. *et al.* (2004). Radiation monitoring with diamond sensors in BaBar. *IEEE Trans. Nucl. Sci.* **51/4** 1808–1811

Einsweiler, K. (1994). private communication

Eremin, V., Verbitskaya, E. and Li, Z. (2002). The origin of double peak electric field distribution in heavily irradiated silicon detectors. *Nucl. Instr. and Meth.* **A476** (2002) 556–564

Fleetwood, D.M. *et al.* (1994) Physical mechanisms contributing to enhanced bipolar gain degradation at low dose rates. *IEEE Trans. Nucl. Sci.* **NS-41** 1871–1883

Gill, K. *et al.* (1992). *Radiation damage by neutrons and photons to silicon detectors. Nucl. Instr. and Meth.* **A322** (1992) 177

Giubellino, P. *et al.* (1992). Study of the effects of neutron irradiation on silicon strip detectors. *Nucl. Instr. and Meth.* **A315** (1992) 156

Graves, R.J. *et al.* (1998). Modeling low dose rate effects in irradiated bipolar oxides.*IEEE Trans. Nucl. Sci.* **NS-45** (1998) 2352–2365

Griffin, P.J. *et al.* (1993). Sandia National Laboratories Report SAND92-0094

Grosse-Knetter, J. for the ATLAS Pixel Collaboration (2004). The ATLAS pixel detector. e-Print Archive: physics/0401068, http://arxiv.org/PS_cache/physics/pdf/0401/0401068.pdf

Grove, A.S. (1967). *Physics and Technology of Semiconductor Devices*. Wiley, New York. ISBN 0-471-32998-3

Holmes-Siedle, A. and Adams, L. (2002). *Handbook of Radiation Effects* (2nd edn). Oxford University Press, Oxford. ISBN 0-19-850733-X, QC474.H59

Huhtinen, M. and Aarnio, P.A. (1993). Pion induced displacement damage in silicon devices. *Nucl. Instr. and Meth.* **A335** (1993) 580–582

Huhtinen, M. (2002). Simulation of non-ionizing energy loss and defect formation in silicon. *Nucl. Instr. and Meth.* **A491** (2002) 194–215

Keil, M. *et al.* (2003). New results on diamond pixel sensors using ATLAS frontend electronics. *Nucl. Instr. and Meth.* **A501** 153–159

Kenney, C. *et al.* (1999). Silicon detectors with 3-D electrode arrays: fabrication and initial test results. *IEEE Trans. Nucl. Sci.* **NS-46/4** (1999) 1224–1236

Kenney, C.J. *et al.* (2001). Observation of beta and X rays with 3D-architecture silicon microstrip sensors. *IEEE Trans. Nucl. Sci.* **NS-48/2** (2001) 189–193

Kipnis, I., Spieler, H. and Collins, T. (1994). An analog front-end bipolar-transistor integrated circuit for the SDC silicon tracker. *IEEE Trans. Nucl. Sci.* **NS-41/4** (1994) 1095–1103

Kondo, T. *et al.* (1984) Radiation damage test of silicon microstrip detectors. *Proceedings of the 1984 Summer Study on the Design and Utilization of the Superconducting Super Collider, Snowmass, CO, Jun 23 – Jul 13, 1984*. (eds. R. Donaldson, J.C. Morfin) pp. 612–614. QCD184 S7 1984

Konobeyev, A. (1992). Neutron displacement cross-sections for structural materials below 800 MeV. *J. Nucl. Mat.* **186** (1992) 117–130

Kramberger, G. *et al.* (2002). Effective trapping time of electrons and holes in different silicon materials irradiated with neutrons, protons and pions. *Nucl. Instr. and Meth.* **A481** (2002) 297–305

Kraner, H.W. (1982). Radiation damage in semiconductor detectors. *IEEE Trans. Nucl. Sci.* **NS-29/3** (1982) 1088–1100

Krasel, O. *et al.* (2004). Measurement of trapping time constants in proton-irradiated silicon pad detectors. *IEEE Trans. Nucl. Sci.* **NS-51/6** (2004) 3055–3062

Kubota, M. *et al.* (1991). Radiation damage of double-sided silicon strip detectors. *Conference Record of the 1991 IEEE Nuclear Science Symposium and Medical Imaging Conference, Nov. 2 – 9, 1991, Santa Fe, New Mexico*. Vol. 1, p. 246

Kuznetsov, V.I. *et al.* (1975). Inversion of the type of conduction and stabilization of the Fermi level in irradiated silicon. *Sov. Phys. Semicond.* **9/4** (1975) 491–492, *Fiz. Tekh. Poluprovodn.* **9** (1975) 749–752

Lemeilleur, F. *et al.* (1982). Neutron-induced radiation damage in silicon detectors. *IEEE Trans. Nucl. Sci.* **NS-39/4** (1992) 551–557

Li, Z. *al* (1992). Investigation of the oxygen-vacancy (A-center) defect complex profile in neutron irradiated high resistivity silicon junction particle detectors. *IEEE Trans. Nucl. Sci.* **39/6** (1992) 1730–1738

Li, Z. (1994). Modeling and simulation of neutron induced changes and temperature annealing of N_{eff} and changes in resistivity in high resistivity silicon detectors. *Nucl. Instr. and Meth.* **A342** (1994) 105–118

Li, Z. *et al.* (2002). Novel prototype Si detector development and processing at BNL. *Nucl. Instr. and Meth.* **A478** (2002) 303–310

Lindström, G., Moll, M. and Fretwurst, E. (1999). Radiation hardness of silicon detectors – a challenge from high-energy physics. *Nucl. Instr. and Meth.* **A426** (1999) 1–15

Lindström, G. (2000). *Displacement Damage in Silicon.* http://sesam.desy.de/members/gunnar/Si-dfuncs.html

Lindström, G. for the RD48 Collaboration (2001). *Radiation hard silicon detectors – developments by the RD48 (ROSE) collaboration.* Nucl. Instr. and Meth. **A466** (2001) 308–326

Messenger, G.C. and Ash, M.S. (1986) *The Effects of Radiation on Electronic Systems.* van Nostrand Reinhold, New York. ISBN 0-442-25417-2, TK7870.M4425

Messenger, S.R. *et al.* (2004). Limits to the application of NIEL for damage correlation. *IEEE Trans. Nucl. Sci.* **NS-51/6** (2004) 3201–3206

Ma, T.P. and P.V. Dressendorfer (1989). *Ionizing Radiation Effects in MOS Devices and Circuits.* Wiley, New York, ISBN 0-471-84893-X, TK7871.99.M44I56

Moll, M. (1999) *Radiation Damage in Silicon Particle Detectors.* PhD thesis, University of Hamburg, 1999. DESY THESIS-1999-040, ISSN-1435-8085. Available at http://mmoll.home.cern.ch/mmoll/thesis/

Moll, M. (2003). Development of radiation hard sensors for very high luminosity colliders – CERN-RD50 project. *Nucl. Instr. and Meth.* **A511** (2003) 97–105

Oldham, T. (1999). *Ionizing Radiation Effects in MOS Oxides.* World Scientific, Singapore. ISBN 981-02-3326-4

Ohsugi, T. *et al.* (1988). Radiation damage in silicon microstrip detectors. *Nucl. Instr. and Meth.* **A265** (1988) 105–111

Palmieri, G.P. *et al.* (1998). Evidence for charge collection recovery in heavily irradiated silicon detectors operated at cryogenic temperatures. *Nucl. Instr. and Meth.* **A413** (1998) 475–478

Palmieri, V.G. *et al.* (2003). Cryogenic semiconductor high-intensity radiation monitors. *Nucl. Instrum. and Meth.* **A510** (2003) 97–100

Parker, S., Kenney, C.J. and Segal, J. (1997). 3-D: a new architecture for solid-state radiation detectors. *Nucl. Instrum. and Meth.* **A395** (1997) 328–343

Pease, R.L. (1983). Total dose effects in recessed oxide digital bipolar microcircuits. *IEEE Trans. Nucl. Sci.* **NS-30** (1983) 4216–4223

Pease, R.L. (2003). Total ionizing dose effects in bipolar devices and circuits. *IEEE Trans. Nucl. Sci.* **NS-50/3** (2003) 539–551

Pitzl, D. *et al.* (1992). Type inversion in silicon detectors. *Nucl. Instr. and Meth.* **A311** (1992) 98–104

Radeka, V. *et al.* (1993). JFET monolithic preamplifier with outstanding noise behavior and radiation hardness characteristics. *IEEE Trans. Nucl. Sci.* **NS-40/4** (1993) 744–749

RD48 (1999), 3rd Status Report. CERN/LHCC 2000-0089. http://rd48.web.cern.ch/rd48/

Rogalla, M. *et al.* (1997). Radiation studies for GaAs in the ATLAS inner detector. *Nucl. Instr. and Meth.* **A395** (1997) 45–48

Saks, N.S. *et al.* (1984). Radiation effects in MOS capacitors with very thin oxides at 80°K. *IEEE Trans. Nucl. Sci.* **NS-31/6** (1984) 1249–1255

Seller, P. *et al.* (1995). RAL-TR-95-055 and ATLAS SCT Technical Proposal Backup Document. ATLAS INDET-NO-085, 1995 (CERN)

Snoeys, W. *et al.* (2000). Integrated circuits for particle physics experiments. *IEEE J. Solid State Circuits* **35/12** (2000) 2018–2030

Spencer, E. *et al.* (1995) A fast shaping low power amplifier-comparator integrated circuit for silicon strip detectors. *IEEE Trans. Nucl. Sci.* **NS-42/4** (1995) 796–802

Spieler, H. (1994). Analog front-end electronics for the SDC silicon tracker. *Nucl. Instr. and Meth.* **A342** (1994) 205–213

Srour, J.R. *et al.* (1979). Radiation damage coefficients for silicon depletion regions. *IEEE Trans. Nucl. Sci.* **NS26/6** (1979) 4784

Srour, J.R. *et al.* (1984). *Radiation Effects on and Dose Enhancement of Electronic Materials.* Noyes Publications, Park Ridge, ISBN 0-8155-1007-1, TK7870.R318

Stephen, J.H. (1985). Low noise field effect transistors exposed to intense ionizing radiation. *IEEE Trans. Nucl. Sci.* **NS-33/6** (1986) 1465–1470

Summers, G.P. *et al.* (1993). Damage correlations in semiconductors exposed to gamma, electron and proton radiations. *IEEE Trans. Nucl. Sci.* **NS-40** (1993) 1372–1379

Tedja, S. *et al.* (1992). Noise spectral density measurements of a radiation-hardened CMOS process in the weak and moderate inversion. *IEEE Trans. Nucl. Sci.* **NS-39/4** (1992) 804–808

Troncon, C. (2004). Radiation hardness performance of ATLAS pixel tracker. *Nucl. Instr. and Meth.* **A530** (2004) 65–70

Tsveybak, I. *et al.* (1992). Fast neutron-induced changes in net impurity concentration of high-resistivity silicon. *IEEE Trans. Nucl. Sci.* **NS-39/6** (1992) 1720–1729

Turala, M. for the ATLAS SCT Collaboration (2001). The ATLAS Semiconductor Tracker. *Nucl. Instrum. and Meth.* **A466** (2001) 243–254

Unno, Y. *et al* (1996). Characterization of an irradiated double-sided silicon strip detector with fast binary readout electronics in a pion beam. *IEEE Trans. Nucl. Sci.* **NS-43/3** (1996) 1175–1179

Unno, Y. *et al* (1997). Beam test of a large area n-on-n silicon strip detector with fast binary readout electronics. it IEEE Trans. Nucl. Sci. **NS-44/3** (1997) 736–742

Weilhammer, P. (1985). Experience with Si detectors in NA32. *Proceedings of the Workshop on New Solid State Devices for High Energy Physics. Lawrence Berkeley Laboratory, Oct. 28–30, 1985*. pp. 83–115. LBL-22778

Wheadon, R. *et al.* (1994). Radiation tolerance studies of silicon microstrip detectors for the LHC. *Nucl. Instr. and Meth.* **A342**(1994) 126–130

Witczak, S.C. *et al.* (1998). Space charge limited degradation of bipolar oxides at low electric fields. *IEEE Trans. Nucl. Sci.* **NS-45** (1998) 2339–2351

Wolf, S. (1990). *Silicon Processing for the VLSI Era, Volume 2 – Process Integration*. Lattice Press, Sunset Beach, ISBN 0-961672-4-5

Wolf, S. (1995). *Silicon Processing for the VLSI Era, Volume 3 – The Submicron MOSFET*. Lattice Press, Sunset Beach, ISBN 0-961672-5-3

Wolf, S. (2002). *Silicon Processing for the VLSI Era, Volume 4 – Deep-Submicron Process Technology*. Lattice Press, Sunset Beach, ISBN 0-9616721-7-X

R. Wunstorf. (1992). *Systematische Untersuchungen zur Strahlenresistenz von Silizium-Detektoren für die Verwendung in Hochenergiephysik-Experimenten.* PhD thesis, University of Hamburg, 1992, DESY FH1K-92-01

Van Lint, V.A.J. *et al.* (1980). *Mechanisms of Radiation Effects in Electronic Materials*. John Wiley and Sons, New York. ISBN 0-471-04106-8, TK7871.M44

Ziock, H.-J. *et al.* (1994). Temperature dependence of the radiation induced change of depletion voltage in silicon PIN detectors. *Nucl. Instr. and Meth.* **A342** (1994) 96–104

8

DETECTOR SYSTEMS

8.1 Conflicts and compromises

System design is invariably the result of balancing conflicting considerations. For example, in a large tracking detector we desire:

1. low mass to reduce scattering,
2. low noise,
3. fast response,
4. low power,
5. radiation tolerance.

To reduce mass one can consider thinning the sensor. Thinning the sensor also improves radiation tolerance at high fluences where type inversion is important. However, thinning the sensor reduces the signal, so electronic noise must be reduced. Lower noise, in turn, requires more power. Increased power incurs more mass in the cooling systems and cabling (to limit ohmic losses), so the reduction in sensor material may well be outweighed by the increased mass incurred in cooling and powering.

Radiation tolerance can also be improved by reducing electronic noise, to maintain signal-to-noise ratio as signal levels decrease. Faster shaping times also reduce the sensitivity to shot noise associated with detector leakage current, but increase the voltage noise contribution, so this also comes at the expense of power when the signal-to-noise ratio is to be maintained. As discussed in Chapter 6 reducing noise can drive up power significantly, so when power is critical, noise requirements must be scrutinized carefully. In these situations it is important to design for adequate noise, rather than minimum noise.

Immunity to external pickup is important in maintaining the overall noise level. A fully shielded system would provide insensitivity to pickup, but the added material would be unacceptable. Techniques exist to reduce susceptibility to extraneous signals without resorting to massive shielding. This is discussed in Chapter 9.

Complex systems require compromises and the choices often carry technical risk, since the technically ideal solutions are hardly ever practical. There is no single way to deal with these conflicting requirements. Detectors that serve the same purpose and work in the same environment often adopt very different designs, as demonstrated by CDF and D∅ at the Tevatron or ATLAS and CMS at the LHC. Individuals' specific experience and personal tastes (or prejudices) often play as much a role as technical considerations. Since the success of large

systems depends on the combination of many components, they are also quite resilient; shortcomings in one area can be mitigated by strengths in another. The following sections describe some large detector systems in high-energy physics. They are not a complete catalog of systems that have been built and operated successfully; instead they have been chosen to illustrate different techniques.

8.2 Design considerations

Several primary considerations enter into the conceptual design of a detector system:

1. detector geometry.
2. efficiency
3. event rate
4. readout
5. support structures, cabling, and cooling
6. cost

These aspects cannot be optimized simultaneously.

8.2.1 *Detector geometry*

How much solid angle should the detector subtend and how is this achieved most efficiently? Since semiconductor sensors cannot be made arbitrarily large, this usually implies some form of tiling. Typically, the edges of the detector incur some dead area, where charge collection is impaired. Furthermore, gaps between individual tiles (or detector modules) reduce coverage. The electronics associated with each tile tend to incur additional dead area, one example being connections. In detectors for high energy charged particles the dead regions and gaps between tiles can be filled by overlapping adjacent tiles. In visible light and x-ray detectors overlapping is not a solution, as photons are absorbed in the dead regions and lost for detection. A common solution in photon imagers is "dithering", where the portions of the image falling on dead regions are shifted to active regions. This incurs additional time, but also provides redundancy for large portions of the image.

8.2.2 *Efficiency*

How much signal is required for efficient detection? Whether one seeks to merely recognize a signal or measure its magnitude, this depends on the required signal-to-noise ratio. The signal charge depends on the thickness of the sensor, the density of the material, and the bandgap. However, the equivalent noise charge depends on the capacitance, so the dielectric constant of the sensor material and the electrode geometry are also important.

8.2.3 *Event rate*

It is tempting to deal with high event rates by increasing the speed of the electronics. However, as shown in Chapter 6 this incurs an increase in power and

often an unavoidable increase in noise. An efficient tool in dealing with high rates is segmentation, which reduces the rate per channel. Segmentation also has other advantages, as will be discussed below.

8.2.4 *Readout*

Techniques to optimize electronic noise and power have been analyzed in previous chapters. However, more important is the choice of readout architecture. Rather than starting with a wish list of nice features, readout design should begin with a requirements document that prioritizes performance requirements and relates them quantitatively to design specifications. "Feature creep" has been the downfall of many designs, which end up burdened with circuitry that isn't really needed, but adds complexity and drives up design time. Then little time remains for initial system tests to verify whether the design is sound. Different approaches often provide comparable results, so good systems design is more important than the "optimum" readout choice.

8.2.5 *Support structures, cooling, and cabling*

Large systems typically include many detector modules with mechanical supports. In imaging systems positional precision and stability are important. Faint light imaging systems, for example, typically operate at temperatures of 80 – 150 K, so mismatches in coefficients of thermal expansion (CTE) and possible hysteresis must be considered when cycling to room temperature and back. Even with power efficient readout circuitry the total power in room temperature operation typically requires cooling systems. Gas cooling invariably comes to mind when reduction of material is crucial, but this is limited to small systems and most large detectors have adopted some form of liquid cooling. Finally, cabling is required for the readout. Data can be read out via metal conductors or optical fibers, but providing power in large systems requires metal conductors with a suitable cross-section to avoid voltage drops with significant power dissipation in the cables. The current draw depends on the required electronic noise level, which is determined by sensor capacitance and the shaping time, so electronic and mechanical design are closely linked (although in reality the two camps tend to work in parallel universes).

8.2.6 *Cost*

The last item, cost, affects all other considerations. In large systems cost containment is a recurring theme. This is closely related to technical robustness. For small detectors a finely tuned design may be quite acceptable. In large systems, however, designs must accommodate parameter variations, as "tweaking" or rework to ensure acceptable performance will require excessive time and drive up engineering and testing costs. The expense of large systems lies not just in the quantity of material and devices, but also in much more demanding engineering. Small systems can usually be made to work with last minute tweaks and all-out efforts when crises arise. In large systems the time and effort required are prohibitive. Robust designs and rigorous testing are crucial for efficient produc-

tion, but require more thorough preparation and "up-front" engineering. When solutions to new requirements have yet to be found, focussed R&D programs have proven to be very effective. Even after laying a sufficient foundation for a new detector, R&D programs should continue, as further innovations can often be incorporated into an ongoing project – oxygenated silicon detectors or "deep-submicron" CMOS being good examples. Detector R&D programs should be experiment oriented, but not project driven, as the schedule pressures of a project militate against exploring new directions.

8.3 Segmentation

Segmentation, subdividing the detector into many readout channels, is a very powerful tool. Highly integrated readout circuitry is a crucial ingredient in applying segmentation, as circuitry can be compact to match small electrodes. Moreover, modern IC technology allows more efficient circuitry than discrete designs. Segmentation improves both rate capability and electronic noise.

If a detector is exposed to a uniform particle rate R, subdividing the detector into N segments reduces the rate per readout channel to R/N. For example, at the LHC a typical detector accepts about 10^{11} tracks per second. Segmenting one layer into 10^6 electrodes reduces the hit rate to $10^5\,\mathrm{s}^{-1}$, which is much more manageable. Segmentation also aids in distinguishing multiple tracks emitted simultaneously into a small solid angle (jets).

Segmentation reduces the area per electrode, which reduces capacitance and electronic noise. This can be exploited to improve sensitivity or energy resolution, but can also be used to improve rate capability. If the achievable electronic noise is lower than required, the shaping time can be reduced, which improves rate capability.

Segmentation also improves radiation resistance. The leakage current flowing into an electronic channel is proportional to electrode area, so increasing segmentation reduces the shot noise. Furthermore, reduced electronic noise by virtue of reduced capacitance allows a greater signal degradation from reduced detection efficiency.

As shown in Section 6.7 the total power dissipation in the front-end can be reduced by segmentation, although this is usually balanced by the additional circuitry per cell. Nevertheless, the power dissipation per unit area of the ATLAS pixel detector with roughly 10^8 channels is about the same as in the ATLAS silicon strip tracker with $6 \cdot 10^6$ channels. Since the power required for the "back-end" circuitry in each channel is roughly independent of segmentation, whereas the frontend power scales inversely with the square of capacitance, the total power assumes a minimum when the two are equal.

The following sections present examples of strip and pixel detectors for different applications. This is not a complete survey. Instead, the examples have been selected to illustrate key design considerations and different design approaches.

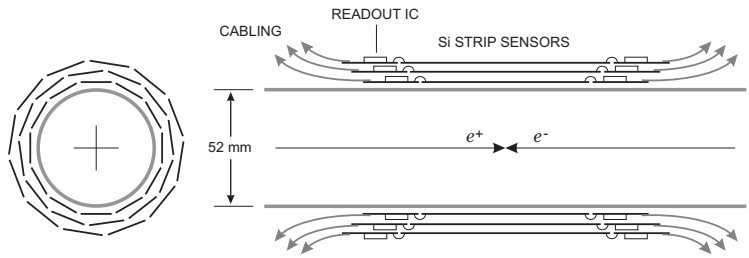

FIG. 8.1. Layout of the Mark II silicon strip vertex detector, showing the collision point and the 52 mm diameter beam pipe.

8.4 Tracking and vertex detectors at e^+e^- colliders

Lepton colliders have the virtue of "simple" event topologies, free of the large backgrounds inherent to hadron interactions. However, the momenta of the particles of interest are frequently quite low, so minimizing material in the active volume is crucial to limit multiple scattering. This also applies to TeV collisions as envisioned for a future Linear Collider. The interesting physics processes tend to feed multijet final states, so the average energy of particles in jets is about $1 - 2$ GeV, not much higher than at LEP, LEP2 or the SLC (Battaglia 2004). Lepton colliders typically have a well-localized interaction region, especially in linear colliders. The interaction volume at the SLC, for example, was only 10 μm in diameter and extended 1.5 mm along the beam axis.

8.4.1 *Layout and detector geometry*

The need to reduce material led to a nearly universally used arrangement, where only the sensors and a low-mass support structure are in the active region. Long rectangular detector modules ("ladders") form "barrels" concentric to the beam axis. The electronics are situated at the outer ends of the barrels, outside the active tracking region. This also places the cooling and cabling outside the active region. The maximum length of a sensor is limited by available wafer diameter, which was 100 mm in the past, but now has increased as 150 mm wafers have become available. Multiple sensors are ganged together to form long ladders of $20 - 30$ cm length or more.

The Mark II silicon strip vertex detector (Adolphsen *et al.* 1992) pioneered silicon vertex detectors at e^+e^- colliders and was the first to utilize custom designed integrated circuits for the readout. The basic arrangement is shown in Figure 8.1. Similar arrangements were utilized by a series of designs at LEP (for some surveys see Schwarz 1994a, 1994b, and Österberg 1999) and in CLEO at Cornell (Kass *et al.* 2003). Different construction techniques and new technologies were explored. Aleph introduced double-sided detectors (Batignani *et al.* 1993) and DELPHI pioneered double-sided detectors with integrated coupling capacitors and bias networks (Chabaud *et al.* 1996). OPAL (Allport *et al.* 1994)

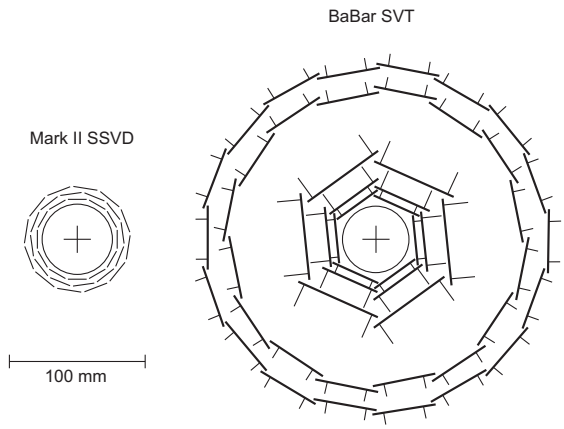

FIG. 8.2. Axial views of the Mark II Silicon Strip Vertex Detector (SSVD) and the BaBar Silicon Vertex Tracker (SVT). The central cross indicates the beam axis.

and L3 (Alpat *et al.* 1992) also contributed to the wealth of experience that current detector systems build on. Weilhammer (1994) gives an overview of silicon vertex detector technology and results at LEP. Meanwhile, vertex detectors have increased in scope to include tracking functions. Figure 8.2 compares the Mark II and the more recent BaBar detector (Bozzi 2000). The Mark II detector had gaps between adjacent modules in a layer. More advanced designs provide full coverage by overlapping adjacent modules. The redundant position information in the overlap region also facilitates *in situ* position calibration using stiff tracks.

Originally, silicon detectors at e^+e^- colliders were designed purely as vertex detectors close to the beam pipe, augmenting wire chambers or other coarse resolution tracking detectors at larger radii that provided tracking and pattern recognition. Development of large silicon tracking detectors for hadron colliders, such as the SDC Silicon Tracker (Seiden 1994), spurred the design of combined vertex-tracking detectors for e^+e^- colliders, for example the SVT in the BaBar Detector at SLAC (Re *al.* 2003).

The primary goal of this detector and its counterpart in Belle (Ushiroda 2003, Taylor 2003) is to measure B mesons from $\Upsilon(4S)$ production. These have very low momentum, so asymmetry in colliding-beam energies (9 GeV e^- on 3.1 GeV e^+ at PEP) is used to provide a relativistic boost ($\beta\gamma = 0.56$) that allows conventional vertex detectors to cope with the short B meson lifetime. In contrast to other collider detectors that require resolution primarily in $r\varphi$, *i.e.* normal to the beam axis, the vertex detector at an asymmetric B-factory must provide resolution in the boost direction, *i.e.* along the beam axis z.

The resolution requirements are not stringent. Less than 10% loss in precision is incurred in the asymmetry measurement if the separation of the B vertices is measured with a resolution of 1/2 the mean separation (250 μm in BaBar). Thus,

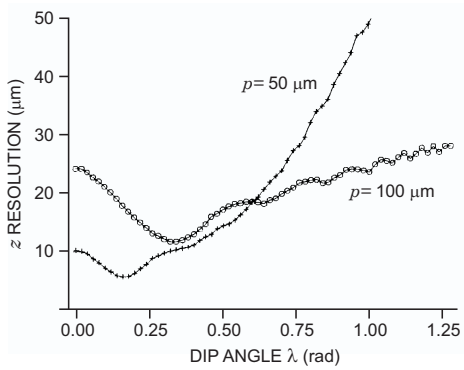

FIG. 8.3. z resolution for analog readout with interpolation *vs.* dip angle λ for strip pitches $p = 50$ and $100\,\mu$m and a signal-to-noise ratio of 20 at normal incidence (Lynch 1993).

a vertex resolution of $80\,\mu$m is adequate for both CP eigenstates and tagging final states. The resolution is multiple-scattering limited; the beam pipe alone introduces $0.6\%\,X_0$.

Both Belle and BaBar use double-sided strip sensors with orthogonal strips, where the z-strips provide vertex resolution and the $r\varphi$ strips are used for pattern recognition. Figure 8.3 shows the calculated resolution *vs.* dip angle for strip pitches of 50 and $100\,\mu$m. For dip angles > 0.6 rad the larger pitch yields better resolution. At large dip angles the signal is distributed over multiple strips, as illustrated in Figure 8.4, so the signal-to-noise ratio suffers. At normal incidence the signal is $Q_s = d \cdot (dE/dx)$, whereas at large dip angles a single strip only subtends a fraction of the track in the sensor, resulting in a smaller signal

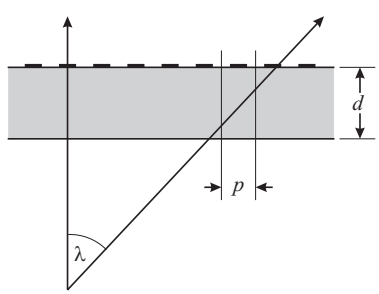

FIG. 8.4. As the dip angle λ increases, the track traverses an increasing thickness of silicon. When the track subtends more than the strip pitch the signal is distributed over multiple strips, so the signal captured by one strip decreases.

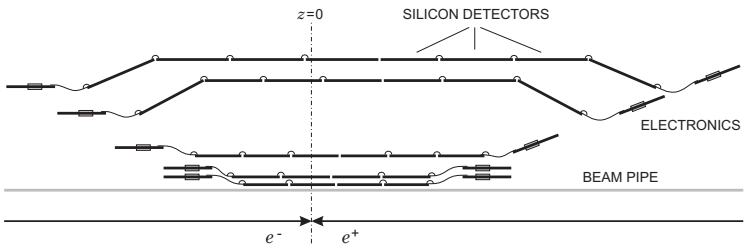

FIG. 8.5. Schematic layout of the layers in the BaBar Silicon Vertex Tracker.

$$Q_s = \frac{p}{\sin\lambda} \cdot \frac{dE}{dx} \qquad (\sin\lambda > p/d) \ . \tag{8.1}$$

For example, in the inner layers BaBar uses a 50 μm pitch, but reads out every second strip. In the central region capacitive interpolation approximates the resolution of a 50 μm pitch, whereas at large angles the pitch is effectively 100 μm.

The BaBar detector uses five layers of silicon to provide both vertexing and tracking information. The geometry is shown in Figure 8.5. Double-sided detectors are used throughout, with integrated coupling capacitors and polysilicon bias resistors (discussed in Appendix A). The inner three layers at 33, 40 and 59 mm radius are simple cylinders, whereas the outer two layers at 127 and 146 mm use a "lampshade" geometry to reduce the angle of incidence and make more efficient use of silicon area. This poses some challenges in wire bonding, but has been implemented successfully. As illustrated in Figure 8.2 the layers are polygons formed of detector modules. Modules in layers 1 and 2 use four sensors and layer 3 uses six. Layer 4 modules have seven sensors and layer 5 modules have eight. Electrically, each module is "split" midway with readouts at the opposite ends of the tracker. Sensors associated with each end are connected together by wire bonds to form contiguous strips. All sensors are rectangular, except for the end sensors in layers 4 and 5, which are trapezoidal. Overall, the detector includes 340 sensors with six different designs (details in Bozzi 2000).

Sensors in the barrel are glued edge-to-edge by dipping the sensor edges into glue and then transferring them to teflon jigs where they are aligned with a 75 μm gap and cured. The sensors are also structural members, with support ribs glued to the outer surface of the sensors to provide rigidity. The ribs are notched to accommodate the wire bonds. The layer 1 and layer 2 modules are joined to form a rigid system; the support ribs of layer 1 are glued to the inner surface of the layer 2 sensors. The ribs are laser cut and made of two carbon-epoxy outer layers with an intermediate layer of kevlar. The carbon-epoxy layers are conductive, so only the inner kevlar layer connects to the silicon. The carbon layers prevent the kevlar from deforming from possible absorption of moisture. Figure 8.6 shows a perspective view of the SVT with the support structure. The low-mass support structure is made of carbon composite tubing and all of the electronics are outside of the active region to minimize material.

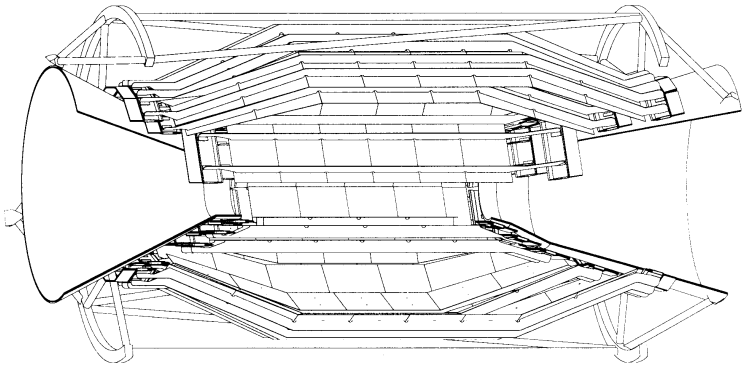

FIG. 8.6. The BaBar SVT with its support structure. The detector ladders are mounted on carbon cones at the two ends and read out from each end with the electronics outside the active volume. A carbon composite space frame provides overall support.

As noted above, the pitch of the z-strips in layers 1 through 3 is $50\,\mu$m, but only every second strip is read out. In the outer layers the pitch of the z-strips is $210\,\mu$m. Since these layers are quite long, the width of a module does not allow a readout line for each z-strip, so a subset of the strips with lower occupancy is ganged. This introduces some ambiguities in z position, which are resolved by overall pattern recognition. Some other detectors use monolithically integrated bussing for the z-strips, which appears more elegant (Chabaud *et al.* 1996). However, the dielectric constant of polyimide is lower then of SiO_2 and the achievable dielectric thickness is much greater, so the external polyimide bussing used in BaBar incurs significantly less capacitance and costs less.

The detector ladders are connected to the readout electronics through short lengths of polyimide flex cable, so that the electronics modules (called High Density Interconnects or HDIs) can be placed outside the active volume.

In each layer the $300\,\mu$m thick silicon detectors contribute 0.3% and the flex circuits and carbon composite support structure add 0.2% of a radiation length to the tracker material, comparable to the $0.6\%\,X_0$ of the beam pipe.

8.4.2 *Electronics*

The inner layers have the shortest strips and hence present the smallest capacitive load, but the occupancy is highest. Conversely, the outer layers present the largest capacitive load, but the smallest occupancy. The shaping times can be chosen appropriately, with smaller shaping times used in the inner than in the outer layers.

Radiation background is mostly due to lost beam particles, which are distributed in a narrow angular range. This drives both the readout speed and the requirement that the readout electronics withstand doses up to 3 Mrad.

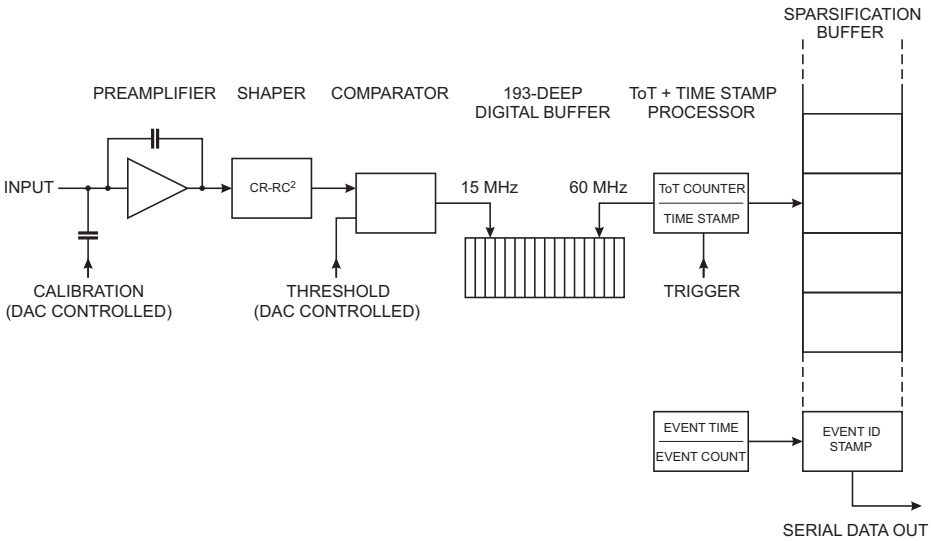

FIG. 8.7. Block diagram of one channel of the AToM readout IC of the BaBar Silicon Vertex Tracker.

Figure 8.7 shows a block diagram of the SVT readout IC, the AToM chip. Designed as a joint effort at LBNL, Pavia, and UCSC it has 128 channels per die and is fabricated in 0.8 μm radiation-hard CMOS (Kipnis *et al.* 1997). Bunches collide every 4.2 ns, much smaller than the required shaping time, so the readout is designed for a DC beam using continuous shaping. Each channel includes a preamplifier with continuous reset, a CR-RC^2 shaper with selectable shaping times of 100, 200, 300 and 400 ns, a threshold comparator, a digital pipeline, and data readout. Since the data field is sparse, the readout logic detects which channels are struck and suppresses the readout of empty channels. The AToM chip includes analog-to-digital conversion of the detector signals and all input and output signals are digital.

Input polarity selection allows the same IC to be used for both n- and p-strips. The capacitance of the n-strips is higher than of the p-strips. To optimize z-resolution in the inner layers the z-strips are on the p-side. In the two outer layers the arrangement is reversed. The measured noise $Q_n = 350\,e + 42\,e/\text{pF}$ at 100 ns, $Q_n = 333\,e + 35\,e/\text{pF}$ at 200 ns, and $Q_n = 306\,e + 28\,e/\text{pF}$ at 400 ns shaping time (Manfredi *et al.* 1999). The maximum strip length is 26 cm with a capacitance of 35 pF, so at 400 ns the noise is about 1300 e. The noise increases by about 10 – 20% after exposure to 2.4 Mrad. Power dissipation is 3.5 mW per channel.

The length of the digital pipeline accommodates the level 1 trigger latency of 12 μs. It is filled at a clock rate of 15 MHz to provide sufficient resolution for time stamping of events. The readout clock operates at 60 MHz to accommodate

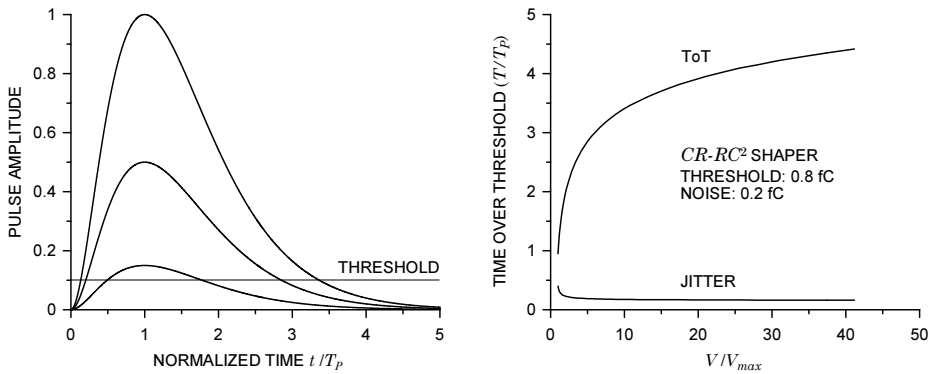

FIG. 8.8. Measuring the time the shaper output pulse exceeds a fixed threshold provides a simple digitization of pulse amplitude (left). The time is normalized to the peaking time of the pulse T_P. The calculated time over threshold for a CR-RC^2 shaper is shown at the right. The time jitter indicates the obtainable resolution.

the required readout time for a module. The input sample rate is also sufficient to digitize the pulse amplitude by recording the width of the comparator output, *i.e.* the time over threshold (ToT).

About three to four bits analog resolution suffice to provide adequate position resolution by interpolation, so BaBar uses the time over threshold to digitize the pulse height. The principle of the ToT measurement is shown in Figure 8.8. For a triangular pulse the time over threshold is a linear function of pulse height, whereas for a simple CR-RC shaper the rounded cusp and gradual decay provide a roughly logarithmic response, which is a good match to this application. The time over threshold measurement incurs no dead time and practically no additional circuitry. The analog circuitry must not accommodate the maximum pulse height, *i.e.* the circuit can limit, as long as the falling edge is correctly reproduced as it crosses the threshold. Figure 8.8 also shows the time jitter, which sets the ultimate resolution of the amplitude measurement. In BaBar the jitter is less than the sampling time, so the latter determines the amplitude resolution.

On-chip calibration circuitry injects charge into the input. A six-bit DAC sets the level and digital masks select which channels receive calibration signals. A second six-bit DAC sets the comparator threshold. The signal level at the comparator input is 100 mV/fC and the threshold can be set in 5 mV steps. The output of the comparator can be masked to suppress noisy or defective channels. Noise measurements are performed by scanning the threshold for a fixed calibration charge. This also calibrates the time-over-threshold measurement.

The 128-channel IC includes $3.3 \cdot 10^5$ transistors. Its dimensions are $5.7 \times 8.3\,\mathrm{mm}^2$ and the power dissipation is 3.5 mW per channel. Connections to the HDI include power (2 V and 5 V analog, 5 V digital, and detector bias), serial

lines for the clock, command, and data signals, and connections for temperature sensors on each HDI. All electrical signals are transferred as balanced lines. On layers 1 and 2 each HDI incorporates 14 readout ICs, Layer 3 has 20 ICs and layers 4 and 5 have five ICs.

The detectors use integrated coupling capacitors. To avoid potential breakdown the electronics are referenced to the detector bias potential, *i.e.* at $-30\,\mathrm{V}$ on the p-side and $+30\,\mathrm{V}$ on the n-side. All electrical cables must be carefully insulated to prevent any connection to "ground".

8.4.3 *"Common mode noise"*

Many detectors observe what is often called "common mode noise", which manifests itself as a baseline shift during data transfer off the detector. Typically, these systems use single-ended transfer links with a common return. During readout the increased current flow leads to a voltage drop on the return, which is superimposed on all output signals. This is especially critical in systems with analog readout. The current spikes associated with the transitions of the digital control signals give rise to a voltage drop between the detector module and the off-detector electronics, which also transfers to the analog signal. Many detector builders unwittingly design this flaw into their system, leading many to believe that it is unavoidable.

Unlike previous collider vertex detectors BaBar operates with an essentially continuous beam, so signal acquisition and readout proceed simultaneously. In the LEP detectors, for example, the time between collisions was ample to read out the detector. Building on prior designs that addressed similar requirements (Kipnis, Spieler, and Collins 1994, Spieler 1994, Ludewigt *et al.* 1994), both the electronics and SVT interconnection scheme (Eisner *et al.* 1997) were carefully designed and the BaBar SVT operates without discernible noise pickup. Figure 8.9 shows the connection scheme, which controls signal return paths and avoids unnecessary ground connections (Nyman 1996). A key criterion is that all electrical signals are transferred as balanced pairs in which the net current remains constant.

Connections to the HDI are made through 40 cm long high-density polyimide cables to a "matching card", which includes some filtering and interfaces to a more rugged ribbon cable. After 3 m an inline connector at the detector boundary facilitates disconnection for installation and maintenance. A further run of 12 m brings the cable to the "MUX Rack", where the digital signals are transferred to a 1.2 Mb/s optical link. Together with the power cables the fibers are routed an additional 30 or 40 m to the power supply racks in the main electronics area. The optical links are not necessary to prevent cross-talk, as the electrical signals traverse 15 m before being converted to optical, but the optical fibers save space.

The entire detector is enclosed in a thin aluminum Faraday shield, which is extended by the cable shields to the power supply racks. The shield also surrounds the beam pipe, implemented as a metallized polyimide layer to insulate the shield from the beam pipe. To provide local potential referencing the com-

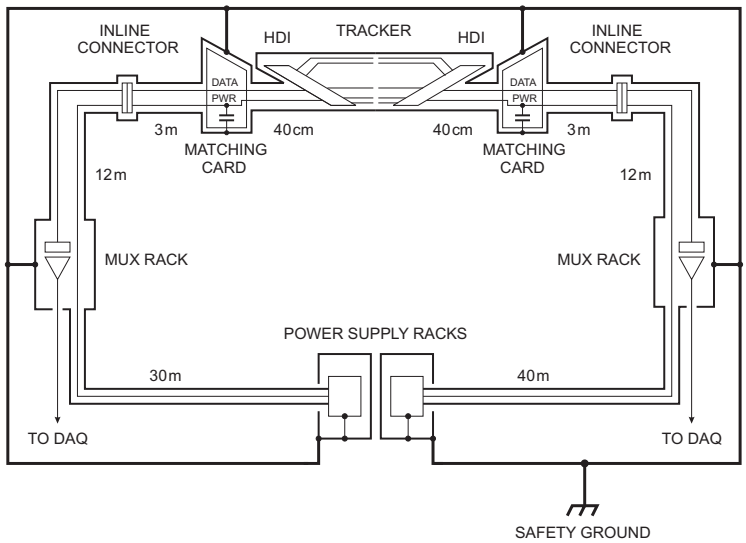

FIG. 8.9. Connection and cabling scheme of the BaBar SVT. (Adapted from Nyman 1996.)

mon connection of the power is capacitively coupled to the Faraday shield at the matching card. Since all currents are balanced, the current flow between the two halves of the detector is negligible, so no noticeable potential drop builds up between the two capacitively coupled reference connections in the matching cards. A safety ground (connecting the Faraday shield to earth ground) is provided through a separate cable, but this is independent of the signal transfer and has no effect on the integrity of the signal transfer. It is there only for safety. Implementing a "clean" system is challenging, not because the necessary techniques are unknown, but because the prevailing "grounding mythology" must be resisted in every design meeting and every step of the way. "Grounding" and its pitfalls are discussed more in Chapter 9.

8.4.4 *Noise limits in long strip detectors*

The desire to reduce mass in the active region has driven designs towards layouts that place the electronics outside the active region. This has been a key consideration at B-factories, where the resolution is multiple scattering limited, but will remain equally important at future linear colliders (Battaglia 2004). Here the active region will be longer than in current e^+e^- detectors, so this raises the question of how long a strip detector could provide the required signal-to-noise ratio. Long strips increase the capacitive load C at the preamplifier input, contributing to the noise charge as

$$Q_n^2 = i_n^2 F_i T_s + \frac{e_n^2}{C^2}\frac{F_v}{T_s} \ . \tag{8.2}$$

The noise level can be maintained by reducing the input noise voltage spectral density e_n, the shaper noise coefficient F_v, or increasing the shaping time T_S. An increase in shaping time T_S is limited by occupancy and by the strip's leakage current, which both grow with strip length. As shown in Chapter 4 the shaper noise coefficients are approximately one, so large improvements will not obtain by utilizing sophisticated shapers. The equivalent input noise voltage is determined by the input amplifier as discussed in Chapter 6 and by the thermal noise associated with the resistance of the strip electrodes, as noted in Chapter 3. Consider an FET input. As the strip capacitance increases the acceptable input capacitance of the amplifier increases, so its contribution to the equivalent input noise voltage decreases. Ultimately the thermal noise of the strip resistance will dominate. With increasing strip length the leakage current also increases, so it is useful to consider the noise when dominated by the sensor parameters leakage current, capacitance, and strip resistance.

The leakage current of the sensor scales with the area subtended by an electrode

$$I_b = I_0 pld \ , \tag{8.3}$$

where p is the strip pitch, l the length of an electrode and d the sensor thickness. I_0 is the leakage current per unit volume. The corresponding shot noise current

$$i_{ns}^2 = 2eI_b \equiv 2elI_b' \ . \tag{8.4}$$

Here and in the following primed quantities are normalized to the strip length. For a strip pitch of 50 μm a typical value is 1 nA per cm strip length, so for 50 cm long strips the shot noise current is 130 fA/$\sqrt{\text{Hz}}$.

The capacitance is the sum of the fringing capacitance and the capacitance to the backplane. The fringing capacitance C_{ss} is a logarithmic function of the ratio of strip width to pitch. Thus, we'll neglect the dependence on strip width and set the capacitance proportional to strip periphery $2(l+p)$. For strips $p \ll l$, $C_{ss} \approx 2lC_{ss}'$. For typical designs the fringing capacitance is about 1 pF per cm strip length, so for a 50 cm long detector the fringing capacitance is 50 pF. The capacitance to the backplane

$$C_b = \varepsilon\varepsilon_0 \frac{pl}{d} \ , \tag{8.5}$$

so the total strip capacitance $C_s = C_{ss} + C_b \equiv lC_s'$. For a strip pitch of 50 μm, a strip length of 50 cm, and a detector thickness of 300 μm the backplane capacitance is 8 pF, yielding a total strip capacitance of about 60 pF.

Increasing the strip pitch will increase the leakage current per channel, but have only a minor effect on the capacitance, until the strip pitch exceeds several hundred microns. Thus, the sensor parameters that determine electronic noise

drive the design towards small strip pitches. The reduction in strip pitch is limited by the width required by an electronic channel, which is about $50\,\mu$m.

The strip resistance introduces a thermal noise voltage $e_{nR}^2 = 4kTR \equiv 4kTlR'$. The strip resistance $R = \rho wtl$, where ρ is the resistivity of the electrode material, w the width of the electrode (not the pitch), and t is the electrode thickness. For sputtered aluminum the resistivity $\rho \approx 4\,\mu\Omega\,\mathrm{cm}$ and a practical limit to the deposition thickness is about $1\,\mu$m. Thus, for a metallization width of $20\,\mu$m the resistance is $20\,\Omega$ per cm length. For a 50 cm strip length the resistance is $1000\,\Omega$ with a thermal noise voltage of $4\,\mathrm{nV}/\sqrt{\mathrm{Hz}}$. This places a lower limit on the total noise voltage. At this voltage noise level both the voltage noise and capacitance contributed by the input amplifier can be relatively small.

Thus, for a 50 cm strip length the sensor parameters that enter into the electronic noise are $I_b = 50\,\mathrm{nA}$ ($i_{nd} = 130\,\mathrm{fA}/\sqrt{\mathrm{Hz}}$), $C_s = 60\,\mathrm{pF}$, and $e_{nr} = 4\,\mathrm{nV}/\sqrt{\mathrm{Hz}}$, so the amplifier input noise is negligible. At the optimum shaping time (equal current and voltage noise contributions)

$$T_S = \frac{e_n}{i_n} C \sqrt{\frac{F_v}{F_i}} \tag{8.6}$$

the equivalent noise charge

$$Q_n^2 = 2 e_n i_n C \sqrt{F_i F_v}\ . \tag{8.7}$$

Since $\sqrt{F_i F_v} \approx \sqrt{F_i/F_v} \approx 1$, we can determine the minimum noise due to sensor parameters alone,

$$Q_n^2 \approx 2 e_{nr} i_{nd} C = 2lC' \sqrt{(4kTlR')(2q_e l I_b')} = 2l^2 C' \sqrt{(4kTR')(2q_e I_b')}\ . \tag{8.8}$$

If the shaping time is adjusted optimally as the strip length increases, the noise charge increases proportionally to strip length. For the example parameters used above the noise due to the sensor alone is $1600\,e$ at a shaping time of $2\,\mu$s.

If, on the other hand, the shaping time is scaled with strip length to maintain occupancy

$$T_S = \frac{T_S'}{l}\ , \tag{8.9}$$

the noise charge

$$Q_n^2 = i_n^2 T_S F_i + \frac{e_n^2 C^2}{T_S} F_v \approx i_n^2 \frac{T_S'}{l} + e_n^2 \frac{(lC')^2}{T_S'/l} = (2q_e l I_b')\,\frac{T_S'}{l} + (4kTlR')\,l^3 \frac{C'^2}{T_S'}$$

$$Q_n^2 \approx 2q_e I_b' T_S' + (4kTR')\,l^4 \frac{C'^2}{T_S'} \tag{8.10}$$

increases with the square of the length.

8.4.5 *CCD detectors at e^+e^- colliders*

Although not essential for the relatively simple event topologies and "clean" background environment at e^+e^- colliders, two-dimensional detectors were introduced early on. Double-sided detectors with orthogonal strips were utilized in ALEPH (Batignani *et al.* 1993). The design placed interconnect structures and electronics in the active region, which added substantially to multiple scattering. However, the two-dimensional position information did provide "x-ray" images showing the capacitors and other components on the hybrid circuit that covered the sensors (Schwarz 1994b).

Despite the prevalence of strip detectors, truly two-dimensional devices were already utilized in the early 1980s. CCDs provide nonprojective two-dimensional sensing without the multihit ambiguities encountered in crossed-strip arrays and can be read out with minimal additional material in the active region. However, they do have to be cooled to about 200 K, so the required cryostat adds mass, albeit at the outer radius. CCDs were first used in fixed target experiments (Damerell 1981, Bailey 1983). Standard commercial CCDs were limited in size ($< 1\,\mathrm{cm}^2$), but as larger devices became available it became practical to construct vertex detectors with full coverage and minimum mass. The most mature and powerful CCD detector, the VXD3, was installed and operated successfully at SLD (Abe *et al.* 1997). Two CCDs form a ladder with an active length of 16 cm, allowing the readout electronics to be placed outside the active region. A three-layer array of 96 CCDs provided $< 6\,\mu$m position resolution in both $r\varphi$ and z. The double track resolution of $20\,\mu$m provided excellent pattern recognition and allowed the detector to cope with the high track density at small radii.

Most CCD structures are designed for optical sensing with shallow sensitive regions. In the VXD3 charge is collected from about $20\,\mu$m depth, so the signal from minimum ionizing particles is about $1200\,e$, much smaller than in the commonly used $300\,\mu$m thick strip detectors. However, as the charge is deposited on a very small capacitance, the resulting voltage signal is large, so a small equiva-

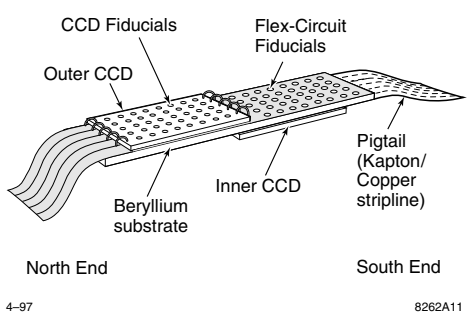

FIG. 8.10. Arrangement of a VXD3 detector ladder. (Figure courtesy of C.J.S. Damerell.)

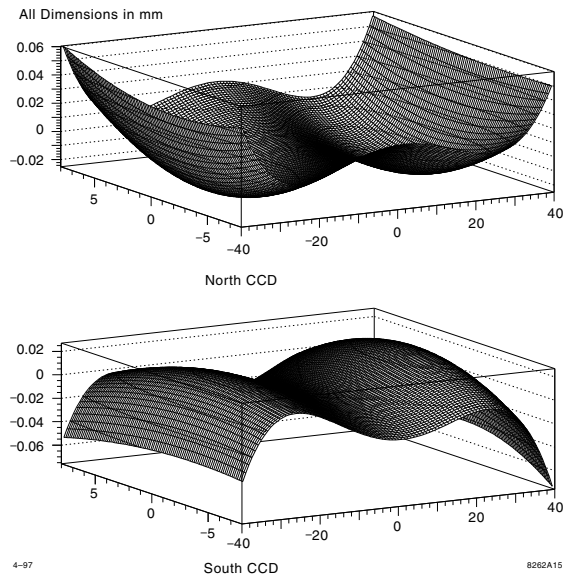

FIG. 8.11. Deviations from planarity of mounted CCDs in VXD3. (Figure courtesy of C.J.S. Damerell.)

lent noise charge can be achieved with rather noisy electronics. In the VXD3 the charge conversion is $3\,\mu\mathrm{V}/e$ and at a readout frequency of 5 MHz the equivalent noise charge is about $100\,e$, yielding an adequate signal-to-noise ratio. The well capacity is about $4 \cdot 10^5\,e$, so the dynamic range is ample. Although the shallow collection depth appears as a handicap with respect to signal charge, it is advantageous for track reconstruction at shallow incidence, as the signal is distributed over fewer pixels and is less sensitive to fluctuations in the spatial distribution of charge clusters. Essentially, the VXD3 CCDs provide $(20\,\mu\mathrm{m})^3$ space points. Thus, a simple cylindrical geometry could be used to provide wide polar angle coverage.

VXD3 used large CCDs with an active area of $80 \times 16\,\mathrm{mm}^2$, so with a pixel size of $20\,\mu$m the die accommodates 800×4000 pixels. The width of the dead area along the edges was $< 300\,\mu$m. A protective layer of $2\,\mu$m thick polyimide coated the top surface of the CCDs. The CCDs were thinned to $180\,\mu$m and supported by a thin beryllium slab, $21\,\mathrm{cm} \times 1.6\,\mathrm{cm} \times 0.38\,\mathrm{mm}$. Figure 8.10 illustrates the ladder structure. The two CCDs were mounted on opposite surfaces, providing a pathway for the polyimide ribbon cables carrying clock and control lines. The total material is $4.05 \cdot 10^{-3} X_0$, with the beryllium slab contributing $1.08 \cdot 10^{-3} X_0$, the CCDs + adhesive $1.60 \cdot 10^{-3} X_0$, the two polyimide cables and adhesive $0.47 \cdot 10^{-3} X_0$, and the $17.8\,\mu$m copper traces, averaged over the whole area,

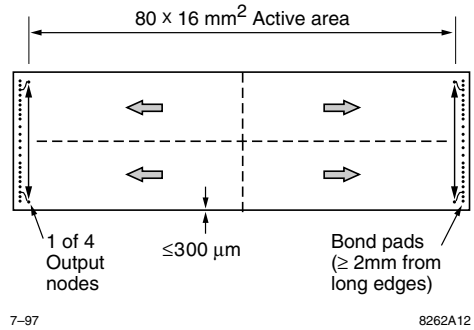

FIG. 8.12. The VXD3 CCDs are read out by quadrants through separate outputs to reduce the readout time. (Figure courtesy of C.J.S. Damerell.)

adding $0.90 \cdot 10^{-3} X_0$. Figure 8.11 shows the measured deviations from planarity of two CCDs in a representative ladder after assembly.

The basic principle of a CCD was described in Chapter 1. Charge deposited in a pixel is transferred by shifting a potential well from one pixel to the next. Thus, the readout is sequential. After reaching the end of a row the signal charge Q_S is deposited in a readout register, which transfers the charge to the readout node, where it is deposited on a capacitance C. The resulting voltage Q_S/C is sensed by an output amplifier. The time required to read out a column of N_{row} pixels is $N_{row} \cdot f_{row}$, where f_{row} is the image transfer frequency. The readout register must be clocked at a higher frequency $> M \cdot f_{row}$, as the charge from all M columns must be transferred to the output before the signal from the next row may be transferred to the readout register. The same process is repeated for all columns, so the total readout time for an $N \times M$ array is $M \cdot N \cdot f_{col}$.

In VXD3 the readout time was reduced by subdividing the structure into four quadrants, so each readout channel serves 400×2000 pixels, as indicated in Figure 8.12. The image columns use three-phase clocking, to allow charge transfer in both directions. The readout register always reads in one direction, so a two-phase clock was used. The readout time for the full image was 200 ms. As the SLD beam crossing period was 8.3 ms, the readout of an entire frame extended over 25 beam crossings, which introduced some background from undesired crossings. The occupancy was low, so the well-localized interaction point coupled with the excellent pattern recognition allowed the desired events to be separated from the background. A full readout was only initiated for trigger events. Since the time required to generate the trigger was 5 ms, whereas the crossing time was 8.3 ms, the readout sequence was initiated at every beam crossing. In the absence of a trigger all digitized events were discarded and the CCD was flushed by a fast clear. Since this does not require clocking the readout registers, one row shift requires only 10 μs, rather than 100 μs with the full readout. Subsequently, the read register is cleared. In the local front-end electronics the CCD outputs were

digitized in parallel in eight-bit flash ADCs and the combined data sent through 960 MHz optical links to the off-detector electronics for image processing (Figure 8.13). Figure 8.14 shows how the vertex detector is embedded in the overall detector. Figure 8.15 shows an assembled half-shell prior to installation.

Next generation CCD detectors are candidates for the proposed International Linear Collider (ILC). This will require a reduction in material and an increase in readout speed. A modest increase in CCD size (*e.g.* to 125 mm length) would allow the ladders to be implemented with two CCDs. The CCD substrates can be thinned to $< 100\,\mu$m and low-mass mounting schemes are being considered that could bring the material down to the range $0.01 - 0.02X_0$ for a five-layer vertex–tracker with $8 \cdot 10^7$ pixels (Damerell 2001). However, for the bunch structure under discussion the readout speed is a major challenge. One obvious approach is to add more readout channels, ultimately with one readout per column. A higher readout rate incurs a reduction in shaping time with a corresponding increase in electronic noise, so optimization of signal-to-noise ratio is important. This depends both on CCD parameters and the electronic readout circuitry.

Similarly to integrated circuits, typical CCDs are fabricated on low resistivity substrates with a thin high quality epitaxial layer of higher resistivity at the surface. The devices used in VXD3 were fabricated on $20\,\mathrm{m}\Omega\,\mathrm{cm}$ p-type substrates with a $20\,\mu$m thick epitaxial layer of $20\,\Omega\,\mathrm{cm}$. Depleting the epitaxial layer would require a reverse bias of 200 V, while clock levels are about 10 V,

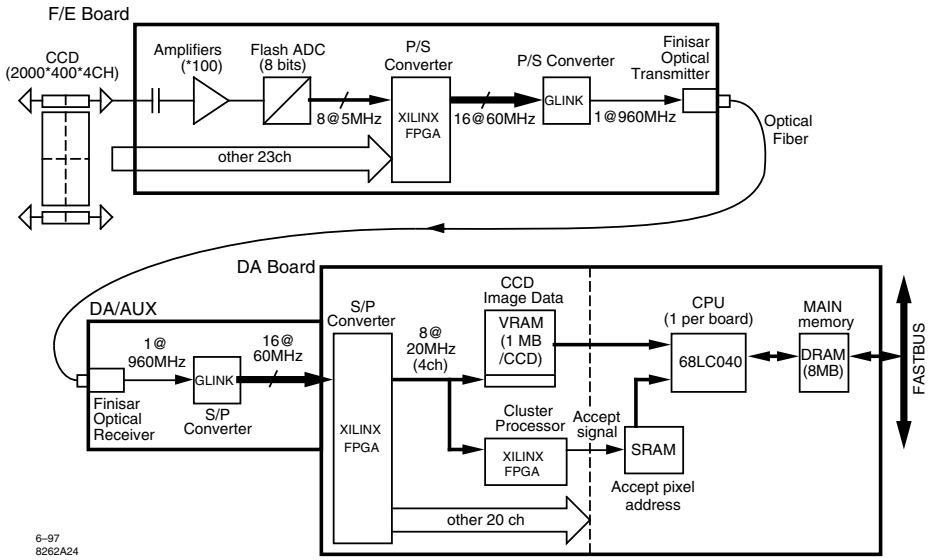

FIG. 8.13. Readout electronics of the VXD3 CCD vertex detector. (Figure courtesy of C.J.S. Damerell.)

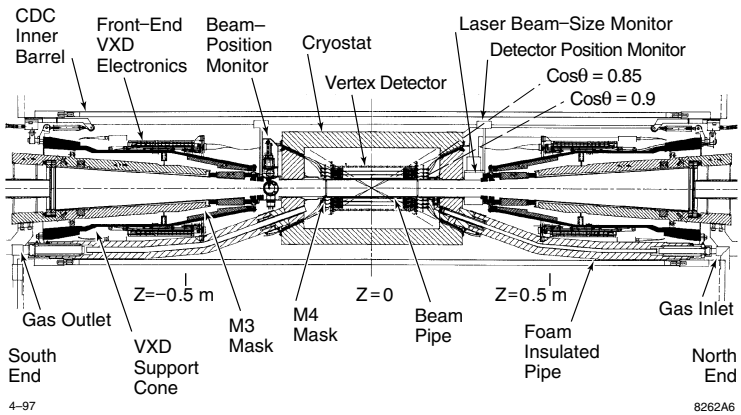

FIG. 8.14. Overall layout of the VXD3 vertex detector. (Figure courtesy of C.J.S. Damerell.)

so charge collection to the n-channel CCD structure is through a combination of diffusion and drift. The field associated with the abrupt doping gradient at the $p - p^+$ interface from the epitaxial layer to the bulk reflects the electrons, so the charge collection efficiency is high. Since the diffusion time is proportional to the square of collection distance, the diffusion component places an upper limit on the collection depth d and clock frequency f, so the clock frequency cannot be increased arbitrarily without a reduction in signal. In addition, the power associated with charging and discharging the capacitances in the CCD structure increases with clock frequency (as discussed for CMOS circuitry in Section

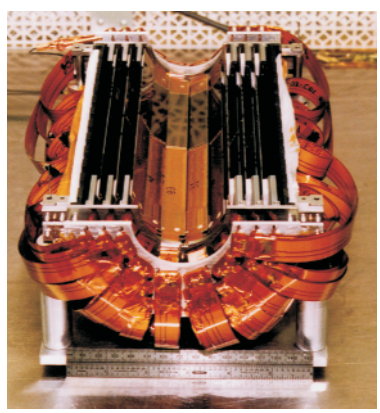

FIG. 8.15. An assembled half-shell of the VXD3 vertex detector prior to installation. (Figure courtesy of C.J.S. Damerell.)

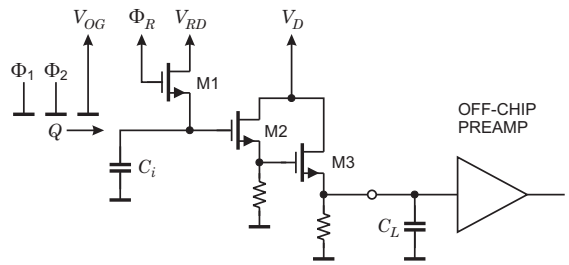

FIG. 8.16. Output circuit of the VXD3 CCD. Signal charge Q is transferred to the input node with capacitance C_i. M1 is the reset switch that sets the input node to potential V_{RD}. M2 and M3 form the cascaded source follower. The output load capacitance C_L represents the total capacitance presented to the output node.

5.1.2). The distributed resistance and capacitance of the clock lines also forms a low-pass filter that limits the clock frequency. Some improvement is gained by using two-phase CCDs with sinusoidal clocks. Simulations indicate that the phase shift and amplitude loss in CCDs suitable for vertex detectors could allow clock frequencies of 50 MHz (Damerell 2001).

The capacitance of the clock lines determines the power dissipation at high clock rates. In the VXD3 CCDs the total capacitance of the image section to the substrate is 16 nF and the interphase capacitance is 6 nF, resulting in a power dissipation of 1.3 W when clocked to provide a pixel transfer rate of 200 kHz with a clock level of 10 V. The readout register and output amplifier dissipate 25 and 45 mW, respectively, so the switching power dominates. The high instantaneous currents exacerbate cross-talk. For example, the peak current on the image clock lines is 1.3 A, so shared current paths can easily contaminate the signal (see Chapter 9). Two-phase clocks are also beneficial in this respect, as they can be configured to balance the currents in the two phases.

Another limit to readout speed is imposed by the readout circuitry. For simplicity CCD designers early on adopted source follower output stages to sense the signal voltage. In voltage mode a source follower has a gain < 1, so to maintain the noise level of the first stage, the second stage must have substantially lower noise. This is feasible, as the first-stage source follower can have sufficient drive capability to cope with the capacitive load of the second stage. Although this is not the optimum configuration, as discussed in Chapter 6, it does work and CCD designers have embraced an elaborately developed lore to justify its further use. However, the source follower does impose a serious speed limitation. To minimize the capacitive load the input source follower typically uses a rather small transistor with a small transconductance (typically of order 100 μS). Consequently, the output resistance $r_o \approx 1/g_m$ is rather high ($r_o = 10^4\,\Omega$ for $g_m = 10^{-4}$). This has led to the use of cascaded source followers, with increased transconductance in the second stage. Figure 8.16 shows the configuration used in VXD3.

In VXD3 the total capacitance C_i at the input node of M2 is 40 fF. The cascaded source followers have a gain of 0.75, resulting in a charge sensitivity of $3\,\mu\mathrm{V}/e$. The output resistance is $260\,\Omega$. Together with the load capacitance $C_L = 40\,\mathrm{pF}$ this results in an upper cutoff frequency of 15 MHz. This limitation can be circumvented by driving a low-impedance load. If the load resistance is much smaller than the output impedance $R_L \ll r_o$, the source follower operates as a transconductance stage. Configuring the external amplifier to present a $50\,\Omega$ input impedance, for example, allows connection through a matched transmission line with no capacitive load component. An FET operated in common base has worked well (Spieler 1982); feedback amplifiers must be designed more carefully to limit their current noise contribution.

The reset introduces "kTC" noise. When the reset switch is closed the input baseline fluctuates with the noise voltage $v_n^2 = kT/C_i$, as derived in Section 4.3.2. When the switch is opened the instantaneous noise voltage remains stored on the input node. As the charge sensitivity is increased by reducing C_i, the reset noise increases. To correct for this the baseline can be sampled immediately following the reset and the baseline subtracted from the subsequent signal samples. This requires additional time, but it is not necessary to reset the input node for each readout cycle. One can simply let signals from successive pixels add to the output and retrieve the signal by successively subtracting the previous pixel signal from the current value. Then resets are only necessary when the cumulative signal nears the maximum allowable input level.

A severe design challenge is the presence of large clock signals. For example, in a typical CCD a 10 V clock pulse precedes a signal of a few millivolts, so baseline recovery must be smooth and reproducible if the noise level is not to be corrupted by baseline fluctuations. Reset pulse transitions must be free of ringing or other artifacts that can couple capacitively to the input even after the reset transistor has switched off. Additional baseline fluctuations are frequently introduced by cross-talk from other clocks. When these fluctuations dominate, reducing the size of the input transistor is beneficial, as its lower input capacitance increases the signal level. Nevertheless, the fundamental rules for optimizing the equivalent noise charge also apply to CCDs and future designs may apply the appropriate design sophistication.

Modern CCDs exhibit excellent charge transfer efficiency with essentially no charge loss even when transferred over thousands of pixels. Trapping in small defects is mitigated by tunneling under the influence of the applied fields (Poole–Frenkel effect). CCDs are rather resistant to ionization damage, but displacement damage leads to degradations in charge transfer efficiency. Traps are filled by the applied clock potentials, which actually improves charge transfer efficiency at high clock rates. CCDs can be implemented both as n-channel devices (the conventional) that transfer electrons, or as p-channel devices utilizing hole transfer (Holland *et al.* 1996, 2003). The two are subject to different defect types. Displacement damage in n-channel devices has been evaluated and indicates usable operation to a fluence of $10^{10}\,\mathrm{cm}^{-2}$ (1 MeV neutrons, Brau and Sinev 2000,

Brau *et al.* 2004), whereas p-channel devices have been irradiated to $10^{11}\,\mathrm{cm}^{-2}$ (12 MeV protons, Bebek *et al.* 2002). Both device structure and readout technology determine the radiation resistance of CCDs and much work remains to determine the practical limits.

8.5 Vertex and tracking detectors at hadron colliders

Experiments at hadron colliders must deal with much higher interaction rates, which increases demands on rate capability. However, most of the interactions are background, so pattern recognition is crucial, which increases the number of layers required for efficient track reconstruction. Furthermore, the interaction region tends to be much more spread out in length (about 50 cm at the Tevatron), so the detectors must be longer. Fully cylindrical geometries become inefficient for full coverage, so combinations of barrels and disks are common, as illustrated in Figure 8.17.

8.5.1 *CDF and DØ*

CDF at the Fermi National Accelerator Laboratory installed a silicon vertex detector early in its operation in 1987 and since then has operated a succession of upgraded detectors. The original SVX had four concentric barrel layers at radii of 2.9, 4.1, 5.4, and 8.1 cm, each 51 cm long (Carithers *et al.* 1990). The barrels were read out at both ends and consisted of individual 8.5 cm long strip detectors wire-bonded together. The strip pitch was 60 μm in the inner three layers and 110 μm

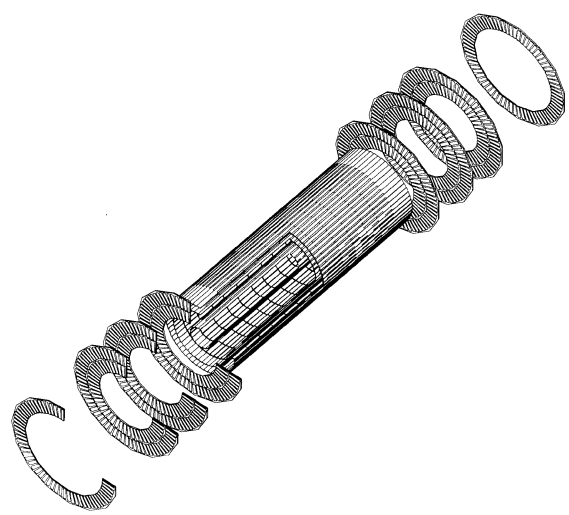

FIG. 8.17. A combination of barrel layers in the central region and disks in the forward regions makes efficient use of silicon area in large silicon detectors. The figure shows the layout of the SDC Silicon Tracker (Seiden 1994, Unno *al.* 1994).

in the outer layer for a total channel count of about 37 000. The readout IC was fabricated in 3 μm CMOS (Kleinfelder *et al.* 1988). Only a small fraction of the strips would be struck in a given event, so SVX pioneered on-chip sparsification (or zero-suppression), which selects only struck channels for readout. With this technology the required readout bandwidth is independent of segmentation, a crucial consideration in the very large detectors already envisaged for the next generation of high-luminosity colliders.

The readout IC included 128 parallel channels of charge-sensitive preamplifier, switched capacitor pulse shaping (correlated double sampling), threshold discrimination and readout logic. The readout logic scanned all comparator outputs on the chip and routed the corresponding signal amplitude to the analog readout bus. For analog interpolation built-in neighbor logic also added the analog information from the adjacent channels to include shared signal components that may have been below threshold. An additional mode allowed the readout of all channels. Charge-injection circuitry allowed testing and monitoring of all channels. The SVX power dissipation was 3 mW per channel. The next version, SVX-H, retained the basic architecture, but implemented it in 1.2 μm radiation-hard CMOS.

For the upgraded Tevatron CDF substantially expanded coverage of the vertex detector and added silicon layers at radii beyond 10 cm to enhance particle tracking (Merkel 2003). A "Layer 00" was also mounted at the smallest possible radius just outside the beam pipe. The layout is shown in Figure 8.18 together with an axial view. Table 8.1 summarizes the detector geometries. Within one layer the detector modules overlap to provide full coverage and facilitate relative position calibration. Layers 0, 1, and 3 use orthogonally crossed strips to provide two-dimensional information. As discussed in Chapter 1, this gives rise to "ghost

Table 8.1 Specifications of the CDF Run II silicon detector.

Layer	Inner/outer radius (cm)	Axial pitch (μm)	Stereo angle	Stereo pitch (μm)
00	1.35/1.62	25	–	–
0	2.5/3.0	60	90°	141
1	4.1/4.6	62	90°	125.5
2	6.5/7.0	60	1.2°	60
3	8.2/8.7	60	90°	141
4	10.1/10.6	65	1.2°	65
5 forward	19.7/20.2	112	1.2°	112
5 central	22.6/23.1	112	1.2°	112
6 forward	28.6/29	112	1.2°	112

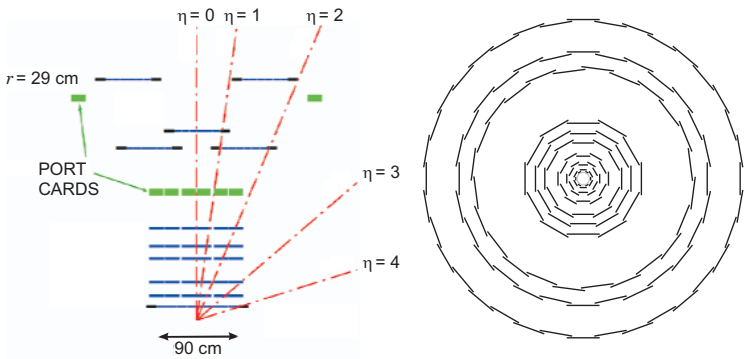

FIG. 8.18. Layout of the SVX upgrade at CDF (left) together with an axial view (right). "Port cards" accommodate interface and driver circuitry that links to the off-detector electronics.

hits" due to the ambiguity of the projective geometry, so layers 2, 4, 5, and 6 use a 1.2° stereo angle.

As coverage is extended to forward angles, a barrel geometry requires inordinate silicon area and resolution suffers for grazing incidence tracks. Figure 8.19 shows a novel layout used at the Tevatron by DØ (Kajfasz 2003). Interspersed barrels and disks in the central region optimize resolution over the full extent of the interaction region ($\sigma_z \approx 25\,\mathrm{cm}$). Additional disks provide forward coverage to $\eta \approx 3$ (Figure 8.19). The material in the four-layer barrel adds up to $0.1\,X_0$, dominated by support and services.

For their detector upgrades CDF and DØ joined forces to develop the readout ICs. Unlike previous SVX chips, which used analog voltage levels for the threshold setting and calibration inputs and read out the signal magnitude in analog form, the new generation includes on-chip digitization of the signal level and on-chip DACs set the threshold and calibration levels. Thus, all communication to and from the chip is by a digital bus. Figure 8.20 shows a block diagram of the SVX2 chip (Zimmerman *al.* 1995). Following a charge sensitive amplifier with adjustable bandwidth, a switched capacitor network provides correlated double-sampling and analog storage. The analog pipeline stores samples up to $5.5\,\mu$s to accommodate the trigger latency time. Following the pipeline is a Wilkinson ADC with a common ramp for all channels. Thus, the ADC circuitry per channel is a comparator, also needed for sparsification, and a counter latch to record the pulse height. The ADC runs at a clock rate of 106 MHz and has a range of 8 bits, so the maximum conversion time is $2.4\,\mu$s. The ADC adds $100\,\mu$m to the length of each channel and $300\,\mu$W to the total power per channel of 3 mW. The IC is fabricated in $0.8\,\mu$m, triple-metal radiation-hard CMOS. The dimensions are $6.3 \times 8.7\,\mathrm{mm}^2$. Figure 8.21 shows a die photo and indicates the size of the main

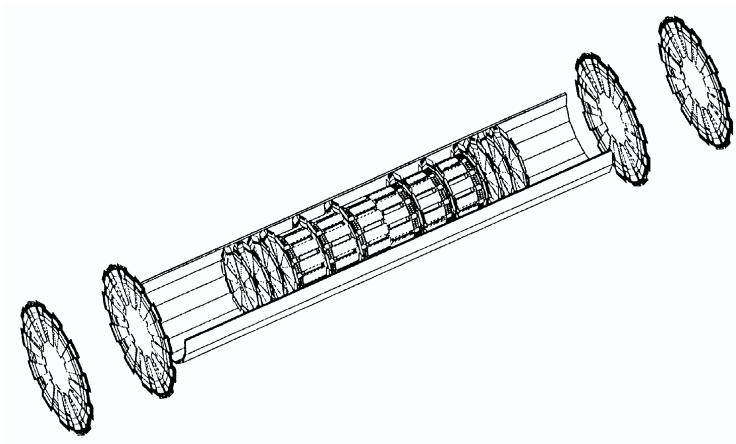

FIG. 8.19. DØ uses interspersed disks and barrels in the central region, with additional disks to provide forward coverage. The DØ barrel is 762 mm long with four layers distributed over 2.7 to 9.4 cm radius. The outer disks have 26 cm radius and the overall length is 2.5 m. The detector has about $8 \cdot 10^5$ readout channels and $3\,\mathrm{m}^2$ of silicon area (Bean 2001). (Figure courtesy of A. Bean.)

circuit blocks. The 128 signal channels are laid out on a 42 μm pitch to provide space at the edges of the chip for bussing and common logic.

Despite what its name implies, the analog pipeline does not shift signals sequentially. Instead the signal is stored in one of the memory locations B1 – B32 and remains there until retrieved. Opening or closing a switch injects charge through the gate–channel capacitance, introducing a baseline shift, which is removed by correlated double sampling. Prior to signal acquisition the baseline voltage V_{BL} is sampled and stored in the pipeline. When a subsequent signal V_S is read, the level applied to the comparator is the sum of the signal and the baseline $V_S + V_{BL}$. When switches S_C and S_{CR} are closed, the voltage $V_S + V_{BL}$ is impressed on the coupling capacitor C_C, as closing S_{CR} establishes a low impedance at the comparator input through shunt feedback. To subtract the baseline, the baseline sample stored in the pipeline is selected, while opening switch S_{CR}. This sets the voltage at the left side of C_C to V_{BL}, while the voltage across C_C is still $V_S + V_{BL}$, so the two sample voltages in series yield V_S at the comparator input. Although implemented differently, this performs the same function as discussed in Section 4.5. The circuit utilizes switched capacitor circuitry for many functions. The operating modes and switching sequences are described in the SVX2 manual (Yarema *et al.* 1994).

The circuit can accept both positive and negative polarity input signals. The preamplifier is designed with sufficient bipolar dynamic range. Switching polarity involves shifting the baseline of the analog pipeline to increase the range for either

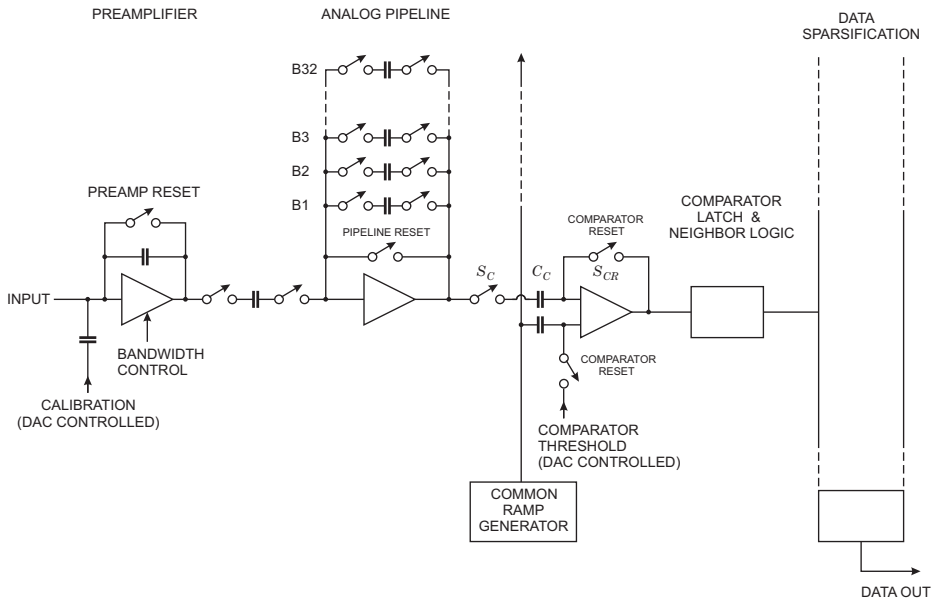

FIG. 8.20. Block diagram of the SVX2 chip. For details of the operating modes and switching sequences see Yarema *et al.* (1994).

positive or negative excursions. In addition, the polarity of the ADC ramp and comparator are reversed.

The neighbor logic allows three modes: read only channels whose level exceeded threshold, read "struck" channels and their two neighbors, and read all channels. The hit threshold is set digitally and is common to all channels on a chip. This setting is rather coarse (about $2000\,e$), but external adjustment of the ADC ramp start voltage provides a threshold resolution of about $400\,e$.

The SVX2 IC (used by DØ) is designed for sequential signal acquisition and readout. A further development, SVX3 (used by CDF) allows concurrent read-write, so that signal acquisition and readout can proceed concurrently (Garcia-Sciveres *et al.* 1999). The floor plan is similar to SVX2, but the die size increased to $6.26 \times 12\,\mathrm{mm}^2$. The measured noise $Q_n \approx 2400\,e$ at an input load of 33 pF, a preamplifier rise time of 60 ns, and a sample time of 120 ns. In the SVX detector design the concurrent readout introduces substantial common mode noise. A compensation scheme was implemented on-chip that uses all channels in a chip to calculate a common pedestal event by event and subtract it during digitization. A subsequent version implemented in $0.25\,\mu$m CMOS, the SVX4, was designed and tested for planned upgrades of CDF and DØ (Krieger *et al.* 2004). At the 106 MHz digitizer speed and 53 MHz readout rate the SVX4 achieves $2000\,e$ noise with a 40 pF input load, while dissipating 2 mW/channel at 2.5 V supply voltage. The IC was tested to > 20 Mrad total dose.

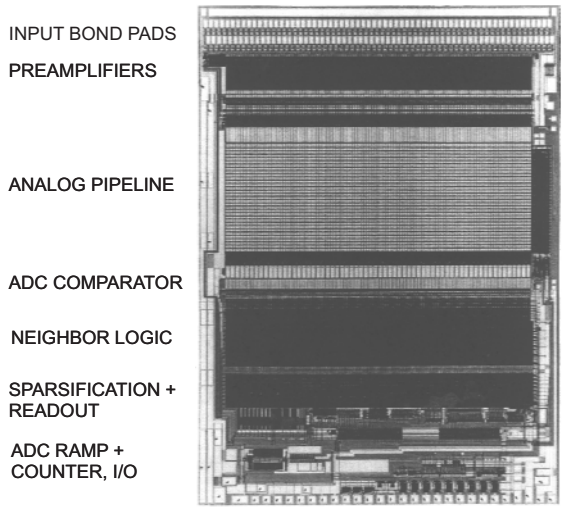

FIG. 8.21. Die photo of the SVX2 chip indicating individual circuit blocks. Note the wafer-probe marks on the lower bond pads from IC testing.

8.6 Silicon trackers at the Large Hadron Collider

The LHC poses unprecedented challenges to detector designers. Work on suitable detector concepts began in the 1980s, culminating in final assembly in 2005–2007. A worldwide detector R&D program was necessary to develop the concepts and technologies, especially in the areas of sensors, microelectronics, and radiation effects. The results of this ongoing work also flowed into preceding experiments such as CDF, DØ, BaBar, and Belle.

Key LHC parameters include colliding proton beams of 7 TeV on 7 TeV to provide 14 TeV center of mass energy at a luminosity of $10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. For comparison, the SSC (Donaldson and Marx 1986) planned 10 TeV on 10 TeV; the higher energy allowed a ten-fold lower luminosity of $10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ for essentially the same physics goals. The higher luminosity at the LHC increases the backgrounds and the radiation damage that the detectors must cope with.

The LHC bunch crossing frequency is 40 MHz with an average of 23 interactions per bunch crossing and about 150 charged particles per unit of pseudorapidity $\eta = -\log\tan(\Theta/2)$, where Θ is the angle relative to the beam axis. Thus, the hit rate

$$n' = \frac{2\cdot 10^9}{r_\perp^2}\left(\mathrm{cm}^{-2}\mathrm{s}^{-1}\right) \ , \tag{8.11}$$

where $r_\perp$ is the distance from beam axis. In a detector that subtends ± 2.5 units of rapidity, the total hit rate is $3\cdot 10^{10}\,\mathrm{s}^{-1}$. At $r_\perp = 14\,\mathrm{cm}$ the rate is about $10^7\,\mathrm{s}^{-1}\mathrm{cm}^{-2}$. This radial dependence is modified in the presence of a magnetic field.

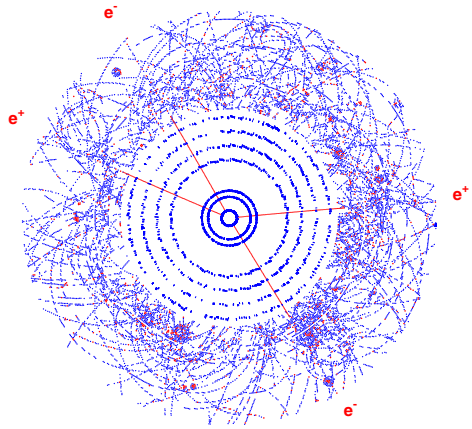

FIG. 8.22. Axial projection of a simulated Higgs to four electrons event at LHC design luminosity. Since the tracks are spread out along the beam axis, the z-resolution of the silicon detectors (inner rings) is essential for pattern recognition, although the high hit density of the outer straw chambers makes the tracks more apparent in this representation. (Figure from ATLAS TDR 1997.)

A general purpose detector includes vertexing for B-tagging, precision tracking in a magnetic field (2 T in ATLAS, 4 T in CMS), calorimetry (electromagnetic + hadronic, for a thorough discussion see Wigmans 2000), and muon detection. Figure 8.22 shows an axial view of a simulated event.

8.6.1 *Coping with high rates*

The high hit rate can easily lead one to believe that very high speed electronics are needed. However, maintaining the required signal-to-noise ratio at fast shaping times drives up the required power. Segmentation, on the other hand, is much more efficient, as subdividing the detector into many small elements reduces the hit rate per element. For example, at $r_\perp = 30\,\text{cm}$ the hit rate on a strip electrode of $50\,\mu$m width and 10 cm length, *i.e.* an area of $5 \cdot 10^{-2}\,\text{cm}^2$, is about $10^5\,\text{s}^{-1}$. This corresponds to an average time between hits of $10\,\mu$s, so longer shaping times are allowable, which translates to lower power for given noise level. As discussed in Chapter 6, with careful design the power requirements don't increase in highly segmented detectors.

An additional problem is the large number of events per crossing. As the interaction region is spread out, vertex reconstruction with the appropriate z-resolution can resolve individual interactions. Again, segmentation helps. If a detector element is sufficiently small, the probability of two tracks striking it within one crossing is negligible. Tracks from different crossings can be separated if the electronics are capable of single-bunch time resolution, that is 25 ns, and time-stamped data are stored.

Semiconductor detectors are well-matched to these requirements and they are key components in all LHC experiments. Patterning the sensor at the "μm-scale" is straightforward and monolithically integrated electronics can be mounted locally with on-chip multiplexing or sparsification to reduce the cable plant. The fast collection times of semiconductor are very advantageous; achieving collection times $< 25\,\mathrm{ns}$ with a $300\,\mu\mathrm{m}$ thick sensor is quite practical.

8.6.2 *Radiation damage*

The high rates bring another problem with them; radiation damage, both in the sensors and the electronics. There are two sources of particles, beam collisions and neutron albedo from calorimeter. Estimated fluences per year at design luminosity expressed in equivalent 1 MeV neutrons are $5 \cdot 10^{13}\,\mathrm{cm}^{-2}$ at $r = 10\,\mathrm{cm}$ and $2 \cdot 10^{13}\,\mathrm{cm}^{-2}$ at $r = 30\,\mathrm{cm}$. The corresponding ionizing doses are 30 kGy (3 Mrad) at 10 cm and 4 kGy (400 krad) at 30 cm. In reality, complex maps are required of the radiation flux, which is dependent on local material distribution (examples in ATLAS TDR 1997).

As discussed in Chapter 7, displacement damage in the sensor leads to an increase in leakage current

$$I_R = I_0 + \alpha \Phi A d \ , \tag{8.12}$$

which in the electronics increases the shot noise

$$Q_{ni}^2 = 2 e I_d F_i T_S \ . \tag{8.13}$$

The leakage current drops exponentially with temperature

$$I_R(T) \propto T^2 e^{-E/2kT} \ , \tag{8.14}$$

so even moderate reductions in temperature bring a significant improvement.

The electronic shot noise can be reduced by choosing short shaping times. Furthermore, reducing the area of a detector element reduces the leakage current per channel, so the shot noise is also smaller. Again, segmentation is advantageous.

Reducing the leakage current by cooling is crucial for another reason. The power due to the leakage current together with the bias voltage leads to self-heating of the detector. Unless the sensor is cooled adequately, the increased power dissipation increases the temperature, which increases the bias current exponentially and leads to thermal runaway (Kohriki *et al.* 1996).

The second effect of displacement damage is an increase in the required operating voltage. Bias voltages can be maintained at reliable levels by thinning the sensor and by allowing for operation below full charge collection. This reduces the signal and requires lower noise to maintain the required signal-to-noise ratio. Again, segmentation helps, as the decreased area of a detector element reduces the capacitance and the achievable noise level.

In the course of developing silicon detector systems many unexpected problems arose and invariably solutions were found. Key is the use of a highly developed technology, which provides performance reserves and design flexibility. It is tempting to favor a new technology because it appears to offer a "silver bullet" against one specific problem, but systems depend on the interplay of many design considerations. For this reason many "advanced" technologies have fallen by the wayside.

8.6.3 *Layout*

General purpose detectors provide full coverage by a combination of barrel and disk layers. Both ATLAS (Turala *al.* 2001, Unno *al.* 2003) and CMS (Abbaneo 2004) use barrels in the central region and disks in the forward regions to provide the required coverage and tracking performance with minimum silicon area. The ATLAS SemiConductor Tracker (SCT) has about $60\,\mathrm{m}^2$ of silicon with $6 \cdot 10^6$ strip detector channels, augmented by a gaseous outer tracking detector. After going rather far in the development of a mixed silicon-gaseous detector system CMS decided to build an all-silicon tracker with about $230\,\mathrm{m}^2$ of silicon and 10^7 strip detector channels. Both ATLAS and CMS use pixel devices covering $1 - 2\,\mathrm{m}^2$ with $50 - 100$ million channels at the inner radii ($< 15\,\mathrm{cm}$) because of their superior pattern recognition at high track densities and radiation resistance. The small capacitance allows low noise $Q_n \approx 200\,e$ at sufficiently fast shaping times, so the system starts out with a very high signal-to-noise ratio. These performance reserves allow greater degradation of signal and noise with radiation damage than in a silicon strip system, which extends the lifetime. Strips take over at larger radii to minimize material and cost.

In ATLAS four layers of silicon strips are used at radii of 30, 37, 44, and 51 cm inside a superconducting solenoid with a 2 T magnetic field (Figure 8.23). Beyond 56 cm radius a 70-layer straw-tube gaseous tracking and transition radiation detector (TRT) provides at least 40 hits per track. The TRT is operating close to its limits at 10^{34} luminosity, but was retained for cost reasons. At both ends an array of 9 disks arranged at distances $|z| = 0.85\,\mathrm{m}$ to $2.7\,\mathrm{m}$ provide coverage at rapidity $|\eta| > 1.2$. Resolution in the barrel is determined by the strip pitch of $80\,\mu\mathrm{m}$ in $r\varphi$ and 40 mrad small angle stereo in z. In the disks the strips go radially, with pitches ranging from 55 to $90\,\mu\mathrm{m}$, using a 40 mrad stereo angle for r-resolution. All modules are double-sided, formed by gluing two single-sided detectors back-to-back under an angle of 40 mrad. Three layers of pixel detectors at 5.1, 9.9, and 12.3 cm in the central region and three pixel disk layers at each end provide the two track resolution and radiation resistance required close to the interaction region. The pixels are $50\,\mu\mathrm{m} \times 400\,\mu\mathrm{m}$, with the long dimension along the beam axis to accommodate inclined tracks. The pixel subsystem is enclosed within the pixel support tube and can be inserted or removed as a unit to facilitate replacement. Services are brought out to patch panels at the cryostat wall through the gaps between the TRT subunits, which leads to significant local increases in the material distribution.

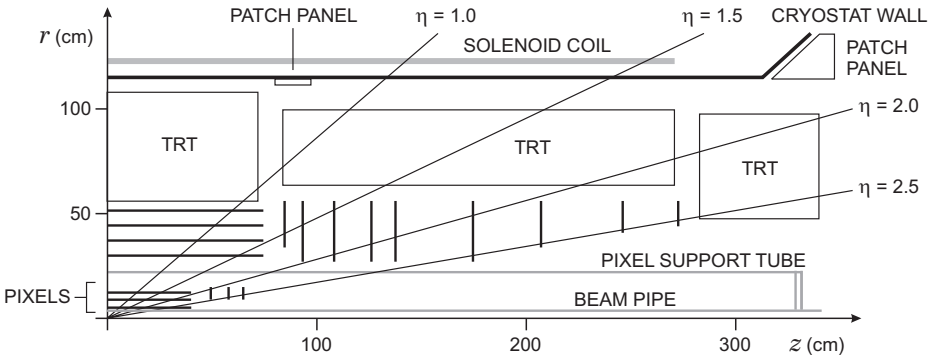

FIG. 8.23. Cross-section through one quadrant of the ATLAS SCT. The silicon tracker is about 1 m in diameter and 5.4 m long. The pixel subsystem is self–contained and can be inserted or removed separately. Intermediate patch panels mounted at the cryostat wall facilitate assembly. Unno *al.* (2003) give detailed dimensions.

Estimated fluences after 10 years of operation are estimated to be $10^{15}\,\mathrm{cm}^{-2}$ (1 MeV neutron equivalent) with a total dose of 50 Mrad at the innermost pixel layer and a fluence $2 \cdot 10^{14}\,\mathrm{cm}^{-2}$ at the inner strip layer. The p^+-on-n sensors used in the SCT must be biased for full charge collection. Type inversion and anti-annealing increase the required detector bias voltage to > 350 V, so the sensors have been designed to sustain 500 V. In production about 10% of the modules exhibited the onset of high bias current well below the 500 V required in the acceptance tests (Unno *al.* 2003). Most of these modules met specifications after operating them with gradually increasing bias voltage over several hours. Both the traditional $\langle 111 \rangle$ and $\langle 100 \rangle$ orientations were studied and $\langle 111 \rangle$ chosen because of easier availability. The pixel modules use oxygenated sensors, which allow full voltage operation over > 5 years (RD48 1999). Unlike the strip sensors, which are p^+-on-n, the pixel sensors are n^+-on-n, which are also usable at voltages below full collection. As a consequence, detector performance deteriorates gradually with radiation damage as the signal-to-noise ratio falls off.

In both the disks and barrels the modules are "shingled" to provide full coverage and facilitate relative position calibration (Figure 8.24). The cant angle of the detectors is chosen to minimize the resolution spread due to Lorentz deflection of the carriers in the 2 T magnetic field (Unno *al.* 1991, Albiol *et al.* 1998). The resolution in $r\varphi$ is 12 μm in the pixel and 16 μm in the strip system. The respective resolutions in z are 66 μm and 580 μm.

CMS uses an all-silicon tracker with 2.4 m diameter and 5.4 m length in a 4 T solenoidal magnetic field (Abbaneo 2004, Biasini 2004). Figure 8.25 shows the layout. Strip detectors are used in all layers except at the smallest radii, where the interaction region is surrounded by two barrel layers of pixel detectors at 4

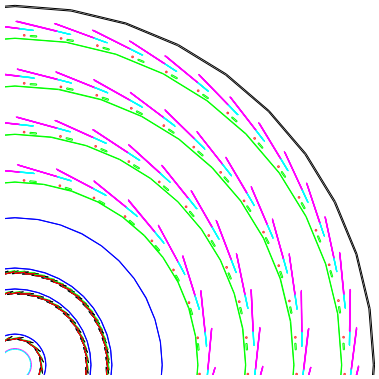

FIG. 8.24. Axial quarter view of the ATLAS Semiconductor Tracker illustrating shingling of detector modules to provide overlap and compensate for Lorentz deflection of the collected charge (ATLAS TDR 1997).

and 7 cm for low luminosity running and at 7 and 11 cm at high luminosity. Two endcap pixel disks cover radii from 6 to 15 cm. Strip pitches range from 80 to 205 μm and the pixel size is 100 μm $\times$ 150 μm (Erdmann 2004).

In the strip detector portion double-sided detectors are used in layers 1, 2, 5, and 6 of the barrel and in rings 1, 2, and 5 of the disks. As in ATLAS the double-sided modules use two single-sided sensors, glued back-to-back to form a small stereo angle. CMS uses a somewhat larger stereo angle of 100 mrad. The endcap disks consist of wedge shaped segments, each covering 1/16 of 2π.

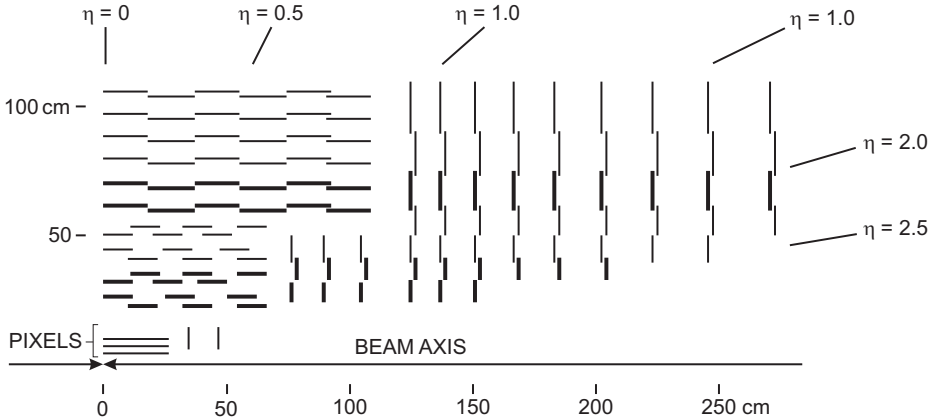

FIG. 8.25. Layout of the CMS tracker, showing one quadrant of the lengthwise cross–section. The thick lines indicate double-sided detectors.

In total the CMS tracker implements 25 000 silicon strip sensors covering an area of 210 m^2. The $9.6 \cdot 10^6$ strips are read out by 75 000 readout ICs. About 25 million wire bonds interconnect strip sensors and readout ICs. Detector segmentation is chosen so that typical channel occupancies are about 1% throughout the detector. Position resolution in $r\varphi$ is strongly affected by the approximately 30° Lorentz angle of the electron drift in the 4 T magnetic field. The barrel strip detectors are tilted by 9° to compensate. The barrel pixel geometry is deliberately chosen so that this large Lorentz angle induces significant sharing of charge across neighboring cells. This yields spatial resolutions $\sigma_{r\varphi} \approx 10\,\mu$m and $\sigma z \approx 15\,\mu$m.

In the inner layers the strip length is about 10 cm, whereas in the outer region the strip length is doubled, which increases the electronic noise. To compensate the signal is increased by using 500 μm thick sensors instead of the normal 320 μm devices used elsewhere. Utilization of 150 mm wafer technology provided the cost savings needed for this huge silicon system.

Sensors use the $\langle 100 \rangle$ orientation to minimize surface damage. This yields a somewhat lower interstrip capacitance after irradiation (Braibant *et al.* 2002). The strips are AC-coupled with integrated polysilicon resistors. The inner region utilizes lower resistivity material (1.25 – 3.25 kΩ cm) to delay type inversion, whereas the 500 μm thick sensors in the outer layers use 3.5 – 7.5 kΩ cm material (Bergauer 2004, Krammer 2004). The lower resistivity sensors start with a higher depletion voltage but end with a lower operating voltage after type inversion and 10 years of LHC operation. The strip pitch is 80 to 183 μm in the barrel and up to 205 μm in the disks with no intermediate strips.

8.6.4 *Readout electronics*

Both ATLAS and CMS distribute the total number of tracks over many detector segments to reduce the rate per channel and reduce the double-hit probability. For example, in ATLAS the occupancy in the pixel system is $4.4 \cdot 10^{-4}$ at 4 cm radius and $6 \cdot 10^{-5}$ at 11 cm radius. In the strip system the occupancies are $6 \cdot 10^{-3}$ at 30 cm and $3.4 \cdot 10^{-3}$ at 52 cm radius. Thus, the choice of shaping time is not driven by rate considerations, but by the requirement for 25 ns single-bunch time resolution.

8.6.4.1 *CMS readout electronics* The CMS readout is a direct descendant of the systems used at LEP and utilizes full CMOS circuitry that exploits switched capacitor techniques. Figure 8.26 shows the block diagram of the readout IC, the APV25 (French *et al.* 2001). Each strip is read out by a charge sensitive amplifier followed by a switchable unity gain inverter to allow both p- or n-strip readout. Subsequently, a 50 ns CR-RC shaper drives a 192-stage analog pipeline to accommodate up to 4 μs trigger latency. On receipt of a trigger a switched-capacitor analog pulse processor applies a weighted sum algorithm to provide the desired single-bunch time resolution.

Development of the pulse processor was originally motivated by the lack of sufficiently fast CMOS processes to allow efficient operation at the shaping times

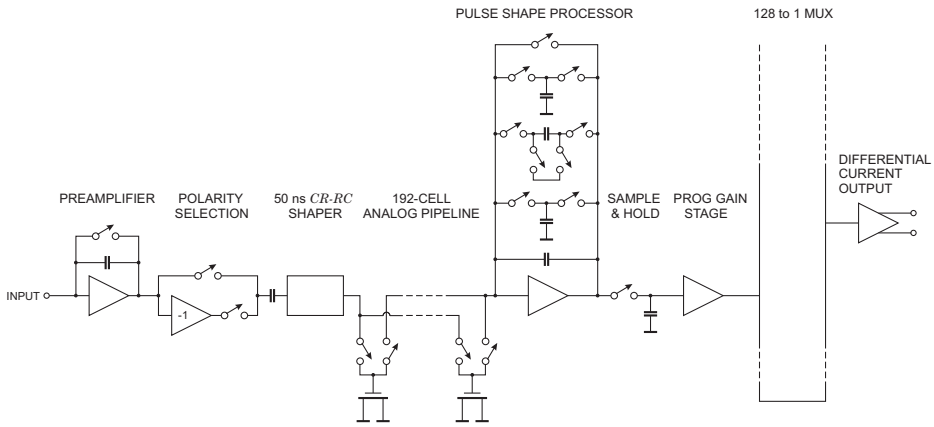

FIG. 8.26. Block diagram of the APV25 readout IC used by CMS.

required for single-bunch time resolution (Bingefors *et al.* 1993). The underlying notion is to use rather slow pulse shaping and then to apply a deconvolution algorithm to reconstruct the fast components of the input signal. For the sampled output of a *CR-RC* shaper with the step response

$$v(t) = \frac{t}{\tau} e^{-t/\tau} \tag{8.15}$$

this can be implemented by forming the weighted sum of three successive samples

$$V_k = w_1 V_k + w_2 V_{k-1} + w_3 V_{k-2} \tag{8.16}$$

with the weights

$$w_1 = \frac{1}{x} e^{x-1}, \quad w_2 = -\frac{2}{x} e^{-1}, \quad w_3 = \frac{1}{x} e^{-(x+1)} \tag{8.17}$$

(Gadomski 1992, Bingefors *et al.* 1993). The weights depend on the sampling interval normalized to the shaping time constant $x = \Delta t/\tau$. For a step input the result of the deconvolution is zero. However, for a finite rise time the result is a short pulse with the duration of the rise time. The APV25 uses a time constant $\tau = 50\,\mathrm{ns}$ in the *CR-RC* filter and samples the output at $40\,\mathrm{MHz}$, so $x = 0.5$ and the weighting factors

$$w_1 = 1.2, \quad w_2 = -1.5, \quad w_3 = 0.45\ .$$

Figure 8.27 shows the APV25 output in peak and deconvolution mode. As this is a crude form of differentiation, the signal is reduced and the noise bandwidth increased, so the noise in deconvolution mode is higher than in peak mode. For example, the noise in peak mode of $246\,e + 36\,e/\mathrm{pF}$ increases to $396\,e + 59.4\,e/\mathrm{pF}$

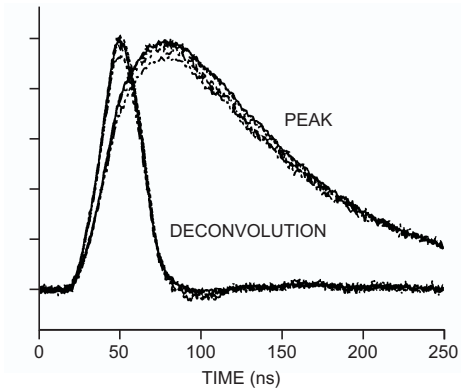

FIG. 8.27. APV25 pulse shapes in peak and deconvolution mode, shown for input capacitive loads ranging from 2 to 20.5 pF. The reduced pulse height at 17.5 and 20.5 pF indicates that the preamplifier bandwidth is becoming marginal. (Adapted from Raymond *al.* 2000). Figure courtesy of M. Raymond.

in deconvolution mode. The latter is required for single-bunch timing resolution at the LHC, so that is the relevant noise in normal operation.

Viewed somewhat differently, the *CR-RC* shaper is the anti-aliasing filter required before sampling the signal into the analog pipeline. The algorithm then produces a near triangular weighting function (Bingefors *et al.* 1993) with a peaking time of Δt and the noise indices $F_i = 0.35$ and $F_v = 1.84$, similar to a conventional $CR\text{-}RC^3$ or $CR\text{-}RC^4$ filter.

The decision to utilize an all CMOS readout IC with the deconvolution processor was made early on and applied in a series of designs using different fabrication processes. Fortuitously, the demonstration of excellent radiation resistance in standard commercial "deep sub-micron" CMOS opened the path to an efficient implementation in 0.25 μm CMOS (French *et al.* 2001). The APV25 chip combines 128 readout channels with an output multiplexer and is 7.1 mm wide and 8.1 mm long (Raymond *al.* 2000).

The analog output signals are transmitted to the off-detector electronics through optical links using edge-emitting semiconductor lasers operating at a standard telecommunications wavelength of 1310 nm. Off-detector the optical signals are received by a photodiode-amplifier on the "front end driver", which digitizes and processes the signals, subtracts pedestals and stores the results in a local memory. When operating at the maximum trigger rate, cluster finding is applied to reduce the data volume.

8.6.4.2 *ATLAS readout electronics* ATLAS chose a readout system that sought to efficiently balance the technology against cost, while meeting the physics requirements. The goal in these large systems is not to provide the best possible

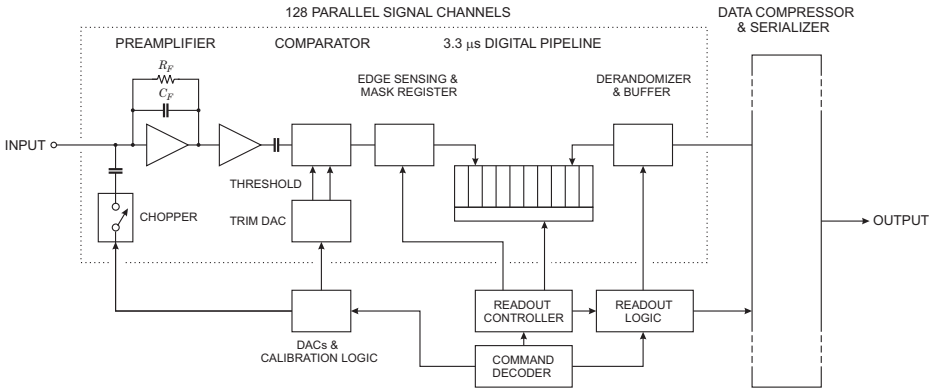

FIG. 8.28. Block diagram of the ABCD readout IC for the ATLAS SCT.

performance, but to maintain adequate performance over the lifetime of the detector. After a lengthy process of testing and comparing options, the collaboration concluded that these goals can be achieved by a binary readout, *i.e.* a system that just records the presence of a hit, so the output only provides a time stamp with a series of hit addresses. This technique also lends itself readily to on-chip zero-suppression, which reduces the cost and space requirements of the readout lines. Threshold scans, as described in Section 4.7.3 and used in the BaBar SVT, allow pulse height measurements for diagnostics, but this is only necessary infrequently to monitor changes in response, *e.g.* due to radiation damage. Figure 8.28 shows the block diagram of the readout IC. The front-end utilizes time-invariant filtering. The bandwidths of the cascaded amplifiers needed to provide the necessary gain are tailored to provide approximately a CR-RC^3 response. A comparator fires when a signal exceeds threshold and the time is stored in a digital pipeline, whose length accommodates the ATLAS level 1 trigger latency. Data sparsification and compression circuitry reduce the data volume that must be read out. More details follow in Section 8.6.4.4.

8.6.4.3 *Required signal-to-noise ratio in a binary readout system* Binary readout systems were discussed in Section 4.7 of Chapter 4. The threshold must be set low enough to capture the desired portion of the amplitude spectrum, but not so low that the rate of noise pulses is too high.

The signal for minimum ionizing particles is a Landau distribution, where for 99% efficiency in a 300 μm thick detector the threshold must be set to about one half the of the most probable charge Q_0. Assume that the minimum signal to be measured is $f_L Q_0$. Tracks passing between two strips will deposit charge on both strips. The ability to distinguish one-hit from two-hit clusters improves the obtainable position resolution, as two-hit clusters are assigned mid-way between two strips. If the fraction of the signal to be detected is f_{sh}, the circuit must be sensitive to signals as low as

$$Q_{\min} = f_{sh} f_L Q_0 \ . \tag{8.18}$$

As derived in Section 4.7, the required threshold-to-noise ratio for a noise occupancy P_n in a time interval Δt is

$$\frac{Q_T}{Q_n} = \sqrt{-2\log\left(4\sqrt{3}_n T_S \frac{P_n}{\Delta t}\right)} \ . \tag{8.19}$$

In the strip system the average hit occupancy is about $5 \cdot 10^{-3}$ in a time interval of 25 ns. If we allow a noise occupancy of 10^{-3} at a shaping time of 20 ns, this corresponds to $Q_T/Q_n = 3.2$.

The threshold uniformity is not perfect. The relevant measure is the threshold uniformity referred to the noise level. For a threshold variation ΔQ_T, the required threshold-to-noise ratio becomes

$$\frac{Q_T}{Q_n} = \sqrt{-2\log\left(4\sqrt{3}_n T_S \frac{P_n}{\Delta t}\right)} + \frac{\Delta Q_T}{Q_n} \ . \tag{8.20}$$

If $\Delta Q_T/Q_n = 0.5$, the required threshold-to-noise ratio becomes $Q_T/Q_n = 3.7$. To maintain good timing, the signal must be above threshold by at least Q_n, so $Q_T/Q_n > 4.7$.

Combining the conditions for the threshold

$$\left(\frac{Q_T}{Q_n}\right)_{min} Q_n \leq Q_{min} \tag{8.21}$$

and signal (eqn 8.18)

$$Q_{min} = f_{sh} f_L Q_0 \tag{8.22}$$

yields the required noise level

$$Q_n \leq \frac{f_{sh} f_L Q_0}{(Q_T/Q_n)_{min}} \ . \tag{8.23}$$

If charge sharing is negligible $f_{sh} = 1$, so with $f_L = 0.5$, $Q_0 = 3.5\,\mathrm{fC}$, and $(Q_T/Q_n)_{min} = 4.7$, the required noise level $Q_n \leq 0.37\,\mathrm{fC}$ or $Q_n \leq 2300\,e$. If the system is to operate with optimum position resolution, *i.e.* equal probability of one- and two-hit clusters, then $f_{sh} = 0.5$ and $Q_n \leq 0.19\,\mathrm{fC}$ or $Q_n \leq 1150\,e$. ATLAS requires $Q_n \leq 1500\,e$.

8.6.4.4 *ATLAS SCT readout implementation* ATLAS adopted a bipolar transistor front-end with CMOS digital circuitry. Initial prototypes used separate bipolar and CMOS ICs. The production device utilizes a BiCMOS process that combines all of the circuitry in a single chip, the ABCD chip (Dabrowski *et al.* 2000, Campabadal *et al.* 2005a). Each chip includes 128 channels, on a $6.4 \times 4.5\,\mathrm{mm}^2$ die, bondable to a $50\,\mu$m pitch. Pitch adapters make the transition to the detector strip pitch of $80\,\mu$m. Designing the ICs for the smaller

pitch provides space between adjacent ICs for bypass capacitors and wire bonds. The analog portion uses continuous shaping, approximating a CR-RC^3 response with a peaking time of 20 ns. At the nominal operating threshold of 1 fC this yields a time walk of 12 ns for signals of 1.2 to 10 fC. The double-pulse resolution for two successive 4 fC pulses is 50 ns. The operating current of the input transistor is adjustable to optimize noise with radiation damage and the total power is 1.3 to 1.8 mW/ch.

On-chip DACs control the threshold and operating point. Trim DACs on each channel fine tune the thresholds to compensate for threshold nonuniformity from channel to channel, bringing the threshold dispersion well below the noise level. This technique is even more important when power and readout area are minimized, as in pixel devices, so the efficacy of trimming will be illustrated in Section 8.6.7.

Each chip also includes digitally controlled calibration circuitry with a DAC-controlled injection level, shown in Figure 8.28. All control and output signals are digital and each module communicates with the off-detector electronics through optical fibers. The serial link and token passing (see Figure 1.31) present single-point failure modes, so redundant readout modes are incorporated. In token passing defective chips within a module can be bypassed. Normally, each module has two readout lines, one for each side. Should the master of one side fail, the other side's master takes over and both sides are read out through one line. For a more detailed description of the ABCD readout IC see Campabadal *et al.* (2005a).

The adopted IC fabrication process was specially designed for LHC applications and lacked a strong commercial base. Consequently, process control was not fully developed and the overall yield was about 20%. The project required about $5 \cdot 10^4$ ICs, so this required a fast testing system, which was custom designed to provide the necessary throughput to match the production schedule (Anghinolfi *et al.* 2002).

8.6.5 *Detector modules*

CMS uses a conventional module configuration with ceramic hybrids connected at the ends of the detectors ATLAS adopted a novel module design, so it will be discussed in more detail. Figure 8.29 shows the module layout. Connecting the electronics at the mid-point of the barrel modules reduces the noise contribution of the strip resistance, as discussed in Section 6.5. The electronics hybrid uses a four-layer polyimide substrate shown in Figure 8.30 (Kondo 2002 *et al.*, Kohriki *et al.* 2002). This reduces material and also allows the electronics for both sides to be placed on the same layer, as the hybrid can be wrapped around the detector (Figure 8.31). A close up of the ICs mounted on the hybrid is shown in Chapter 1 (Figure 1.33).

Two single-sided p-on-n sensors are glued back to back to form a 40 mrad stereo angle. Each sensor has 784 strips on an 80 μm pitch. Two sensors are butted to provide an overall strip length of 126 mm. An intermediate baseboard of

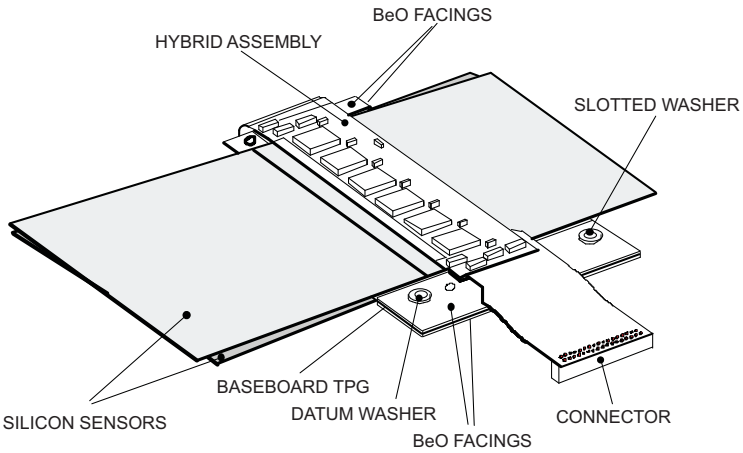

FIG. 8.29. ATLAS SCT barrel detector module. Two single-sided sensors are glued back-to-back with an intermediate TPG heat spreader. The "ear" extending from the module attaches to the support/cooling stave of the SCT barrel structure. (Figure courtesy of T. Kondo.)

thermalized pyrolytic graphite (TPG) provides support and a high-conductivity cooling path to the mounting stave of the barrel support structure. The hybrid is mounted on spacers to prevent direct heat transfer from the readout ICs to the sensors. This bridge configuration and the high thermal conductivity of the TPG heat spreader between the sensors ensures that the sensors are cooled sufficiently to limit anti-annealing and thermal runaway after radiation damage. Results of a finite element thermal simulation are shown in Figure 8.32 (Kondo *et al.* 2002). The disk modules use a similar structure, but the electronics are end mounted (Moorhead 2002, Feld 2003, Nisius 2004). Table 8.2 lists contributions to the material of a barrel module. At normal incidence the detector modules,

FIG. 8.30. SCT front-end ICs and associated components are mounted on a flex hybrid that wraps around the module. The hybrid integrates electronics, interconnections and the connector for both sides. Bypass capacitors are visible adjacent to each IC and off the ends of the two arrays of readout ICs. (Photograph courtesy of T. Kondo.)

Table 8.2 Material in an SCT module, expressed in percent of a radiation length X_0.

Silicon sensors and adhesive	0.612
Baseboard and BeO facings	0.194
ICs and adhesive	0.063
Cu/polyimide hybrid	0.221
Passive components	0.076
Total	1.17% X_0

the cooling and support structure, and the cabling in the four-layer tracker add up to $0.1\,X_0$.

Electrical signals are transmitted by fully balanced LVDS links. Noise pickup from the digital electronics to the sensors is negligible and modules operating in systems mock-ups show negligible common mode noise and no cross-talk (Ferrari 2004). Figure 8.33 shows a photograph of an assembled module. For a summary of test beam results see Campabadal *et al.* (2005b).

The scale of these projects does not allow the improvisation and last minute crash programs that characterize smaller projects. Small projects can make last minute changes and implement them rather quickly. Small systems also tend to be more accessible, so after some initial running it is common to take them out for rework. In huge detectors like ATLAS and CMS removing the silicon systems is a major effort that necessitates significant downtime. Thus the reliability requirements are similar to systems in space. Both ATLAS and CMS have adopted

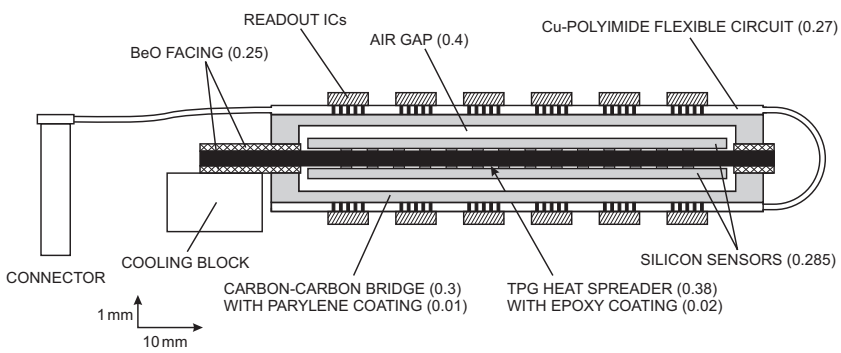

FIG. 8.31. Schematic cross-section of an SCT barrel module (vertical scale exaggerated). The hybrid is glued to a bridge to reduce heat transfer to the sensors. The height of the bridge still allows reliable wire bonding to the sensors. Thicknesses (in parenthesis) are in mm. (Figure courtesy of T. Kondo.)

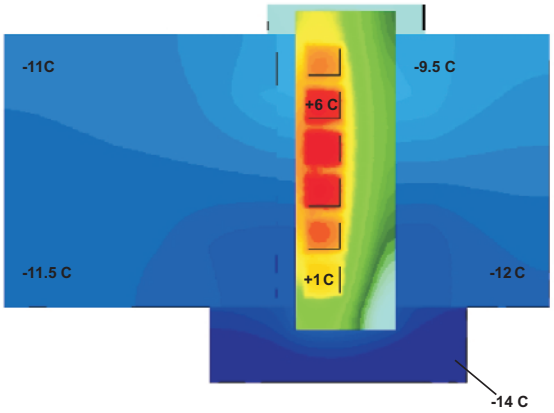

FIG. 8.32. Temperature distribution of an SCT detector module. (Figure courtesy of T. Kondo.)

extensive test and quality control procedures that track components through the production process and record test data at every step in a database (Anghinolfi *et al.* 2002, Krammer 2003, Macchiolo 2004). Assembly and testing are distributed over multiple institutions (Turala *al.* 2001, Biasini 2004), so uniform acceptance criteria must be established and enforced.

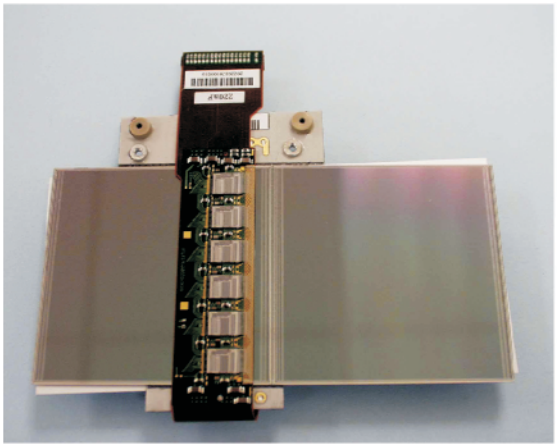

FIG. 8.33. Photograph of an assembled SCT barrel module. The attached bar code allows component tracking during tracker assembly. (Figure courtesy of T. Kondo.)

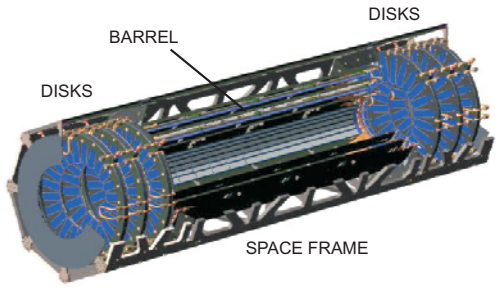

FIG. 8.34. The ATLAS pixel detector. The length of the detector is 1.4 m and the radius of the outermost pixel layer is 12 cm. (Figure courtesy of M.G.D. Gilchriese.)

8.6.6 *Pixel detectors*

The high occupancy at the inner layers of semiconductor tracking detectors at high luminosity colliders precludes the use of strip detectors. Pixel detectors are key components of ATLAS, CMS (Schnetzer 2003, Erdmann 2004), ALICE (Kuijer 2004, Stefanini 2004), and other systems. Unlike CCDs, these devices allow the selective readout of individual pixels, so they are often called random access pixel devices.

Small-scale two-dimensional segmentation allows pattern recognition at high track densities. The low capacitance provides a high signal-to-noise ratio, which allows degradation of both detector signal and electronic noise due to radiation damage. With the small detector elements the detector bias current per element is still small after radiation damage. The drawback is that the engineering complexity is at least an order of magnitude greater than in strip systems, so the path towards devices suitable for the LHC has been arduous. Along the way devices have been used successfully in DELPHI (Becks *et al.* 1997) and WA97 (Heijne 1995 *et al.*), but the LHC detectors are a significant step up in complexity and scale. Heijne (2001) gives an overview of these early systems. Designs suitable for high-luminosity colliders began in the late 1980s (Spieler 1988, Barkan *et al.* 1991, Kramer *et al.* 1991) and have come to fruition for the LHC. The ATLAS pixel device will be described to illustrate the design techniques.

8.6.7 *ATLAS pixel detector*

Figure 8.34 shows the layout of the ATLAS pixel detector (Gemme 2003). The overall pixel detector has about $2\,\mathrm{m}^2$ of sensor area and 10^8 channels. The pixels are $50 \times 400\,(\mu\mathrm{m})^2$, oriented along the beam axis to reduce distribution of the signal from inclined tracks over multiple pixels. Figure 8.35 shows a module, which consists of a $6 \times 1.6\,\mathrm{cm}^2$ silicon sensor wafer, onto which two rows of eight readout ICs with a total of 46 080 pixels are bonded by a two-dimensional array of solder bumps (Rossi 2003, for an overview of high-density interconnect technology see John *et al.* 2004). On a readout IC the pixels are arranged in

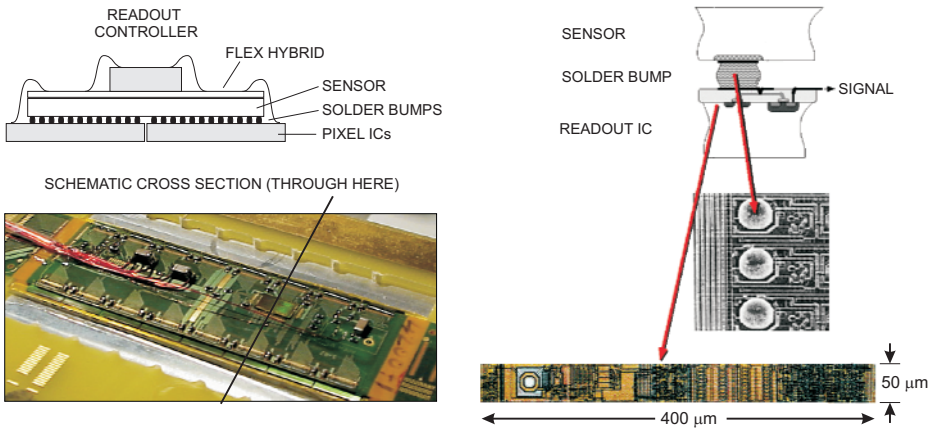

FIG. 8.35. An ATLAS pixel module. In the cross-section view the pixel ICs are at the bottom, bonded to the sensor above through a two-dimensional array of solder bumps. A polyimide flex-hybrid on top of the sensor has traces for bussing, bypass capacitors, and a readout controller IC. Connections from the pixel ICs to the flex hybrid are by wire bonds. A pixel cell and bump bond are shown at the right. The bump bond pad is a $20\,\mu$m diameter octagon with a $12\,\mu$m opening in the passivation for the solder bump. (Figure courtesy of M. Garcia-Sciveres.)

18 columns of 160 pixels. Each pixel cell contains a full electronic channel with control circuitry, described below. Tiling is accomplished without dead area by making the sensor pixels at the readout chip boundaries longer to bridge the gap. The electronic noise averages $170\,e$, obtained during simultaneous readout at a 40 MHz rate.

As shown in Figure 8.35 the output pads of the readout ICs extend beyond the edge of the sensor to allow wire bonding to a flex-hybrid, which accommodates bypass capacitors, a readout controller IC, and power, control and readout bussing. Communication is through three links, input, output, and clock. The module communicates through an optical package, up to 1 m distant, that connects to the off-detector electronics.

The sensors utilize oxygenated n-type silicon bulk with n^+ electrodes (Wunstorf 2001, Gorelov *et al.* 2002). Interelectrode isolation is provided by a contiguous "p-spray" (Richter *et al.* 1996). The n^+-on-n structure still provides good efficiency when operated below the voltage required for full charge collection, so it extends the overall detector lifetime after radiation damage. Sensors are $250\,\mu$m thick. The pixels are direct coupled to the amplifiers, as space constraints do not allow the bias structures required per pixel with AC coupling. However, some form of common biasing is required for sensor testing and also to maintain a uniform potential distribution around a faulty bond. The ATLAS sensors implement a bias grid that provides punch-through biasing, which is inactive during oper-

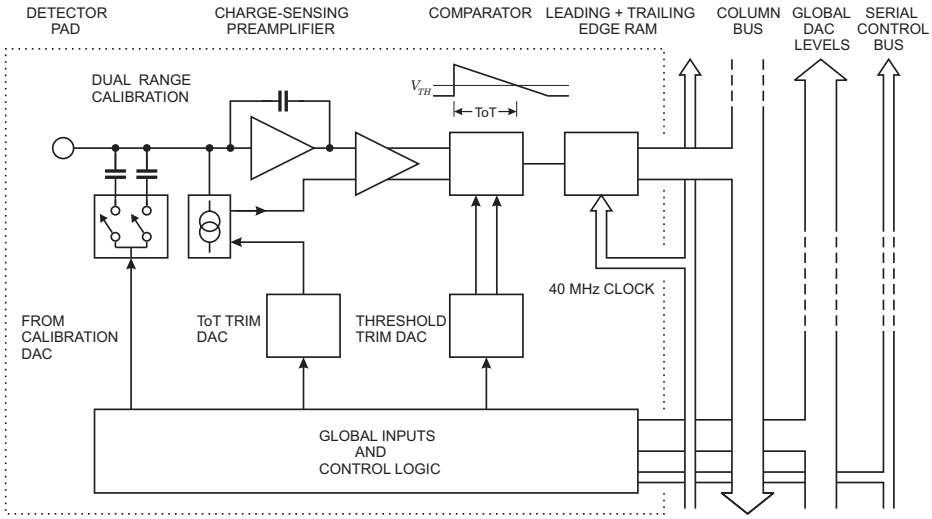

FIG. 8.36. Block diagram of an ATLAS pixel cell. Each cell contains a full analog chain: a preamplifier, shaper, and threshold comparator. A buffer records time stamps of the leading and trailing edge of each pulse and immediately sends it to a buffer at the foot of each column.

ation. Tests indicate that the voltage required for full charge collection should remain below 600 V throughout ten years of LHC running (Wunstorf 2001, see Figure 7.6). Beyond this the signal degrades gradually, both due to incomplete charge collection and trapping, so the system still remains functional, albeit with reduced signal-to-noise ratio. The readout IC is fabricated in a standard "deep submicron" 0.25 μm process and has shown only minor degradation after irradiation to 100 Mrad (Einsweiler 2004).

A block diagram of the circuitry in each cell is shown in Figure 8.36. Each pixel cell contains a preamplifier, pulse shaping with a 30 ns rise time, a threshold comparator, a trim-DAC for pixel-by-pixel fine adjustment of the threshold, time stamp logic, and event buffering (Blanquart *et al.* 2004). Current feedback in the charge-sensitive amplifier provides a linear 500 ns – 1 μs discharge and also compensates for the sensor bias current. Pulse heights are digitized by measuring the time over threshold (ToT). Unlike the BaBar AToM IC, this is designed to provide a linear response. The preamplifier is direct coupled to a two-stage differential amplifier. The reference level is generated in the preamplifier discharge block to track the baseline at the preamplifier output. The amplifier drives a differential comparator whose threshold is set by adjusting the baseline of the second gain stage. The PMOS input transistor with $W = 25.2\,\mu$m and $L = 0.6\,\mu$m operates at 8 μA drain current. The feedback capacitor is 6 fF, yield-

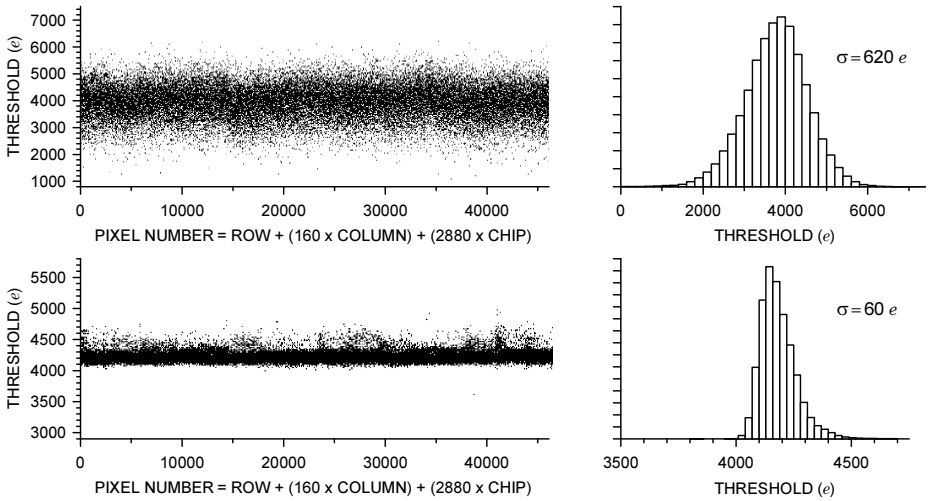

FIG. 8.37. Threshold distribution of an ATLAS pixel module before and after trimming. A signal charge of 4000 e is injected and the trim DACs in each pixel cell adjusted to minimize the threshold dispersion. (Data courtesy of M. Garcia-Sciveres.)

ing a 15 ns risetime with the sensor connected. Charge interpolation using the ToT yields a spatial resolution of 7 to 10 μm in $r\varphi$.

The small transistors required for this design do not provide adequate matching to keep threshold variations well below the noise level, so a fine adjustment is incorporated. Each pixel cell includes a seven-bit threshold trim DAC and a three-bit DAC to trim the ToT. Figure 8.37 shows the threshold dispersion before and after trimming. Individual pixel cells can be selected for charge injection, masking the output of noisy pixels, or shutdown in case of cell failure.

Globally all critical bias currents and voltages on the chip are controlled by DACs (11 DACs with eight-bit resolution). A ten-bit DAC controls charge injection and another ten-bit DAC modifies the input current discharge to provide a measurement of the leakage current of each individual pixel. Two charge injection capacitors are incorporated in each pixel to provide a low range for noise and threshold measurements and a high range for calibration, time walk and cross-talk measurements.

The analog supply voltage is 1.6 V with a total current drain of 75 mA. Each pixel cell consumes 40 μW. The digital supply is 2 V at 40 mA, so the total power dissipation of the pixel IC is 200 mW.

A 40 MHz differential time stamp bus routes timing signals to all pixels. This bus is 8 bits wide and uses a Gray code so that the number of high and low levels remains constant. During signal acquisition each pixel cell records the leading and trailing edge timing. As soon as a trailing edge is recorded a hit signal

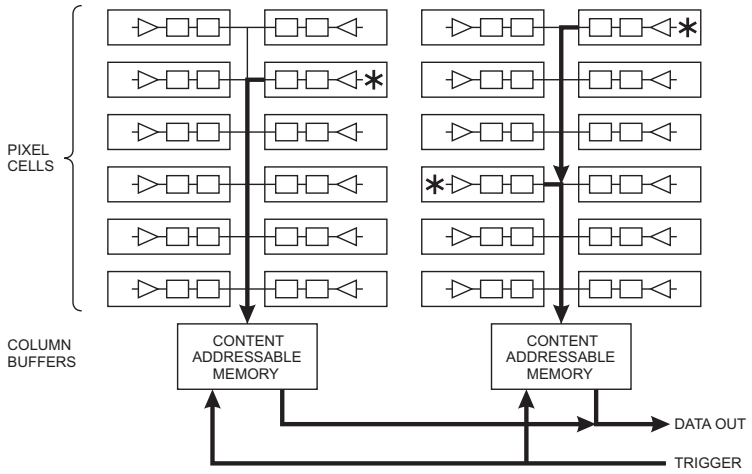

FIG. 8.38. Pixel columns are arranged as pairs of "back-to-back" cells, so the digital control and readout lines are not adjacent to the input pads. When a pixel is struck it sends its address and time stamp to its respective column buffer, configured as a content addressable memory. A trigger selects valid time stamps, which are then sent sequentially to the readout buffer.

(address plus leading and trailing edge timing) is sent to the column periphery (Figure 8.38). This operates at transfer rates up to 20 MHz. Differential drivers and receivers are used to minimize cross-talk. At the end of each column pair a content addressable buffer memory with 64 locations is available (one location for each five pixels). Upon receipt of a level 1 trigger, the buffers are checked for valid events and hits from rejected crossings are cleared. A readout sequencer stores up to 16 events pending readout. Data are "pushed" off the front-end chip to the readout chip without handshaking. All column pairs operate independently and in parallel,. The readout operations incur no deadtime, so the overall rate is limited by the buffer capacity (Mandelli *et al.* 2002).

All control data are stored in a 231-bit control register with full readback capability. Configured as a shift register with a shadow latch it utilizes triple redundancy for single event upset tolerance, as it holds critical configuration data. In addition to a broadcast mode, each chip can be uniquely addressed; its identity is controlled by external wire bonds. Intermediate metal layers and special layout limit cross-talk from the digital signals to the sensors and input nodes (Blanquart *et al.* 2004). Figure 8.39 shows the measured noise. Fischer (2003a) summarizes design and layout considerations for pixel readout ICs.

Figure 8.40 shows a reticle containing two pixel ICs, a readout controller, and support and test devices. On the pixel ICs the upper 75% are the pixel cells, whereas the lower 25% are readout logic and output drivers. Higher density processes would reduce this area and also allow smaller pixel cells. The 0.25 μm

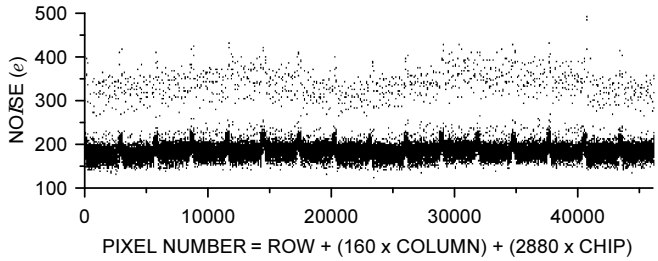

FIG. 8.39. Noise distribution of an ATLAS pixel module. Three groups are visible: the nominal pixels, the extended pixels that bridge columns between ICs (visible as spikes every 2880 pixels), and ganged pixels to bridge rows between ICs. The distributions of pixel noise levels are Gaussian, with $Q_n \pm \sigma = (184 \pm 11)\,e$, $(204 \pm 13)\,e$, and $(336 \pm 31)\,e$. This module has 7 "bad" pixels. (Data courtesy of M. Garcia-Sciveres.)

process used in ATLAS production would allow half the pixel size, but the sensor design was determined for an earlier design. Die size is $7.2 \times 10.8\,\mathrm{mm}^2$. The edges of the die may not extend beyond $100\,\mu$m from the active area. External contacts are by 30 wire bonds, with $100 \times 200\,\mu\mathrm{m}^2$ pads. Diagnostics are provided on 17 additional bond pads. ICs are thinned to $180\,\mu$m. Starting thickness of the wafers is $\sim 500\,\mu$m. The chip boundaries are grooved to a depth of $\sim 200\,\mu$m and then

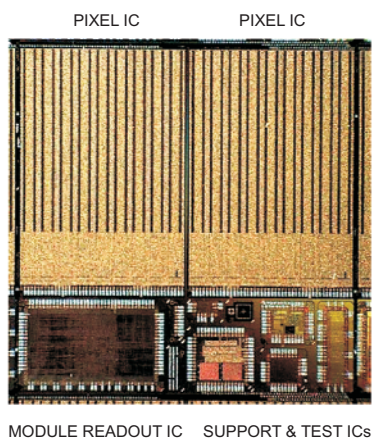

FIG. 8.40. A reticle of the pixel IC wafer, showing how multiple ICs are accommodated in one reticle. Reticles are copied by a step-and-repeat process to fill the entire 200 mm wafer. Two pixel ICs are at the top, with the readout controller and test/support ICs at the bottom. The pixel IC is $7.3 \times 10.9\,\mathrm{mm}^2$ and contains 2880 pixels. (Die photo courtesy of K. Einsweiler.)

the wafer is ground to a thickness of $\sim 180\,\mu$m, releasing the chips with very smooth edges. Material in a module is about $0.8\%\,X_0$. Including the support structure and services the material adds up to $7.5\%\,X_0$ at normal incidence, increasing to about $35\%\,X_0$ at $\eta = 2$.

Other pixel designs incorporate many of the techniques described above. CMS also uses a bump-bonded pixel structure, but the pixel size is $100\,\mu\text{m} \times 150\,\mu\text{m}$ after migrating to a $0.25\,\mu$m process for the readout IC (Erdmann 2004). A column-based readout is also used, but with a fully analog readout (Barbero *et al.* 2004). ALICE uses $50\,\mu\text{m} \times 425\,\mu\text{m}$ pixels with a binary readout. A fast OR output can be used in the level 0 trigger. The proposed BTeV experiment took the readout a step beyond all existing silicon detector systems by reading out at the full beam crossing rate of 7.6 MHz (Kwan *et al.* 2003, Wang 2003).

Design and construction of these large-scale pixel systems pose formidable challenges. The ATLAS pixel IC contains nearly four million transistors, so simulation and design verification are crucial. As in all large-scale semiconductor detector systems, electro–mechanical integration – combining sensors, electronics, cabling, cooling, and mechanical support systems – is a major part of the project. The complexity of integrating these systems is usually not appreciated by those who haven't done it. Furthermore, these systems are chronically underfunded, as funding agencies, reviewers, and project managers tend to underestimate the required effort.

8.7 Monolithic active pixel devices

The hybrid structure has the drawback of requiring bump bonding. Currently, only a few vendors provide this service at the fine pitches required. Furthermore, the cost and technical overhead are barriers for small projects. A fully monolithic sensor that utilizes mainstream IC technology would simplify construction and reduce costs.

8.7.1 *CMOS imagers*

Monolithic pixel devices are in widespread use in optical imaging ("active pixel arrays" or "CMOS imagers") and are being developed for charged particle detection (Deptuch *et al.* 2003, Turchetta *al.* 2003, Kleinfelder *et al.* 2004). As described in Chapter 1, the epitaxial layer in conventional CMOS processes is used for detection (Figure 1.23). Devices have also been implemented on $10\,\Omega\,\text{cm}$ substrates (Dulinski 2004). This provides an interesting alternative, as modern processes tend to reduce the epi-layer thickness, which decreases the signal. However, since charge collection is primarily by diffusion, charge reflection at the epi-layer boundary adds to the recovered signal, as discussed in Section 8.4.5. Signals tend to be small, of order $10^3\,e$, so low-noise pixel circuitry is essential.

A challenge in implementing large CMOS imager arrays is achieving full coverage. In mainstream processes the size of an array is limited by the maximum reticle size. For tracking of high-energy charged particles double-layered structures with overlapping dice are feasible, but for the imaging of visible light or

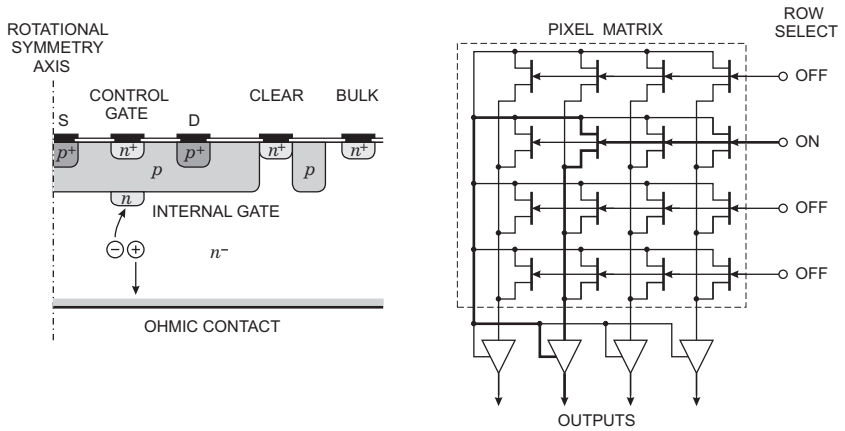

FIG. 8.41. DEPFET structure (left) and readout matrix (right). The second row is selected for readout and all outputs are available. The control and signal paths are shown for the second pixel.

soft x-rays the dead area of the front IC blocks the active area of the one behind. The gap between dice can be reduced by including multiple reticles in one die. With some compromises in layout, circuitry might also be "stitched" to form contiguous multireticle ICs. However, both techniques are limited by yield.

This is a very active area and the limits of the technology are still being explored. Adequate signal-to-noise ratios for particle tracking have been demonstrated and designs for full systems are underway, especially targeting linear collider applications. In those applications that require sparsified readout, the circuit complexity will be similar to the ATLAS pixel readout described above.

The charge collection mechanism limits the radiation resistance of these devices. As charge collection is primarily by diffusion, the magnitude of the signal is sensitive to minority carrier lifetime (Appendix F), especially since the injected charge concentration is low. Displacement damage will reduce the minority carrier lifetime to 10 ns after exposure to a hadron fluence of order $10^{13}\,\mathrm{cm}^{-2}$, scaling inversely proportional to fluence (Messenger and Ash 1986). Since the diffusion time is of order 100 ns, this limits the application of these devices to less severe environments.

8.7.2 *DEPFET pixel detectors*

The simplest active pixel device is an array of transistors. A novel implementation that integrates the sensor and transistor monolithically is the "DEPFET" structure (Kemmer and Lutz 1987, Richter *al.* 2003). The DEPFET combines the functions of a charge collecting electrode and an FET. The structure is shown in the left panel of Figure 8.41. The p-channel JFET can be controlled by two gate electrodes. The upper gate is used to activate the device. The buried gate senses

the signal charge. Through a combination of doping and biasing a potential well is formed at the buried gate so that signal electrons from the associated pixel volume will collect there, regardless of whether the FET is activated or not.

For readout the DEPFETs are configured as a matrix, as shown in the second panel of Figure 8.41. All control gates in a row are bussed, as are all collectors in a column. By switching the gates on, all FETs in the corresponding row become the first stage of a readout amplifier that drives the external second amplifier stage through the common output bus. The first and second stages are conveniently configured as a cascode. After readout the internal gate is discharged, by forward biasing either the clear electrode or the control gate. Discharge is not always complete and the reset itself can introduce baseline shifts, so the residual baseline must be recorded and subtracted from the next readout (Fischer *et al.* 2003b).

The absence of external connections to the input node reduces stray capacitance and the low input capacitance yields low equivalent noise charge. In matrices operated with 50 kHz line rates a noise level of 2.2 eV has been achieved with 131 eV resolution for 5.9 keV x-rays. The key parameter is the conversion of signal charge q_s to output current $g_q = \Delta I_D/\Delta q_s$. Values of 200 pA/$e$ are typical (Fischer *et al.* 2003b) and as high as 400 pA/e have been reported (Wermes 2004). This is to be compared to a MOSFET's g_m/C_i, which in modern devices can be significantly higher, even for low-power operating points (see Chapter 6).

Obtainable noise levels and full frame readout rates are comparable to CCDs. Similarly, since charge is stored when the device is inactive, applications with low frame rates require very low power. Common to both is that the shot noise depends on the exposure time between readouts. During an exposure time T the sensor dark current I_d accumulates I_dT/e electrons, so the shot noise

$$Q_{ni} = \sqrt{\frac{I_dT}{e}} \tag{8.24}$$

is independent of the external pulse shaper. However, the voltage noise contributions from the front-end transistor, series resistances, *etc.* depend on the bandwidth of the signal processing chain.

Both DEPFETs and CCDs can be implemented on fully depleted substrates with hundreds of μm thickness (fully depleted CCDs are described below), so achievable signal levels are comparable. Both utilize specialized processes that are not constrained by the die size of commercial CMOS, so large devices are possible, as described in Section 8.4.5. Unlike CCDs, DEPFET readout is nondestructive and pixels can be addressed individually, so selected fields can be read out rapidly while continuing to integrate the full field. Readout also proceeds without the data moving through the image frame. DEPFET structures have been studied for x-ray astronomy (Holl 2000), tritium autoradiography in biomedical applications (Ulrici 2004), and vertex detectors for linear colliders (Trimpl 2003).

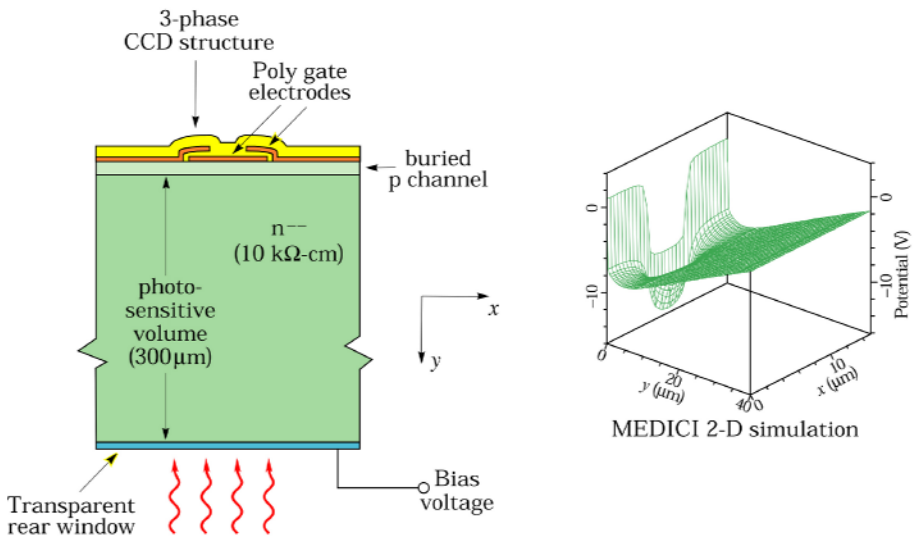

FIG. 8.42. Fully depleted CCD structure and potential distribution into the bulk. Light impinges on the rear window (bottom left). (Figure courtesy of S. Holland.)

8.8 Astronomical imaging

Astronomical imaging of visible light today relies on CCDs. Unlike the applications described for high energy physics, which emphasize fast readout, exposures in astronomy can be very long, ranging up to hours, so low dark current is crucial. A few electrons per hour per pixel are achievable. The photon flux from weak objects is very low. A typical system on a ground-based 4 m telescope has a sensitivity of 1 photon/s for a magnitude 26.7 object. To set the scale, humans can recognize stars of magnitude 5 to 6 with the naked eye, where 5 units of magnitude correspond to a flux ratio of 100. Since the signal transferred through the CCD structure is very small, CCDs for astronomical imaging are very sensitive to traps that are unimportant in applications such as digital cameras, where the brightness of illumination fills the traps rapidly. Electronic noise in the readout amplifier sets the ultimate sensitivity. Today noise levels of a few electrons at sample times of $\sim 5\,\mu$s are typical. As illustrated in Section 8.4.5, multiple readout amplifiers are used to increase the frame rate.

Frontside illumination of CCDs limits the quantum efficiency, because of absorption in the metallization and charge transfer structures, so devices are back-illuminated. However, since the substrate is field-free, the position resolution is limited by transverse diffusion, which is roughly equal to the thickness of the material the carriers must traverse by diffusion (Groom *et al.* 1999 give a detailed analysis). For use in astronomy devices are thinned to about $15\,\mu$m, which greatly increases the cost and incurs a host of other problems. Janesick

(2000) gives a comprehensive description of CCD technology and systems for astronomical imaging, so only one novel development will be described here to indicate how disparate fields can benefit from one another.

The CCDs used in high energy physics are adaptations of devices developed for other applications. Conversely, technologies developed in the course of generic detector R&D for high energy physics have enabled a new class of CCDs for astronomy and other applications.

The first ingredient was the monolithic integration of high-quality detectors and electronics on high-resistivity, fully depleted substrates (Holland and Spieler 1990, Holland 1992). Although portrayed in Chapter 1 as an evolutionary "dead end" for detectors with complex readouts, this technology is well-matched to CCDs.

The second ingredient was to adapt the technology to photodiode arrays. The backside gettering layer provided the low dark current required, but the backside dead layers had to be thinned substantially to obtain high quantum efficiency in the visible. The impetus for this development came from medical imaging and was implemented in the course of developing the photodiode arrays in the PET system whose readout was described in Chapter 4 (Choong *et al.* 2002, Holland, Wang, and Moses 1997).

Unlike conventional CCDs that use a p-substrate and transport electrons, this device uses an n-substrate and transports holes. The substrate is fully depleted by an applied bias (Holland *et al.* 2003). Figure 8.42 shows the structure and potential distribution. The applied field speeds up collection time, which limits transverse diffusion; at 30 V bias voltage the transverse diffusion is about $10\,\mu$m rms (Karcher *et al.* 2004).

For astronomical observations the $300\,\mu$m depletion depth has the very important advantage of improving the red response, as shown in Figure 8.43. Since the interstellar dust absorbs in the blue, the extended red response significantly enhances imaging sensitivity (Groom 2000, Holland *et al.* 2003). Radiation resistance is also good; devices have been tested to fluences of $10^{11}\,\mathrm{cm}^{-2}$ 12 MeV protons (Bebek *et al.* 2002).

8.9 Emerging applications

8.9.1 *Space applications*

CCDs are at the heart of the Hubble Space Telescope and larger arrays will be used in future faint light imagers in space. The fully depleted CCDs described above are the enabling technology for a proposed satellite observatory, the SuperNova Acceleration Probe (SNAP) (Linder 2002, Aldering 2004). A typical focal plane design includes 36 $2\mathrm{K} \times 2\mathrm{K}$ HgCdTe near infrared sensors with $18\,\mu$m pixels and 36 $3.5\mathrm{K} \times 3.5\mathrm{K}$ CCDs with $10.5\,\mu$m pixels to cover the visible spectrum. The complete focal plane would be cooled to 140 K. Packaging is a major challenge. The $200-300\,\mu$m fully depleted CCDs are self-supporting, so single devices can be mounted on a "window frame" made of aluminum-nitride, which closely

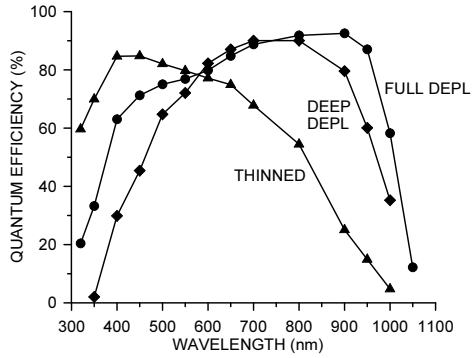

FIG. 8.43. Quantum efficiency of a thinned CCD, a deep depletion CCD (high resistivity substrate, partially depleted), and a fully depleted CCD with 300 μm sensitive thickness. (Data courtesy of S. Holland)

matches silicon's thermal expansion. However, the width of the frame is not suitable for a densely packed array. In a higher density package the CCD is glued with the circuit side down to a contiguous aluminum-nitride carrier. The CCD extends beyond the carrier, so the bond pads are accessible to make connections to traces on the backside of the carrier, which also accommodates bypass capacitors, heater resistors, and a connector for the readout cable. (Stover *al.* 2004). The large-scale readout utilizes custom-designed ICs and builds on experience from high energy physics.

X-ray detectors are important for both cosmological surveys and the study of black holes. Some desired characteristics are summarized by Remillard (2004). In the soft and medium x-ray band (0.5 – 15 keV), for example, the detector goals include a pixel size of 20 μm, spectral resolution of 2 keV at 6 keV, a time resolution of 100 μs, and a count rate capability of $10^4\,\mathrm{s}^{-1}$. No single detector technology offers the prospect of meeting all of these goals. Gas counters and CCDs are commonly used, but enhanced CCDs or hybrid pixel arrays can improve substantially on current capabilities. However, this will require substantial development.

Large silicon arrays are integral to a new space-based gamma-ray telescope. The Gamma-ray Large Area Space Telescope (GLAST) utilizes stacks of alternating conversion foils and silicon strip detectors to track gamma-rays in the range of 20 – > 300 GeV (Attwood 1994, Morselli 2004). Figure 8.44 shows the principle. The tracker consists of 16 square towers about 37 cm on a side and 60 cm high. The modular design utilizes individual trays that are stacked to form a tower (Bellazini *et al.* 2003). As shown in Figure 8.45 each tray includes two layers of single-sided strip sensors with a 90° stereo angle, a tungsten converter foil, and a multichip module with 24 readout ICs and a readout controller. An

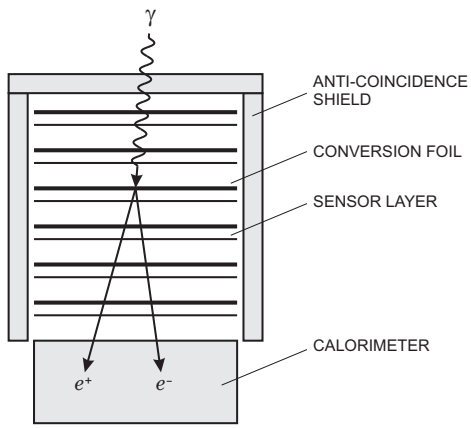

FIG. 8.44. Incident gamma-rays produce e^+e^- pairs in converter foils and the trajectories of the particles are measured in a silicon strip tracker.

aluminum honeycomb with carbon-carbon facings provides rigidity. The total weight of the 16 tracker towers is 500 kg.

By the standards of high energy physics the technology is conventional, but it is quite novel for NASA. The detectors are fabricated on 150 mm wafers. Four sensor wafers are ganged to form a 36 cm long ladder read out at one end. Four ladders comprise the active area of a layer and 36 layers (18 x-y pairs) per tower provide sufficient redundancy to tolerate the loss of a few layers. Building on experience from BaBar, the readout consumes only 190 μW per channel at a noise level that ensures a noise occupancy $< 10^{-4}$ per trigger. The total power dissipation of the 24 front-end ICs and the readout controller on the readout hybrid is 0.25 W. With a strip pitch of 228 μm the tracker has nearly $9 \cdot 10^5$ channels and consumes less than 160 W.

The silicon tracker is an international collaboration with key fabrication and assembly steps in Italy (tower assembly), Japan (sensors), and the U.S. (electronics and final integration). Sensor design builds on extensive experience from high energy physics (Ohsugi *et al.* 1999). Detector performance is excellent; measurements on 600 ladders after assembly show average bias currents of 600 nA per ladder with an rms spread of 200 nA (Latronico 2004). Only 105 of $1.1 \cdot 10^7$ strips didn't meet specifications, *i.e.* a rejection rate of 10^{-5} (Bellazini *et al.* 2003). All communication between ICs and with the external readout electronics is by current balanced LVDS, which yields very low cross-talk and no measurable electromagnetic radiation, even without shielding (Nelson 2004).

8.9.2 *X-ray imaging and spectroscopy*

X-ray imaging and energy spectroscopy are important in many areas. However, their requirements differ significantly from high energy physics or x-ray astron-

omy. Unlike tracking devices in high energy physics, which require low noise to obtain a high detection efficiency of a broad energy distribution, x-ray spectroscopy demands precision pulse-height measurements, often at high fluctuating random rates that require excellent instantaneous baseline control. This also sets them apart from x-ray astronomy, which requires high energy resolution, but at rather low rates. Image fields are filled, so the frame rate cannot be increased by on-chip sparsification. The required trade-off between energy selectivity and rate capability depends greatly on the application. Digital imaging, *i.e.* photon counting, lessens the requirements on energy precision, but still requires energy windowing to reduce backgrounds. Integrating detectors are also widely used. This is one area where amorphous silicon arrays are useful, as the total signal is larger than in single photon counting. In contrast, precision x-ray spectroscopy often requires resolution of weak lines adjacent to strong lines, so full charge collection in the sensor is essential to reduce low-energy "tails", as is precision baseline control under random rates. Figure 4.18 in Chapter 4 illustrates the required performance. Hybrid pixel arrays offer great promise in this application, as the sensor and electronics can be optimized individually. Small systems are common in these areas, so the effort required for dedicated ICs is prohibitive and developing ICs that can serve multiple applications will make the technology more accessible. The Medipix chip is an example of this type of development (Campbell *et al.* 1998, Llopart *et al.* 2002, 2003).

Silicon is limited to energies $< 30\,\mathrm{keV}$ or so because of the rapid decrease in photoelectric cross-section (Figure 1.20). Figure 8.46 shows the required thick-

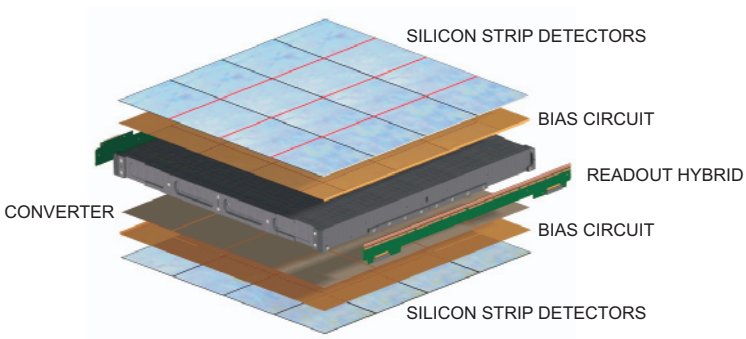

FIG. 8.45. The GLAST tracker towers consist of stacked trays, which include two layers of single-sided silicon strip detectors, the converter, and the front-end electronics module. In the middle an aluminum honeycomb with carbon-carbon facing provides structural rigidity. Connections from the sensors to the front-end electronics make the 90° bend through a flexible pitch adapter. The two silicon strip layers are arranged orthogonally to provide x and y coordinates. (Figure courtesy of R. Johnson.)

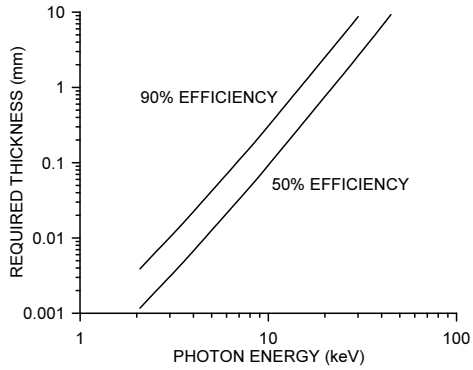

FIG. 8.46. Required silicon detector thickness for 50% and 90% full energy efficiency.

ness *vs.* energy for efficiencies of 50 and 90%. Efficiency improves rapidly with atomic number, which has prompted extensive studies on high-Z detector materials (Owens and Peacock 2004). However, as discussed in Chapter 2, silicon and germanium still provide the best energy resolution. Rossington (1992) compares Si and Ge for x-ray fluorescence spectroscopy in the 2 – 20 keV range. Despite germanium's virtues, it must be cooled to maintain low bias currents, so silicon is still the material of choice when it provides adequate efficiency.

The development of planar detector structures that allow reliable operation at high bias voltages has extended the practical thickness of economical silicon material. A reverse bias of 500 V will provide 1 mm depletion depth with $10^4\,\Omega\,\mathrm{cm}$ n-type silicon. The thickness can be extended by stacking multiple sensors, as illustrated in Figure 8.47. Sensors can be spaced to accommodate the electronics. Clearly, this "brute-force" technique has its limits, but once an integrated sensor module has been developed, replicating many modules is fairly straightforward. Stacking six modules with 500 μm sensors would provide $> 50\%$ efficiency at 30 keV. Fully depleted CCDs with 650 μm sensitive depth have been

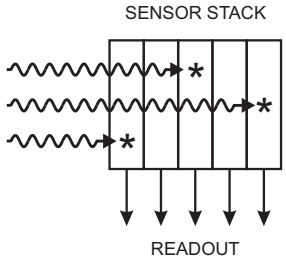

FIG. 8.47. Sensors can be stacked to extend photoelectric detection efficiency to higher energies.

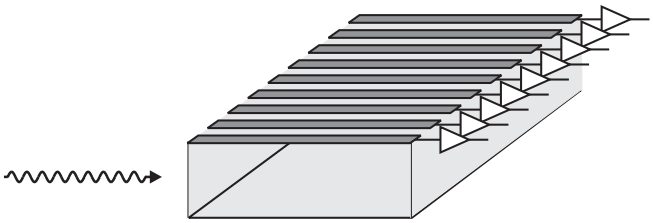

FIG. 8.48. Edge-on illumination of a thin silicon sensor provides a larger absorption depth with position resolution elements determined by the strip pitch and sensor thickness.

demonstrated (Holland 2005). A stack of fully depleted CCDs would provide nonprojective imaging coupled with a measurement of the interaction depth. Another technique to extend the efficiency of silicon strip detectors to higher energy is edge-on illumination, illustrated in Figure 8.48 (for example Arfelli 1997, Beuville 1998). In this configuration stacking multiple detectors increases the area. The readout electronics can be connected through flat ribbon cables fanned out to accommodate the thickness of the electronics modules, but this has limits, so edge-on illumination is best suited to flat beam profiles or scanned slit applications. Systems have been designed and evaluated under conditions representative of mammography (Lundqvist *et al.* 2000).

A complete survey of imaging applications goes beyond the scope of this book. Nevertheless, the basic concepts discussed in previous chapters apply. Surveys by Wermes (2003) and Mikulec (2003) illustrate various applications with numerous references to more detailed discussions.

8.10 Design, assembly and test

Although listed sequentially, design, assembly, and test are intermingled. Designing a system without thinking through the assembly procedure and how components will be tested invites surprises, from which one may not recover. Conversely, testing without understanding the weak points of the design or technology can easily become a major effort that misses key points.

8.10.1 *Design*

The most important step in the design is determining what is needed, rather than "what would be nice". Many designs fail because of "feature creep", adding "enhancements" in mid-stream that are not really needed, but add complexity or just clutter. Often ideas appear clever or elegant, but don't add much to functionality. Complex systems offer great potential for surprises, so it is prudent to adopt well-understood techniques unless there are valid reasons to adopt something new. Innovations can bring significant advantages without increasing risk, but making this decision requires knowledge and technical understanding.

Superficial judgments can kill innovations that are good or accept innovations that lead to grief. A little bit of knowledge can be a bad thing.

Sensor technology is very well developed. For applications that don't require the utmost in radiation resistance, silicon sensors are available from several vendors without requiring extensive development programs. Again, the most important point is determining what is really needed.

IC design commonly is a problem area. On the one hand, the availability of "free" design software and inexpensive access to multiproject fabrication runs allows just about anyone to design an IC and obtain some chips. Some groups have used this approach very effectively, but many have failed. Designing a successful IC requires more than expertise in circuit design. Since these are special purpose ICs, the application environment is not as well understood as for mainstream devices. Furthermore, the ICs of interest in this context tend to be rather complex subsystems, which in turn must be integrated into a system. This requires more breadth than found in typical university engineering departments, which tend to emphasize specialization.

The other side of the IC design problem occurs in groups of professionals. Highly sophisticated design tools assist in cycles of

1. establishing an initial design,
2. simulating its electrical performance,
3. laying out the silicon IC,
4. extracting parasitic resistances, capacitances, and inductances,
5. incorporating parasitics into the circuit simulation,
6. revising the design,
7. another cycle?

Simulations are performed at the circuit level, block level, and for the complete IC. They must evaluate performance over the range of parameter variations specified for the IC fabrication process ("corner parameters"). MOSFET models in the weak and moderate inversion regime often show significant disagreement with measured device data, so models should be checked against test devices. Output resistance is another problem area. Often different simulators claiming the same functionality will find (or miss) different problems, so multiple simulator runs are necessary. On a complex IC this can be quite time consuming. After the design and layout was completed, the presubmission simulations and design verifications of the ATLAS pixel IC took several months, but when received the IC performed to design specifications. Unfortunately, the better simulation and verification packages are expensive.

A common pitfall is the last minute change that is deemed so simple that a new verification is not necessary. Deadlines often lead to premature submissions. Schedules are necessary to keep a project on track, but the time lost by submitting and then troubleshooting a flawed design is much greater than delaying submission to complete the full suite of verifications.

An important aspect in any design is the ease of prototyping. System-like tests should be performed early in the design. This does not require a full system, but incorporates key components to assess how well they perform together. Testing a detector module is one step, operating several modules together the next. In developing the design several cycles of prototyping and test will be necessary. If prototyping requires a technology that doesn't allow quick turnaround, development will be delayed. One project relied on bump bonding where it wasn't necessary and the schedule of a major detector project was jeopardized when problems arose without time for test prototypes. Adoption of a less ambitious design allowed timely completion of the detector and provided the required performance. A similar concern applies to many "modern and elegant" solutions, for example schemes incorporating three-dimensional packaging that stack multiple chips and interconnect layers.

Designs should be scrutinized for "single-point failure modes", where a single flaw will affect large portions of the system. Examples are power connections and daisy-chained control busses. Doubling the wire bonds on critical connections is good practice. Often circuit redundancy can be incorporated without a significant increase in die size, especially in control and communication blocks.

8.10.2 *Assembly*

Assembly of basic detector modules is possible with rather modest facilities. Specialized equipment includes a wire bonder, probe station, and an oven for gluing. Special clean rooms are not necessary, as both detectors and ICs have protective layers against chemical contamination, so the major problem is dust. Existing laboratory space can be upgraded with minor changes. Repainting with nonshedding paint may be all that is needed. Ventilation systems are a prolific source of dust, so mounting filters on the air supply vents is good practice. An alternative is to put critical assembly steps inside local clean areas, frames with polyethylene curtain walls, for example. High humidity levels can be a problem, but standard air conditioning systems are usually sufficient to maintain acceptable levels. Personal discipline is most important. Introduction of debris is reduced by requiring caps, gowns, and booties over shoes. This is not always absolutely necessary, but it sets a cleanliness standard and serves as a reminder that good work practices are important.

Despite the presence of passivation layers on most detectors and ICs, they are still sensitive to contamination. Devices should not be exposed to solvents and electronic grade glues should be used for mounting. Sodium content is especially critical. Although silicon is remarkably resilient, thermal expansion coefficients of mounts should be matched to silicon, especially when curing glue joints at elevated temperature.

CMOS devices are extremely sensitive to electrostatic discharge (ESD), primarily due to breakdown of the gate oxide. They should be placed in conductive bags or containers and all objects that come into contact must be at the same potential. Conductive work surfaces are easy to implement and conductive pads

can be placed on the floor. Conductivity need not be high; even wood or cardboard typically absorb enough moisture to drain off charge, although one should not rely on that. Plastic work surfaces and garments of synthetic fabrics should be avoided. Assembly staff commonly wear wrist straps, but they must be connected to an appropriate reference point. Severe ESD creates small "craters" on the device and is easy to diagnose, but even without causing breakdown the damage to the gate oxide affects device characteristics. One widely reported case of "mysterious" IC failures nearly led to a requirement for retesting ICs at every handling step. Asking the people who actually did the work provided the explanation – ESD due to poor handling procedures. Unnecessary retesting of dice and detector wafers should be avoided, as they are most vulnerable before mounting.

Large-scale systems are more demanding than small prototype runs, as recovering from failures requires more time and effort. Production assembly occurs at many levels and can be distributed efficiently across multiple institutions, as it often must because of multiple funding sources. Specialized tooling must be developed and maintained and metrology systems are essential to check and maintain mechanical precision. This does require temperature control and for the assembly of large systems specially prepared spaces are essential.

8.10.3 *Testing*

Testing means many things to many people. Incomplete testing is common, from the practitioner's "I switched it on and it works" to the IC engineer who after detailed measurements proclaims a chip to be "fully functional", although it is unusable because it doesn't work in a system. Tests must have specified goals and the reach of the test must be well-understood. To avoid unnecessary delays, test strategies and techniques must be developed during the design phase (they usually aren't). The functional requirements of detector readout ICs often provide built-in test capabilities. For example, on-chip charge injection circuitry provides a test of analog performance at the wafer probe level and appropriate command sets can be developed that exercise digital functions. Preparing a test system requires time, so it is wise not to wait until the chips arrive (although that's what usually happens).

Although we all like to plan for success, testing should anticipate failure. Testing detectors and ICs prior to combining them in a detector module makes it easier to localize problems. Detector yields can be quite high, but with IC yields ranging from 20 to 90% testing is essential. Probe stations can be equipped with probe cards that allow contacting many pads at once (64 or 128 probes are common). Semi-automatic probe stations will step from die to die on a wafer and automated measurement systems can record test conditions and measured results for future reference (see Anghinolfi *et al.* 2002, for example). Identifying yield-critical fabrication steps is very important. For example, chip failures on assembled modules require delicate rework, which can easily cause new problems, so screening of individual ICs prior to assembly must be thorough to ensure a

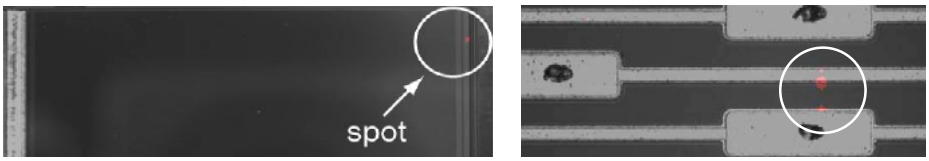

FIG. 8.49. IR image of a microdischarge site in a silicon strip detector. (Figures courtesy on Y. Unno)

very low defect rate. An ATLAS pixel module includes 16 pixel ICs, so if the yield of a module is to be $> 85\%$, individual readout ICs must be tested to $> 99\%$ reliability.

A common pitfall is to test for potential flaws that are known, rather than casting a net for the unexpected. This often requires testing for things that "can't be wrong". In developing test procedures it is important to determine what a given test will accomplish. "Burn in", *i.e.* operating a system for a few days at elevated temperature, is commonly overrated, as it typically only uncovers gross defects that a quick functionality check would find. The concept of "infant mortality" applies primarily to flaws in workmanship. Indeed, the oft-cited "bathtub curve" with early infant mortality and late wearout failure doesn't apply to ICs, where reliability is ensured by process design and control. For a detailed discussion see Lall, Pecht, and Hakim (1997). Vibration and thermal cycling are useful techniques. Thermal cycling should also be considered at the component level, as power dissipation is not uniformly distributed and can vary with operating modes. Here "burn in" can be useful, but only if the system is operated over the range of operating modes occurring in the final system.

Some failures come by complete surprise. One example is the failure of wire bonds in the CDF silicon system (Bolla 2004). During data taking pulsed currents in IC wire bonds excited motion in the 1.4 T magnetic field. The clock timing approximately matched the mechanical resonances, which led to premature failure of the wire bonds. Changing the data taking procedures alleviated the problem. This test would be difficult to perform in the lab, but the problem is avoided in designs with small transient currents.

Some techniques are not suited for production testing, but provide diagnostics. Ohsugi *et al.* (1994) have applied IR imaging to the investigation of microdischarges in detectors. Figure 8.49 shows an individual microdischarge site. The rate of microdischarges typically decays, so after about two hours of conditioning, *i.e.* gradually increasing the bias voltage, the leakage current assumes its stable value. In one module where this was not the case, a hot spot was found where a wire shaving had landed on the detector surface (Unno *al.* 2003).

Large systems require well-developed test procedures at every step of the assembly process to avoid excessive yield losses. Target parameters must allow sufficient margins to accommodate parameter variations and the uncertainty of

radiation load estimates. Component screening is typically conducted at room temperature, whereas some systems must operate cold, so the temperature dependence of operating parameters must also be accounted for. ATLAS and CMS have both developed detailed assembly and test procedures with extensive data tracking of individual components. Since practically all major steps are distributed over multiple institutions, performance specifications must be monitored and enforced at multiple sites. This allows some latitude in the choice of equipment and both experiments appear to have dealt successfully with fabrication processes that span the globe.

8.11 Summary

High energy physics experiments have driven the technology of large-scale semiconductor detector systems and we can expect this to continue. Although not comprehensive, the examples shown above illustrate the diversity of implementations. Since the overall utility of a complex system accrues from many components, there is no single "correct" design. Use of a mature and well-developed technology provides flexibility in choosing implementations and usually allows "work arounds" to solve unanticipated problems. The choice of technology depends on access to expertise. For example, the success of the flex hybrids used in the ATLAS SCT was critically dependent on access to an experienced and reliable vendor willing to take on a rather small project. Although the choice of technology is important, thorough design and implementation are crucial. Some technologies are only available commercially for large volume orders, so they are impractical for small-scale R&D and could severely impede the design and prototyping effort in larger systems.

Silicon strip technology is quite mature. Nevertheless, as in the past, it appears that acceptable radiation limits are set by the sensor, rather than the electronics, so improvements in sensor technology will increase radiation resistance. As the size of silicon trackers increases, technologies for signal and power bussing as for cooling and support systems will gain increased importance. Reduced feature sizes in ICs raise the same issue. Device performance depends on current, but operating voltages are dropping with feature size. Power bussing is becoming much more critical, as voltage drops must be controlled more tightly. The ability to supply power at elevated voltage and reduced current will become more important in controlling both voltage drops and material, but efficient power conversion techniques must be developed that will function in high magnetic fields and sustain radiation damage.

Pixel systems remain the frontier. CCD technology now offers larger sensitive depths in fully depleted devices, which extends x-ray detection to higher energies and increases the signal in particle tracking. The increased signal maintains signal-to-noise ratio at higher readout rates. Overall frame rates can be increased by adopting column-parallel readout. DEPFET arrays offer similar characteristics and performance, with the additional feature of selective read-

out. The fabrication facilities used for both CCDs and DEPFETs can provide "wafer-scale" devices and are geared to special requirements.

Active pixel sensors utilizing mainstream CMOS foundry processes offer the important benefits of mechanical simplicity and access to fabrication. High-density circuitry is possible, so that the complexity of readout cells can go beyond the simple readout matrices used in digital cameras. However, the signal originates in an ancillary component of the process, the doping and thickness of the epitaxial layer, so industry trends are not well aligned with the requirements of detectors. As these devices are limited to standard IC die sizes, tiling is an important design consideration in large arrays. The devices can be thinned to reduce material, but this must be balanced against the required support structures together with the signal and power bussing. Although limited in design flexibility, this technology offers significant benefits in those applications that can find an acceptable compromise in performance.

The hybrid pixel array, which combines separate sensor and readout units, has many advantages. It opens up the choice of sensor material, be it for x-ray detection or increased radiation resistance, and the readout can exploit future advances in IC technology. Tiling in large arrays is facilitated, as the sensor can "bridge" the gaps between the readout ICs. Gaps between arrays can be reduced by advanced sensor designs, for example "3D sensors" that can practically eliminate inactive edge regions. The drawback of the hybrid array is the complexity of bump bonding. Industry is moving towards smaller bonding pitches (John *et al.* 2004), so this situation could change and the technology should become more accessible as high-density packaging in the electronics industry proliferates.

References

Abbaneo, D. (2004). Layout and performance of the CMS Silicon Strip Tracker. *Nucl. Instr. and Meth.* **A518** (2004) 331–335

Abe, K. *et al.* (1997). Design and performance of the SLD vertex detector: a 307 Mpixel tracking system. *Nucl. Instr. and Meth.* **A400** (1997) 287–343

Adolphsen, C. *et al.* (1992). The Mark-II silicon strip vertex detector. *Nucl. Instr. and Meth.* **A313** (1992) 63–10

Albiol, F. *et al.* (1998). Performance of the ATLAS silicon strip detector modules. *Nucl. Instr. and Meth.* **A403** (1998) 247–255

Allport, P.P. *et al.* (1994). The OPAL silicon strip microvertex detector with two coordinate readout. *Nucl. Instr. and Meth.* **A346** (1994) 476–495

Alpat, B. *et al.* (1992). The design of the L3 silicon microvertex detector. *Nucl. Instr. and Meth.* **A315** (1992) 197–200

Anghinolfi, F. *et al.* (2002). ASIC wafer test system for the ATLAS Semiconductor Tracker front-end chip. *IEEE Trans. Nucl. Sci.* **49/3** (2002) 1080–1085

ATLAS TDR (1997). *ATLAS Inner Detector Technical Design Report.* CERN/LHCC/97-16 and 97-17. ISBN 92-9083-102-2 and 92-9083-103-0. available at http://atlas.web.cern.ch/Atlas/internal/tdr.html

Attwood, W.B. (1994). Gamma Large Area Silicon Telescope (GLAST) – applying silicon strip technology to the detection of gamma rays in space. *Nucl. Instr. and Meth.* **A342** (1994) 302–307

Barbero, M. *et al.* (2004). Design and test of the CMS pixel readout chip. *Nucl. Instr. and Meth.* **A517** (2004) 349–359

Barkan, O. *et al.* (1991). Development of a customized SSC pixel detector readout for vertex tracking. *Proceedings of the Symposium on Detector Research and Development for the Superconducting Super Collider, October 15–18, 1990.* eds T. Dombeck, V. Kelley, G. Yost, World Scientific, Singapore

Batignani, G. *et al.* (1993). Operational experience with a large detector system using silicon strip detectors with double sided readout. *Nucl. Instrum. and Meth.* **A326** (1993) 183–188

Battaligia, M. (2004). The Vertex Tracker at future e^+e^- linear colliders. *Nucl. Instr. and Meth.* **A530** (2004) 33–37

Bean, A. (2001). Status of the DØ Silicon Microstrip Tracker. *Nucl. Instr. and Meth.* **A466** (2001) 262–267

Bebek, C.J. *et al.* (2002). Proton radiation damage in high-resistivity n-type silicon CCDs. *SPIE* **4669** (2002) 161–171 and LBNL-49316

Becks, K.H. *et al.* (1997). The DELPHI pixels. *Nucl. Instr. and Meth.* **A386** (1997) 11–17

Bellazini, R. *et al.* (2003). The silicon-strip tracker of the Gamma ray Large Area Space Telescope. *Nucl. Instr. and Meth.* **A512** (2003) 136–142

Bergauer, T. (2004). Tests of the silicon strip sensors for the CMS tracker. *Nucl. Instr. and Meth.* **A518** (2004) 317–320

Biasini, M. (2004). The construction of the Silicon Strip Tracker for the CMS experiment. *Nucl. Instr. and Meth.* **A530** (2004) 17–22

Bingefors, N. *et al.* (1993). A novel technique for fast pulse shaping using a slow amplifier at LHC. *Nucl. Instr. and Meth.* **A326** (1993) 112–119

Blanquart, L. *et al.* (2004). FE-I2: A front-end readout chip designed in a commercial 0.25 μm process for the ATLAS pixel detector at LHC. *IEEE Trans. Nucl. Sci.* **NS-51/4** (2004) 1358–1364

Braibant, S. *et al.* (2002). Investigation of design parameters for radiation hard silicon microstrip detectors. *Nucl. Instr. and Meth.* **A485** (2002) 343–361

Brau, J.E. and Sinev, N. (2000). Operation of a CCD particle detector in the presence of bulk neutron damage. *IEEE Trans. Nucl. Sci* **47/6** (2000) 1898–1901

Brau, J.E. *et al.* (2004). Investigation of radiation damage in the SLD CCD vertex detector. *IEEE Trans. Nucl. Sci* **51/4** (2004) 1742–1746

Campabadal, F. *et al.* (2005a). Design and performance of the ABCD3TA ASIC for readout of silicon strip detectors in the ATLAS Semiconductor Tracker. *Nucl. Instr. and Meth.*, to be published

Campabadal, F. *et al.* (2005b). Beam tests of ATLAS SCT silicon strip detector modules. *Nucl. Instr. and Meth.* **A538** (2005) 384–407

Campbell, M. *et al.* (1998). A readout chip for a 64×64 pixel matrix with 15-bit single photon counting. *IEEE Trans. Nucl. Sci.* **NS-45/3** (1998) 751–753

Carithers, W.C. *et al.* (1990). The CDF SVX: a silicon vertex detector for a hadron collider. *Nucl. Instr. and Methods* **A289** (1990) 388–399

Chabaud, V. *et al.* (1996). The DELPHI silicon strip microvertex detector with double sided readout. *Nucl. Instr. and Meth.* **A368** (1996) 314–332

Choong, W.-S. *et al.* (2002). A compact 16-module camera using 64-pixel CsI(Tl)/Si p-i-n photodiode imaging modules. *IEEE Trans. Nucl. Sci.* **NS-49/5** (2002) 2228–2235

Dabrowski, W. *et al.* (2000). Design and performance of the ABCD chip for the binary readout of silicon strip detectors in the ATLAS Semiconductor Tracker. *IEEE Trans. Nucl. Sci.* **47** (2000) 1843–1850

Damerell, C.J.S. (2001). A CCD-based vertex detector for TESLA. LC-DET-2001-023, DESY, Feb. 2001

Deptuch, G. *et al.* (2003). Development of monolithic active pixel sensors for charged particle tracking. *Nucl. Instr. and Meth.* **A511** (2003) 240–249

Donaldson, R. and Marx, J.N. (eds) (1986). *Physics of the Superconducting Supercollider*. Proceedings, 1986 Summer Study, Snowmass, June 23 – July11, 1986. APS, New York 1986

Einsweiler, K. (2004). private communication

Eisner, A. *et al.* (1997). *The Data Transmission System for the SVT*. SCIPP 97/47, Univ. of California, Santa Cruz Institute for Particle Physics, September 1997. http://scipp.ucsc.edu/ nora/preprint/1997/scipp-97-47.pdf

Erdmann, W. (2004). Development of the CMS pixel detector. *Nucl. Instr. and Meth.* **A518** (2004) 324–327

Feld, L. (2003). Detector modules for the ATLAS SCT endcaps. *Nucl. Instr. and Meth.* **A511** (2003) 183–186

Ferrari, P. (2004). The ATLAS Semiconductor Tracker system test results. *Nucl. Instr. and Meth.* **A530** (2004) 38–43

Fischer, P. (2003a). Design considerations for pixel readout chips. *Nucl. Instr. and Meth.* **A501** (2003) 175–182

Fischer, P. *et al.* (2003b). Readout concepts for DEPFET pixel arrays. *Nucl. Instr. and Meth.* **A512** (2003) 318–325

French, M.J. *et al.* (2001). Design and results from the APV25, a deep submicron CMOS front-end chip for the CMS tracker. *Nucl. Instr. and Meth.* **A466** (2001) 359–365

Garcia-Sciveres, M. *et al.* (1999). The SVX3D integrated circuit for deadtimeless silicon strip readout. *Nucl. Instr. and Meth.* **A435** (1999) 58–64

Gemme, C. (2003). The ATLAS pixel detector. *Nucl. Instr. and Meth.* **A501** (2003) 87–92

Gorelov, I. *et al.* (2002). Electrical characteristics of silicon pixel detectors. *Nucl. Instr. and Meth.* **A489** (2002) 202–217

Groom, D.E. *et al.* (1999). Point-spread function in depleted and partially depleted CCDs. *Proc. 4th ESO Workshop on Optical Detectors for Astron-*

omy, Garching, Sept. 13–16, 1999. pp. 205–206. Available for download at http://www-ccd.LBL.gov

Groom, D.E. (2000). Recent progress on CCDs for astronomical imaging. *SPIE* **4008** (2000) 634–645 and LBNL-45277

Heijne, E.H.M. *et al.* (1995). Construction and characterization of a 117 cm^2 silicon pixel detector. *IEEE Trans. Nucl. Sci.* **NS-42/4** (1995) 413–418

Heijne, E.H.M. (2001). Semiconductor micropattern pixel detectors: a review of the beginnings. *Nucl. Instr. and Meth.* **A465** (2001) 1–26

Holland, S. and Spieler, H. (1990). A monolithically integrated detector–preamplifier on high-resistivity silicon. *IEEE Trans. Nucl. Sci.* **NS-37/2** (1990) 463–468

Holland, S. (1992). Properties of CMOS Devices and circuits fabricated on high-resistivity, detector-grade silicon. *IEEE Trans. Nucl. Sci.* **NS-39/4** (1992) 809–813

Holland, S.E. *et al.* (1996). A 200×200 CCD image sensor fabricated on high-resistivity silicon. *IEDM Tech. Digest* (1996) 911–914

Holland, S.E., Wang, N.W., and Moses, W.W. (1997). Development of low-noise, back-side illuminated silicon photodiode arrays. *IEEE Trans. Nucl. Sci.* **NS-44/3**, 443–447

Holland, S.E. *et al.* (2003). Fully depleted, back-illuminated charge-coupled devices fabricated on high-resistivity silicon. *IEEE Trans. Electron. Dev.* **ED-50/1** (2003) 225–238 and LBNL-49992

Holland, S.E. (2005). private communication

Janesick, J.R. (2000). *Scientific Charge-Coupled Devices.* SPIE Press Monograph Vol. PM83, SPIE Press, Bellingham, 2000. ISBN 0-8194-3698-4, TK7871.99.C45.J36

John, J. *et al.* (2004). High-density hybrid interconnect methodologies. *Nucl. Instr. and Meth.* **A531** (2004) 202–208

Kajfasz, E. (2003). The DØ silicon microstrip tracker for Run IIa. *Nucl. Instr. and Meth.* **A511** (2003) 16–19

Karcher, A. *et al.* (2004). Measurement of lateral charge diffusion in thick, fully depleted, back-illuminated CCDs. *IEEE Trans. Nucl. Sci.* **NS51/5** (2004) 2231–2237

Kass, R. *et al.* (2003). The CLEO III silicon vertex detector. *Nucl. Instr. and Meth.* **A501** (2003) 32–38

Kemmer, J. and G. Lutz (1987). New detector concepts. *Nucl. Instr. and Meth.* **A253** (1987) 365–377

Kipnis, I., Spieler, H. and Collins, T. (1994). An analog front-end bipolar-transistor integrated circuit for the SDC silicon tracker. *IEEE Trans. Nucl. Sci.* **NS-41/4** (1994) 1095–1103

Kipnis, I. *et al.* (1997). A time-over-threshold machine: the readout integrated circuit for the BaBar Silicon Vertex Tracker. *IEEE Trans. Nucl. Sci.* **NS-44/3** (1997) 289–297

Kleinfelder, S.A. *et al.* (1988). A flexible 128 channel silicon strip detector instrumentation integrated circuit with sparse data readout. *IEEE Trans. Nucl. Sci.* **NS-35** (1988) 171–175

Kleinfelder, S.A. *et al.* (2004). Novel integrated CMOS sensor circuits. *IEEE Trans. Nucl. Sci.* **NS-51/5** (2004) 2328–2336

Kohriki, T. *et al.* (1996). First observation of thermal runaway in the radiation damaged silicon detector. *IEEE Trans. Nucl. Sci.* **NS43/3** (1996) 1200–1202

Kohriki, T. *et al.* (2002). Development of the hybrid structure for the barrel module of the ATLAS silicon-microstrip tracker. *IEEE Trans. Nucl. Sci.* **NS-49/6** (2002) 1378–3283

Kondo, T. *et al.* (2002). Construction and performance of the ATLAS silicon microstrip barrel modules. *Nucl. Instr. and Meth.* **A485** (2002) 27–42

Kramer, G. *et al.* (1991). Development of pixel detectors for SSC vertex tracking. *Supercollider 3 (J. Nonte, ed). Proc. 3rd Annual International Industrial Symposium on the Super Collider, Atlanta, Georgia, Mar 13–15, 1991.* pp. 953–964. Plenum Press, New York, 1991. ISBN 0-3064-4037-7, QC787.P7.I57 1991

Krammer, M. (2003). Experience with silicon sensor performance and quality control for a large-area detector. *Nucl. Instr. and Meth.* **A511** (2003) 136–144

Krammer, M. (2004). The silicon sensors for the inner tracker of the Compact Muon Solenoid experiment. *Nucl. Instr. and Meth.* **A531** (2004) 238–245

Krieger, B. *et al.* (2004). SVX4: A new deep-submicron readout IC for the Tevatron collider at Fermilab. *IEEE Trans. Nucl. Sci.* **NS-51/5** (2004) 1968–1973

Kuijer, P. (2004). The inner tracking system of the Alice experiment. *Nucl. Instr. and Meth.* **A530** (2004) 28–32

Kwan, S. *et al.* (2003). The BTeV silicon pixel and microstrip detectors. *Nucl. Instr. and Meth.* **A511** (2003) 48–51

Lall, P., Pecht, M.G. and Hakim, E.B. (1997). *Influence of Temperature on Microelectronics and System Reliability.* CRC Press, Boca Raton. ISBN 0-8493-9450-3, TK7870.25.L35

Latronico, L. (2004). Quality control on the silicon sensors of the GLAST tracker. *Nucl. Instr. and Meth.* **A530** (2004) 163–167

Llopart, X. *et al.* (2002). Medipix2: a 64-k pixel readout chip with 55-μm square elements working in single photon counting mode. *IEEE Trans. Nucl. Sci.* **NS-49/5** (2002) 2279–2283

Llopart, X. and Campbell, M. (2003). First test measurements of a 64k pixel readout chip working in single photon counting mode. *Nucl. Instr. and Meth.* **A509** (2003) 157–163

Ludewigt, B. *et al.* (1994). A high rate, low noise, x-Ray detector system. *IEEE Trans. Nucl. Sci.* **NS-41/4** (1994) 1037–1041

Lundqvist, M. *et al.* (2000). Computer simulations and performance measurements on a silicon strip detector for edge-on imaging. *IEEE Trans. Nucl. Sci.* **NS-47/4** (2000) 1487–1492

Lynch, G. (1993). Figure 2.4 in *Status Report on the Design of a Detector for the Study of CP Violation at PEP-II at SLAC*. SLAC-419, June, 1993

Macchiolo, A. (2004). Control of the fabrication process for the sensors of the CMS silicon tracker. *Nucl. Instr. and Meth.* **A530** (2004) 54–58

Mandelli, E. *et al.* (2002). Digital column readout architecture for the ATLAS pixel 0.25 μm front end IC. *IEEE Trans. Nucl. Sci.* **NS-49/4** (2002) 1774–1777

Manfredi, P.F. *et al.* (1999). Functional characteristics and radiation tolerance of AToM, the front-end chip of BaBar Silicon Vertex Tracker. *IEEE Trans. Nucl. Sci.* **NS-46/6** (1999) 1865–1870

Merkel, P. (2003). The CDF silicon detector upgrade and performance. *Nucl. Instr. and Meth.* **A501** (2003) 1–6

Messenger, G.C. and Ash, M.S. (1986). *The Effects of Radiation on Electronic Systems*. Van Nostrand Reinhold, New York. ISBN 0-442-25417-2, TK7870.M4425

Moorhead, G.F. (2002). Detector modules for the endcaps of the ATLAS semiconductor tracker. *Nucl. Instr. and Meth.* **A485** (2002) 43–53

Morselli, A. (2004). The GLAST tracker. *Nucl. Instr. and Meth.* **A530** (2004) 158–162

Nelson, D. (2004). Private communication

Nisius, R. (2004). End-cap modules for the ATLAS SCT. *Nucl. Instr. and Meth.* **A530** (2004) 44–49

Nyman, M. (1996). Private communication

Ohsugi, T. *et al.* (1994). Microdischarges of AC-coupled silicon strip sensors. *Nucl. Instr. and Meth.* **A342** (1994) 22–26

Ohsugi, T. *et al.* (1999). Design optimization of radiation-hard, double-sided, double-metal, AC-coupled silicon sensors. *Nucl. Instr. and Meth.* **A436** (1999) 272–280

Österberg, K. (1999). Performance of the vertex detectors at LEP2. *Nucl. Instr. and Meth.* **A435** (1999) 1–8

Owens, A. and A. Peacock (2004). Compound semiconductor radiation detectors *Nucl. Instr. and Meth.* **A531** (2004) 18–37

Raymond, M. *al.* (2000). The CMS tracker APV25 0.25 μm CMOS readout chip. *6th Workshop on Electronics for LHC Experiments*, Krakow, September 2000

RD48 (1999). *3rd RD48 Status Report* (1999). CERN/LHCC 2000-0089. http://rd48.web.cern.ch/rd48/

Re, V. *et al.* (2003). Performance of the BaBar Silicon Vertex Tracker. *Nucl. Instrum. and Meth.* **A501** (2003) 14–21

Richter, R.H. *et al.* (1996). Strip detector design for ATLAS and HERA-B using two-dimensional device simulation. *Nucl. Instr. and Meth.* **A377** (1996) 412–421

Richter, R.H. *et al.* (2003). Design and technology of DEPFET pixel sensors for linear collider applications. *Nucl. Instr. and Meth.* **A511** (2003) 250–256

Rossi, L. (2003). Pixel detectors hybridisation. *Nucl. Instr. and Meth.* **A501** (2003) 239–244

Schnetzer, S. (2003). The CMS pixel detector. *Nucl. Instr. and Meth.* **A501** (2003) 100–105

Schwarz, Andreas S. (1994a). Silicon strip vertex detectors at LEP. *Nucl. Instr. and Meth.* **A342** (1994) 218–232

Schwarz, Andreas S. (1994b). Heavy flavor physics at colliders with silicon strip vertex detectors. *Phys. Rept.* **238** (1994) 1–133

Seiden, A. (1994). The SDC Silicon Tracker. *IEEE Trans. Nucl. Sci.* **41/4** (1994) 779–784

Spieler, H. (1982). Unpublished note. The CCD output ($r_o \approx 3\,\mathrm{k}\Omega$) drove an external common-gate JFET amplifier (gain ≈ 4) with an input impedance of $100\,\Omega$. A subsequent limiting amplifier (cascade of bipolar transistor differential pairs) constrained the amplitude of the clock pulse feedthrough and allowed retrieval of MIP level signals at a clock rate of 20 MHz.

Spieler, H. (1994). Analog front-end electronics for the SDC Silicon Tracker, *Nucl. Instr. and Meth.* **A342** (1994) 205–213

Spieler, H. (1988). Integrated microsystems as a driving force in modern detector designs. in *International Conference on the Impact of Digital Microelectronics and Microprocessors on Particle Physics*. pp. 228–238. M. Budinich, E. Castelli, and A. Colavita (eds). World Scientific, Singapore. ISBN 9971-50-742-0

Stefanini, G. (2004). Progress in the ALICE silicon pixel detector. *Nucl. Instr. and Meth.* **A530** (2004) 77–81

Stover, R.J. *et al.* (2004). Packaging design for Lawrence Berkeley National Laboratory high resistivity CCDs. *SPIE* **5499** 58

Taylor, G. (2003). The Belle silicon vertex detector: present performance and upgrade plans. *Nucl. Instr. and Meth.* **A501** (2003) 22–31

Turala, M. *et al.* (2001). The ATLAS semiconductor tracker. *Nucl. Instr. and Meth.* **A466** (2001) 243–254

Turchetta, R. *et al.* (2003). Monolithic active pixel sensors (MAPS) in a VLSI CMOS technology. *Nucl. Instr. and Meth.* **A501** (2003) 251–259

Unno, Y. *et al.* (1996). Beam tests of a double-sided silicon strip detector with fast binary readout electronics before and after proton-irradiation. *Nucl. Instr. and Meth.* **A383** (1996) 211–222

Unno, Y. *et al.* (2003). ATLAS silicon microstrip detector system (SCT). *Nucl. Instr. and Meth.* **A511** (2003) 58–63

Ushiroda, Y. (2003). Belle silicon vertex detectors. *Nucl. Instr. and Meth.* **A511** (2003) 6–10

Weilhammer, P. (1994). Double-sided Si strip sensors for LEP vertex detectors. *Nucl. Instr. and Meth.* **A342** (1994) 1–15

Wigmans, R. (2000). *Calorimetry*. Oxford University Press, Oxford. ISBN 0-19-850296-6

Wunstorf, R. (2001). Radiation tolerant sensors for the ATLAS pixel detector. *Nucl. Instr. and Meth.* **A466** (2001) 327–334

Yarema, R. *et al.* (1994). *A Beginner's Guide to the SVXII.* Fermilab TM-1892, June, 1994. http://library.fnal.gov/archive/test-tm/1000/fermilab-tm-1892.pdf

Zimmerman, T. *et al.* (1995). The SVX2 readout chip. *IEEE Trans. Nucl. Sci.* **NS-42/4** (1995) 803–807.

9

WHY THINGS DON'T WORK

After assembling a detector system a common experience is that it doesn't work as expected. Apart from trivial problems such as disconnected cables, common maladies are spurious signals, self-oscillation, or excessive noise. This chapter discusses some common problems in detector systems and how to avoid them.

Throughout the previous lectures it was assumed that the only sources of noise were random, known, and in the detector, preamplifier, or associated components. In practice, the detector system will pick up spurious signals that are not random, but not correlated with the signal, so with reference to the signal they are quasi-random. These lead to baseline fluctuations superimposed on the desired signal, which increase the detection threshold and degrade resolution.

It is important to distinguish between pickup of spurious signals, either from local or remote sources (clock generators, digital circuitry, readout lines), and self-oscillation, where the system provides a feedback path that causes sustained oscillation due to a portion of the output reaching the input. The same mechanisms that make a system sensitive to external pickup can also form a parasitic feedback path.

This is not an exhaustive treatment; pickup mechanisms comprise a very complex system and just one weak area is sufficient to cause problems, so improving one aspect will not always improve the situation, even if this improvement is essential. Texts with more details are listed in the bibliography at the end of this chapter.

9.1 Reflections on transmission lines

Signals are transmitted from one unit to another through transmission lines, often coaxial cables or ribbon cables. When transmission lines are not terminated with their characteristic impedance, the signals are reflected. As a signal propagates along the cable, the ratio of instantaneous voltage to current equals the cable's characteristic impedance $Z_0 = \sqrt{L/C}$, where L and C are the inductance and capacitance per unit length. Typical impedances are 50 or $75\,\Omega$ for coaxial cables and $\sim 100\,\Omega$ for ribbon cables. If at the receiving end the cable is connected to a resistance different from the cable impedance, a different ratio of voltage to current must be established. This occurs through a reflected signal. If the termination is less than the line impedance, the voltage must be smaller and the reflected voltage wave has the opposite sign. If the termination is greater than the line impedance, the voltage wave is reflected with the same polarity. Conversely, the current in the reflected wave is of like sign when the termination is less than the line impedance and of opposite sign when the termination

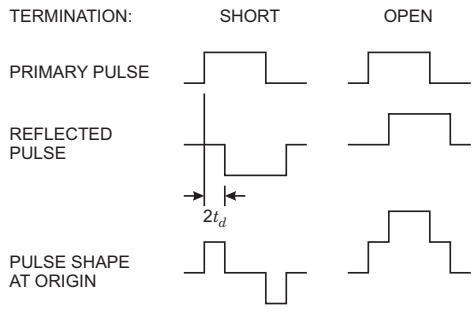

FIG. 9.1. Voltage pulse reflections on a transmission line terminated either with a short (left) or open circuit (right). Measured at the sending end, the reflection from a short at the receiving end appears as a pulse of opposite sign delayed by the round trip delay of the cable. If the total delay is less than the pulse width, the signal appears as a bipolar pulse. Conversely, an open circuit at the receiving end causes a reflection of like polarity.

is greater. Voltage reflections are illustrated in Figure 9.1. At the sending end the reflected pulse appears after twice the propagation delay of the cable. Since in the presence of a dielectric the velocity of propagation $v = c/\sqrt{\varepsilon}$, in typical coaxial and ribbon cables the delay is 5 ns/m.

Cable drivers often have a low output impedance, so the reflected pulse is reflected again towards the receiver, to be reflected again, *etc.* This is shown in Figure 9.2, which shows the observed signal when the output of a low-impedance pulse driver is connected to a high-impedance amplifier input through a 4 m long 50 Ω coaxial cable. If feeding a counter, a single pulse will be registered multiple

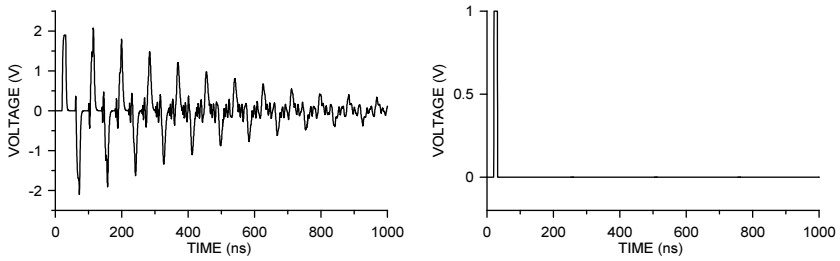

FIG. 9.2. Left: Signal observed in an amplifier when a low-impedance driver is connected to an amplifier through a 4 m long coaxial cable. The cable impedance is 50 Ω and the amplifier input appears as 1 kΩ in parallel with 30 pF (a typical input impedance for oscilloscopes or nuclear instrumentation modules). When the receiving end is properly terminated with 50 Ω, the reflections disappear (right).

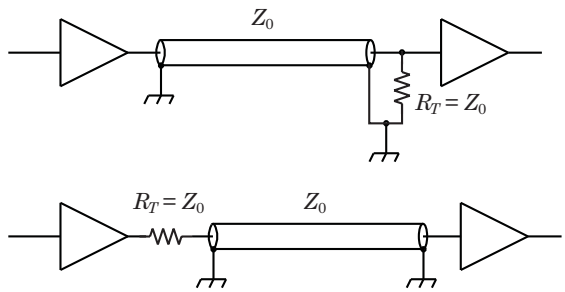

FIG. 9.3. Cables may be terminated at the receiving end (top, shunt termination) or sending end (bottom, series termination).

times, depending on the threshold level. When the amplifier input is terminated with 50 Ω, the reflections disappear and only the original 10 ns wide pulse is seen.

There are two methods of terminating cables, which can be applied either individually or – in applications where pulse fidelity is critical – in combination. As illustrated in Figure 9.3 the termination can be applied at the receiving or the sending end. Receiving end termination absorbs the signal pulse when it arrives at the receiver. With sending-end termination the pulse is reflected at the receiver, but since the reflected pulse is absorbed at the sender, no additional pulses are visible at the receiver. At the sending end the original pulse is attenuated two-fold by the voltage divider formed by the series resistor and the cable impedance. However, at the receiver the pulse is reflected with the same polarity, so the superposition of the original and the reflected pulses provides the original amplitude.

This example uses voltage amplifiers, which have low output and high input impedances. It is also possible to use current amplifiers, although this is less common. Then, the amplifier has a high output impedance and low input impedance, so shunt termination is applied at the sending end and series termination at the receiving end.

Terminations are never perfect, especially at high frequencies, where stray capacitance becomes significant. For example, the reactance of 10 pF at 100 MHz is 160 Ω, which would severely alter a 50 or 100 Ω termination. Thus, critical applications often use both series and parallel termination, although this does incur a 50% reduction in pulse amplitude. In the μs regime, amplifier inputs are usually high impedance, whereas timing amplifiers tend to be internally terminated, but one should always check if this is the case. As a rule of thumb, whenever the propagation delay of cables (or connections in general) exceeds a few percent of the signal risetime, terminations are required.

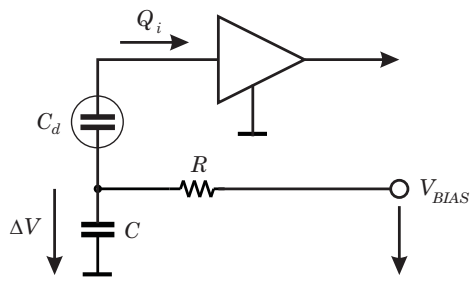

FIG. 9.4. Noise on the detector bias line is coupled through the detector capacitance to the amplifier input.

9.2 Common pickup mechanisms

9.2.1 *Noisy detector bias supplies*

The sensor is the most sensitive node in the system and a common mistake is the use of a noisy power supply to bias the sensor. Figure 9.4 shows how voltage disturbances at the bias connection inject charge into the input. Any disturbance ΔV on the detector bias line will induce charge in the input circuit

$$Q_i = C_d \Delta V \ . \tag{9.1}$$

$\Delta V = 10\,\mu$V and 10 pF detector capacitance yield $Q_i = 0.1$ fC, corresponding to about $620\,e$ or 2.2 keV (Si).

Especially when the detector bias is low (< 100 V), it is tempting to use a general laboratory power supply. However, power supplies are often very noisy. The RC circuits in the bias line provide some filtering, but usually not enough for a typical power supply. Modern power supplies often use switching regulators to provide high efficiency. Well-designed switching regulators can be very clean, but most switchers are very noisy. Power supply vendors often specify rms output noise. This is usually a useful specification for "linear" supplies (using analog regulators), but with switching supplies spikes on the output can be quite large, but short, so that the rms noise specification may appear adequate. Linear supplies often include digital output metering, which can also inject digital noise into the output. Unless specifically designed and specified for low-noise detector applications, the power supply output should be inspected with an oscilloscope or spectrum analyzer to verify low noise (use AC coupling to allow adequate input sensitivity). The relevant frequency range is determined by the overall pulse shaping.

9.2.2 *Light pickup*

Systems susceptible to light pickup include photomultiplier tubes and semiconductor detectors (all semiconductor detectors are photodiodes!). Typical sources are room lighting (light leaks) and vacuum gauges. Interference from room lighting is correlated with the power line frequency (60 Hz in the U.S., 50 Hz in Europe

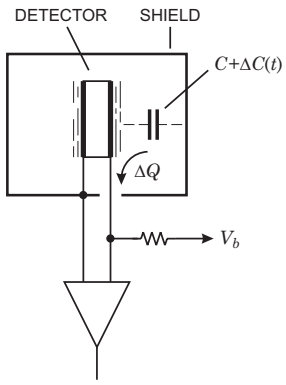

FIG. 9.5. An electrode at a potential different from its surroundings will experience charge flow when the capacitance changes due to vibration ("microphonics").

and Japan). Since incandescent lamps respond to power, the light pulsates at twice the line frequency, as the light intensity is proportional to voltage squared.

To diagnose the problem, inspect the signal output with an oscilloscope set to the trigger mode "line" (which triggers off the AC line) and look for stationary structure on the baseline. Discerning the interference in the random noise is sometimes difficult. The averaging and pattern recognition capability of the human eye is very useful, so analog oscilloscopes tend to be better than digital. If room lighting is suspected, a simple test is to switch off the light. Another useful technique is to cover the system with a black cloth (preferably felt, or very densely woven fabric – check if you can see through it).

9.2.3 *Microphonics*

Another common source of interference is microphonics. The mechanism is illustrated in Figure 9.5. If the electrode at potential V_b vibrates with respect to the shield enclosure, the stray capacitance C is modulated by $\Delta C(t)$, inducing a charge ΔQ in the detector signal circuit.

Typically, vibrations are excited by motors (vacuum pumps, blowers), so the interference tends to be correlated with the line frequency. Again. one can check with an oscilloscope on line trigger. Vibrations can be detected by hand, or one can exacerbate vibrations by banging against sensitive parts of the system (but don't overdo it). Switching off pumps or other vibration sources can also localize the source.

This type of pickup only occurs between conductors at different potentials, so it can be reduced by shielding the relevant electrode. Introducing an additional shield at electrode potential that doesn't vibrate with respect to the detector is an effective cure. In coaxial detectors, one can operate the outer electrode at 0 V, although there are sometimes other reasons against this.

9.2.4 *RF pickup*

RF pickup is the most popular suspect. Whenever interference is observed, talk of radio transmitters and wires acting as antennas is soon to follow. There is some justification for this, as all detector electronics are sensitive to RF signals. However, the frequency of the interference is important. Antennas are a useful concept in the far field, *i.e.* typically several wavelengths from the source, where an electromagnetic wave has formed. At 1 MHz the wavelength in free space is 300 m, so at power line frequencies and even in the MHz range the dimensions of the typical laboratory place the system in the near field and for typical connection lengths the wires act primarily as capacitive probes (or loops as inductive probes).

The critical frequency range depends on the shaping time. The gain of the system peaks at

$$f \approx \frac{1}{2\pi T_P} \ , \tag{9.2}$$

where T_P is the peaking time. Since the half-power bandwidth is several octaves the system will be sensitive over a wide range of frequencies around the peaking frequency.

Typical sources are radio and TV stations. The source can be identified by frequency; AM broadcast stations are in the range 0.5 – 1.7 MHz, FM broadcast stations are at about 100 MHz, and TV stations transmit in the range 50 – 800 MHz. Local sources of RF interference in the laboratory are induction furnaces, which typically operate at legally prescribed "industrial frequencies" of 13.6, 27, or 40.7 MHz. Detector systems are frequent companions to accelerators, but in most facilities these are very well shielded and usually not a problem (although exceptions exist).

RF sources generate sine waves. Since proximity is a major factor, systems close to the detector tend to be most troublesome. Computer and microprocessor clocks operate in the range of tens to hundreds of MHz, as do clocks internal to the readout system. Video displays commonly introduce interference in the range 10 – 100 kHz. Unlike broadcast stations or induction furnaces, which generate sine waves, digital circuitry emanates pulses, which distribute power over a wide spectrum (the repetition rate determines the intensity, not the extent of the spectrum). Frequently pulses excite resonances, so one sees damped high-frequency oscillations repeating at the pulse frequency. As a result, low frequency disturbances lead to high frequency interference.

Pulsed UHF or microwave emissions, from radar stations for example, can affect low-frequency circuitry by driving it beyond linearity, as the bandwidth of the preamplifier can be much greater than of the subsequent shaper.

Again, one of the most powerful diagnostic techniques is to inspect the analog signal on an oscilloscope. Check with different trigger levels and deflection times and look for periodic structure on the baseline. The frequency of the interference is a key diagnostic. Pickup levels as low as 10% of the noise level can be serious, so careful adjustment of the trigger and judicious squinting of the eye is often

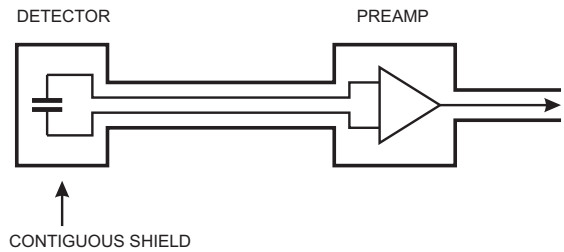

FIG. 9.6. A sensitive system can be protected from external pickup by surrounding it by a contiguous shield.

necessary to see periodic structure superimposed on the random noise. Again, an "old fashioned" analog oscilloscope is best.

An alternative is to inspect the output with a spectrum analyzer. This is a very sensitive technique. Indeed, for some it may be too sensitive, as it tends to show signals that are so small that they are irrelevant. To avoid "wild goose chases" it is important to ascertain quantitatively what levels of interfering signals and frequencies are important. Typically, the overall input noise voltage is of order μV, so interfering signals should be well below this level. At frequencies well above the peaking frequency higher levels of interference are tolerable, as they will be attenuated by the roll-off in gain.

9.3 Pickup reduction techniques

9.3.1 *Shielding*

The classic remedy against RF pickup is shielding. One of the key requirements for a shield is that it be contiguous, as illustrated in Figure 9.6. The shield encloses not just the sensor and preamplifier, but extends to enclose the output line.

A conducting shield attenuates an incident electromagnetic wave through both reflection and absorption. An incident wave is reflected by virtue of the discontinuity relative to free space. The amplitude of the reflected wave

$$E_{0r} = E_0 \left(1 - \frac{Z_{shield}}{Z_0}\right) \tag{9.3}$$

where $Z_0 = \sqrt{\mu/\varepsilon} = 377\,\Omega$ is the impedance of free space. Since the impedance of just about any metallic shield is much smaller, the magnitude of the wave absorbed by the conductor is highly attenuated.

In the metal the absorbed wave gives rise to a local current, whose field counteracts the primary excitation. The net current decreases exponentially as the wave penetrates deeper into the medium

$$i(x) = i_0 e^{-x/\delta} \ , \tag{9.4}$$

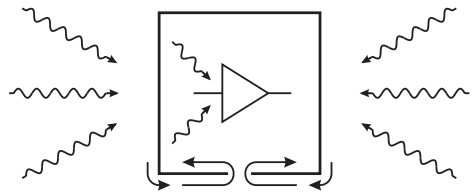

FIG. 9.7. Currents induced on the outside surface of a shield by can flow to the inside through holes or slots.

where i_0 is the current at the surface of the conductor and

$$\delta = \frac{1}{2 \cdot 10^{-4}} \left[\sqrt{\frac{\mathrm{cm}}{\mathrm{s}}} \right] \sqrt{\frac{\rho}{\mu_r f}} \tag{9.5}$$

is the penetration depth or "skin depth". μ_r and ρ are the permeability and resistivity of the conductor and f is the frequency of the incident wave. In aluminum, $\rho = 2.8\,\mu\Omega\,\mathrm{cm}$ and $\mu_r = 1$, so at $f = 1\,\mathrm{MHz}$ the skin depth $\delta = 84\,\mu\mathrm{m} \approx 100\,\mu\mathrm{m}$.

The skin depth decreases with the square root of increasing frequency and decreasing resistivity (increasing conductivity). If the shield is sufficiently thick, the skin effect isolates the inner surface of a shielding enclosure from the outer surface. However, this isolation only obtains if no openings in the shield allow the primary current to flow from the outside to the inside. The shape of the opening is less important than its maximum dimension. For example, a slot acts like a dipole antenna and has zero attenuation at the frequency where the slot's length is a half wavelength. For slots whose length l is less then a half wavelength the attenuation of penetrating currents is

$$A \approx \left(\frac{2l}{\lambda} \right)^2 . \tag{9.6}$$

When N small openings are closely spaced, as in a perforated shield, the leakage increases by $\sqrt{N}$. Since leakage depends primarily on the maximum linear dimension, rather than the area, a linear array of small holes has less leakage than a slot of the same length and width. Pickup is larger than would be expected for a plane wave impinging on the aperture, as the whole outside surface is the capture area. The induced current is then transferred through the opening, as illustrated in 9.7, albeit with attenuation.

The leakage through an opening can be reduced substantially by configuring it as a waveguide below cutoff. A circular tube of diameter d acts as a waveguide at frequencies above a cutoff frequency

$$f_c = \frac{2.7 \times 10^9\,[\mathrm{Hz} \cdot \mathrm{cm}]}{d} . \tag{9.7}$$

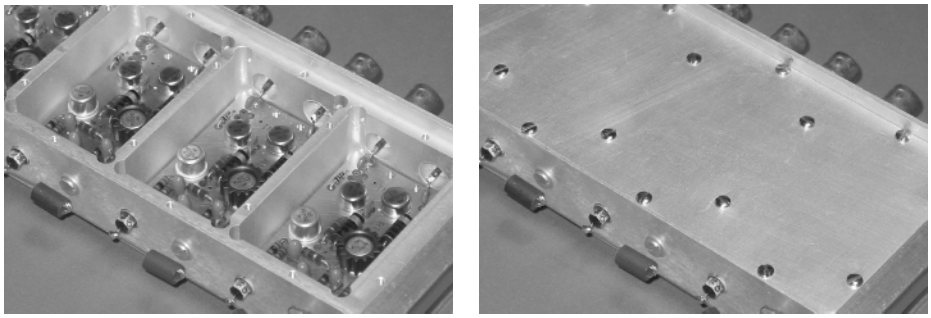

FIG. 9.8. "RF tight" shielding applied to a multichannel subnanosecond amplifier. Amplifiers are in individual compartments and power enters through C-L-C π filters in threaded sleeves mounted in the shield wall.

For a rectangular wave guide whose cross section has a largest dimension d the cutoff frequency is 15% lower. At frequencies well below cutoff a waveguide whose length is three times the diameter provides about 100 dB of attenuation.

To maintain the integrity of the shield, covers must fit tightly with good conductivity at the seams (beware of anodized aluminum!). Screw fasteners must be closely spaced, conforming to eqn 9.6. All signal input and output cables must have good shield connections, the shield coverage of coax or other cables must be $> 90\%$, and connectors must maintain the integrity of the shield connection. The latter two points are worthy of special attention. Many shielded cables have only partial shield coverage. Some use spiral rather than woven shields, which are convenient to connect, but provide poor shielding at high frequencies. Low-cost RF cables of ostensibly the same type as more expensive counterparts often have only $50 - 70\%$ shield coverage. Furthermore, some connectors leave gaps in the shield contact or do not maintain a sufficiently tight fit. Figure 9.8 shows an example of "RF-tight" construction. Clearly this is not practical for low-mass systems, so alternative design techniques are required.

9.3.2 *"Field line pinning"*

Full shielding is not always practical. In vertex detectors where material must be minimized, absorptive shielding is prohibitive. Fortunately, full shielding is not always necessary. Nor are all parts of the circuit equally sensitive, so it is possible to apply local measures to reduce local coupling to sensitive nodes.

Consider a conductor carrying an undesired signal current, with a corresponding signal voltage. Capacitive coupling will transfer interference to an adjacent circuit node, as shown in Figure 9.9. The coupling can be reduced by introducing an intermediate conductor with a large capacitance to the interference source and to ground. The intermediate conductor will "capture" the field lines and effectively "shield" the critical node. The coupling can be reduced even more by introduction of a ground plane that localizes the field between the conductor and

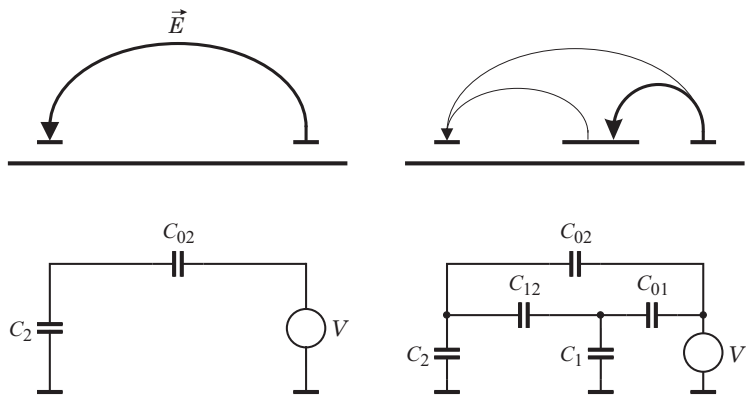

FIG. 9.9. Coupling between adjacent conductors can be reduced by introducing an intermediate conductor that "pins" the field lines. The equivalent circuit shows the capacitive divider network that attenuates the coupled signal.

the ground plane. "Field line pinning" is also the operative mechanism of "Faraday shields", *i.e.* thin electric shields or grids that block capacitive coupling. Guard rings are another implementation, where sensitive nodes are enclosed by a grounded circuit trace. If the guard ring is to be effective at high frequencies, its inductance must be sufficiently small to provide a low impedance relative to the circuit to be protected. The vias providing the ground connection contribute to the inductance, so multiple vias are usually required.

9.3.3 *"Self-shielding" structures*

Another technique to reduce sensitivity to external sources exploits the fact that electric fields concentrate in volumes of high dielectric constant. The magnitude of capacitive coupling depends on the dielectric constant of the intermediate medium. Consider the ensemble of electrodes shown in Figure 9.10. The medium between electrode sets 1 and 2 has a dielectric constant $\varepsilon_r = 1$, whereas the

FIG. 9.10. Capacitive coupling between electrode sets 2 and 3 is ε_r times larger than between sets 1 and 2. If electrode sets 2 and 3 subtend the sensitive volume of a sensor, most of the field lines will be "captured" by the region of high dielectric constant.

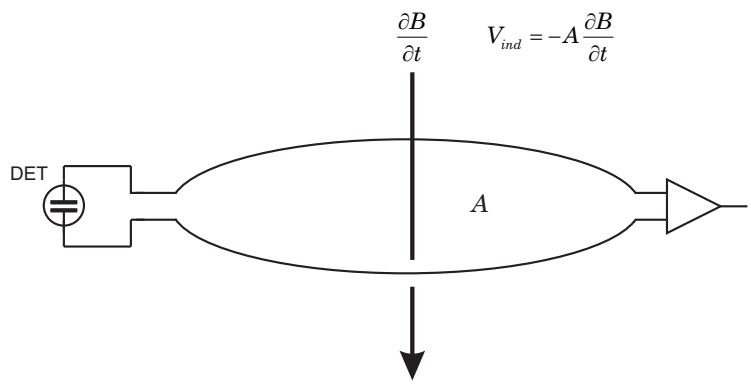

FIG. 9.11. A varying magnetic flux through a conducting loop induces a potential, which gives rise to a current whose magnitude depends on the loop's impedance. Reducing the area of the loop reduces the induced signal.

volume between sets 2 and 3 is filled with $\varepsilon_r > 1$. The capacitance between electrode sets 2 and 3 is ε_r times larger than between sets 1 and 2.

If electrode sets 2 and 3 subtend a silicon sensor with $\varepsilon_r = 11.9$. then 92% of the field lines originating from electrode set 2 terminate on set 3, *i.e.* are confined to the Si bulk, whereas 8% terminate on set 1. With $\varepsilon_r = 1$, for comparison, 50% of the field lines originating from electrode set 2 terminate on set 1. As a result, the high dielectric constant reduces coupling of electrode sets 2 and 3 to external sources.

If the interference source is represented by electrode set 1 and sets 2 and 3 represent a detector, a Si detector is 6.5 times less sensitive to capacitive pickup then a detector with $\varepsilon_r = 1$ (*e.g.* a gas-filled chamber with the same geometry).

9.3.4 *Inductive coupling*

Although interfering signals are most commonly introduced by capacitive coupling, another mechanism that couples interfering currents into a signal loop is induction. Any varying magnetic flux in a conducting loop will induce a potential, which – depending on the impedance of the loop – gives rise to a current. This is illustrated in exaggerated form in Figure 9.11. Clearly, the area A enclosed by any loops should be minimized. This can be accomplished by routing the signal line and return as a closely spaced pair. Better yet is a twisted pair, where the voltages induced in successive twists cancel. Minimizing the area of the input signal loop can be challenging when alternating detector electrodes are read out at opposite ends, which is sometimes done because of mechanical constraints. We'll return to this problem later.

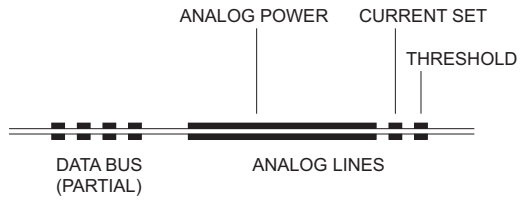

FIG. 9.12. Broadside coupled strip-line cables reduce radiation by confining the field between conductor pairs. The geometry shown also provides a fully balanced power feed.

9.3.5 *"Self-shielding" cables*

Shielding techniques can also be applied at the source, for example cables used for signal transmission. In mixed analog-digital systems, radiation from cables, especially digital signal cables, is a concern. The best approach is to utilize balanced structures, where the fields from supply and return currents cancel. For both the magnetic and electric fields to cancel, both the currents and voltages must be balanced. Grouping the forward and return conductors as closely spaced pairs constrains the extent of the field.

Twisted-pair lines are one approach. A superior geometry utilizes broadside-coupled differential lines with a thin intermediate dielectric. The extent of the fringing field beyond the conductor edge is about equal to the thickness of the dielectric. Figure 9.12 shows an example using a $50\,\mu$m thick dielectric. The data lines are broadside coupled pairs of $150\,\mu$m wide conductors, $50\,\mu$m thick with a $150\,\mu$m gap between conductors. A spacing of three times the dielectric thickness provides > 40 dB isolation per meter. Power connections are made substantially wider ($1 - 5$ mm), forming a low impedance transmission line with high distributed capacitance.

This example is for short runs in the inner region of a tracker, where reduction of material is crucial. Dimensions can be scaled proportionally to achieve lower resistance and signal dispersion in longer cable runs farther from the active region (see Figure 8.9 for an example).

9.3.6 *Shielding summary*

Tight shielding is most important in systems with ns risetimes, where bandwidths extend to hundreds of MHz. In semiconductor arrays the relevant frequencies commonly range from several hundred kHz to tens of MHz. Semiconductor sensors are inefficient antennas, for one because of the concentration of field lines in the sensor, but also because the the sensors are rather small. For example, to act as an efficient antenna, the length of a strip sensor should be about 1/4 wavelength. In the presence of a dielectric the wavelength $\lambda = c/f\sqrt{\varepsilon}$. Even for a relatively fast system with a 10 ns peaking time, the maximum gain is at about 15 MHz, where for $\varepsilon = 11.9$ the wavelength in Si is about 6 m. A 12 cm long

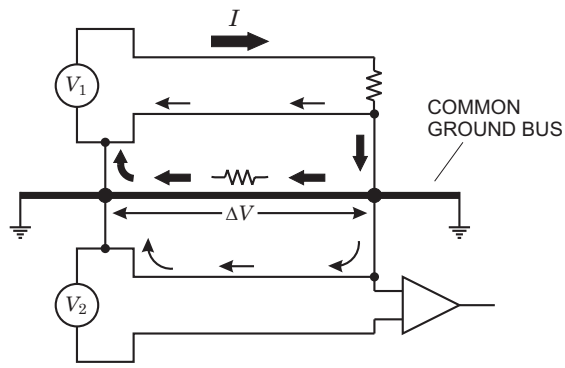

FIG. 9.13. Shared current paths introduce common voltage drops to different circuits.

strip is only 2% of a wavelength and a very inefficient antenna. Wire chambers, for example "straw tube" chambers with several meters length, are much more susceptible in this regard.

Thus, the most important consideration in semiconductor detector systems is not shielding against external sources of electromagnetic radiation, but protection against local fields and currents. Pickup is in the near field where signals are coupled either by capacitance or mutual inductance. One source is the pulsed beam itself, against which the beam tube must provide sufficient shielding. Other sources are local clocks and – most importantly – the current pulses associated with the data readout. The most common problems are associated with currents from downstream stages leaking into the input circuit.

9.4 Shared current paths – grounding and the power of myth

9.4.1 *Shared current paths ("ground loops")*

Although capacitive or inductive coupling cannot be ignored, the most prevalent mechanism of undesired signal transfer is the presence of shared signal paths. The mechanism is illustrated in Figure 9.13.

The upper circuit driven by V_1 has a large circulating current I. Following the prevailing lore, one leg of the circuit is connected to a massive ground bus. Although the circuit associated with generator V_1 has a dedicated current return, the current seeks the past of least resistance, which is the massive ground bus.

The lower circuit is a sensitive signal transmission path. Again, adhering to the common lore, it is connected to ground at both the source and receiver. The large current flowing through the ground bus causes a voltage drop ΔV, which is superimposed on the low-level signal loop associated with V_2 and appears as an additional signal component.

The common ground bus couples the two circuits, which is why this form of cross-coupling is commonly referred to as a "ground loop". The popularity of this term notwithstanding, the cross-coupling has nothing to do with grounding

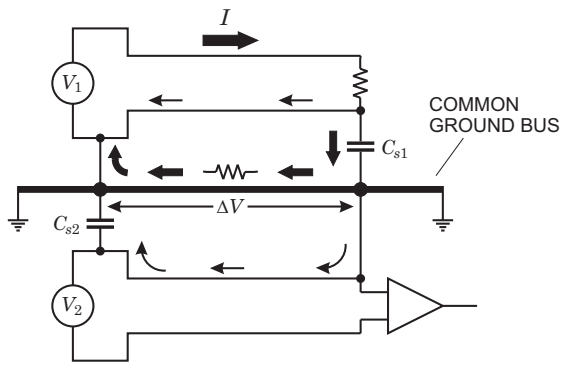

FIG. 9.14. For AC signals current paths can be established by capacitive coupling.

per se, but is due to the common return path. However, the common ground caused the problem by establishing the shared path. The cross-coupling is due to shared current paths, whether grounds are involved or not.

In systems that respond to transients (*i.e.* time-varying signals) rather than DC signals, secondary loops can be closed by capacitance (Figure 9.14). The loops in this figure are the same as shown before, but the loops are closed by the capacitances C_{s1} and C_{s2}. Frequently, these capacitances are not formed explicitly by capacitors, but are the stray capacitance formed by a power supply to ground or by a detector to its support structure (as represented by C_{s2}), *etc.* A DC path is not necessary. This is important to keep in mind, as on circuit boards and hybrids analog and digital grounds are often assigned to separate layers, which can have substantial capacitance with respect to one another.

For high-frequency current components the inductance of the common current path can increase the impedance substantially beyond the DC resistance. The combined effects of skin depth and inductance both increase the impedance with frequency, so cross-coupling is typically worse at the leading or trailing edge of pulses.

The loop connecting the sensor to the preamplifier tends to be the most sensitive part of the circuit, so as a general criterion one should inspect that current path very carefully and keep other currents from flowing through any part of the input loop. We'll scrutinize some examples later.

Up to now we've discussed cross-coupling through voltage drops in shared current paths. However, interference does not cross-couple by voltage alone, but also via current injection. Current spikes originating in logic circuitry, for example, propagate through the bussing system as on a transmission line. Individual connection points will absorb some fraction of the current signal, depending on the relative impedance of the node. This is illustrated in Figure 9.15.

Current spikes originate in the switching stage. Most of the current flows through the low-impedance loop towards the right, but current can also propa-

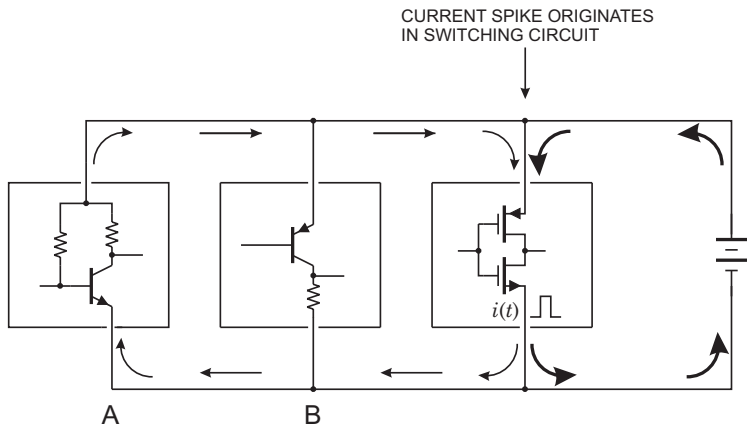

FIG. 9.15. Current spikes originating in a switching circuit propagate through all available current paths. Although most of the current flows through the path of least resistance towards the power supply, a portion of the current flows towards the input. Point B connects to a collector, *i.e.* a high-impedance node, so little current flows into that transistor. Point A, however, connects to the emitter, which presents a low impedance, so the current will take that path and drive the input transistor.

gate towards the left. The left-most stage (node A) appears as a common-base amplifier, so the emitter presents a low impedance that closes the current loop. Node B, on the other hand, is connected to the collector, which presents a high impedance, so the current flow into this leg is small.

9.4.2 *Remedial techniques*

9.4.2.1 *Reduce impedances ("improve grounding")* The most common response to cross-coupling via shared current paths is to reduce the impedance of the shared path. This leads to the "copper braid syndrome", where massive conductors are connected between various parts of the circuit until the interference improves. Colloquially this is called "improving the ground", although grounding may have nothing to do with it. Sometimes these cross-connections introduce an out-of-phase component of the interference, leading to cancellation. The problem with the "copper braid syndrome" is that it tends to be a haphazard approach, which is poorly controlled. Changes to the system can substantially change the current distribution, requiring some more tinkering. This leads to continual surprises, which some relish as the challenge of doing science. Rather than applying *ad hoc* fixes, it is better to avoid the problem in the first place.

9.4.2.2 *Avoid grounds* The most important principle to remember is that signals are transferred via closed loops and don't rely on "grounds". This is true

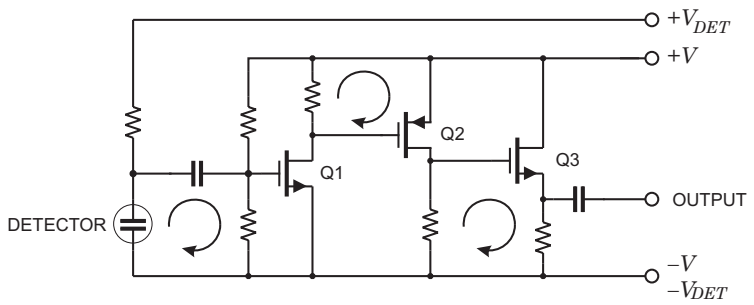

FIG. 9.16. The signal is transferred from the sensor to the input stage and from stage to stage via local current loops.

whether the signals are voltages or currents. Figure 9.16 illustrates this for a sensor and multistage amplifier with output driver.

1. At the input the detector signal is applied between the gate and source of Q1.
2. At the output of Q1 the signal is developed across the load resistor in the drain of Q1 and applied between the gate and source of Q2.
3. The output of Q2 is developed across the load resistor in its drain and applied across the gate and source resistor and load.

Note that – except for the input voltage divider that biases Q1 – varying either the positive or negative supply voltage does not affect the local signals. The circuit does not rely on "grounding", although this circuit would commonly be implemented with the negative voltage rail as common or "ground". However, note that signal transfer involves both the positive and negative voltage rails, so there is nothing special about the negative rail. Also note that the circuit is arranged sequentially, so that current from the output stage does not flow through the input loop. Circuits rely on current return paths, not ground connections!

9.4.2.3 *Control signal paths* Let's extend the circuit in Figure 9.16 to include an external data acquisition system, as shown in Figure 9.17. The output driver is an example of an extended local loop. To provide sufficient current drive for the transmission line the driver is configured as a source follower. The source resistor is chosen to be large compared to the termination resistor in the DAQ system and also large compared to the output resistance of the source follower. For the signal, the source resistor is largely irrelevant; it is only there for DC biasing. Thus, the output signal current does not return through the source resistor. Instead, it returns to the drain of Q3. Figure 9.18 shows the current path. The "bypass capacitor" provides a local return path to the drain of Q3. Without this capacitor the high-frequency current components would have to return through the power supply, or poorly controlled stray capacitances. Figure 9.18 also shows the conventional implementation of the output driver, using a ground for common

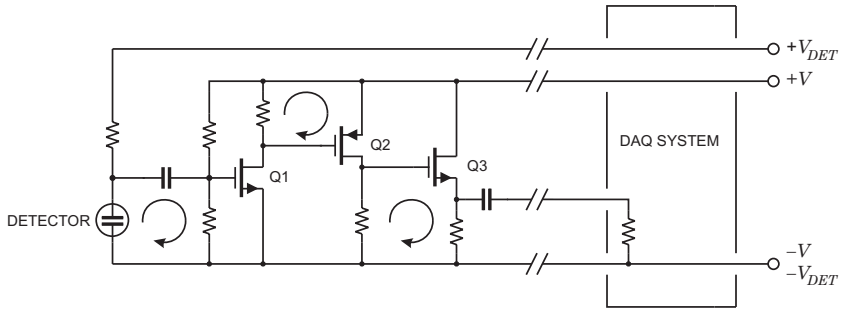

FIG. 9.17. Simple detector readout connected to an external data acquisition system through a multiconductor cable.

connections. In this representation the return path is not obvious, but the circuit diagram is less cluttered. In practically all circuits, analog or digital, the "bypass capacitor" is an integral part of the circuit and must provide a well-controlled current return path, whose impedance is often critical. We'll discuss capacitor properties in Section 9.6.

It is customary to use the negative rail as a common "ground", but as one can see in Figs. 9.16 and 9.18 the positive rail is equally important for the signal return loop. Usually the "common" is configured as a "ground plane", which has a lower inductance than a circuit trace, so the "common" should be chosen to accommodate the most sensitive loops. For example, the recommended configuration for ECL digital circuitry uses the positive rail as a "ground" plane.

In Figure 9.17 the load is connected through a transmission line. The "bypass capacitor" in Figure 9.18 should be located near the output driver stage to localize the current return path. For the signal it is actually a coupling capacitor, so the return leg of the transmission line should have a direct path to the "ground" leg of the bypass capacitor, preferably connected to the same point. In multichannel systems that transmit signals through a multiconductor cable to common data acquisition circuitry, it is very easy to mix current paths, so

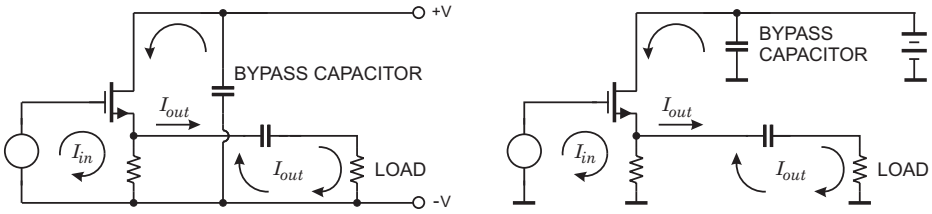

FIG. 9.18. Current return path of a source follower output driver. The right-hand figure shows the circuit as implemented with ground connections.

the output loop must be controlled very carefully. For example, grounding both the "low" end of the sensor and the negative pole of the power supply in Figure 9.17 would provide an alternate current path for the output driver involving the sensitive input loop.

9.4.3 *Potential distribution on ground planes*

Since the "ground" is a large area conducting surface – often a chassis or a ground plane – with a "low" impedance, it is often considered to be an equipotential surface. This assumption is not always justified. As discussed above, at high frequencies current flows only in a thin surface layer ("skin effect"). The skin depth in aluminum is about 100 μm at 1 MHz. A pulse with a 3 ns rise time will have substantial Fourier components beyond 100 MHz, where the skin depth is 10 μm.

Even large area conductors can have substantial resistance. For example, a strip of aluminum, 1 cm wide and 5 cm long has a resistance of about 20 mΩ at 100 MHz (single surface, typical Al alloy). A current pulse of 50 mA will cause a voltage drop of 1 mV, which can be much larger than the input signal. A current pulse of 50 mA is quite common for an output driver (2.5 V into 50 Ω or 5 V into 100 Ω). A 1 V pulse with a 10 ns risetime applied to a 10 nF capacitor requires an average charging current of 1 A ($i = C \cdot \Delta V/\Delta t$). The resistance of a strip is determined by the ratio of length to width, *i.e.* a strip 1 mm wide and 5 mm long, or 1 μm wide and 5 μm long, will have the same resistance. Inductive effects will increase the impedances much beyond the DC value.

Consider a current loop closed by two connections to a ground plane. The current is typically injected into a small area and then spreads out before concentrating at the collection point. The current flow is illustrated in Figure 9.19 (left) together with the equipotential contours. Next we'll place an integrated circuit on this ground plane, together with its bypass capacitor (Figure 9.20). When connected as shown in the left panel, a voltage drop of 50 mV will be introduced into the IC's return path. When the positive supply and ground pads

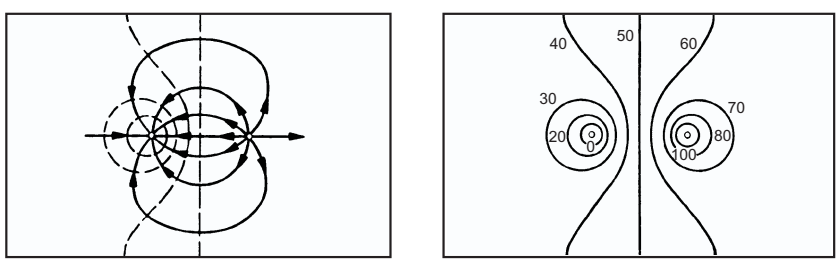

FIG. 9.19. Current distribution (left) and equipotential contours (right) for a current passing through a ground plane with a total voltage drop of 100 mV.

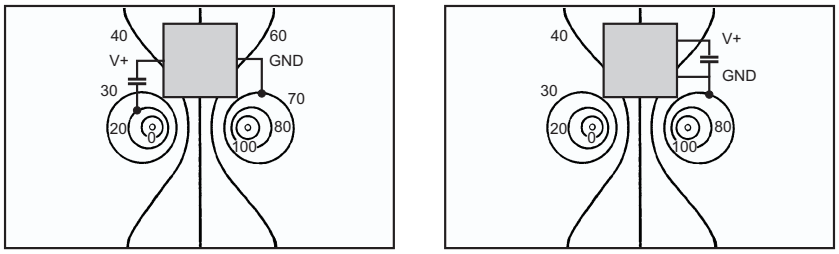

FIG. 9.20. An IC mounted on the ground plane with a bypass capacitor mounted at the left will incorporate a 50 mV drop into the ICs return path (left). The configuration in the second panel avoids this problem.

are located at the same edge of the IC, the bypass capacitor can be connected readily without tapping into voltage drops on the ground plane.

Earlier in this chapter it was pointed out that apertures in shields are signal leaks. In this context high transmission through an aperture is desirable. In multilayer printed circuit boards and hybrids ground connections are frequently made through vias. In this case one must make the diameter of the via d sufficiently large compared to its length l so that high frequencies can pass without significant attenuation. The via inductance (essentially the inductance of a short wire)

$$L \approx 0.2 \left[\frac{\text{nH}}{\text{mm}}\right] \cdot l \left(1 + \log \frac{4l}{d}\right) \ . \tag{9.8}$$

Since the dependence on diameter is logarithmic, multiple vias or a slot are most effective at reducing via impedance.

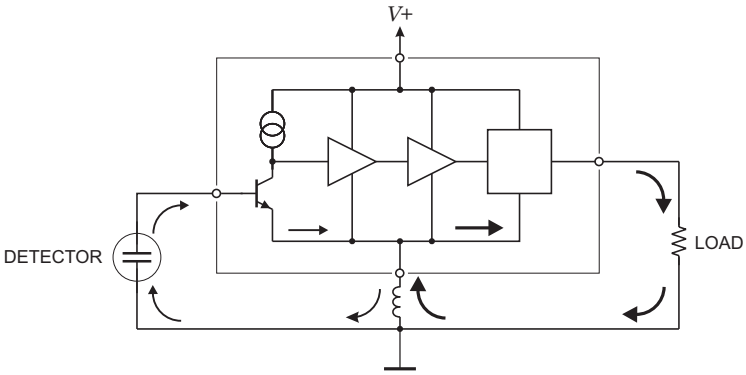

FIG. 9.21. A common ground connection couples low-level and high-level signal return paths.

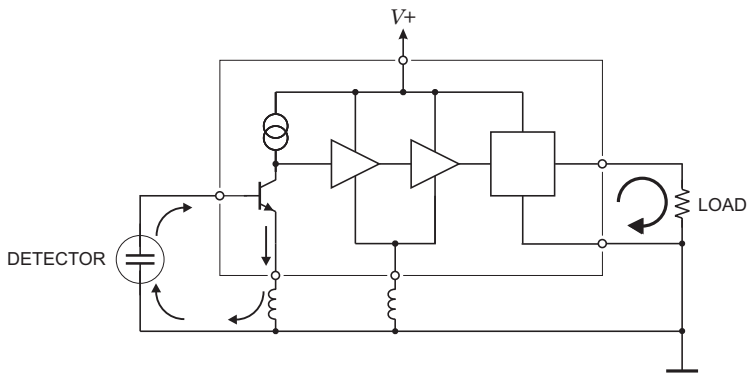

FIG. 9.22. Multiple "ground" connections separate low-level and high-level current return paths.

9.4.4 *Connections in multi-stage circuits*

Figure 9.21 shows an IC combining a preamplifier, gain stages, and an output driver. The negative rail is brought out on a single bond pad. This forces both the input and output current loops to share a common impedance. The output current is typically orders of magnitude greater than the input current, due to amplifier gain and the lower load impedance. Even a short bond wire has an inductance L of order nH, so a 1 mA current transient ΔI at 10 – 20 MHz will cause a voltage change $\Delta I \cdot \omega L$ of about 100 μV. This changes the voltage across the sensor. For a 10 pF sensor capacitance, this would correspond to a signal charge of about 6000 e. Figure 9.22 shows how separating the "ground" connections for sensitive and high current loops avoids cross-coupling and constrains the extent of the output loop, which tends to carry the highest current.

Figure 9.21 also illustrates the use of a popular technique – the "star" ground – and its pitfalls. This circuit has two star connections coupled by the bond wire impedance.

Circuits cannot always be implemented with a cleanly sequential signal path and sometimes one simply has to live with flawed designs. One tool against cross-coupling in these systems is breaking parasitic current paths to isolate sensitive loops.

9.5 Breaking parasitic current paths

Multichannel systems require multiple connections between sensitive analog circuitry and the readout and control systems. Many potential problems associated with shared signal paths can be avoided by good circuit design. However, compromises are always necessary, so additional techniques must be applied to arrive at a functioning system.

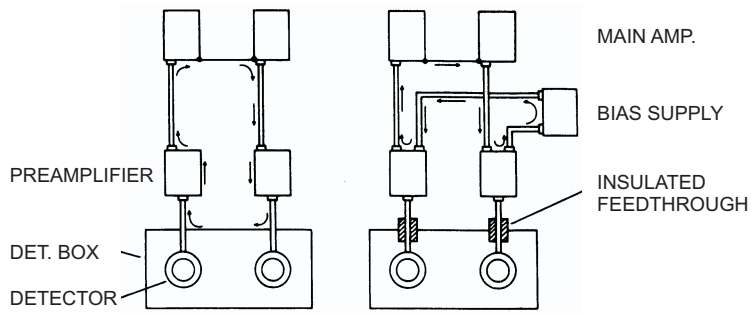

FIG. 9.23. Detectors mounted in a vacuum chamber are connected to external preamplifiers through vacuum feedthroughs. The preamplifier outputs are connected to main amplifiers, often mounted some distance away. A shared current path is formed by the common connection at the feedthroughs. Insulated feedthroughs (right) break the current path. The second panel also show a detector bias supply common to both detectors, which forms a new loop. However, the shared path no longer includes the preamplifier inputs, which are the most sensitive part of the system.

9.5.1 *Isolate sensitive loops*

Multiple connections are usually unavoidable, but not all are equally sensitive. This is illustrated in Figure 9.23. The configuration at the left has a loop that includes the most sensitive part of the system – the detector and preamplifier input. By introducing insulated feedthroughs, as shown at the right, the input loop is broken. The right hand figure also shows a new loop, introduced by the common detector bias supply. However, this loop is restricted to the output circuit of the preamplifier, where the signal has been amplified, so it is less sensitive to interference.

Note that the problem is not caused by loops *per se*, *i.e.* enclosed areas, but by the multiple connections that provide entry paths for interference. Although not shown in the schematic illustrations above, both the "detector box" (*e.g.* a vacuum chamber) and the main amplifiers (*e.g.* in a NIM bin or VME crate) are connected to potential interference sources, so it is important to isolate the input signal path to close it off to interfering currents.

9.5.2 *Differential signal transmission*

In a preceding section the current return path of a single-ended driver was discussed. Since single-ended transmission circuits utilize common returns, interactions between multiple circuits through the common ground are unavoidable. Differential (balanced) transmission systems eliminate this problem. Differential receivers have balanced inputs (inverting and non-inverting) and respond to the difference signal between the two inputs, regardless of a "common mode" component. Assume that the levels at the output of the driver are $V_0 + \Delta V$

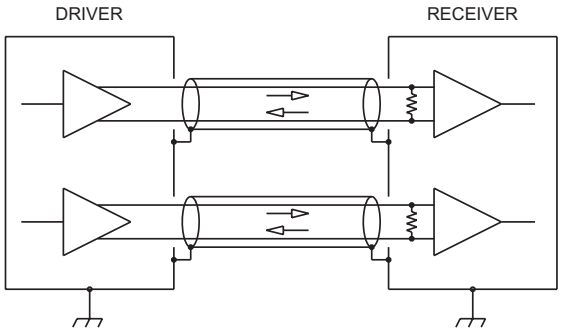

FIG. 9.24. Fully balanced signal transmission.

and $V_0 - \Delta V$. Furthermore, let an interfering signal introduce the same voltage V_{CM} in both legs of the transmission line, so that the levels at the inputs of the receiver are $V_{R1} = V_{CM} + V_0 + \Delta V$ and $V_{R2} = V_{CM} + V_0 - \Delta V$. Since the differential receiver responds to the difference between the two inputs, the output is $V_{R1} - V_{R2} = 2\Delta V$.

Besides providing common mode noise rejection, differential receivers also allow "ground free" connections. Figure 9.24 shows a fully balanced connection, using both balanced transmitters and receivers. Since the common mode range over which the receiver functions properly is limited, some referencing between the transmitter and receiver is necessary, here provided by a shield connection.

Sometimes only single-ended drivers are available or the number of available bond-pads on an IC precludes a fully balanced system, but differential receivers can still be utilized to provide common-mode rejection, as shown in Figure 9.25. No shield is included here, representative of ribbon cable or twisted pair trans-

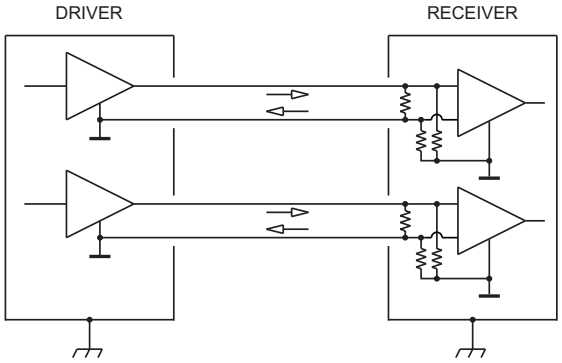

FIG. 9.25. Use of balanced receivers with single-ended drivers also provides common mode rejection, but doesn't ensure balanced currents in the line.

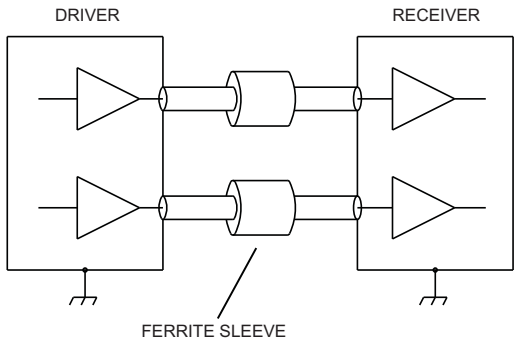

FIG. 9.26. Threading coax cables through ferrite sleeves suppresses common mode currents without affecting the desired differential current.

mission lines. To provide DC referencing a pair of matched resistors is connected from the inputs to local ground. These could also form the termination resistor for the transmission line, but it can be advantageous to use a large referencing resistance to isolate the two grounds, so a separate termination resistance is included. The value of the referencing resistors is limited by the common mode current flow.

For a balanced system to work optimally, the current in both legs should be balanced, *i.e.* instantaneous levels are equal, but of opposite sign. This is difficult to ensure with single-ended drivers. Single-ended drivers also have the side-effect of injecting current spikes into the power supply lines and common "grounds". Balanced drivers can operate in current steering mode, which maintains constant current and also ensures balanced currents in both output legs. In both implementations it is important that the differential receiver maintain its common mode rejection up to the highest frequencies where the system is sensitive. This should not be taken for granted, especially at high frequencies.

The Low Voltage Differential Signaling (LVDS) standard is now widely used and drivers and receivers are commercially available. The current version of the standard is capable of gigabit rates. The original standard was tailored to a single point $100\,\Omega$ load, but variants address the needs of multipoint bussing. The voltage swing of 350 mV reduces the levels of circulating current. Custom readout ICs for detectors have utilized LVDS very successfully, utilizing current steering drivers (which ensures current balance in the transmission line) with commercial receivers.

9.5.3 *Blocking Common Mode Currents*

Coaxial transmission lines provide excellent shielding between lines, but the common ground provides a path for common mode signals. These can be blocked by threading the cable through ferrite sleeves, as shown in Figure 9.26. The ferrite sleeves block common mode currents without affecting the desired signal currents

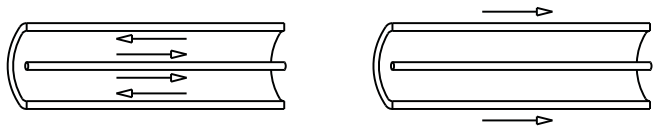

FIG. 9.27. At high frequencies differential currents flow on the outside of the inner conductor and the inside of the shield (left). In addition, a common mode shield current can flow on the outside of the shield (right).

because the net magnetic field of the differential signal current is zero outside the cable. The common mode current has the same polarity on both the inner and outer conductors, so there is a net field outside the coax cable upon which the high permeability ferrite acts.

The effectiveness of this technique depends on the frequency range where attenuation is needed and the properties of the ferrite. The permeability of ferrites depends on frequency, and introduces both an inductive and resistive term $Z = \mathbf{i}\omega L + R$. At low frequencies the permeability yields a purely inductive reactance, but at higher frequencies the losses increase and the inductance rolls off. At yet higher frequencies the loss term rolls off and the ferrite becomes ineffective. Where the loss term dominates, the impedance appears purely resistive, rather than inductive. This provides a high impedance at high frequencies (order 100 MHz) where stray capacitances could introduce series resonances that obviate the blocking power of the network. High permeability ferrites have useful characteristics at low frequencies ($\mu \sim 10^3$ up to about 1 MHz), whereas lower permeability materials extend the frequency range ($\mu \sim 10$ up to about 1 GHz).

The common mode choke also suppresses shield currents. At high frequencies a coax cable becomes a three-conductor line, because the skin effect separates the inner surface of the shield from the outer surface. As a result, the cable can carry a shield current in addition to the desired mode (Figure 9.27). For the desired differential currents the external magnetic field cancels, whereas for the shield current the field is that of a single wire.

This technique is most useful in suppressing high frequency currents and can also be applied to twisted pair (or flat) ribbon cables. Ferrite sleeves are available in both cylindrical and flat geometries. Toroid cores can be used to obtain higher impedance by looping the cable through the toroid multiple times.

9.5.4 *Isolating parasitic ground connections by series resistors*

Sensors require bias voltages and the connection of bias voltage supplies can introduce current paths in the most sensitive part of the system. Figure 9.28 shows a sensor with preamplifier connected to a remote shaper (or data acquisition system) and bias supply mounted in a common rack. Since both the shaper and bias supply are connected in the rack, the output return current of the preamplifier could take a parasitic path through the bias supply and input circuit. This path is blocked by inserting isolation resistors R_2 and R_3 in both legs of the bias

supply. Since the bias currents are small, these resistors can be relatively large ($\sim$ kΩ) compared to the impedance of the output drive circuitry. Capacitor C_2 together with the isolation resistors forms a low-pass filter and provides a direct return path for any differential interference propagating through the line from the bias supply. R_1 provides additional filtering to the detector backplane and C_1 closes the signal return path for the detector. Although C_1 is commonly referred to as a bypass capacitor, which implies a connection to ground, it really is a coupling capacitor, which should provide a direct path from the detector backplane to the input reference of the preamplifier.

When debugging an existing system the isolation resistors can also be mounted in an external box that is looped into the bias line. Either use an insulated box or be sure to isolate the shells of the input and output connectors from one another.

A simple check for noise introduced through the detector bias connection is to use a battery. "Ground loops" are often formed by the third wire safety ground in the AC power connection. Avoid voltage differences in the "ground" connection by connecting all power cords associated with low-level circuitry into the same outlet strip.

9.5.5 *Directing the current flow away from sensitive nodes*

A multichannel timing discriminator was built on a PC board and mounted in a NIM module. All inputs and outputs were mounted on the front panel. The outputs drove about 20 mA into 50 Ω cables. Whenever an output fired, the unit broke into oscillation. The oscillation was traced to a portion of the output

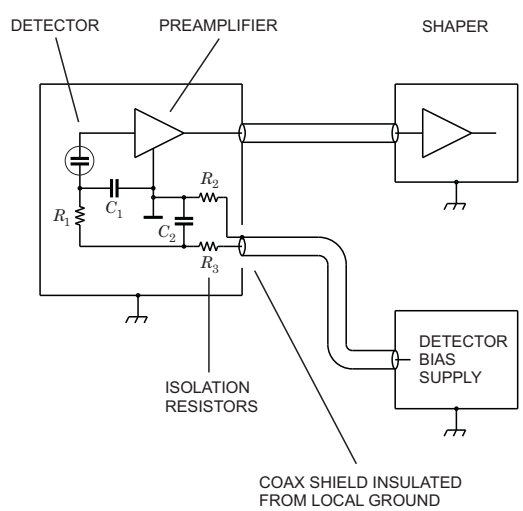

FIG. 9.28. Series resistors R_2 and R_3 in both legs of the sensor bias line isolate the ground of the detector bias supply and break the parasitic current return path through the shaper and bias supply.

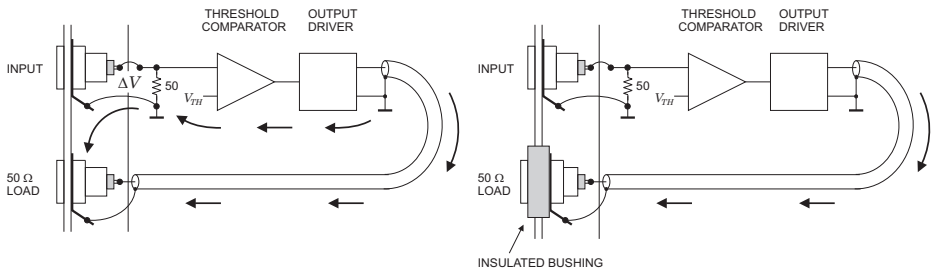

FIG. 9.29. Self-oscillation due to a portion of the output signal flowing through the input circuit (left) was eliminated by insulating the output connector from the front panel (right).

current that flowed through the input ground connection, as shown in Figure 9.29. The voltage drop ΔV was sufficient to fire the comparator. Insulating the output connector from the front panel broke the loop and removed the problem. This form of self oscillation is quite common in mixed analog-digital systems. For example, in a large strip detector system cross-talk from the digital readout to the analog input can increase the occupancy, which in turn increases the digital readout activity.

Often it is convenient to replace the coax cable at the output by a strip line integrated on the PC board. For an ideally conducting ground plane the return current is concentrated adjacent to the strip line ("image current"), but it spreads when the impedance of the ground plane is appreciable. In this implementation the current paths can be controlled by patterning the ground plane, as shown in Figure 9.30. This technique can be applied advantageously in mixed analog-digital systems to steer digital current spikes from the analog circuitry.

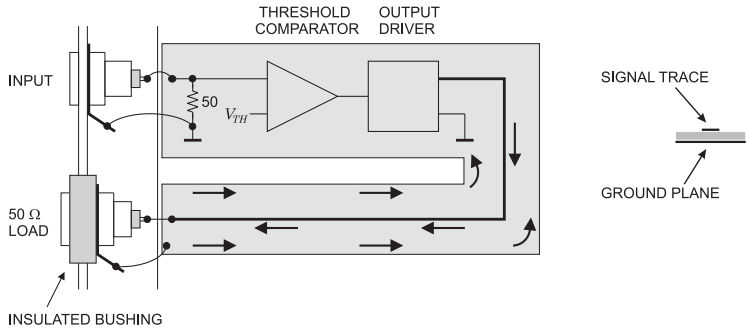

FIG. 9.30. When using strip lines (right detail) or circuit board traces where the current return is via a ground plane, signal paths can be isolated by patterning the ground plane.

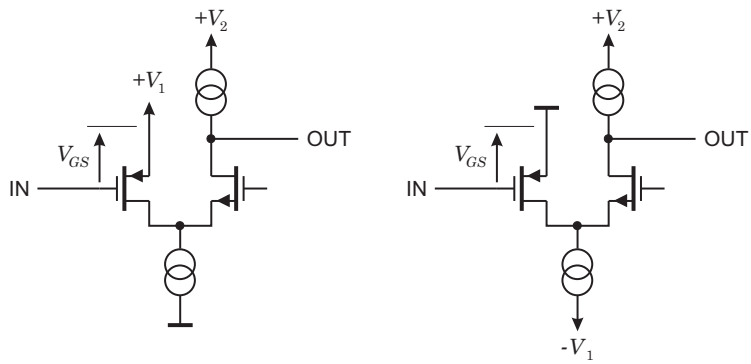

FIG. 9.31. Folded cascode with a PMOS transistor and a single supply voltage (left) places a voltage supply in series with the input signal. Use of a split supply(right) places the input transistor's source at DC ground, which as usually implemented tends to be cleaner.

9.5.6 *The folded cascode*

The folded cascode is frequently used in preamplifiers optimized for low power. Summarizing the description in Chapter 6, the cascode combines two transistors to obtain the high transconductance and low noise of a wide transistor combined with the high output resistance (increased by local feedback) and small output capacitance of a narrow transistor. Furthermore it reduces the capacitance between output and input, as the gate–drain capacitance of the input transistor "sees" a low impedance at the drain. Since the input transistor determines the noise level, its current requirement tends to dominate. In a conventional linear cascode the current required for the input transistor must flow through the whole chain. The folded cascode splits the DC path and allows the (second) cascode transistor to operate at a lower current and, as a result, higher output resistance. It also allows a smaller supply voltage, with an overall reduction in power dissipation.

Since PMOS transistors tend to have lower "$1/f$" noise than NMOS devices, the adaptation shown in the left panel of Figure 9.31 is often used. The problem with this configuration is that the supply V_1 becomes part of the input signal path. Unless the V_1 supply bus is very carefully configured and kept free of other signals, interference will be coupled into the input. Thus, it is better to use a split supply and "ground" the source of the input transistor, as shown in the right panel of Figure 9.31.

Figure 9.32 shows the conventional variant of the PMOS folded cascode connected to a strip detector. Unless the connection points of the bypass capacitors from the FET source and the detector backplane are chosen carefully, interference will be introduced into the input signal loop. Figure 9.33 shows the same amplifier redrawn to provide a direct capacitive connection from the backplane

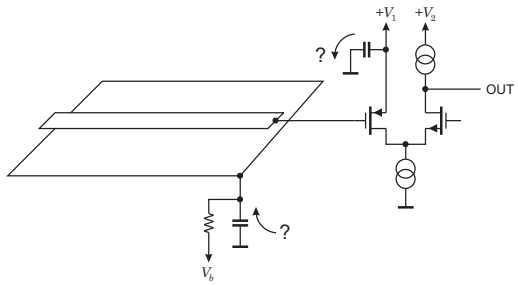

FIG. 9.32. When not carefully implemented the PMOS folded cascode forms a distributed signal return path that is prone to interference.

to the source of the input transistor. This also illustrates the concept that the power bussing should be treated in the same manner as "ground".

It is much better to "ground" the FET source to a local signal reference and use a negative second supply, as shown in the second panel of Figure 9.33. Connected to a strip detector, this configuration provides a direct input return loop. For some (mythical?) reason positive supplies are more popular. Proper connection of the detector can still provide a direct input path. However, in most implementations power and ground lines are not treated equally, so the ground has a lower impedance. As a result, positive and negative voltage lines are more susceptible to pickup, so the implementation with the grounded source is safer, as deeply entrenched habits tend to prevail.

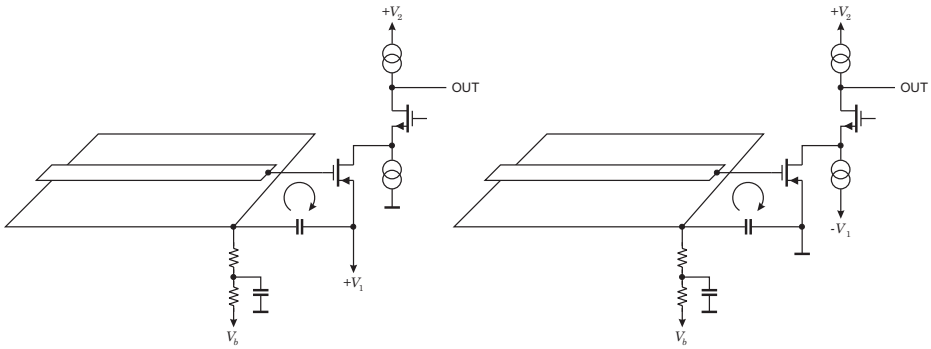

FIG. 9.33. Implementing the folded cascode with "clean" input signal return paths. When only positive supplies are used, it is important to provide a direct path from the backplane to the source of the input transistor (left). Using positive and negative supplies allows the source to be grounded (right).

9.6 Capacitors

The preceding discussions have underscored the key role of "bypass" capacitors in the signal path. When used to close low-impedance loops, as in bypass capacitors at IC power connections, the impedance must be small. Although drawn as an ideal capacitor in circuit diagrams, it is important to remember that every capacitor also has a series inductance and resistance. The capacitance and inductance form a series resonant circuit. At resonance the impedance is limited by the series resistance and above resonance the capacitor appears inductive. The frequency response is complicated by the fact that the equivalent series resistance is frequency dependent. When in the signal path, the capacitor's series resistance can introduce noise.

The inductance depends on the geometry. Foil capacitors are commonly made of thin metallized layers of plastic, often rolled into a cylinder, but also available as rectangular stacks. These capacitors tend to have a rather high inductance and their use is limited to frequencies up to hundreds of kHz or so. In tantalum electrolytic capacitors the dielectric is formed as an oxide on a rough surface or sponge-like electrode. These devices provide capacitance up to hundreds of μF and are usable into the MHz range. However, they have poor tolerances and large temperature coefficients. The lowest inductance is provided by multilayer ceramic chip capacitors, where many layers of metallized ceramic are fused together. Essentially, these are many small capacitors connected in parallel, so the inductance is reduced correspondingly. Multilayer ceramic capacitors are the devices of choice in high frequency applications, especially when used as chip capacitors, which practically eliminate the lead inductance. However, there are significant differences between the dielectrics used in these devices.

The dielectrics are $BaTiO_5$ based, diluted with rare earth oxides. Z5U or Y5V dielectrics have dielectric constants up to $2 \cdot 10^4$, so they provide high capacitance in a small volume. For example, a $0.1\,\mu$F/16 V capacitor has a footprint of $1.6 \times 0.8\,\text{mm}^2$ with a thickness of 0.9 mm. Available capacitances range into the 10s of μF. These dielectrics have a large voltage coefficient. Operated at their rated voltage the capacitance of Y5V capacitors can be as small as 20% of the value at zero volts. The capacitance also shows a strong frequency dependence.

X7R has a dielectric constant of about 10^3. The capacitance variation with frequency is about 20% from 1 kHz to 10 MHz and the voltage coefficient is small. Maximum capacitance is several μF.

The highest quality ceramic dielectric is NP0 ("negative positive zero") with a temperature coefficient of ± 30ppm/°C and practically no dependence of capacitance *vs.* frequency. The capacitance is limited to about 10 nF.

Since the dielectrics are piezoelectric, ceramic capacitors are microphonic. Application of a 1 kHz alternating voltage can make a capacitor "sing". Conversely, vibration will translate into voltage changes. Z5U and Y5U are most susceptible.

Figure 9.34 shows the impedance *vs.* frequency for various capacitors. The inductance is determined by the size of the package and is of order nH. For a given

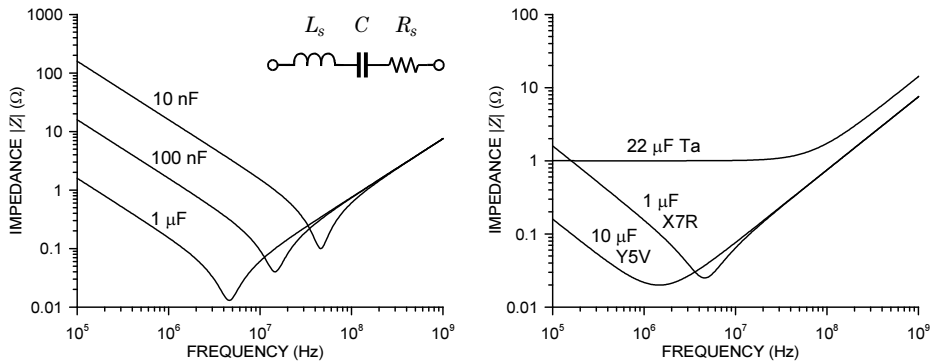

FIG. 9.34. Impedance *vs.* frequency for various surface mount capacitors. The left plot shows the impedance for three values of X7R capacitors in 1206 packages (3.2 mm × 1.6 mm footprint). The second panel compares a 1 μF X7R with a 10 μF Y5V capacitor and the third curve is for a 22 μF Ta capacitor. Although the Y5V capacitance is ten times larger, the impedance at high frequencies is practically the same. The Ta capacitor has a significantly higher series resistance, so its higher capacitance only manifests itself at low frequencies. The curves were calculated using vendor data.

package the equivalent series resistance increases with decreasing capacitance. The second panel of Figure 9.34 compares a 1 μF X7R capacitor with a 10 μF Y5V device. At high frequencies the impedance is practically the same. However, with an applied DC voltage the Y5V capacitor suffers from its strong voltage dependence. The plot also shows data for a 22 μF Ta chip capacitor. The Ta capacitor has a significantly higher series resistance, so its higher capacitance only becomes effective at low frequencies.

The capacitor providing the signal return from the amplifier to the detector backplane must sustain the full bias voltage, which may require a voltage rating of 1 kV in a strip detector subject to high radiation levels. A typical 1 kV capacitor with 100 nF in X7R has a footprint of $6.4 \times 4.5\,\mathrm{mm}^2$ and is 2.5 mm thick. As the capacitor materials have high atomic number and density, capacitors can contribute significantly to the material in a tracking detector. The volume of ceramic capacitors with a given capacitance and voltage rating has decreased significantly over the years, as the dielectric strength (breakdown field) has increased. Mixtures with high dielectric constants and reduced voltage coefficients will further reduce capacitor size.

9.7 System considerations

9.7.1 *Choice of shaper*

Aside from considering the noise performance of shapers, it is also useful to compare their immunity to interference. Figure 9.35 compares the frequency

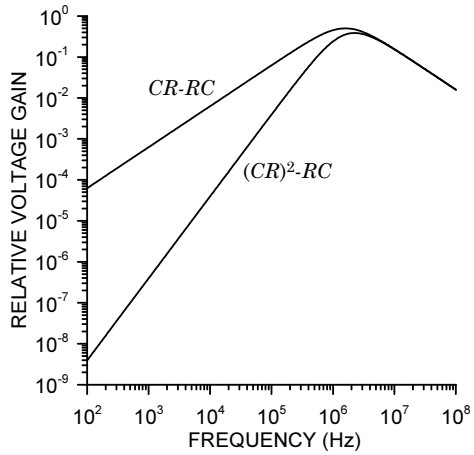

FIG. 9.35. Frequency response of CR-RC and CR-CR-RC pulse shapers. The bipolar shaper provides much greater attenuation of low frequencies.

response of a simple CR-RC shaper and a bipolar CR-CR-RC shaper, both with the same peaking time. The CR-CR-RC shaper provides significantly better low-frequency attenuation then the CR-RC shaper, which can be critical if power line pickup is a problem. Correspondingly, the cascaded low-pass filters of a CR-nRC shaper provide a more rapid high-frequency roll-off than the simple CR-RC shaper. Although a bipolar shaper has slightly inferior noise performance than a unipolar shaper ($Q_{n,opt} = 1.406\sqrt{i_n e_n C}$ *vs.* $Q_{n,opt} = 1.355\sqrt{i_n e_n C}$), it may provide superior results in the presence of significant low-frequency noise or interference. The bipolar shaper also avoids rate-dependent baseline shifts.

9.7.2 *Local referencing*

The preceding discussion has emphasized cross-coupling through shared current paths. An equally important consideration is capacitive coupling of a detector or detector module to its neighbors and environment. Figure 9.36 illustrates how noise on a support structure can couple into a detector module. Clearly, the stray capacitance from the module to the support structure should be minimized and spurious signals on the mounting structure must be controlled. This is another variation on controlling current paths, but also requires local potential referencing. If stray currents are small, local potential referencing may be accomplished through high-value resistors. Another approach is to make any interference into the module a common mode disturbance. This means that the detector backplane and the local module ground change by the same potential. The inductance of long power cables may provide adequate impedance to allow the module to "float" with respect to the interference source, especially if the coupling capacitance is small. Common mode chokes in the power and data cabling can be

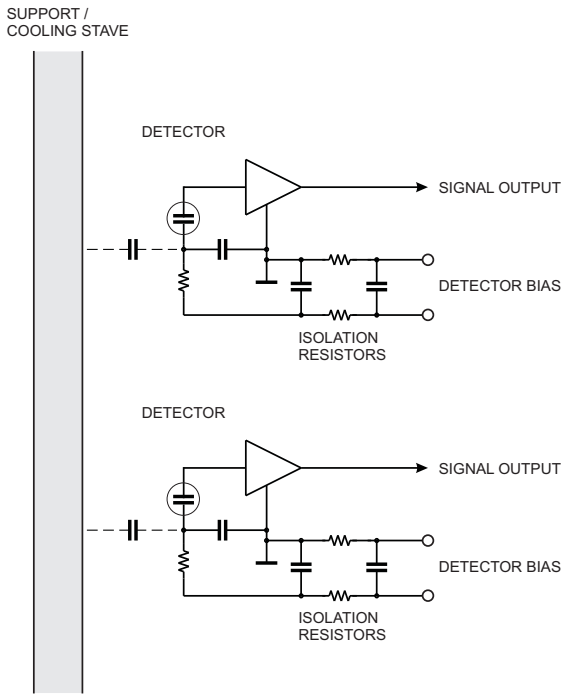

FIG. 9.36. Capacitive coupling between detectors or detector modules and their environment introduces interference when relative potentials and stray capacitance are not controlled.

used to increase the impedance with respect to the external readout system and balance the coupling of interference. This is one of the most difficult problems in designing a system. Because of the many possible variations, there are no hard and fast solutions.

Bibliography

Johnson, H. and Graham, M. (1993). *High-Speed Digital Design*. Prentice Hall PTR, Upper Saddle River. ISBN 0-13-395724-1, TK7868.D5J635

Johnson, H. and Graham, M. (2002). *High-Speed Signal Propagation*. Prentice Hall PTR, Upper Saddle River, ISBN 0-13-084408-X, TK5103.15.J64 2002

Morrison, R. (1998). *Grounding and Shielding Techniques* (4th edn). Wiley, New York. ISBN 0-471-24518-6, TK7878.4.M66 1998

Ott, H.W. (1988). *Noise Reduction Techniques in Electronic Systems*(2nd edn). Wiley, New York. ISBN 0-471-85068-3, TK7867.5.087 1988

Paul, C.R. (1992). *Electromagnetic Compatibility*. Wiley, New York. ISBN 0-471-54927-4, TK 7867.5.P38

APPENDIX A

SEMICONDUCTOR DEVICE TECHNOLOGY

Fabrication of semiconductor devices is a complex technology that uses many different techniques. This is only a brief survey of the key steps to provide a better understanding of the implications in utilizing devices. Many process tutorials can be found on university web-sites. For a detailed description of modern fabrication techniques see a recent compilation by Nishi and Doering (2000), for example. Lutz (1999) describes simulations and some aspects of detector fabrication.

The key ingredients of a semiconductor device fabrication process are

1. Bulk material, *e.g.* Si, Ge, or GaAs.
2. Dopants to create p- and n-type regions.
3. Deposition of contact and insulator films.
4. Patterning to selectively introduce dopants and remove material.
5. Passivation to protect the semiconductor surfaces from damage and contaminants.

Practically all semiconductor devices are fabricated in a planar geometry. Notable exceptions are large volume detectors for gamma-ray spectroscopy.

A.1 Bulk material

The starting point is a semiconductor wafer, *i.e.* a slice of silicon from a large cylindrical single crystal (boule). Here detectors differ significantly from integrated circuits. Electronic circuitry resides in a thin layer at the wafer surface, whereas in detectors the diode depletion region extends into or throughout the bulk. This requires both high purity and a low density of mechanical microdefects, *e.g.* dislocations. This is achieved by growing the crystals using a float zone process, which through zone refining removes impurities and yields essentially dislocation-free crystals. A polysilicon rod is passed through a heating coil, typically using RF induction heating. Impurities segregate from the molten zone and the melt solidifies as a single crystal. Since the silicon boule is self-supporting and only in contact with a gas ambient, crystal growth can proceed without introducing significant contamination. Resistivity levels in the range 10 – 50 kΩ are obtainable, although routine levels are around 5 kΩ for n-type material. Float zone wafers with 100 mm diameter were standard for many years, but 150 mm wafers are now common. The major market drivers are high voltage MOSFETs and rectifiers. Typical wafer thicknesses for detector applications have been around 300 μm, as this would yield depletion voltages < 100 V with typical

resistivities, while yielding an adequate signal-to-noise ratio with low-power electronics. Refinement of detector structures to operate at voltages $> 500\,\mathrm{V}$ after radiation damage has extended the range of usable thicknesses.

Wafers for integrated circuits, on the other hand, are grown using the Czochralski method, which starts with polysilicon fragments placed in a silica crucible that is heated for crystal growth. Typical resistivities are substantially lower ($< 50\,\Omega\mathrm{cm}$), although higher resistivity material in the range adequate for detectors is now available. As crystal growth proceeds in a crucible, impurity control is more difficult. However, Czochralski-grown silicon has some major advantages for IC fabrication. The rather high density of microdefects provides efficient gettering of impurities and the high oxygen content strengthens the material, reducing breakage of wafers during fabrication. Since IC fabrication involves many more processing steps than detectors, resistance to breakage is crucial. Czochralski wafers are readily available up to 300 mm diameter. The need for high-purity silicon for detectors is relaxed when radiation damage is significant. Furthermore, as discussed in Chapter 7 oxygenation ameliorates the effects of displacement damage, so Czochralski-grown silicon is being investigated for detectors.

Traditionally, detector wafers have used the $\langle 111 \rangle$ orientation. This goes back to an analysis in the early days of semiconductor radiation detectors that showed a smaller probability of channeling, *i.e.* the passage of particle through the "free channels" of the crystal without interaction. The semiconductor industry, however, has used $\langle 100 \rangle$ because of the smaller number of bonds per unit area, $6.8 \cdot 10^{14}\,\mathrm{cm}^{-2}$ for $\langle 100 \rangle$ compared to $11.8 \cdot 10^{14}\,\mathrm{cm}^{-2}$ for $\langle 111 \rangle$ (Sze 1981), which is advantageous in MOS structures, as it reduces the density of interface traps. In most detector applications the small probability of channeling is not very relevant, so both orientations are suitable.

A.2 Introduction of dopants

Dopants are introduced to make the silicon n- or p-type. Typical n-dopants are phosphorus and arsenic, whereas boron is the most common p-dopant. Dopants can be introduced either by diffusion or ion implantation. For doping by diffusion the wafer is exposed to a gaseous ambient of the desired dopant at temperatures of $800 - 1000\,^{\circ}\mathrm{C}$. Higher temperatures accelerate the diffusion process, so the concentration depends on both the temperature and duration of exposure. High doping concentrations lead to deep concentration profiles. Since the diffusion constants tend to depend on concentration, the doping profile often deviates from the complementary error function distribution expected from simple theory. Arsenic, for example, shows a more "box-like" profile (Fair 1981).

The second technique is ion implantation, which today is most commonly used, as it allows more precise control of doping levels and depth. Accelerators are used to bombard the wafers with the desired dopant ions at energies ranging from keV to MeV. Since the beam spot is smaller than the wafer, it is scanned with a carefully designed pattern to ensure the desired uniformity (Ryssel and Glawischnig 1982). After implantation the ions are in interstitial sites, so they

are electrically inactive. Thermal annealing is required to move the dopant atoms into bound lattice sites and electrically activate them. During thermal annealing the dopants diffuse, so the final doping profile extends deeper into the bulk than the range of the implanted ions. Implanted wafers are commonly annealed in furnaces of the same type used for diffusion, although flash annealing with brief bursts of intense infrared radiation – typically from tungsten-halogen lamps – is also used to minimize diffusion. Figure 2.36 in Chapter 2 shows a practical doping distribution.

Doping by diffusion ensures a high level of activation, whereas formation of shallow layers by ion-implantation requires careful control of the annealing process to limit diffusion while ensuring good activation. Since the doping profiles from diffusion are more gradual, they are more conducive to high voltage operation, as the fields at the edges of doped regions are smaller by virtue of the greater radius of curvature than usually attained in ion implanted devices. Ion implantation generates interstitials, which migrate beyond the doped region, so defect zones tend to accompany ion-implanted layers. Frequently these are not of great importance, but they can induce low-frequency noise in transistors.

A.3 Deposition

Making electrical contact to doped regions requires metallization. In a strip detector the metallization must extend along the whole length of the strip to reduce the series resistance in the signal loop, which introduces both thermal noise and signal dispersion. The Fermi level of the contact must be compatible with the semiconductor to ensure a non-rectifying ("ohmic") contact. This is often a problem with large bandgap semiconductors. Aluminum is commonly used in silicon detectors. In integrated circuits doped polysilicon and a variety of refractory metals and alloys are used. Metallization is applied either in thermal evaporators or in sputtering systems, where atoms are ejected from a cathode through bombardment by inert energetic ions, typically argon. Atoms are ejected with an approximately cosine distribution, which is directed towards the sample. Deposition by evaporation is susceptible to "shadowing", whereas sputtering systems can be designed to provide better step coverage and fill trenches. Evaporators are still found in R&D facilities, whereas IC fabrication uses sputtering exclusively. Multistation chambers allow multilayer depositions without breaking vacuum. This is also important in the deposition of antireflective coatings for photodiodes.

Chemical vapor deposition (CVD) and plasma enhanced CVD are used to form films of polysilicon (doped and undoped), silicon nitride, and silicon dioxide (both low and high temperature). Doped polysilicon is used for the integrated bias resistors in strip detectors and gate electrodes of MOSFETs. Some applications of deposited oxides will be shown below.

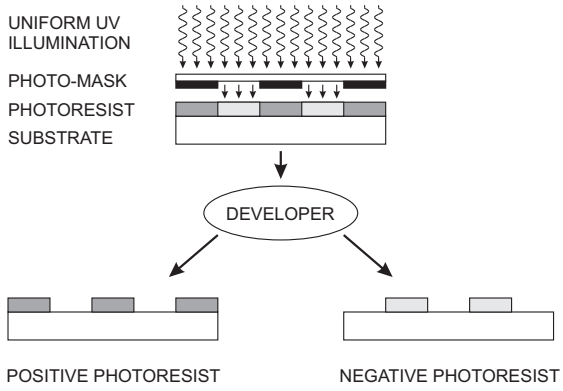

FIG. A.1. Layers are patterned by exposing a photosensitive layer (photoresist) through a mask. After immersion in a developer bath the illuminated areas either dissolve (positive photoresist) or remain (negative photoresist).

A.4 Patterning

The shapes and sizes of device structures are defined either by selective doping or layer removal. Both utilize a technique called photolithography, illustrated in Figure A.1. A photo-sensitive material ("photoresist") is applied to the surface. The photoresist is illuminated through a mask that is transparent in the areas that are to be removed. The illumination changes the structure of the photoresist, so that when immersed in a developer solution the exposed areas dissolve. This is a "positive photoresist". In a "negative photoresist" the illuminated fields polymerize and the unexposed areas dissolve in the developer. This requires a photomask with the negative of the desired pattern. This is useful when patterning a metallization layer where the photoresist remains on the areas to be preserved and protects them from etching. Photoresist is applied and distributed across the wafer by spinning. The desired resist thickness results from a combination of centrifugal force and resist viscosity. Typical resist thicknesses are 1 – 2 μm.

Photolithography is used to pattern layers that are used as protective masks. Silicon is impervious to etchants that attack silicon-dioxide, so an oxide layer can be patterned to block diffusion or ion-implants and only dope selected areas.

Material is removed by etching, either by immersing the wafer in a bath ("wet etching") or through plasma etching, also called "dry etching". Dry etching removes material with gaseous reactants that produce gaseous products. A radio frequency or microwave discharge dissociates and ionizes the reagent to produce a plasma of highly reactive molecular and atomic free radicals. The processes involved in plasma etching are very complex and the reagents and operating conditions must be tailored to specific applications.

Wet etching tends to undercut thick layers, so it is poorly suited to etching trenches, for example. Plasma etching can be very directional and local electric fields direct the reagent ions to shadowed surfaces. In both wet and dry etching one frequently exploits the selectivity of the etch to different materials. Wet etch rates depend strongly on temperature and etchant concentration, so etchant selectivity is used to control the thickness of material removal. As already noted above, silicon is impervious to most wet etches, so when etching oxide the underlying silicon provides an etch stop. Etching silicon requires an oxidizing solution, that oxidizes the silicon atomic layer by layer and then dissolves the oxide. Doping can also affect etch rates, so highly doped layers are also used as etch stops to limit the thickness of material to be removed during etching. Plasma etching is the prevailing technique in IC processing. Wet etching is adequate for many detector applications, although plasma etching is a key ingredient in fabricating high performance CCDs.

A.5 Surface passivation

The silicon surface must be protected by a layer that establishes a well-controlled termination for the "dangling bonds" where the crystal lattice is truncated. Furthermore, the thermal coefficient of expansion should be reasonably well matched to silicon. Thermally grown silicon dioxide has proven to be ideal for this purpose, so it is used for the gate insulator in MOSFETs and as surface passivation between strips or pixels in radiation detectors. In these respects SiO_2 on Si is unequalled – indeed this is probably the key ingredient that allows Si ICs to achieve a circuit density that is at least an order of magnitude greater than in any other semiconductor technology.

Oxides can be thermally grown or deposited by a process called chemical vapor deposition. In both instances the wafers are exposed to gaseous ambients in high temperature furnaces, typically in quartz tubes surrounded by heating elements. The appropriate gases flow through the furnace tube, often using specially designed quartz manifolds to maintain local gas concentrations. Both thermally grown – where oxygen diffuses into the silicon and forms SiO_2 layer by layer – and deposited oxides are included in the example detector process flow described below. Silicon dioxide does not block sodium, which can strongly affect device characteristics, so an additional layer of phosphosilicate glass (PSG) is frequently applied. This also getters mobile ions from the underlying oxide.

A.6 Detector fabrication

Many variations of the basic process ingredients can be combined to yield very good detectors. Processes must avoid introduction of deleterious impurities (those that introduce mid-gap states) and provide surface passivation. A major change in detector fabrication was initiated by Kemmer (1980, 1984), who developed the first oxide-isolated silicon detector process that yielded good results ("planar process"). A drawback of that specific recipe was that high-temperature steps were limited severely to maintain low reverse bias currents. Consequently, ion

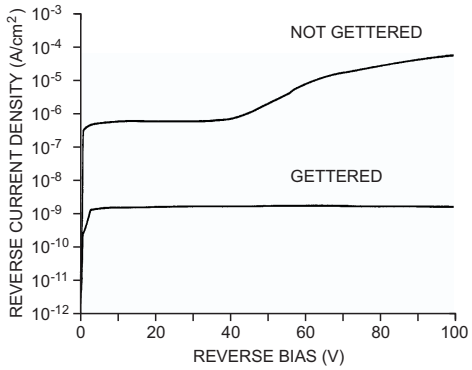

FIG. A.2. Comparison between two diodes fabricated together on the same wafer, but one with gettering and one without. (Data courtesy of S.E. Holland.)

implants were only partially annealed. This low-temperature process was instrumental in providing detectors for the first generation of silicon vertex detectors. Subsequently, many variations of the planar process have been developed. One example is the process developed by Holland (1989a, 1989b). This process differs from others by setting compatibility with mainstream CMOS fabrication as a key criterion. It explicitly utilizes gettering to remove deleterious impurities from the active region and exploits high temperature process steps to improve device performance. For a discussion of gettering techniques see Wolf (2002). The process was used to monolithically integrate high-quality detectors with low-noise electronics (Holland and Spieler 1990) and was extended to a full CMOS implementation (Holland 1992). It also formed the basis of monolithic random-access pixel detectors (Snoeys *et al.* 1992) and the full-depletion CCDs (Holland 2003) described in Chapter 8. Room-temperature reverse bias currents of 100 pA/cm^2 have been achieved in production.

Gettering is a key ingredient in most semiconductor fabrication processes, although it is not always explicit. Disordered material tends to capture mobile contaminants, so polysilicon is an effective getterer (in early devices abrading the back surface was a common technique). Chemical gettering is the second mechanism. Phosphorus is good gettering agent for deleterious impurities in silicon and phosphorus-doped polysilicon is a very effective getterer. Figure A.2 shows the reverse bias current for two diodes, both fabricated simultaneously on the same wafer, but one without and one with a doped polysilicon layer. The flat current *vs.* voltage dependence of the gettered devices indicates the low concentration of deleterious impurities in the active region.

A.7 Detector process flow

The following is an abbreviated process description for silicon strip detectors (courtesy of S.E. Holland). It lists numerous process steps to indicate the com-

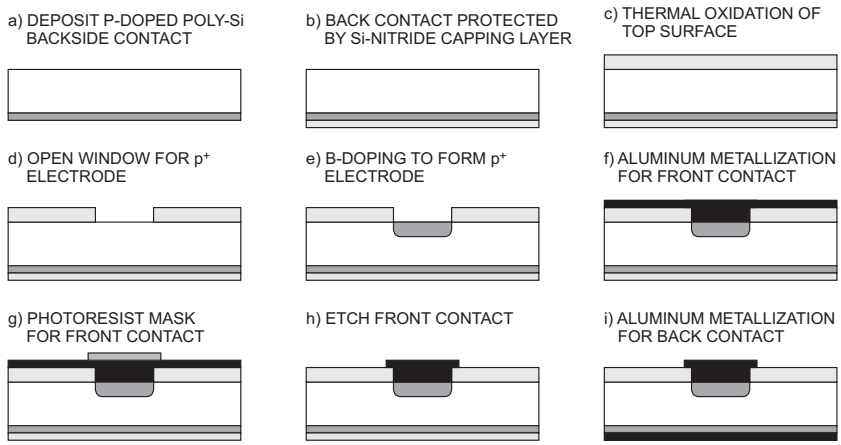

FIG. A.3. Some key steps in a detector fabrication process. See text for details.

plexity involved in fabricating rather simple devices. Intermediate steps are shown in Figure A.3. Various process ingredients interact in complex ways, notably doping and thermal cycles. Detailed process simulators are available to evaluate the cumulative effect of all process steps on the final device.

1. First the entire wafer is coated with a doped polysilicon layer, which is then removed from the front of the wafer. On the backside the doped polysilicon remains as both a gettering layer and the ohmic electrode of the diode.
   (a) Deposition of $\sim 0.5\,\mu$m low-temperature oxide (LTO) ~ 1 hr at 400 °C.
   (b) Apply photoresist to wafer frontside.
   (c) Etch backside oxide. This leaves a clean silicon surface.
   (d) Deposit $\sim 1.2\,\mu$m phosphorus-doped polysilicon (Figure A.3a).
   (e) Deposit silicon nitride to protect wafer backside. This acts as a backside capping layer during subsequent processing.
   (f) Dry etch frontside nitride.
   (g) Dry etch frontside polysilicon.
   (h) Wet etch LTO (Figure A.3b).
2. Grow thermal oxide (steam, 4 hrs at 900 °C to form a 400 nm thick oxide). This will ultimately remain as the interstrip passivation (Figure A.3c).
3. Deposit photoresist.
4. Expose photoresist through mask with strip pattern, develop.
5. Etch exposed oxide (Figure A.3d).
6. Introduce p-dopant, typically boron, to form strip electrodes (Figure A.3e). Two methods are possible:
   (a) thermal diffusion: expose to B_20_3 source for 30 min at 900 °C.
   (b) ion implantation: 30 keV B ions at a dose of $2 \cdot 10^{15}\,\mathrm{cm}^2$.

7. Drive-in and thermal anneal of implant, combined with oxidation: 40 min at 900 °C in steam followed by 80 min at 900 °C in N_2.
8. Deposit 500 nm aluminum-silicon alloy for contacts (Figure A.3f).
9. Spin on photoresist.
10. Expose through a mask to form strip metallization and bonding pads and then develop resist (Figure A.3g).
11. Etch metal (Figure A.3h).
12. Coat front side with photoresist.
13. Etch away backside nitride capping layer.
14. Deposit 100 nm aluminum on back side to form ohmic contact (Figure A.3i).
15. Forming gas anneal to reduce density of interface states at Si–SiO_2 interface: 20 min at 400 °C in 80% H_2 + 20% N_2.

Additional intermediate steps are necessary: Wafers must be cleaned prior to each furnace step (except the post-metallization anneal). "DI water" refers to high purity water with sub-ppb contamination levels. With regard to contamination water is one of the most critical process ingredients, as wafers are exposed to many more water molecules than air. An "HF etch" immerses the wafer in diluted hydrofluoric acid to strip oxide off of the wafer surface.

1. Immerse in 5:1:1 solution of $H_2O : H_2S0_4 : H_20_2$ at 120 °C.
2. HF etch.
3. Rinse in DI water.
4. Immerse in 5:1:1 solution of $H_2O : NH_4OH : H_20_2$ at 65 °C.
5. Dilute HF etch.
6. Rinse in DI water.
7. Immerse in 5:1:1 solution of $H_2O : HCl : H_20_2$ at 65°C.
8. Rinse in DI water.

The overall process takes several weeks. Processing typically proceeds in clean rooms of class 100 or better, *i.e.* with fewer than 100 particles $> 0.5\,\mu$m in size per cubic foot. For comparison, hospital operating rooms are class $\sim 10^5$ and the "normal" air in a big city has $\sim 10^8$ particles per cubic foot. Wafers are typically processed in batches of 25 – 50. Handling of individual wafers is minimized by using wafer holders, for example to immerse multiple wafers at once in a bath. Semirobotic systems automate the application of photoresist and baking, which must be well-controlled at small feature sizes. Furnaces are computer controlled to ensure that temperature ramping and gas flows follow the required recipe. Dedicated furnace tubes are used and gas lines are purged before and after each process step to prevent cross-contamination.

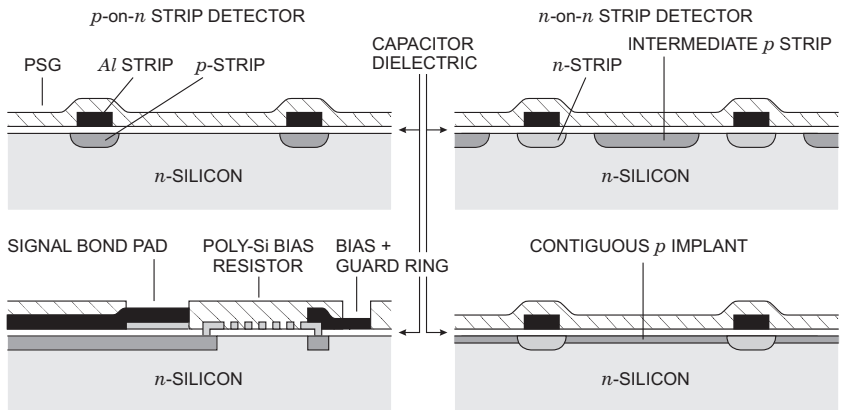

FIG. A.4. Cross-sections (vertical scale exaggerated) of the electrode structure of an AC coupled p-on-n strip detector (left) and an n-on-n detector right. The lower left shows schematically the arrangement of the signal bonding pad and the polysilicon bias resistor, typically implemented as a meander, so it appears as a sequence of conductors. The n-on-n panels show two isolation techniques, an intermediate p-strip (top) and a contiguous p-layer (right).

A.8 Strip detector structures

Figure A.4 (top left) shows a basic strip detector structure with integrated coupling capacitors, formed by the oxide layer between the doped strip electrode and the metal contact. Several considerations enter into the detailed geometry. The depletion voltage depends on the strip pitch and width. A higher voltage than in an extended plane geometry is required to "fill in" the volume adjacent to the surface between the strips. For example, Barberis *et al.* (1994) found that for a strip width of $9\,\mu$m the depletion voltage increased by 1.08, 1.19, and 1.39 for pitches of 30, 60, and $90\,\mu$m.

The choice of strip width is subject to several conflicting requirements. A small ratio of electrode width to pitch reduces the interstrip capacitance. However, when strip resistance is important, a wider strip is desired, as the maximum thickness of the contact metallization is $1 - 2\,\mu$m. Furthermore, studies of the electric field at the electrode and contact edges place constraints on the geometry. Ohsugi *et al.* (1994, 1996a, 1996b) have performed extensive studies and measured the location of "microdischarges" caused by local high fields using infrared imaging. Microdischarge occurs in the bulk at the strip edges. It is a reversible phenomenon and is not associated with dielectric breakdown. Under the condition that all electrode contours are rounded ($r > 10\,\mu$m) to limit the local electric field, three mechanisms affect the onset of microdischarge.

1. The reverse bias voltage required to deplete the detector. This increases with decreasing strip width and a steep increase in the microdischarge rate

is observed for ratios of strip width to pitch $w/p \geq 0.2$ (Ohsugi *et al.* 1999). A deep diffusion is beneficial as it increases the radius of curvature.

2. The potential across the integrated coupling capacitor. When operated with the full bias voltage across the capacitor this is the dominant field around the strip, so the total field is highest at the junction side. However, this is also significant at the ohmic side when the detector is operated well beyond depletion. The field increases steeply when the contact width exceeds the implant width. The contact edge should be kept $2-3\,\mu$m inside the strip electrode.

3. Charge trapped in the dielectric oxide or at the oxide–silicon interface. This contribution to the field is exacerbated at the p-side, as on the n-side it is partially cancelled by the field across the capacitor.

Moving the high-field regions from the bulk into the dielectric is advantageous, as the breakdown field in the oxide is much greater than in the bulk ($10^3\,\mathrm{V}/\mu\mathrm{m}$ *vs.* $30\,\mathrm{V}/\mu\mathrm{m}$). This can be accomplished by thickening the oxide beyond the implant edges and extending the metallization to provide an overhang beyond the implant width (Ohsugi *et al.* 1996a, Passeri *et al.* 2000). Another limit to the maximum bias voltage is determined by the guard ring structure, which must ensure that the potential is appropriately graded to avoid high fields at the edge of the sensor. Multiple guard rings are often used to control the potential gradient (for an example see Andricek *et al.* 2000), although other techniques can be applied (Ohsugi *et al.* 1999).

The design of optimum structures for double-sided detectors is discussed by Ohsugi *et al.* (1999). Passeri *et al.* (2001) discuss the interstrip capacitance *vs.* metal overhang. Figure A.4 (left) shows a strip detector with integrated coupling capacitors. For simplicity, this does not include the metal overhang. Since the dielectric is deposited, it is subject to pinholes depending on the quality of the underlying surface. This limits the breakdown voltage and overall reliability of the capacitor. Multilayer dielectrics are effective against pinholes. Silicon-nitride has a higher dielectric constant and breakdown field, so oxide-nitride layers are commonly used (Ohsugi *et al.* 1996a). Holland (1995) discusses oxide-nitride-oxide layers and equivalent coupling circuits. Nevertheless, the integrated capacitors often develop defects, so some detectors "float" their readout electronics at the detector bias level to maintain minimal voltage across the capacitors. Their sole purpose then is to block the detector bias current from flowing into the amplifier input.

With AC coupling each strip requires a DC return path, which is best implemented with resistors, implemented as narrow strip meanders of lightly doped polysilicon. These are placed at the ends of the strips between the strip electrodes and the guard ring, which isolates the detector's active area from the wafer saw cut. A common bias connection feeds all strips. This is shown in the lower left cross-section of Figure A.4. Some detector designs utilize MOSFET structures to bias the individual strips (often called FOXFET biasing), but as

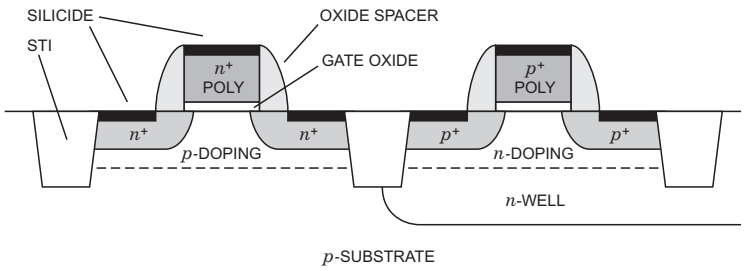

FIG. A.5. Cross-section of a CMOS structure, combining an NMOS transistor (left) and a PMOS transistor (right). An n-well isolates the PMOS device from the p-substrate. The gate electrodes are doped polysilicon and silicide layers provide connections to the gate, drain, and source. Device boundaries are set by shallow trench isolation (STI).

they operate in thermionic emission they introduce additional shot noise, which becomes noticeable after radiation damage. This was first observed in CDF (Azzi *et al.* 1996) and subsequent tests have buttressed these observations. The metal bond pad for the signal connection is placed on an intermediate polysilicon pad to reduce the risk of damage to the underlying silicon when wire bonding.

In double-sided detectors one set of strips is on the ohmic side, so together with the bulk the strip electrodes form n^+-n-n^+ structures. Positive oxide charge also forms an electron accumulation layer at the adjacent silicon surface. To break this conductive path some form of interstrip isolation is required. Several techniques have been applied (Weilhammer 1994), but intermediate p-implants are most commonly used. This is shown schematically in Figure A.4. The interstrip capacitance is higher than on the p-side, as the width of the n-strips is effectively increased by the electron accumulation layer in the gap between the strip and the isolation implant. Wider p-implants yield lower interstrip capacitance, as the implant charge is fixed. The ATLAS SCT and pixel subsystems utilize "p-spray" isolation, which utilizes a contiguous p-implant with carefully controlled concentration to limit the local electric fields (Richter *et al.* 1996). With this technique electric fields at the electrode edges decrease with ionizing radiation dose, whereas with discrete p-stops they increase. Prior to irradiation, however, the p-spray results in higher fields than with an intermediate p-strip, so for applications without severe radiation damage the latter may be preferable. More complex structures than shown in A.4 have been developed to reduce the interstrip capacitance (Hopman *et al.* 1996, Iwata *et al.* 1998).

A.9 CMOS devices

Figure A.5 shows a typical CMOS structure. Descriptions of the process flows can be found in numerous texts, for example by Taur and Ning (1998), Baker, Li, and Boyce (1998), and Wolf (1995, 2002). In Figure A.5 a p-substrate is

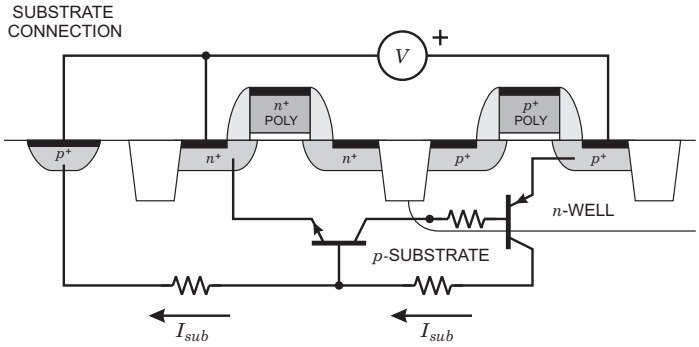

FIG. A.6. Parasitic bipolar transistors in a CMOS structure can be turned on by substrate currents and "latch up" in a stable state.

used, so the PMOS device must be embedded in an n-well. Channel implants set the threshold voltage. Isolation between devices in modern devices is provided by shallow trench oxide isolation, which is formed by etching a trench into the silicon, oxidizing the surface to control the silicon surface, and then filling the trench with insulating material (see Wolf 2002).

The gate electrodes are polysilicon, p^+ doped in PMOS and n^+ doped in NMOS devices. Silicide (a mixture of a refractory metal and polysilicon) layers on the source, gate, and drain electrodes provide a low-resistance path that avoids voltages drops due to the distributed current flow in the source and drain regions. $TiSi_2$ yields roughly equal barrier heights on both n^+ and p^+ material. The gate structure is protected by a sidewall of deposited oxide or nitride. In an integrated circuit many additional layers are formed on top of the basic transistors to provide interconnects, insulation, and planarization. Rather substantial differences in height accrue, so surfaces must be smoothed for good step coverage of traces. As described for detectors, a passivation of PSG or other impermeable material covers the whole chip, with openings for bond and probe pads.

Bulk CMOS structures also form parasitic bipolar transistors, as shown in Figure A.6. In many circuit configurations these form positive feedback structures that under high transient current conditions switch to an undesired state that can be destructive or only be reset by removing power (latch up). This can be ameliorated by both device design (Troutman 1986) and layout (see Baker, Li, and Boyce 1998, for an example).

References

Andricek, L. *et al.* (2000). Design and test of radiation hard p^+n silicon strip detectors for the ATLAS SCT. *Nucl. Instr. and Meth.* **A439** 427–441

Azzi, P. *et al.* (1996). Radiation damage experience at CDF with SVX′. *Nucl. Instr. and Meth.* **A383** 155–158.

Baker, R.J., Li, H.W., and Boyce, D.E. (1998). *CMOS Circuit Design, Layout, and Simulation*. IEEE Press, New York. ISBN 0-7803-3416-7, TK7871.99.M44B35 1997

Fair, R.B. (1981). Concentration profiles of diffused dopants in silicon. in *Impurity Doping Processes in Silicon*, F.F.F. Wang (ed). North Holland, Amsterdam. ISBN 0-444-86095-9, TK7871.85.I48

Holland, S.E. (1989a). An IC-compatible detector process *IEEE Trans. Nucl. Sci.* **NS-36/1** (1989) 283–288

Holland, S.E. (1989b). Fabrication of detectors and transistors on high-resistivity silicon. *Nucl. Instr. and Meth.* **A275** (1989) 537–541

Holland, S. and Spieler, H. (1990). A monolithically integrated detector–preamplifier on high-resistivity silicon. *IEEE Trans. Nucl. Sci.* **NS-37/2** (1990) 463–468

Holland, S. (1992). Properties of CMOS devices and circuits fabricated on high-resistivity, detector-grade silicon.*IEEE Trans. Nucl. Sci.* **NS-39/4** (1992) 809–813

Holland, S. (1995). An oxide-nitride-oxide capacitor dielectric film for silicon strip detectors. *IEEE Trans. Nucl. Sci.* **NS-42/4** 423–427

Holland, S.E. *et al.* (2003). Fully depleted, back-illuminated charge-coupled devices fabricated on high-resistivity silicon. *IEEE Trans. Electron. Dev.* **ED-50/1** (2003) 225–238 and LBNL-49992

Hopman, P.I. *et al.* (1996). Optimization of silicon microstrip detector design for CLEO III. *Nucl. Instr. and Meth.* **A383** (1996) 98–103

Iwata, Y. *et al.* (1998). Optimal p-stop pattern for the n-side strip isolation of silicon microstrip detectors. *IEEE Trans. Nucl. Sci.* **NS-45/3** (1998) 303–309

Kemmer, J. (1980). Fabrication of low noise silicon radiation detectors by the planar process. *Nucl. Instr. and Meth.* **169** (1980) 499–502

Kemmer, J. (1984). Improvement of detector fabrication by the planar process. Nucl. Instr. and Meth. **226** (1984) 89–93

Lutz, G. (1999). *Semiconductor Radiation Detectors*. Springer Verlag, Berlin, 1999. ISBN 3-5406-4859-3

Nishi, Y. and Doering, R. (eds) (2000). *Handbook of Semiconductor Manufacturing Technology*. Marcel Dekker, New York, 2000. ISBN 0-8247-8783-8

Ohsugi, T. *et al.* (1994). Microdischarges of AC-coupled silicon strip sensors. *Nucl. Instr. and Meth.* **A342** (1994) 22–26

Ohsugi, T. *et al.* (1996a). Micro-discharge at strip edge of silicon microstrip sensors *Nucl. Instr. and Meth.* **A383** (1996) 116–122

Ohsugi, T. *et al.* (1996b). Micro-discharge noise and radiation damage of silicon microstrip sensors *Nucl. Instr. and Meth.* **A383** (1996) 166–173

Ohsugi, T. *et al.* (1999). Design optimization of radiation-hard, double-sided, double-metal, AC-coupled silicon sensors. *Nucl. Instr. and Meth.* **A436** (1999) 272–280

Passeri, D. *et al.* (2000). Optimization of overhanging-metal microstrip detectors: test and simulation *IEEE Trans. Nucl. Sci.* **NS-48/3** (2001) 249–253

Passeri, D. *et al.* (2001). Physical modeling of silicon microstrip detectors: influence of the electrode geometry on critical electric fields. *IEEE Trans. Nucl. Sci.* **NS-47/4** (2000) 1468–1473

Richter, R.H. *et al.* (1996). Strip detector design for ATLAS and HERA-B using two-dimensional device simulation. *Nucl. Instr. and Meth.* **A377** (1996) 412–421

Ryssel, H. and Glawischnig, H. (eds) (1982). *Ion Implantation Techniques.* Springer Verlag, Berlin, 1982. ISBN 3-540-11878-0

Sze, S.M. (1981). *Physics of Semiconductor Devices.* Wiley, New York 1981. ISBN 0-471-05661-8, TK7871.85.S988 1981

Snoeys, W. *et al.* (1992). A new integrated pixel detector for high energy physics *IEEE Trans. Nucl. Sci.* **NS-39/5** (1992) 1263–1269

Taur, Y. and Ning, T.H. (1998). *Fundamentals of Modern VLSI Devices.* Cambridge University Press, Cambridge. ISBN 0-521-55056-6, TK7871.99.M44T38

Troutman, R.R. (1986). *Latch-up in CMOS Technology.* Kluwer, Boston

Weilhammer, P. (1994). Double-sided Si strip sensors for LEP vertex detectors. *Nucl. Instr. and Meth.* **A342** (1994) 1–15

Wolf, S. (1995). *Silicon Processing for the VLSI Era, Volume 3 – The Submicron MOSFET.* Lattice Press, Sunset Beach, ISBN 0-961672-5-3

Wolf, S. (2002). *Silicon Processing for the VLSI Era, Volume 4 – Deep-Submicron Process Technology.* Lattice Press, Sunset Beach, ISBN 0-9616721-7-X

APPENDIX B

PHASORS AND COMPLEX ALGEBRA IN ELECTRICAL CIRCUITS

Consider the RLC circuit shown in Figure B.1. The total voltage developed across the circuit

$$\begin{aligned} V &= V_R + V_L + V_C \\ V &= IR + L\frac{dI}{dt} + \frac{Q}{C} \\ \frac{dV}{dt} &= \frac{dI}{dt}R + L\frac{d^2I}{dt^2} + \frac{I}{C} \ . \end{aligned} \tag{B.1}$$

The desired result can be found by solving the differential equation, but since similar problems occur repeatedly, simpler techniques have been developed. If we write $V(t)$ and $I(t)$ in the form

$$\begin{aligned} V(t) &= V_0 e^{\mathbf{i}\omega t} \\ I(t) &= I_0 e^{\mathbf{i}(\omega t + \varphi)} \ , \end{aligned} \tag{B.2}$$

then eqn B.1 becomes

$$\begin{aligned} \mathbf{i}\omega V_0 e^{\mathbf{i}\omega t} &= \mathbf{i}\omega R I_0 e^{\mathbf{i}(\omega t - \varphi)} - \omega^2 L I_0 e^{\mathbf{i}(\omega t - \varphi)} + \frac{1}{C} I_0 e^{\mathbf{i}(\omega t - \varphi)} \\ \frac{V_0}{I_0} e^{\mathbf{i}\varphi} &= R + \mathbf{i}\omega L - \mathbf{i}\frac{1}{\omega C} \ . \end{aligned} \tag{B.3}$$

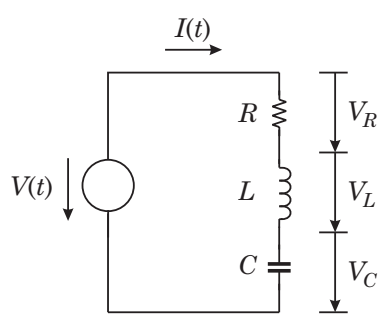

FIG. B.1. A resistor, inductor, and capacitor connected in series and driven by a voltage source.

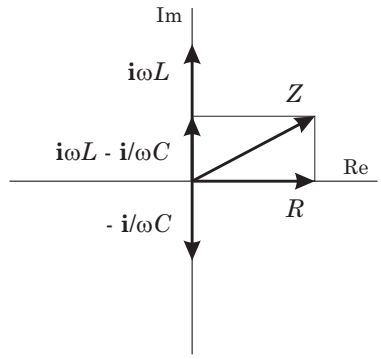

FIG. B.2. The series RLC circuit is represented in the complex plane by "phasors" that characterize the magnitude and phase of the impedance.

Thus, we can express the total impedance $Z \equiv (V_0/I_0)e^{\mathbf{i}\varphi}$ of the circuit as a complex number with the magnitude $|Z| = V_0/I_0$ and phase φ. In this representation the equivalent resistances (reactances) of L and C are imaginary numbers.

Plotted in the complex plane (Figure B.2) the components on the right side of eqn B.3 are represented as "phasors", whose magnitude and phase characterize the impedance. Relative to V_R, the voltage across the inductor V_L is shifted in phase by $+90°$. The voltage across the capacitor V_C is shifted in phase by $-90°$.

The total impedance has the magnitude

$$|Z| = \sqrt{[\mathrm{Re}(Z)]^2 + [\mathrm{Im}(Z)]^2} = \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2} \tag{B.4}$$

and phase

$$\tan\varphi = \frac{\mathrm{Im}(Z)}{\mathrm{Re}(Z)} = \frac{\omega L - \dfrac{1}{\omega C}}{R} \ . \tag{B.5}$$

From eqn B.4 one sees immediately that the impedance Z assumes a minimum at

$$\omega = \frac{1}{\sqrt{LC}} \ , \tag{B.6}$$

the resonant frequency of the tuned circuit. The impedance *vs.* frequency yields the resonance curve. At resonance the phase φ becomes zero.

At frequencies above resonance the inductive reactance dominates (as apparent in Figure B.2) and the asymptotic phase is $+90°$. Below resonance the capacitive reactance dominates and the asymptotic phase is $-90°$. This representation can be used for any element that introduces a phase shift, *e.g.* an amplifier. A phase shift of $+90°$ appears as $+\mathbf{i}$, $-90°$ as $-\mathbf{i}$.

APPENDIX C

EQUIVALENT CIRCUITS

Equivalent circuits are a valuable tool in analyzing circuit behavior. The purpose is to strip the actual circuit of all elements not relevant to the specific question being addressed. Removing unnecessary elements helps to bring out the salient features. However, this also means that a given circuit can have several equivalent circuits, depending on the purpose of the analysis.

Figure C.1 shows a simple voltage amplifier. Its basic features are that it provides gain with a high input impedance and low output impedance. Whenever only these characteristics are relevant, the detailed circuit is commonly represented by the symbol at the right of Figure C.1. Note that the common connection shared by the input and output is not shown, but it is implicit to the symbol.

The primary purpose of an amplifier is to provide gain, but this is not restricted to voltage input and voltage output. For example, an amplifier can be driven by an input current and provide an output voltage, as in a charge-sensitive preamplifier. Any combination of current or voltage input or output is possible, as shown in Table C.1.

Of course, when analyzing the details of circuit performance, the role of individual components does have to be considered, but this too can be simplified. As a starting point take the circuit in Figure C.2.

First, just consider the DC operating point of the circuitry between C_1 and C_2. The transconductance of the MOSFET depends on the standing current flowing through the device. The signal is superimposed on this as a differential change. Thus, we distinguish two voltages at the input, a constant bias voltage that sets the operating point and a small signal voltage v_s superimposed on

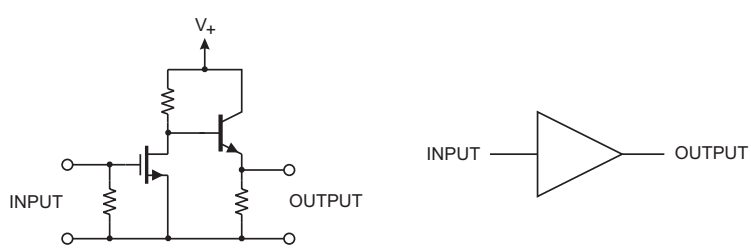

FIG. C.1. An amplifier – regardless of its complexity – is represented by a simple symbol (right) when only its role as a gain stage is relevant.

Table C.1 Amplifier types.

Input	Output	Gain	Type
v_i	v_o	$A_v = v_o/v_i$	Voltage
i_i	i_o	$A_i = i_o/i_i$	Current
v_i	i_o	$A_g = i_o/v_i$	Transconductance
i_i	v_o	$A_r = v_i/i_o$	Transresistance

it. Input divider $R_1 - R_2$ sets the required gate voltage, which for an n-type MOSFET is positive,

$$V_{GS} = \frac{R_2}{R_1 + R_2} V_B \ . \tag{C.1}$$

Assume a drain current I_D. Then the quiescent voltage (*i.e.* absent an input signal) at the drain is

$$V_{DS} = V_B - I_D R_3 \ . \tag{C.2}$$

Next, consider the time-varying signal v_S introduced by the signal source. The signal at the gate G is dV_{GS}/dt. The current flowing through R_2 is

$$\frac{dI_{R_2}}{dt} = \frac{dV_{GS}}{dt} \cdot \frac{1}{R_2} \tag{C.3}$$

and the current flowing through R_1

$$\frac{dI_{R_1}}{dt} = \frac{1}{R_1} \cdot \frac{d}{dt}(V_B - V_{GS}) \ . \tag{C.4}$$

The battery voltage V_B is constant, $dV_B/dt = 0$, so

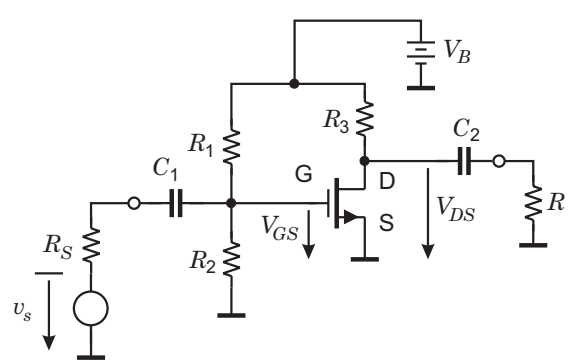

FIG. C.2. A simple MOSFET amplifier showing all bias components.

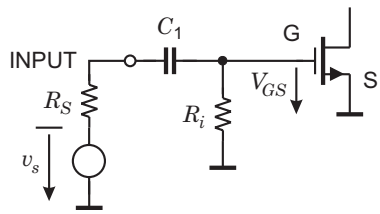

FIG. C.3. Equivalent input circuit.

$$\frac{dI_{R_1}}{dt} = \frac{1}{R_1} \cdot \frac{dV_G V_{GS}}{dt} \ . \tag{C.5}$$

The total time-dependent input current is

$$\frac{dI}{dt} = \frac{dI_{R_1}}{dt} + \frac{dI_{R_2}}{dt} = \left(\frac{1}{R_1} + \frac{1}{R_2}\right) \cdot \frac{dV_{GS}}{dt} \equiv \frac{1}{R_i} \cdot \frac{dV_{GS}}{dt} \ , \tag{C.6}$$

where

$$R_i = \frac{R_1 \cdot R_2}{R_1 + R_2} \tag{C.7}$$

is the parallel connection of R_1 and R_2. Consequently, for the time-varying input signal the circuit behaves like the equivalent circuit in Figure C.3.

At the output, the voltage signal is formed by the current of the transistor flowing through the combined output load formed by R and R_3. For the moment, assume that $R \gg R_3$. Then the output load is dominated by R_3. Referring back to Figure C.2, the quiescent voltage at the drain is

$$V_{DS} = V_B - I_D R_3 \ . \tag{C.8}$$

When the gate voltage is varied, the transistor drain current changes, with a corresponding change in output voltage

$$\frac{dV_{DS}}{dI_D} = \frac{d}{dI_D}(V_B - I_D R_3) = -R_3 \ . \tag{C.9}$$

The constant supply voltage V_B does not directly affect the magnitude of the output signal and the output voltage is 180° out of phase with the drain current.

If we remove the restriction $R \gg R_3$, the total load impedance for time-variant signals is the parallel connection of R_3 and $X_{C_2} + R$, yielding the equivalent output circuit shown in Figure C.4. A similar circuit is present at the input. The coupling capacitor C_1 together with the source resistance R_S and the input resistance R_i form a high-pass filter with the corner frequency

$$f_c = \frac{1}{2\pi(R_S + R_i)C_1} \ . \tag{C.10}$$

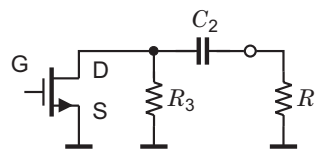

FIG. C.4. Equivalent output circuit.

If the signal source resistance $R_S \ll R_i$, as desired for efficient voltage drive, the source signal v_s will suffer negligible attenuation at frequencies

$$f \gg \frac{1}{2\pi R_i C_1} \ , \tag{C.11}$$

so the coupling capacitor C_1 can be neglected. Correspondingly, at the output, if the impedance of the output coupling capacitor $\omega C_2 \ll R$, the signal across R is the same as across R_3. Thus, at high frequencies circuit behavior is described by the equivalent circuit in Figure C.5, where R_L is the parallel combination of R_3 and R. With knowledge of the MOSFET's transconductance $g_m = dI_D/dV_{GS}$ this circuit allows a good prediction of the amplifier's voltage gain $A_v = g_m R_L$. The accuracy of this circuit is adequate for many applications, but at high gains and high frequencies the equivalent circuit of the MOSFET must be included. As shown in Figure C.5 the MOSFET's output resistance shunts R_L and at high frequencies the channel resistance couples significantly to the input through the gate–channel capacitance, so the FET input does not appear purely capacitive.

As already noted in the introductory comments, equivalent circuits are an invaluable tool in analyzing systems, as they remove extraneous components and show only the components and parameters essential for the problem at hand. Equivalent circuits are often tailored to very specific questions and include simplifications that are not generally valid. However, focussing on a specific question with a restricted model may be the only way to analyze a complicated situation.

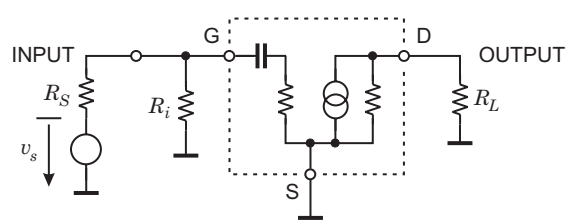

FIG. C.5. Equivalent circuit of an amplifier including the equivalent circuit of the MOSFET.

APPENDIX D

FEEDBACK AMPLIFIERS

The most basic amplifier is a gain device itself, *e.g.* a bipolar transistor, JFET, or MOSFET. None of these is perfectly linear, so any amplifier utilizing these devices will also exhibit nonlinearity. Furthermore, the overall gain depends on device parameters that can vary among nominally identical devices.

Feedback amplifiers can provide predictable gain and significantly improve linearity by making the gain dependent primarily on linear components, *i.e.* resistors or capacitors, whose values are independent of signal amplitude.

D.1 Gain of a feedback amplifier

Figure D.1 shows the principle. A portion of the output is fed back to a summing circuit at the input. The net voltage applied to the amplifier input

$$v_{ia} = v_i + v_{fb} \ . \tag{D.1}$$

The feedback signal

$$v_{fb} = A_{fb} v_o \ . \tag{D.2}$$

Assume an inverting amplifier, so the ratio of output to input voltage

$$\frac{v_o}{v_{ia}} = -A_v \ . \tag{D.3}$$

This relationship always applies, whether feedback is present or not. Thus, the voltage at the amplifier input

$$v_{ia} = -\frac{v_o}{A_v} \ . \tag{D.4}$$

Inserting eqns D.2 and D.4 into eqn D.1 yields

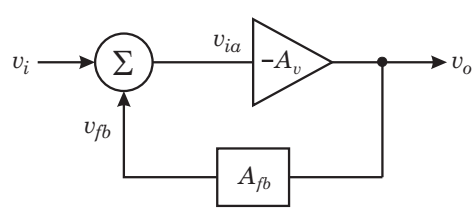

FIG. D.1. In a feedback system a portion of the output signal is fed back to a summing junction at the input.

$$v_o = -\frac{v_i}{\dfrac{1}{A_v} + A_{fb}} \tag{D.5}$$

and the closed loop gain, *i.e.* the gain with feedback

$$A_{CL} = \frac{v_o}{v_i} = -\frac{1}{\dfrac{1}{A_v} + A_{fb}} = -\frac{A_v}{1 + A_{fb}A_v} \ . \tag{D.6}$$

When the amplifier gain (the "open loop gain") is sufficiently large that $A_v \gg 1/A_{fb}$, the gain of the overall system ("closed loop gain")

$$A_{CL} \equiv \frac{v_o}{v_i} \approx -\frac{1}{A_{fb}} \tag{D.7}$$

is independent of the open loop amplifier gain A_v and determined by the feedback network alone. If the feedback network is an attenuator ($A_{fb} < 1$), *e.g.* formed by resistors, the overall gain is set by the resistor values and is independent of the amplifier gain.

However, note that when the amplifier gain A_v is marginal it cannot be ignored. For example, when $A_v = 1/A_{fb}$ (*i.e.* the amplifier gain equals the nominal closed loop gain)

$$A_{CL} = -\frac{1}{2}\frac{1}{A_{fb}} \ , \tag{D.8}$$

only half the naively expected value.

D.2 Linearity

Feedback linearizes the amplifier response. For a small deviation ΔA_v in open loop gain

$$\frac{A_v + \Delta A_v}{1 + (A_v + \Delta A_v)A_{fb}} \approx \frac{A_v + \Delta A_v}{1 + A_{fb}A_v} = A_{CL} + \frac{\Delta A_v}{1 + A_{fb}A_v} \ , \tag{D.9}$$

where A_{CL} is the nominal closed loop gain. Thus, deviations from linearity are reduced by the factor

$$\frac{1}{1 + A_{fb}A_v} \ . \tag{D.10}$$

D.3 Bandwidth

Similarly, the bandwidth is also improved. The frequency-dependent gain of a single-pole amplifier with an upper cutoff frequency f_u is

$$A_v(f) = \frac{A_{v0}}{1 + \mathbf{i}(f/f_u)} \ . \tag{D.11}$$

Inserting this into eqn D.6 yields

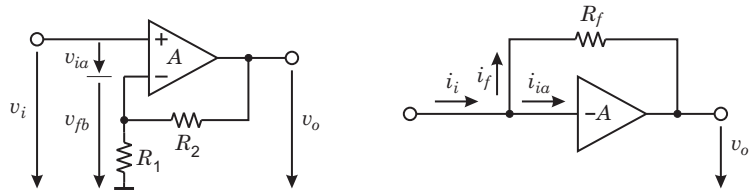

FIG. D.2. Feedback can be applied as a voltage in series with the input signal (left) – series feedback – or in parallel as a current (right) – shunt feedback.

$$A_{CL} = \frac{A_v(f)}{1 + A_v(f)A_{fb}} = \frac{\dfrac{A_{v0}}{1 + \mathbf{i}(f/f_u)}}{1 + \dfrac{A_{v0}}{1 + \mathbf{i}(f/f_u)}} = \frac{A_{v0}}{1 + A_{fb}A_{v0} + \mathbf{i}(f/f_u)} \ . \qquad \text{(D.12)}$$

By dividing the numerator and denominator by $(1 + A_{fb}A_{v0})$ this result can be rewritten as

$$A_{vf} = \frac{A_{v0f}}{1 + \mathbf{i}(f/f_{uf})} \ , \qquad \text{(D.13)}$$

where

$$A_{v0f} \equiv \frac{A_{v0}}{1 + A_{fb}A_{v0}} \quad \text{and} \quad f_{uf} \equiv f_u(1 + A_{fb}A_{v0}) \ . \qquad \text{(D.14)}$$

The gain and phase response of the fedback amplifier are of the same form as for an open loop single-pole amplifier. The midband gain is as predicted by eqn D.6 and the frequency response is extended by the factor $(1 + A_{fb}A_{v0})$. The gain–bandwidth product and unity gain frequency remain unchanged whether or not feedback is applied. By a similar calculation, a lower cutoff frequency is reduced by the factor $1/(1 + A_{fb}A_{v0})$.

D.4 Series and shunt feedback

Feedback can be applied as either voltage or current, as illustrated in Figure D.2. In series feedback the feedback signal is applied as a voltage in series with the input. The feedback signal is commonly applied to the inverting input of a differential amplifier input. Shunt feedback adds currents, so the feedback signal connects directly to the input to form a current summing node.

D.5 Input and output impedance

Although for simplicity in analysis we often use amplifier models with infinite input resistance (*i.e.* no current ever flows into the amplifier input), in reality all amplifying devices have a finite input impedance (impedance because the input generally does not appear purely resistive). A MOSFET input, for example, appears capacitive at low frequencies.

D.5.1 *Series feedback*

Consider the configuration in the left panel of Figure D.2. Without feedback the input current $i_i = v_i/Z_{i0}$, where Z_{i0} is the input impedance of the amplifier without feedback. With feedback the voltage applied to the amplifier input is reduced

$$v_{ia} = v_i - v_{fb} \ . \tag{D.15}$$

Thus, the input current

$$i_i = \frac{v_i - v_{fb}}{Z_{i0}} \tag{D.16}$$

is reduced relative to the open loop configuration and the input impedance becomes

$$Z_{if} = \frac{v_i}{i_i} = Z_{i0}(1 + A_{fb}A) \ , \tag{D.17}$$

where

$$A_{fb} = \frac{R_1}{R_1 + R_2} \ . \tag{D.18}$$

Series (voltage) feedback increases the input impedance.

D.5.2 *Shunt feedback*

Again consider an inverting amplifier with an infinite input impedance and a voltage gain A_v, but now the output is fed back directly to the input through an impedance Z_f, as shown in the right panel of Figure D.2. Since the amplifier has an infinite input impedance, any input current i_i must flow through the feedback impedance Z_f. Thus, the input current

$$i_i = \frac{v_i - v_o}{Z_f} \ . \tag{D.19}$$

The output voltage

$$v_o = -A_v v_i \ , \tag{D.20}$$

so the input current

$$i_i = \frac{v_i(1 + A_v)}{Z_f} \ . \tag{D.21}$$

The input impedance

$$Z_{if} = \frac{Z_f}{1 + A_v} \tag{D.22}$$

and for large gains $A_v \gg 1$

$$Z_{if} \approx \frac{Z_f}{A_v} \ . \tag{D.23}$$

Shunt negative feedback reduces the input impedance. Thus, a large amplifier gain can yield a small input impedance, depending on the feedback impedance.

This is the mechanism that leads to the notion of "virtual ground". For example, an amplifier with a gain of 10^5 and a feedback resistor of $10^4\,\Omega$ yields

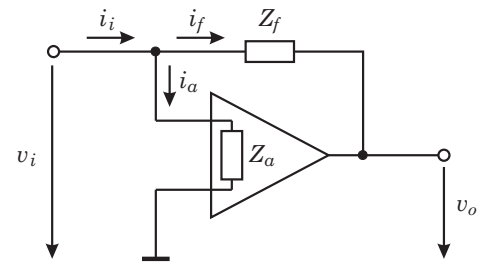

FIG. D.3. In a shunt feedback amplifier the input impedance of the non-fedback amplifier Z_a is in parallel with the active input impedance.

an input resistance is $0.1\,\Omega$. This is quite practical at low frequencies. However, at high frequencies the amplifier gain may be only 100, and the input resistance is $100\,\Omega$, which may qualify as a low impedance, but not a virtual ground.

It is straightforward to extend the result to an amplifier with a finite input impedance Z_a. As indicated in Figure D.3 the input current splits into two components, each corresponding to a component of the input impedance,

$$i_i = i_f + i_a = \frac{v_i}{Z_{if}} + \frac{v_i}{Z_a} \equiv \frac{v_i}{Z_i} \ . \tag{D.24}$$

The total input impedance is the parallel combination of the feedback and amplifier input impedance

$$\frac{1}{Z_i} = \frac{1}{Z_{if}} + \frac{1}{Z_a} \ . \tag{D.25}$$

D.5.3 *Output impedance*

The same reasoning can be applied to the output impedance. In this case shunt feedback takes the voltage directly from the output, as in Figure D.2. Any reduction in the output signal with increasing load current due to a finite output

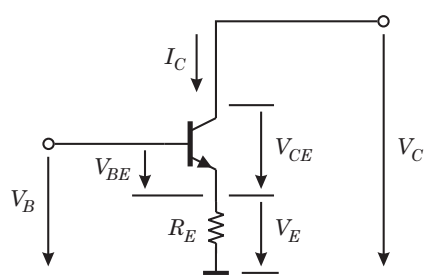

FIG. D.4. An emitter resistor R_E introduces local series feedback into a common emitter stage. The same principle applies to other devices.

impedance will also reduce the magnitude of the fedback signal, compensating for the output drop. This effectively reduces the output impedance to

$$Z_{of} = \frac{Z_o}{1 + A_{fb}A_v} \ . \tag{D.26}$$

Conversely, series feedback will increase the output impedance. Figure D.4 shows an example of local series feedback. The emitter resistor introduces a voltage drop into the input circuit that depends on the output current. If the output voltage sags because of the inherent output resistance of the device, the current decreases, which in turn reduces the voltage drop across the emitter resistor and increases the base–emitter voltage to counteract the decrease in output current.

For a constant input voltage V_B applied to the base, the collector voltage

$$V_C = I_C R_E + V_{CE}(V_{BE}, I_C) \ . \tag{D.27}$$

The total differential

$$dV_C = dI_C R_E + dV_{BE}\frac{dV_{CE}}{dV_{BE}} + dI_C\frac{dV_{CE}}{dI_C} \ . \tag{D.28}$$

Since $dV_{BE} = -dI_C R_E$ and the output resistance of the transistor $r_o = dV_{CE}/dI_C$, the change in output voltage

$$dV_C = dI_C R_E - dI_C R_E\frac{dV_{CE}}{dV_{BE}} + dI_C r_o \ . \tag{D.29}$$

Using $dI_C/dV_{BE} = g_m$ and $dV_{CE} = -dI_C R_E$, the output resistance with feedback becomes

$$r_{of} \equiv \frac{dV_C}{dI_C} = r_o + R_E(1 + g_m R_E) \ . \tag{D.30}$$

The source resistor also introduces local series feedback into the input, increasing the input resistance as discussed for the emitter follower in Chapter 6.

D.6 Loop gain

The quantity $A_{fb}A_v$ that improves linearity, extends bandwidth, and affects the input or output impedance is called the loop gain. It can be measured by breaking the feedback loop at any convenient point and measuring the total gain between the break, as illustrated in Figure D.5. The original input signal source should remain connected, but with zero signal.

Since the benefits that accrue from negative feedback depend on the loop gain, as the loop gain decreases with $A_v(f)$ beyond the either the lower or upper corner frequency, the effects of negative feedback decrease. For small loop gains the overall response follows the open loop response.

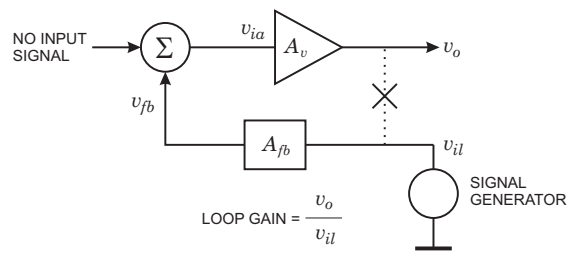

FIG. D.5. The loop gain can be determined by splitting the feedback loop, injecting a signal v_{il}, and measuring the output v_o. The loop can be split at any convenient point.

D.7 Stability

The amount of feedback is limited by stability considerations. In reality, amplifiers have multiple corner frequencies, so eventually the additional phase shift attains 180° and negative feedback turns into positive feedback, leading to self-oscillation. Figure D.6 illustrates the frequency response of an amplifier in open loop and closed loop operation. Two values of closed loop gain are shown. For A_{CL1} the corner frequency is in the regime where the open loop phase shift is 90°, so the total phase shift of the inverting amplifier is $180° + 90° = 270°$. At frequencies beyond the second corner frequency the additional phase shift is 180°, so the amplifier is noninverting and potentially unstable if too much feedback is applied. The stability criterion is that the loop gain may not exceed unity at the frequency where the additional phase shift is 180°. However, to ensure a safety margin, a phase margin of 45°, *i.e.* an additional phase shift of 135° is generally accepted to be the minimum. This is shown for closed loop gain A_{CL2}.

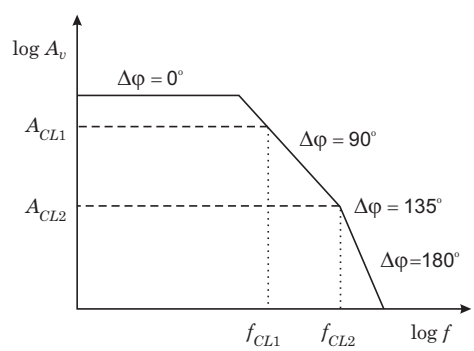

FIG. D.6. Open loop gain (solid line) and closed loop gains (dashed) for two values of closed loop gain A_{CL1} and A_{CL2}.

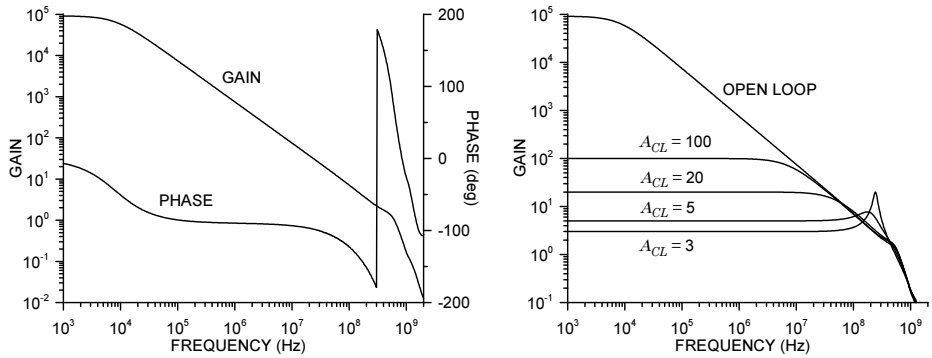

FIG. D.7. Open loop gain and phase *vs.* frequency (left) of a commercial operational amplifier. The phase wraps at values exceeding 180°. The right panel shows the gain *vs.* frequency for various closed loop gains with the open loop gain for comparison. Pronounced peaking for closed loop gains A_{CL} of 3 and 5 is due to reduced phase margin.

Figure D.7 shows the open loop gain and phase *vs.* frequency (left) of a commercial operational amplifier. Two corner frequencies with an extended 90° phase shift regime are apparent. The right panel shows the gain *vs.* frequency for various closed loop gains with the open loop gain for comparison. Beyond their respective corner frequencies the closed loop gain curves follow the open loop response. Reduced phase margin at closed loop gains of 3 and 5 results in pro-

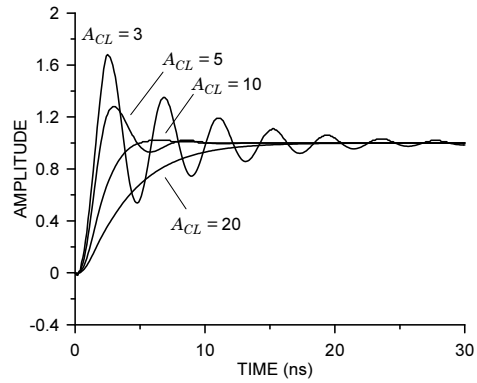

FIG. D.8. Pulse response of the amplifier for closed loop gains $A_{CL} = 3$, 5, 10, and 20. The peaking in the frequency response for $A_{CL} = 3$ and 5 translates into ringing in the pulse response. The output pulse for $A_{CL} = 10$ shows a very slight overshoot and the pulse for $A_{CL} = 20$ exhibits a fully monotonic response.

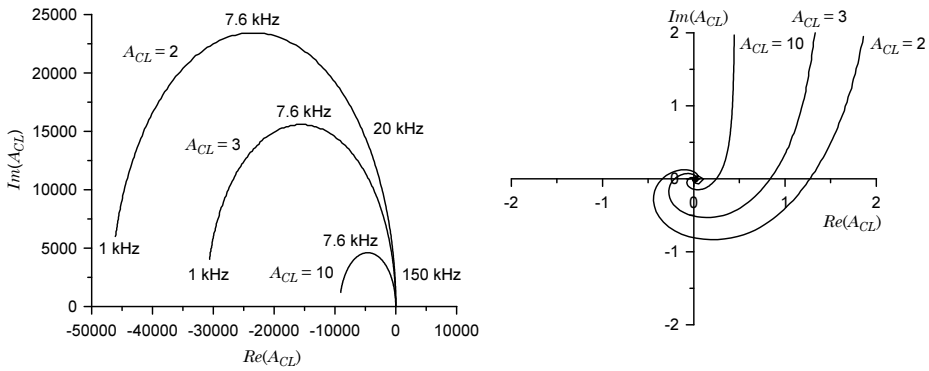

FIG. D.9. Plotting the imaginary part of the loop gain *vs.* the real part shows whether a feedback amplifier is stable. If the resulting curve encloses the point (1,0), the system will oscillate, as shown in the right panel for $A_{CL} = 2$. A closed loop gain of 3 fulfills the stability criterion, but exhibits severe peaking in the frequency response and ringing in the time response.

nounced peaking near the corner frequency. Figure D.8 shows the pulse response for closed loop gains $A_{CL} = 3$, 5, 10, and 20. The peaking in the frequency response for $A_{CL} = 3$ and 5 translates into ringing in the time response.

The stability of a feedback circuit can be assessed either by inspection of the open loop gain and phase (Bode 1940) or by plotting the imaginary part of the loop gain *vs.* the real part (Nyquist 1932). If the resulting curve encloses the point (1,0) – *i.e.* one on the real axis – the circuit is unstable. This is illustrated in Figure D.9. A closed loop gain $A_{CL} = 2$ leads to self-oscillation, whereas reducing the feedback to obtain $A_{CL} = 3$ achieves stability, but the small phase margin leads to pronounced peaking in the frequency response (Figure D.7) and ringing in the pulse response (Figure D.8).

References

Bode, H.W. (1940). Relations between attenuation and phase in feedback amplifier design. *Bell System Tech. Journal* **19** (1940) 421–454

Nyquist, H. (1932). Regeneration theory. *Bell System Tech. Journal* **11** (1932) 126–147

APPENDIX E

THE DIODE EQUATION

Before entering into a quantitative analysis of the diode characteristics, it is useful to review the relationships that determine carrier concentrations. To avoid confusion with exponential functions the symbol q_e will be used for the electronic charge instead of e. Furthermore, E_i denotes the intrinsic energy level, instead of the ionization energy.

E.1 Carrier concentrations in pure semiconductors

In thermal equilibrium the probability that an electron state in the conduction band is filled is given by the Fermi–Dirac distribution

$$f_e(E) = \frac{1}{e^{(E-E_F)/kT} + 1} \ . \tag{E.1}$$

The parameter E_F is the Fermi level (or chemical potential). Figure E.1 shows the distribution for several temperatures. The density of atoms in a Si or Ge crystal is about $5 \cdot 10^{22}$ atoms/cm^3. Since the minimum carrier density of interest in practical devices is of order 10^{10} to 10^{11} cm^{-3}, very small occupancies are quite important.

In silicon the bandgap is 1.12 eV. If the Fermi level is at midgap, the band-edges will be 0.56 eV above and below E_F. As is apparent from Figure E.1, relatively large deviations from the Fermi level, *i.e.* extremely small occupancies, will still yield significant carrier densities.

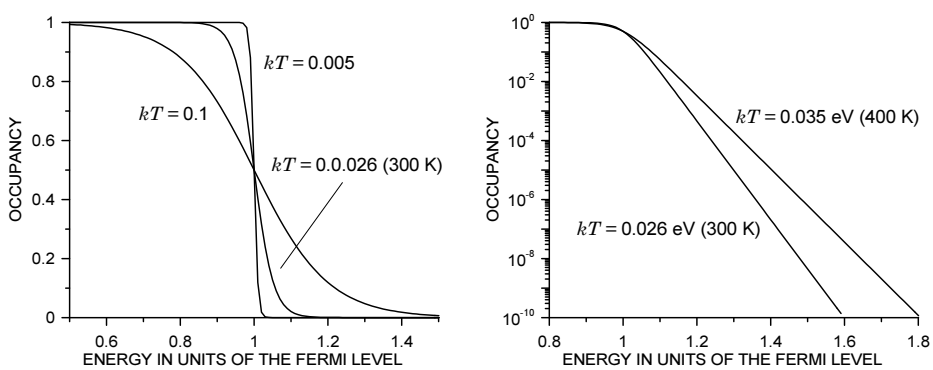

FIG. E.1. The Fermi–Dirac distribution plotted for various temperatures.

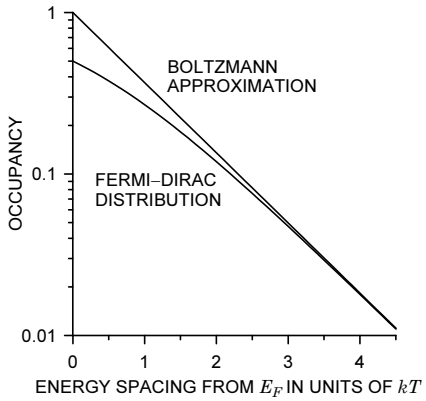

FIG. E.2. Comparison between the Fermi–Dirac and Boltzmann distributions.

The number of occupied electron states N_e is determined by summing over all available states multiplied by the occupation probability for each individual state:

$$N_e = \sum_i m_i f(E_i) \tag{E.2}$$

Since the density of states near the band edge tends to be quite high, this can be written as an integral

$$N_e = \int_{E_c}^{\infty} f(E) g(E) dE \ , \tag{E.3}$$

where $g(E)$ is the density of states. Solution of this integral requires knowledge of the density of states. Fortuitously, to a good approximation the density of states near the band edge has a parabolic distribution

$$g(E)dE \propto (E - E_c)^{1/2} \ . \tag{E.4}$$

As the energy increases beyond the band edge, the distribution will deviate from the simple parabolic form, but since the probability function decreases very rapidly, the integral will hardly be affected.

The second obstacle to a simple analytical solution of the integral is the intractability of integrating over the Fermi distribution. Fortunately, if $E - E_F$ is at least several times kT, the Fermi distribution can be approximated by a Boltzmann distribution, as shown in Figure E.2.

$$1 + e^{(E-E_F)/kT} \approx e^{(E-E_F)/kT}$$

$$f(E) \approx e^{-(E-E_F)/kT} \ . \tag{E.5}$$

At energies $2.3\,kT$ beyond the Fermi level the difference between the Boltzmann approximation and the Fermi distribution is $< 10\%$, for energies $> 4.5\,kT$ it is less than 1%.

Applying the approximation to the occupancy of hole states, the probability of a hole state being occupied, *i.e.* a valence state being empty, is

$$f_h(E) = 1 - f_e(E) = \frac{1}{e^{(E_F-E)/kT}+1} \approx e^{-(E_F-E)/kT} \ . \tag{E.6}$$

The conditions for the Boltzmann approximation are fulfilled for excitation across the bandgap, as the bandgap is of order 1 eV and kT at room temperature is 0.026 eV.

With these simplifications the concentration of electrons in the conduction band in thermal equilibrium

$$n \propto (kT)^{3/2} e^{-(E_c-E_F)/kT} \tag{E.7}$$

or

$$n = N_c e^{-(E_c-E_F)/kT} \ , \tag{E.8}$$

where N_c is the effective density of states at the band edge. Correspondingly, the hole concentration

$$p = N_v e^{-(E_F-E_v)/kT} \ . \tag{E.9}$$

In an ideal semiconductor the only source of mobile carriers is thermal excitation across the bandgap (additional impurity atoms or crystal imperfections that would allow other excitation mechanisms are absent), so the concentrations of electrons and holes are equal,

$$n = p = n_i \ . \tag{E.10}$$

n_i is called the intrinsic carrier concentration. In silicon ($E_g = 1.12$ eV) the intrinsic concentration $n_i = 1.45 \cdot 10^{10}\,\mathrm{cm}^{-3}$ at 300 K and in germanium ($E_g = 0.66$ eV), $n_i = 2.4 \cdot 10^{13}\,\mathrm{cm}^{-3}$. For comparison, the purest semiconductor material that has been fabricated is Ge with active impurity levels of about $3 \cdot 10^{10}\,\mathrm{cm}^{-3}$.

Using the above results

$$n_i = N_c e^{-(E_c-E_F)/kT} = N_v e^{-(E_F-E_v)/kT} \ , \tag{E.11}$$

which one can solve to obtain

$$E_F = E_i = \frac{E_c+E_v}{2} - \frac{kT}{2} \log\left(N_c/N_v\right) \ . \tag{E.12}$$

If the band structure is symmetrical ($N_c = N_v$), the intrinsic energy level E_i lies near the middle of the bandgap. Even rather substantial deviations from

a symmetrical band structure will not affect this result significantly, as N_c/N_v enters logarithmically and kT is much smaller than the bandgap.

A remarkable result is that the product of the electron and hole concentrations

$$np = n_i^2 = N_c N_v e^{-(E_c - E_v)/kT} = N_c N_v e^{-E_g/kT} \tag{E.13}$$

depends only on the bandgap E_g and not on the Fermi level.

This result, the law of mass action, is very useful in semiconductor device analysis. It requires only that the Boltzmann approximation holds. Qualitatively, it says that if one carrier type exceeds this equilibrium concentration, recombination will decrease the concentrations of both electrons and holes to maintain $np = {n_i}^2$, a relationship that also holds in doped crystals.

E.2 Carrier concentrations in doped crystals

The equality $n = p$ only holds for pure crystals, where all of the electrons in the conduction band have been thermally excited from the valence band. In practical semiconductors the presence of impurities tips the balance towards either electrons or holes.

Impurities are an unavoidable byproduct of the crystal growth process, although special techniques can achieve astounding results. For example, as noted above, in the purest semiconductor crystals – "ultrapure" Ge – the net impurity concentration is about $3 \cdot 10^{10}\,\mathrm{cm}^{-3}$.

In semiconductor device technology impurities are introduced intentionally to control the conductivity of the semiconductor. Let ${N_d}^+$ be the concentration of ionized donors and ${N_a}^-$ the concentration of ionized acceptors. Overall charge neutrality is preserved, as each ionized dopant introduces a charged carrier and an oppositely charged atom, but the net carrier concentration is now

$$\Delta n = n - p = N_d^+ - N_a^- \tag{E.14}$$

or

$$p + N_d^+ = n + N_a^- \ . \tag{E.15}$$

Assume that the activation energy of the donors and acceptors is sufficiently small so that they are fully ionized. Then ${N_d}^+ = N_d$ and ${N_a}^- = N_a$, so

$$p + N_d = n + N_a \ , \tag{E.16}$$

which, using $np = {n_i}^2$, becomes

$$\begin{aligned} p + N_d &= \frac{n_i^2}{p} + N_a \\ \frac{p}{N_a} + \frac{N_d}{N_a} &= \frac{n_i}{p}\frac{n_i}{N_a} + 1 \ . \end{aligned} \tag{E.17}$$

If the acceptor concentration $N_a \gg N_d$ and $N_a \gg n_i$, the hole and electron concentrations

$$p \approx N_a \quad \text{and} \quad n \approx \frac{n_i^2}{N_a} \ll N_a \ , \tag{E.18}$$

i.e. the conductivity is dominated by holes. Conversely, if the donor concentration $N_d \gg N_a$ and $N_d \gg n_i$ the conductivity is dominated by electrons.

If the conductivity is dominated by only one type of carrier, the Fermi level is easy to determine. If, for example, $n \gg p$ then eqn E.16 can be written as

$$\begin{aligned} n &= N_d - N_a \\ N_c e^{-(E_c - E_F)/kT} &= N_d - N_a \ , \end{aligned} \tag{E.19}$$

yielding

$$\frac{E_c - E_F}{k_B T} = \log\left(\frac{N_c}{N_d - N_a}\right) \ . \tag{E.20}$$

If $N_d \gg N_a$, then $E_c - E_F$ must be small, *i.e.* the Fermi level lies close to the conduction band edge.

In reality the impurity levels of common dopants are not close enough to the band edge for the Boltzmann approximation to hold, so the calculation must use the Fermi distribution and solve numerically for E_F. Nevertheless, the qualitative conclusions derived here still apply.

It is often convenient to refer all of these quantities to the intrinsic level E_i, as it accounts for both E_c and E_v. Then

$$\begin{aligned} n &= N_c e^{-(E_c - E_F)/kT} = n_i e^{-(E_F - E_i)/kT} \\ p &= N_v e^{-(E_F - E_v)/kT} = n_i e^{-(E_i - E_F)/kT} \end{aligned} \tag{E.21}$$

and the Fermi level

$$E_F - E_i = -k_B T \log \frac{N_a - N_d}{n_i} \ . \tag{E.22}$$

E.3 *pn*-junctions

A *pn*-junction is formed at the interface of a *p*- and an *n*-type region. Since the electron concentration in the *n*-region is greater than in the *p*-region, electrons will diffuse into the *p*-region. Correspondingly, holes will diffuse into the *n*-region. As electrons and holes diffuse across the junction, a space charge due to the ionized donor and acceptor atoms builds up. The field due to this space charge is directed to impede the flow of electrons and holes.

The situation is dynamic. The concentration gradient causes a continuous diffusion current to flow, whereas the field due to the space charge drives a drift current in the opposite direction. Equilibrium is attained when the two currents are equal, *i.e.* the sum of the diffusion and drift currents is zero. The net hole current density is

$$J_p = -q_e D_p \frac{dp}{dx} + q_e p \mu_p E_p \ , \tag{E.23}$$

where D_p is the diffusion constant for holes and E_p is the electric field in the *p*-region.

To solve this equation we make use of the following relationships: The hole concentration is

$$p = n_i e^{(E_i - E_F)/kT} \ , \tag{E.24}$$

so its derivative

$$\frac{dp}{dx} = \frac{p}{kT}\left(\frac{dE_i}{dx} - \frac{dE_F}{dx}\right) \ . \tag{E.25}$$

Since the force on a charge q_e due to an electric field E is equal to the negative gradient of the potential energy,

$$q_e E = -\frac{dE_c}{dx} = -\frac{dE_v}{dx} = -\frac{dE_i}{dx} \ . \tag{E.26}$$

As only the gradient is of interest and E_c, E_v, and E_i differ only by a constant offset, any of these three measures can be used. We'll use the intrinsic Fermi level E_i, since it applies throughout the sample.

The remaining ingredient is the Einstein relationship, which relates the mobility to the diffusion constant

$$\mu_p = \frac{q_e D_p}{kT} \ . \tag{E.27}$$

Using these relationships the net hole current becomes

$$J_p = q_e p \frac{D_p}{kT}\frac{dE_F}{dx} = \mu_p p \frac{dE_F}{dx} \ . \tag{E.28}$$

Accordingly, the net electron current

$$J_n = -q_e n \frac{D_n}{kT}\frac{dE_F}{dx} = -\mu_n n \frac{dE_F}{dx} \ . \tag{E.29}$$

Since, individually, the net hole and electron currents in equilibrium must be zero, the derivative of the Fermi level

$$\frac{dE_F}{dx} = 0 \ . \tag{E.30}$$

In thermal equilibrium the Fermi level must be constant throughout the junction region.

For the Fermi level to be flat, the band structure must adapt, since on the p-side the Fermi level is near the valence band, whereas on the n-side it is near the conduction band (Figure 2.19). If we assume that the dopants are exclusively donors on the n-side and acceptors on the p-side, the difference in the respective Fermi levels is

$$\Delta E_F = -kT \log \frac{N_a N_d}{n_i^2} \ . \tag{E.31}$$

This corresponds to an electric potential

$$\Delta V_F = \frac{1}{q_e}\Delta E_F \equiv V_{bi} \ , \tag{E.32}$$

often referred to as the "built-in" voltage of the junction.

As either N_a or N_d increases relative to n_i, the respective Fermi level moves closer to the band edge, increasing the built-in voltage. With increasing doping levels the built-in voltage approaches the equivalent potential of the bandgap E_g/q_e.

E.4 The forward-biased pn-junction

Applying an external bias leads to a condition that deviates from thermal equilibrium, *i.e.* the Fermi level is no longer constant throughout the junction. If a positive voltage is applied to the p-electrode relative to the n-electrode, the total variation of the electric potential across the junction will decrease (Figure 2.20). Since this reduces the electric field across the junction, the drift component of the junction current will decrease. Since the concentration gradient is unchanged, the diffusion current will exceed the drift current and a net current will flow.

This net current leads to an excess of electrons in the p-region and an excess of holes in the n-region. This "injection" condition leads to a local deviation from equilibrium, *i.e.* $pn > {n_i}^2$. Equilibrium will be restored by recombination.

Note that a depletion region exists even under forward bias, although its width is decreased. The electric field due to the space charge opposes the flow of charge, but the large concentration gradient overrides the field.

Consider holes flowing into the n-region. They will flow through the depletion region with small losses due to recombination, as the electron concentration is small compared with the bulk. When holes reach the n-side boundary of the depletion region the concentration of electrons available for recombination increases and the concentration of holes will decrease with distance, depending on the cross-section for recombination, expressed as a diffusion length. Ultimately, all holes will have recombined with electrons. The required electrons are furnished through the external contact from the power supply.

On the p-side, electrons undergo a similar process. The holes required to sustain recombination are formed at the external contact to the p-region by electron flow toward the power supply, equal to the electron flow toward the n-contact. The following derivation follows the discussions by Shockley (1949, 1950) and Grove (1967).

The steady-state distribution of charge is determined by solving the diffusion equation,

$$D_n \frac{d^2 n_p}{dx^2} - \frac{n_p - n_{p0}}{\tau_n} = 0 \ . \tag{E.33}$$

Electrons flowing into the p-region give rise to a local concentration n_p in excess of the equilibrium concentration n_{p0}. This excess will decay with a recombination time τ_n, corresponding to a diffusion length L_n.

The first boundary condition required for the solution of the diffusion equation is that the excess concentration of electrons vanish at large distances x,

$$n_p(\infty) = n_{p0} \ . \tag{E.34}$$

The second boundary condition is that the carriers are injected at the origin of the space charge region $x = 0$ with a concentration $n_p(0)$. This yields the solution

$$n_p(x) = n_{p0} + (n_p(0) - n_{p0})\, e^{-x/L_n} \ . \tag{E.35}$$

From this we obtain the electron current entering the p-region

$$J_{np} = -q_e D_n \left.\frac{dn_p}{dx}\right|_{x=0} = q_e D_n \frac{n_p(0) - n_{p0}}{L_n} \ . \tag{E.36}$$

This says that the electron current is limited by the concentration gradient determined by the carrier density at the depletion edge $n_p(0)$ and the equilibrium minority carrier density n_{p0}. Determining the equilibrium density n_{p0} is easy,

$$n_{p0} = n_i^2/N_a \ . \tag{E.37}$$

The problem is that $n_p(0)$ is established in a non-equilibrium state, where the previously employed results do not apply.

To analyze the regions with non-equilibrium carrier concentrations Shockley introduced a simplifying assumption by postulating that the product pn is constant. In this specific quasi-equilibrium state this constant will be larger than ${n_i}^2$, the pn-product in thermal equilibrium. In analogy to thermal equilibrium, this quasi-equilibrium state is expressed in terms of a "quasi-Fermi level", which is the quantity used in place of E_F that gives the carrier concentration under non-equilibrium conditions.

The postulate $pn = \text{const}$ is equivalent to stating that the non-equilibrium carrier concentrations are given by a Boltzmann distribution, so the concentration of electrons is

$$n = n_i e^{(E_{Fn} - E_i)/kT} \ , \tag{E.38}$$

where E_{Fn} is the quasi-Fermi level for electrons, and

$$p = n_i e^{(E_i - E_{Fp})/kT} \ , \tag{E.39}$$

where E_{Fp} is the quasi-Fermi level for holes. The product of the two carrier concentrations in non-equilibrium is

$$pn = n_i^2 e^{(E_{Fn} - E_{Fp})/kT} \ . \tag{E.40}$$

If pn is constant throughout the space-charge region, then $E_{Fn} - E_{Fp}$ must also remain constant.

Using the quasi-Fermi level and the Einstein relationship, the electron current entering the p-region becomes

$$J_{np} = -q_e D_n \left.\frac{dn_p}{dx}\right|_{x=0} = -q_e D_n \frac{d}{dx}(n_i e^{(E_{Fn} - E_i)/k_B T}) = -\mu_n n \frac{dE_{Fn}}{dx} \ . \tag{E.41}$$

These relationships describe the behavior of the quasi-Fermi level in the depletion region. How does this connect to the neutral region?

In the neutral regions the *majority* carrier motion is dominated by drift (in contrast to the injected *minority* carrier current, which is determined by diffusion). Consider the n-type region. Here the bulk electron current that provides the junction current

$$J_{nn} = -\mu_n n \frac{dE_i}{dx} \ . \tag{E.42}$$

Since the two electron currents must be equal

$$J_{nn} = J_{np} \ , \tag{E.43}$$

it follows that

$$\frac{dE_{Fn}}{dx} = \frac{dE_i}{dx} \ , \tag{E.44}$$

i.e. the quasi-Fermi level follows the energy band variation. Thus, in a neutral region, the quasi-Fermi level for the majority carriers is the same as the Fermi level in equilibrium. At current densities small enough not to cause significant voltage drops in the neutral regions, the band diagram is flat, and hence the quasi-Fermi level is flat.

In the space charge region, pn is constant, so the quasi-Fermi levels for holes and electrons must be parallel, *i.e.* both will remain constant at their respective majority carrier equilibrium levels in the neutral regions.

If an external bias V is applied, the equilibrium Fermi levels are offset by V, so it follows that the quasi-Fermi levels are also offset by V,

$$E_{Fn} - E_{Fp} = q_e V \ . \tag{E.45}$$

Consequently, the pn-product in non-equilibrium

$$pn = n_i^2 e^{(E_{Fn} - E_{Fp})/kT} = n_i^2 e^{q_e V/kT} \ . \tag{E.46}$$

If the majority carrier concentration is much greater than the concentration due to minority carrier injection ("low-level injection"), the hole concentration at the edge of the p-region remains essentially at the equilibrium value. Consequently, the enhanced pn-product increases the electron concentration.

$$n_p(0) = n_{p0} e^{q_e V/kT} \ . \tag{E.47}$$

Correspondingly, the hole concentration in the n-region at the edge of the depletion zone becomes

$$p_n(0) = p_{n0} e^{q_e V/kT} \ . \tag{E.48}$$

Since the equilibrium concentrations

$$n_{p0} = \frac{n_i^2}{N_a} \quad \text{and} \quad p_{n0} = \frac{n_i^2}{N_d} \ , \tag{E.49}$$

the components of the diffusion current due to holes and electrons are

$$J_n = q_e D_n \frac{n_i^2}{N_a L_n} \left(e^{q_e V/kT} - 1 \right)$$

$$J_p = q_e D_p \frac{n_i^2}{N_d L_p} \left(e^{q_e V/kT} - 1 \right) \ . \tag{E.50}$$

The total current is the sum of the electron and hole components

$$J = J_n + J_p = J_0 (e^{q_e V/kT} - 1) \ , \tag{E.51}$$

where

$$J_0 = q_e n_i^2 \left(\frac{D_n}{N_a L_n} + \frac{D_p}{N_d L_p} \right) \ . \tag{E.52}$$

This is the diode equation (or Shockley equation), which describes the current–voltage characteristic both under forward and reverse bias. Under forward bias the current increases exponentially. Under reverse bias (negative V), the exponential term vanishes when the bias exceeds several kT/q_e and the current becomes the reverse saturation current $J = -J_0$. For a uniform junction cross-section the current densities J_n, J_p, and J_0 can be replaced by their respective currents.

Note that in the diode equation:

1. The bandgap does not appear explicitly (only implicitly in J_0 via n_i).
2. The total current has two distinct components, due to electrons and holes.
3. The electron and hole currents are generally not equal. The ratio

$$\frac{J_n}{J_p} = \frac{N_d}{N_a} \qquad \text{if} \qquad \frac{D_n}{L_n} = \frac{D_p}{L_p} \ . \tag{E.53}$$

4. Current flows for all values of V. However, when plotted on a linear scale, the exponential appears to have a knee, often referred to as the "turn-on" voltage.
5. The magnitude of the turn-on voltage is determined by J_0. Diodes with different bandgaps will show the same behavior if J_0 is the same.

Figure E.3 shows measured I–V curves for commercial Si and Ge junction diodes (1N4148 and 1N34A). On a linear scale the Ge diode "turns on" at 200 – 300 mV, whereas the Si diode has a threshold of 500 – 600 mV. However, on a logarithmic scale it becomes apparent that both diodes pass current at all voltages > 0.

The reverse current (Figure E.4) shows why the Ge diode shows greater sensitivity at low voltages. The smaller bandgap leads to increased n_i. The Si diode shows a "textbook" exponential forward characteristic at currents > 10 nA, whereas the Ge diode exhibits a more complex structure.

The discrepancies in the forward current between the measured results and the simple theory require the analysis of all processes in the depletion zone:

1. Generation-recombination in the depletion region (see Appendix F).
2. Diffusion current (as just calculated for the ideal diode).

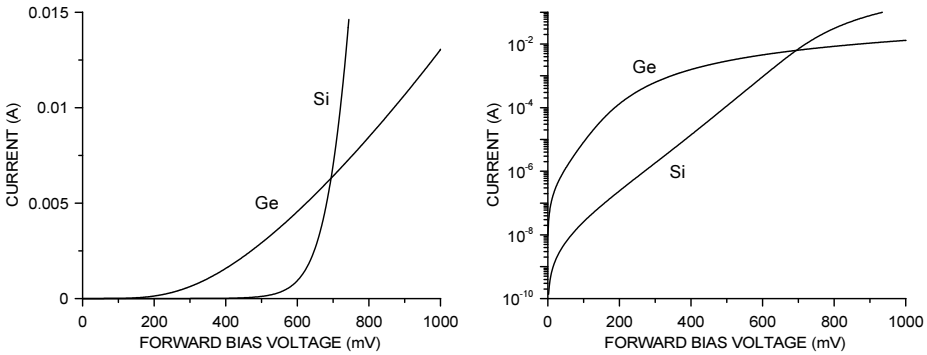

FIG. E.3. Current *vs.* voltage for forward biased Si and Ge diodes.

3. High-injection region where the injected carrier concentration affects the potentials in the neutral regions.
4. Voltage drop due to bulk series resistance.

For a discussion of these effects see Sze (1981).

The reverse current is increased due to generation and recombination currents in the depletion zone, as discussed in Appendix F. In optimized photodiodes reverse bias currents of about $100\,\mathrm{pA/cm^2}$ have been achieved, which is about 3 times the theoretical value (Holland 2004).

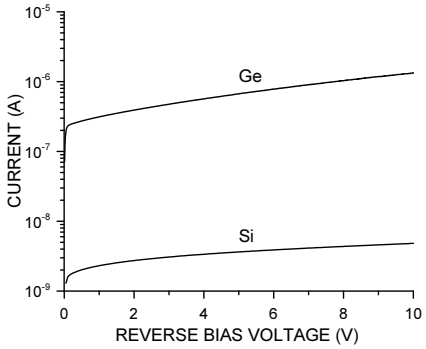

FIG. E.4. Reverse current of Si and Ge diodes at room temperature.

References

Grove, A.S. (1967). *Physics and Technology of Semiconductor Devices*. Wiley, New York. ISBN 0-471-32998-3

Holland, S.E. (2004). private communication

Shockley, W. (1949). The Theory of $p-n$ junctions in Semiconductors and $p-n$ Junction Transistors, *Bell System Tech. Journal* **28** (1949) 435–489

Shockley, W. (1950). *Electrons and Holes in Semiconductors*. van Nostrand, Princeton

Sze, S.M. (1981). *Physics of Semiconductor Devices* (2nd edn). Wiley, New York. ISBN 0-471-05661-8, TK7871.85.S988 1981

APPENDIX F

ELECTRICAL EFFECTS OF IMPURITIES AND DEFECTS

As in Appendix E, the symbol q_e will be used for the electronic charge instead of e and E_i is the intrinsic Fermi level.

Although derivation of the diode equation in Appendix E proceeded under the title "The forward-biased *pn*-junction", nothing in the assumptions and algebraic manipulations restricted the sign of the applied voltage. If a negative bias is applied to the junction, the minority carrier concentrations at the junction edges decrease with respect to thermal equilibrium and reverse the concentration gradient.

Setting a reverse voltage $V \gg kT$ in the diode equation yields

$$J = -J_0 \ , \tag{F.1}$$

where

$$J_0 = q_e n_i^2 \left(\frac{D_n}{N_a L_n} + \frac{D_p}{N_d L_p} \right) \ . \tag{F.2}$$

In this ideal case the diode current at large reverse bias voltage would be determined by the

- doping concentrations,
- diffusion constants, and
- recombination lengths.

In reality, the measured currents are often orders of magnitude larger.

Whereas the diode equation predicts the saturation of the reverse diode current at voltages greater than order 100 mV ($\sim 4kT$), one frequently observes a monotonically increasing current, which increases linearly with depletion width. This implies the presence of imperfections in the crystal that increase the reverse leakage current. For a uniform distribution of imperfections, the number of active sites will increase with the depletion volume.

F.1 Emission and capture processes

Figure 7.2 in Chapter 7 summarizes the emission and capture processes:

1. Hole emission: The process of hole emission from a defect can also be viewed as promoting an electron from the valence band to the defect level, as shown in (a).
2. Electron emission: In a second step (b) this electron can proceed to the conduction band and contribute to current flow, generation current.

3. Electron capture: Conversely, a defect state can capture an electron from the conduction band (c), which in turn can capture a hole (d). This "recombination" process reduces current flowing in the conduction band.
4. Trapping: Defect levels close to a band edge will capture charge and release it after some time, a process called "trapping" (e).

All of these processes are governed by Fermi statistics (or the Boltzmann approximation), as discussed in Appendix E. We will now examine these processes quantitatively. The discussion follows Grove (1967).

Assume a concentration of centers N_t whose energy level E_t lies within the bandgap. The probability of a center being occupied is

$$f = \frac{1}{1 + e^{(E_t - E_F)/kT}} \ , \tag{F.3}$$

so the concentration of vacant centers is

$$N_{t0} = N_t(1 - f) \ . \tag{F.4}$$

F.1.1 *Electron capture*

The rate of electron capture is proportional to the concentration of unoccupied centers

$$\frac{dN_{nc}}{dt} = v_{th}\sigma_n n N_{t0} = v_{th}\sigma_n n N_t(1 - f) \ , \tag{F.5}$$

where v_{th} is the thermal velocity of an electron (about 10^7 cm/s at 300 K), σ_n is the capture cross-section, and n is the concentration of electrons in the conduction band. The velocity enters because the capture centers are localized and an electron has to move near the center to be captured. The thermal velocity is superimposed on the much slower motion due to drift or diffusion, so the thermal velocity determines the number of defect sites scanned per unit time.

F.1.2 *Electron emission*

The rate of electron emission is proportional to the concentration of occupied centers $N_{nc} = N_t f$. If the emission probability is e_n, the rate of electron emission

$$\frac{dN_{ne}}{dt} = e_n N_t f \ . \tag{F.6}$$

F.1.3 *Hole capture and emission*

The rates of hole capture and emission can be expressed analogously to electrons. The rate of hole capture

$$\frac{dN_{pc}}{dt} = v_{th}\sigma_p p N_t f \ , \tag{F.7}$$

since hole capture corresponds to the transition of an electron from the center to the valence band, this process is proportional to the concentration of centers occupied by electrons $N_t f$.

The rate of hole emission is proportional to the concentration of centers not occupied by electrons $N_t(1-f)$, so

$$\frac{dN_{pe}}{dt} = e_p N_t (1-f) \ . \tag{F.8}$$

F.1.4 *Emission probabilities*

In equilibrium, the rates of the two processes that move electrons to and from the conduction band, capture and emission, must be equal. This seemingly trivial statement reflects the more profound principle in statistical mechanics of detailed balance, which states that under equilibrium conditions every process and its reverse must proceed at exactly equal rates.

Thus, for electrons and holes, respectively,

$$\begin{aligned} v_{th}\sigma_n n N_t (1-f) &= e_n N_t f \\ v_{th}\sigma_p p N_t f &= e_p N_t (1-f) \ . \end{aligned} \tag{F.9}$$

From this, the emission probability for electrons

$$e_n = v_{th}\,\sigma_n\,n\,\frac{1-f}{f} \ . \tag{F.10}$$

The concentration of electrons in the conduction band

$$n = n_i e^{(E_F - E_i)/kT} = N_c e^{-(E_c - E_F)/kT} \ . \tag{F.11}$$

Since

$$f = \frac{1}{1 + e^{(E_t - E_F)/kT}} \ , \tag{F.12}$$

the fourth factor of eqn F.10 becomes

$$\frac{1-f}{f} = \frac{1}{f} - 1 = e^{(E_t - E_F)/kT} \tag{F.13}$$

and the emission probability

$$e_n = v_{th}\sigma_n n_i e^{(E_t - E_i)/kT} = v_{th}\sigma_n N_c e^{(E_c - E_i)/kT} \ . \tag{F.14}$$

Similarly, the emission probability of holes

$$e_p = v_{th}\sigma_p n_i e^{(E_i - E_t)/kT} = v_{th}\sigma_n N_v e^{(E_t - E_v)/kT} \ . \tag{F.15}$$

As intuitively expected, the emission probability grows exponentially as the energy level of the center approaches the band edge.

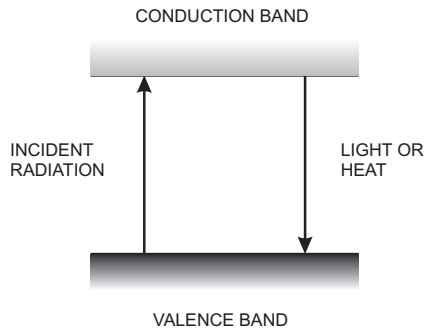

FIG. F.1. An electron can be excited to the conduction band by radiation and subsequently decay to the valence band to recombine with a hole and release light or heat.

F.2 Recombination

Recombination is important in detectors as it incurs a loss of signal charge.

F.2.1 *Band-to-band recombination*

Incident radiation excites electrons from the valence to the conduction band, forming an electron–hole pair. The simplest recombination mechanism would be for electrons in the conduction band to recombine with holes in the valence band (Figure F.1). The energy released in the recombination could be emitted as light or converted into heat. Direct transitions from the conduction to the valence

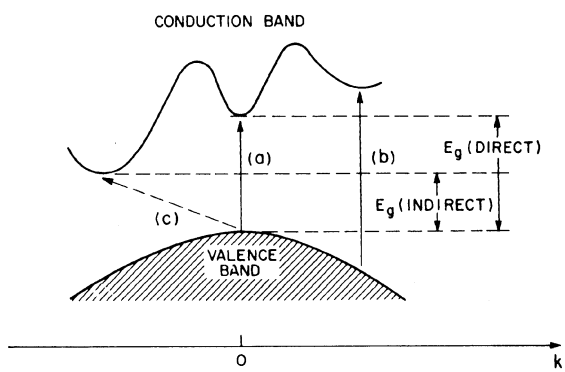

FIG. F.2. When the minimum of the conduction band and the maximum of the valence coincide in k space an electron can be promoted to the conduction band without momentum transfer (direct transition). When the band extrema are offset, momentum must be transferred (indirect transition). (From Sze 1981. ©John Wiley & Sons, reprinted with permission.)

band in Si or Ge are extremely improbable, as Si and Ge are "indirect bandgap" semiconductors, where the minimum of the conduction band and the maximum of the valence band are offset in wavevector k, *i.e.* momentum (Figure F.2 and Figure 2.9 in Chapter 2). Direct transitions in Ge and Si are possible, but only at the larger bandgaps with exponentially lower population densities.

Indirect transitions can be facilitated by intermediate states that provide "stepping stones" for electrons and holes traversing the forbidden band.

F.2.2 *Recombination via intermediate states*

Consider a steady flux of radiation, for example light, leading to a uniform generation rate per unit volume G_L.

To determine the effectiveness of centers as recombination sites, the charge due to the radiation will not be removed by an external circuit, but allowed to decay by recombination alone. In the steady state the rate at which electrons enter the conduction band must equal the rate at which they leave it:

$$\frac{dn_n}{dt} = G_L - \left(\frac{dN_{nc}}{dt} - \frac{dN_{ne}}{dt}\right) = 0 \ . \tag{F.16}$$

Similarly, for the holes

$$\frac{dn_p}{dt} = G_L - \left(\frac{dN_{pc}}{dt} - \frac{dN_{pe}}{dt}\right) = 0 \ . \tag{F.17}$$

Incident radiation takes the system out of thermal equilibrium, so none of the equilibrium carrier concentrations are valid, nor is the occupancy determined by the Fermi distribution. Instead, the concentrations n and p and the fractional occupancy f depend on the radiation flux G_L.

By equating G_L from the two expressions above,

$$\begin{aligned} \left(\frac{dN_{pc}}{dt} - \frac{dN_{pe}}{dt}\right) &= \left(\frac{dN_{nc}}{dt} - \frac{dN_{ne}}{dt}\right) \\ v_{th}\sigma_p p N_t f - e_p N_t(1-f) &= v_{th}\sigma_n n N_t(1-f) - e_n N_t f \ , \end{aligned} \tag{F.18}$$

and inserting the emission probabilities calculated above one can extract the steady state fractional occupancy

$$f = \frac{\sigma_n n + \sigma_p n_i e^{(E_i - E_t)/k_B T}}{\sigma_n(n + n_i e^{(E_t - E_i)/k_B T}) + \sigma_p(p + n_i e^{(E_i - E_t)/k_B T})} \ . \tag{F.19}$$

This occupancy depends implicitly on the generation flux G_L, which determines n and p.

Electrons are continually captured and emitted by the center, and so are holes. If an electron and a hole recombine, this leads to a deficit in the emission rates of both electrons and holes. In the steady-state the emission deficit for

electrons and holes must be equal, so the net rate of recombination is the capture rate minus the emission rate

$$\frac{dN_R}{dt} = \left(\frac{dN_{nc}}{dt} - \frac{dN_{ne}}{dt}\right) = \left(\frac{dN_{pc}}{dt} - \frac{dN_{pe}}{dt}\right) \ . \tag{F.20}$$

This is the same expression as above, so one can determine the recombination rate by inserting the steady state fractional occupancy into either side of eqn F.18,

$$\frac{dN_R}{dt} = \frac{\sigma_p \sigma_n v_{th} N_t (pn - n_i^2)}{\sigma_n (n + n_i e^{(E_t - E_i)/k_B T}) + \sigma_p (p + n_i e^{(E_i - E_t)/k_B T})} \ . \tag{F.21}$$

To simplify the equation and facilitate the interpretation of this result, assume (somewhat arbitrarily) that the capture cross-sections σ_n and σ_p are equal. Then

$$\frac{dN_R}{dt} = \sigma v_{th} N_t \frac{pn - n_i^2}{n + p + 2n_i (e^{(E_t - E_i)/k_B T} + e^{-(E_t - E_i)/k_B T})} \ . \tag{F.22}$$

From this expression one can see that the driving force of the recombination process is the excess carrier concentration pn beyond the equilibrium concentration ${n_i}^2$.

The third term in the denominator describes the relative occupancies of electrons and holes. A center close to the conduction band will have a higher occupancy of electrons than holes, so the recombination rate is limited by the hole population. Conversely, a center close to the valence band will have an excess of holes, so the population of electrons limits the recombination rate. The recombination rate is maximum when $E_t = E_i$, *i.e.* when the energy of the recombination center is at mid-gap.

A special case of recombination is minority carrier injection. Consider holes injected into an n-type region, as in a forward biased diode. In this case

$$n_n \gg p_n \ . \tag{F.23}$$

Furthermore, since efficient recombination centers are far from the band edge, the Boltzmann approximation holds, so the equilibrium electron concentration

$$n_n \gg n_i e^{(E_t - E_i)/k_B T} \ . \tag{F.24}$$

Then the above expression for the recombination rate simplifies to

$$\frac{dN_R}{dt} = \frac{\sigma_p \sigma_n v_{th} N_t (n_n p_n - n_i^2)}{\sigma_n n_n} = \sigma_p v_{th} N_t (p_n - p_{n0}) \ . \tag{F.25}$$

Expressed as a lifetime

$$\frac{dN_R}{dt} = \frac{p - p_{n0}}{\tau_p} \ , \tag{F.26}$$

so the carrier lifetime

$$\tau_p = \frac{1}{\sigma_p v_{th} N_t} \ . \tag{F.27}$$

The lifetime of the holes in the n-bulk is independent of the concentration of the electrons.

This is due to an abundance of electrons, so as soon as a hole is captured, an electron is available for immediate recombination. Hence, the hole concentration is the rate-limiting parameter. Conversely, if electrons are injected into n-type material where they are majority carriers, their lifetime will be significantly greater, since few holes are available for recombination.

Minority carrier injection is the worst case with respect to recombination, so "minority carrier lifetime" is a figure of merit used to characterize the presence of defects in semiconductors.

Recombination is important whenever the carrier concentration deviates from thermal equilibrium

$$pn > n_i^2 \ . \tag{F.28}$$

This occurs

- with incident radiation,
- in a forward biased diode.

F.3 Carrier generation

F.3.1 *Generation in the depletion region*

In a diode operated with reverse bias, *e.g.* a radiation detector, with

$$V_R \gg \frac{kT}{q_e} \tag{F.29}$$

all of the free carriers are swept from the depletion region, so there are no free carriers available for capture and recombination. In this configuration only emission processes are important.

Emission, in the absence of capture, can only proceed by alternating hole and electron emission, *i.e.* generation of electron–hole pairs. The rate of generation of electron–hole pairs can be determined from the previously derived expressions eqns F.20 and F.21 for the difference between capture and emission rates

$$\frac{dN_c}{dt} - \frac{dN_e}{dt} = \frac{\sigma_p \sigma_n v_{th} N_t (pn - n_i^2)}{\sigma_n (n + n_i e^{(E_t - E_i)/k_B T}) + \sigma_p (p + n_i e^{(E_i - E_t)/k_B T})} \ . \tag{F.30}$$

Since $dN_c/dt = 0$ and $p \ll n_i$, $n \ll n_i$,

$$\frac{dN_e}{dt} = \frac{\sigma_p \sigma_n v_{th} N_t n_i}{\sigma_n e^{(E_t - E_i)/k_B T} + \sigma_p e^{(E_i - E_t)/k_B T}} \ . \tag{F.31}$$

This is often written as

$$\frac{dN_e}{dt} \equiv \frac{n_i}{2\tau_g} \ , \tag{F.32}$$

where τ_g is called the generation lifetime.

Again consider the simplified case of equal cross-sections $\sigma_p = \sigma_n = \sigma$. Then the generation rate becomes

$$\frac{dN_e}{dt} = \frac{\sigma v_{th} N_t n_i}{e^{(E_i - E_t)/k_B T} + e^{-(E_i - E_t)/k_B T}} \ , \tag{F.33}$$

which again shows that only states near the intrinsic Fermi level E_i, *i.e.* mid-gap states, contribute significantly to the generation rate.

Intuitively, this is easy to see in the "stepping stone" picture. Since the emission probabilities for electrons and holes increase exponentially with the separation from their respective band edges, the probability for sequential hole and electron emission is maximum at mid-gap.

The emission rate of carriers leads to an electrical current, the generation current, which increases with the density of centers. If the emission centers are distributed uniformly throughout the depletion width W, the generation current density

$$J_{gen} = q_e \frac{dN_e}{dt} W = q_e W \frac{\sigma v_{th} N_t n_i}{e^{(E_i - E_t)/k_B T} + e^{-(E_i - E_t)/k_B T}} \ . \tag{F.34}$$

F.3.2 *Generation in the neutral region*

In the neutral region the absence of a significant electric field means that any excess carriers due to generation move only by diffusion. Charges generated near the transition to the depletion region can reach the influence of the electric field and will be swept to the opposite electrode. This additional contribution to the reverse diode current is called the diffusion current.

The starting point of the calculation is the steady-state diffusion equation for minority carriers. Consider electrons generated in the p-region:

$$D_n \frac{d^2 n_p}{dx^2} - \frac{n_p - n_{p0}}{\tau_n} = 0 \ . \tag{F.35}$$

Far from the space charge region the carrier concentration attains the thermal equilibrium value

$$n_p(\infty) = n_{p0} \ . \tag{F.36}$$

At the edge of the depletion region all carriers will be swept away by the electric field, so

$$n_p(0) = 0 \ . \tag{F.37}$$

The solution of the diffusion equation for these boundary conditions is

$$n_p(x) = n_{p0}(1 - e^{-x/L_n}) \ , \tag{F.38}$$

where

$$L_n \equiv \sqrt{D_n \tau_n} \tag{F.39}$$

is the diffusion length of electrons in the p-region.

This gives rise to an electrical current

$$J_{diff,n} = -q_e \left(-D_n \left. \frac{dn_p}{dx} \right|_{x=0} \right) = q_e D_n \frac{n_{p0}}{L_n} = q_e \frac{D_n}{L_n} \frac{n_i^2}{N_a} \ . \tag{F.40}$$

Similarly, for holes in the n-region

$$J_{diff,p} = q_e D_p \frac{p_{n0}}{L_p} = q_e \frac{D_p}{L_p} \frac{n_i^2}{N_d} \ . \tag{F.41}$$

The diffusion current increases with the square of the intrinsic carrier concentration, in contrast to the generation current in the depletion zone, which increases linearly with n_i.

The generation rate in a neutral region depleted of minority carriers can be drastically different from the depletion region. For simplicity, assume that the diffusion lifetime is equal to the generation lifetime. Then the ratio of the two generation currents

$$\frac{J_{diff,n}}{J_{gen}} = \frac{\frac{n_{p0}}{\tau} L_n}{\frac{n_i}{2\tau} W} = 2 \frac{n_{p0}}{n_i} \frac{L_n}{W} = 2 \frac{n_i}{N_a} \frac{L_n}{W} \ . \tag{F.42}$$

In an n-bulk Si radiation detector with a thin p-electrode, the diffusion length is limited by the electrode thickness, *i.e.* $\sim 1\,\mu$m. For $n_i \approx 10^{10}\,\mathrm{cm}^{-3}$, $N_a \approx 10^{15}\,\mathrm{cm}^{-3}$, and $W \approx 300\,\mu$m

$$\frac{J_{diff,n}}{J_{gen}} \approx 3 \cdot 10^{-8} \ .$$

In high-quality radiation detectors the generation current dominates.

By contrast, in a symmetrical Ge small signal rectifier diode with $L_n \approx 100\,\mu$m, $W \approx 1\,\mu$m, $n_i \approx 10^{13}\,\mathrm{cm}^{-3}$, and $N_a \approx 10^{15}\,\mathrm{cm}^{-3}$,

$$\frac{J_{diff,n}}{J_{gen}} \approx 2 \ .$$

At higher temperatures the exponential increase in n_i can increase the diffusion current so much that the generation current is negligible.

F.4 The origin of recombination and generation centers

Recombination and generation centers can be introduced by

1. impurity atoms,
2. structural imperfections,
3. radiation damage (displacement of atoms from lattice sites).

Figure F.3 shows the energy levels of impurities in Si and GaAs crystals. Mn, Cd, Zn, Au, Co, V and Fe are effective "lifetime killers" in Si. Au is commonly introduced intentionally in devices where short lifetimes are desirable (fast switching diodes and transistors).

All three defect mechanisms can create states distributed throughout the bandgap Since only mid-gap states can contribute significantly to generation and recombination, in a continuum of states statistics automatically select the states near mid-gap. Indirect bandgap materials show reduced transition probabilities, as the mismatch in k-space must be bridged.

GaAs appears favorable, since relatively few impurity states are near mid-gap. However, a structural defect tends to dominate: The "anti-site" defect, where a Ga atom occupies an As site (or vice versa) and introduces a mid-gap state that effectively pins the Fermi level at mid-gap. This is why GaAs commonly appears intrinsic, or "semi-insulating", so that large depletion widths can be obtained with small reverse bias voltages. Unfortunately, recombination is also high, although for some applications acceptable in thin ($\sim 10^2 \mu$m) detectors.

F.5 The diode equation revisited

F.5.1 *Reverse Current*

Both the generation and diffusion currents invariably override the ideal reverse saturation current

$$J_0 = q_e n_i^2 \left(\frac{D_n}{N_a L_n} + \frac{D_p}{N_d L_p} \right) << J_{diff} + J_{gen} \ , \tag{F.43}$$

so the diode equation becomes

$$J = J_R (e^{q_e V/kT} - 1) \ . \tag{F.44}$$

The reverse current J_R for voltages $> 3kT/q_e$ is the sum of the diffusion and generation currents

$$J_R = q_e n_i^2 \left(\frac{1}{N_a} \sqrt{\frac{D_n}{\tau_n}} + \frac{1}{N_d} \sqrt{\frac{D_p}{\tau_p}} \right) + q_e \frac{n_i}{2\tau_g} W \ . \tag{F.45}$$

Whether the generation or diffusion current dominates can be determined from the temperature coefficient. The diffusion current scales with ${n_i}^2$, so

$$\frac{dJ_R}{dT} = J_R \frac{E_g}{kT^2} \ , \tag{F.46}$$

whereas the generation current scales with n_i, yielding

$$\frac{dJ_R}{dT} = J_R \frac{E_g}{2kT^2} \ . \tag{F.47}$$

In practice, a plot of $\log J_R$ *vs.* $1/kT$ will yield a slope of $-E_g$ for diffusion and approximately $-E_g/2$ for generation dominated operation. At sufficiently high temperatures diffusion will always dominate.

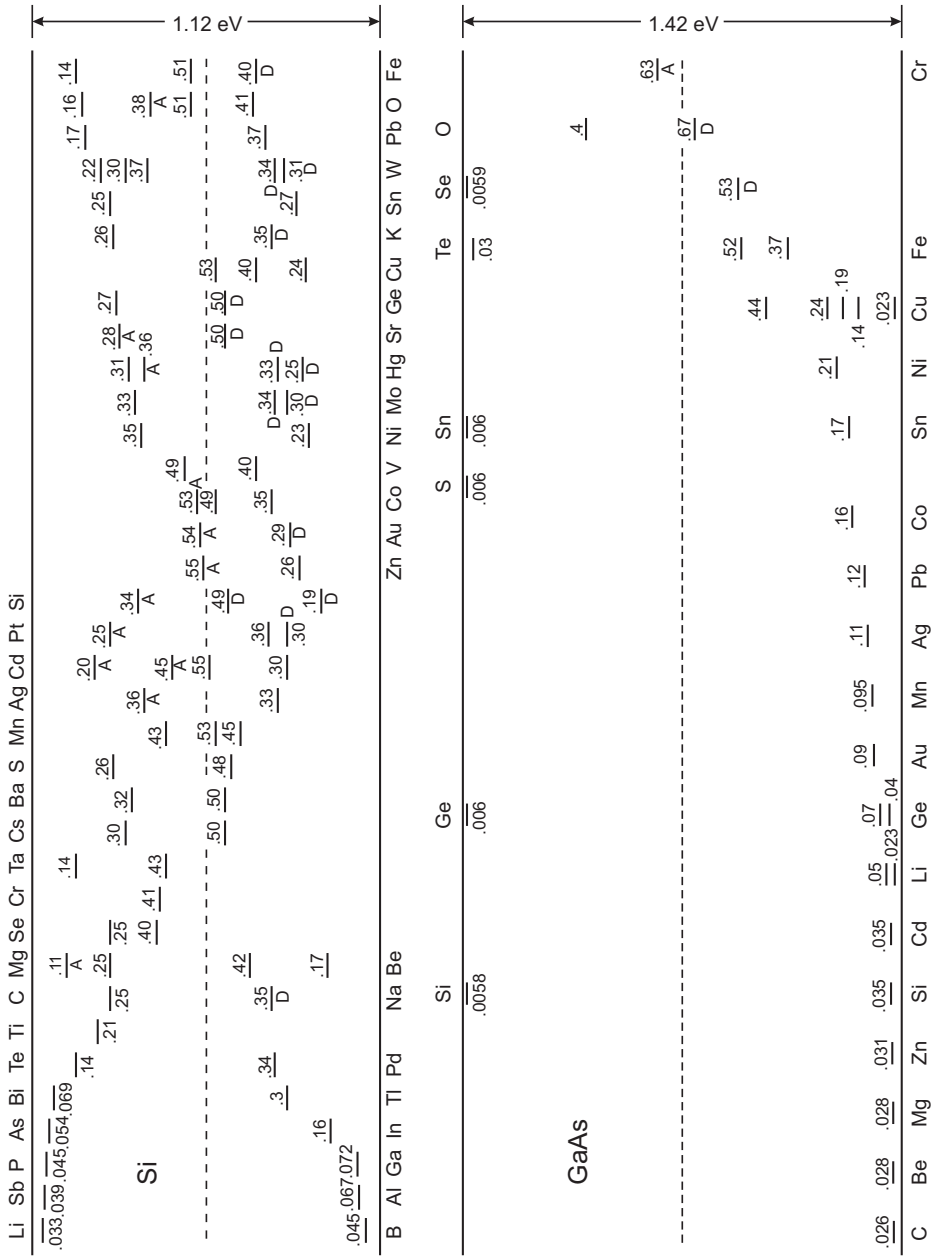

FIG. F.3. Energy levels introduced by impurities in Si and GaAs. The levels in the bottom half are measured from the valence band edge and in the top half from the conduction band. Acceptor and donor levels beyond mid-gap are denoted by A or D. (Adapted from Sze 1981. ©John Wiley & Sons, reprinted with permission.)

In radiation damaged diodes the generation current dominates even after rather low fluences, so the reverse bias current

$$I_R(T) \propto v_{th} n_i \propto v_{th}\sqrt{N_c N_v} e^{-E/2kT} \propto T^2 e^{-E/2kT} \ , \tag{F.48}$$

where $E \approx E_g$, depending on the impurity or defect energy level. The ratio of currents at two temperatures T_1 and T_2 is

$$\frac{I_R(T_2)}{I_R(T_1)} = \left(\frac{T_2}{T_1}\right)^2 \exp\left[-\frac{E}{2k}\left(\frac{T_1 - T_2}{T_1 T_2}\right)\right] \ . \tag{F.49}$$

Cooling to 0 °C typically reduces the reverse bias current to 1/6 of its value at room temperature. As discussed in Chapter 7, after irradiation the leakage current initially decreases with time. Pronounced short term and long term annealing components are observed and precise fits to the annealing curve require a sum of exponentials.

In practice, the variation of leakage current with temperature is very reproducible from device to device, even after substantial doping changes due to radiation damage. The leakage current can be used for dosimetry and diodes are offered commercially specifically for this purpose.

F.5.2 *Forward current*

Recombination in the depletion region also affects the forward diode characteristic. Experimental results can generally be described by introducing an "ideality factor" n

$$J = J_R(e^{q_e V/nkT} - 1) \ , \tag{F.50}$$

where $n = 2$ when the recombination current dominates and $n = 1$ when the current is dominated by diffusion in the neutral regions. In practical diodes n lies between 1 and 2.

At very low currents the generation currents dominate. Since these currents are opposite to the forward injection current, one observes a change of sign in the current flow at low voltage.

F.5.3 *Comments*

In radiation detectors the reverse current is of primary interest, as it is a source of shot noise. Nevertheless, the forward current–voltage characteristic can provide useful diagnostic information. Since recombination and generation are both maximized for mid-gap states, one commonly observes that devices with large generation currents also exhibit high recombination rates.

This has promoted a tendency to characterize both phenomena by one parameter, the minority carrier lifetime. However, generation and recombination are two distinct phenomena. First, their temperature dependencies differ, and second, it is not at all assured that a state is equally effective at generation as it is at recombination.

References

Grove, A.S. (1967). *Physics and Technology of Semiconductor Devices*. Wiley, New York. ISBN 0-471-32998-3

Sze, S.M. (1981). *Physics of Semiconductor Devices* (2nd edn). Wiley, New York. ISBN 0-471-05661-8, TK7871.85.S988 1981

APPENDIX G

BIPOLAR TRANSISTOR EQUATIONS

The basic processes in an *npn* bipolar junction transistor are summarized below.

1. Holes are injected into the base through the external contact. The potential distribution drives them towards the emitter. Since they are majority carriers in the base, few will recombine. Holes entering the base–emitter depletion region will either
   (a) pass through the depletion region into the emitter or
   (b) be lost due to recombination.
2. As shown in the discussion of the *pn*-junction (Appendix E), coupled with the hole current is an electron current originating in the emitter. This electron current will flow towards the collector, driven by the more positive potential. These electrons either
   (a) enter the collector to become the collector current or
   (b) recombine in the base region. The holes required for the recombination are furnished by the base current.
3. Thus, the base current is the sum of the
   (a) hole current entering the emitter,
   (b) hole losses due to recombination in the base–emitter depletion region, and
   (c) electron losses due to recombination in the base during transport to the collector.

The goal is to maximize 2(a) and 3(a).

As in Appendix E, the discussion follows Grove (1967) and the symbol q_e will be used for the electronic charge instead of e to avoid confusion with exponential functions.

The transport of minority carriers in the base is driven by diffusion, so

$$D_n \frac{d^2 n_p}{dx^2} - \frac{n_p - n_{p0}}{\tau_n} = 0 \ . \tag{G.1}$$

At the boundary to the base–emitter depletion region

$$n_p(0) = n_{p0} e^{q_e V_{BEe}/kT} \ . \tag{G.2}$$

The equilibrium concentration of electrons in the base is determined by the base acceptor doping level N_{aB}

$$n_{p0} = \frac{n_i^2}{N_{aB}} \ . \tag{G.3}$$

At the collector boundary all minority carriers will be immediately swept away by the reverse bias field, so that the boundary condition becomes

$$n_p(W_B) = 0 \ . \tag{G.4}$$

Then the solution of the diffusion equation is

$$n_p(x) = n_{p0}\left(1 - \frac{\sinh \dfrac{x}{L_n}}{\sinh \dfrac{W_B}{L_n}}\right) + (n_p(0) - n_{p0})\frac{\sinh \dfrac{W_B - x}{L_n}}{\sinh \dfrac{W_B}{L_n}} \ . \tag{G.5}$$

If $V_{BE} \gg 4kT/q_e$ the non-equilibrium concentration will dominate,

$$n_p(0) \gg n_{p0} \ , \tag{G.6}$$

so this simplifies to

$$n_p(x) = n_p(0)\frac{\sinh \dfrac{W_B - x}{L_n}}{\sinh \dfrac{W_B}{L_n}} \ . \tag{G.7}$$

Since the base width W_B in good transistors is much smaller than the diffusion length L_n, the concentration profile can be approximated by a linear distribution

$$n_p(x) = n_p(0)\left(1 - \frac{x}{W_B}\right) \ . \tag{G.8}$$

Now we can evaluate the individual current components. In this approximation the diffusion current of electrons in the base region becomes

$$J_{nB} = -q_e D_{nB}\frac{dn_p(x)}{dx} = q_e D_{nB}\frac{n_i^2}{N_{aB}W_B}e^{q_e V_{BE}/kT} \ , \tag{G.9}$$

where D_{nB} is the diffusion constant of electrons in the base. Similarly, the diffusion current of holes injected into the emitter, under the assumption that the emitter depth W_E is much smaller than the diffusion length,

$$J_{pE} = q_e D_{pE}\frac{n_i^2}{N_{dE}W_E}e^{q_e V_{BE}/kT} \ . \tag{G.10}$$

For the moment, we'll neglect recombination of holes in the base–emitter depletion region. Under this assumption, the base current is

$$I_B = J_{pE}A_{JE} \ , \tag{G.11}$$

where A_{JE} is the area of the emitter junction.

The collector current is primarily the electron current injected into the base, minus any losses due to recombination during diffusion. The collector transport factor

$$\alpha_T = \frac{\text{electron current reaching collector}}{\text{electron current injected from emitter}} = \frac{\left.\dfrac{dn_p}{dx}\right|_{x=W_B}}{\left.\dfrac{dn_p}{dx}\right|_{x=0}} \ . \tag{G.12}$$

Using the above result this becomes

$$\alpha_T = \frac{1}{\cosh \dfrac{W_B}{L_n}} \ . \tag{G.13}$$

Recalling that the base width is to be much smaller than the diffusion length, this expression can be approximated as

$$\alpha_T \approx 1 - \frac{1}{2}\left(\frac{W_B}{L_n}\right)^2 \ . \tag{G.14}$$

In modern devices $\alpha \approx 1$. The resulting collector current is the diffusion current of electrons in the base times the transport factor

$$I_C = J_{nB} A_{JE} \alpha_T \ . \tag{G.15}$$

One of the most interesting parameters of a bipolar transistor is the direct current gain (DC gain), which is the ratio of collector current to base current

$$\beta_{DC} = \frac{I_C}{I_B} \ . \tag{G.16}$$

Using the above results

$$\beta_{DC} = \frac{q_e D_{nB} \dfrac{n_i^2}{N_{aB} W_B} e^{q_e V_{BE}/kT}}{q_e D_{pE} \dfrac{n_i^2}{N_{dE} W_E} e^{q_e V_{BE}/kT}}$$

$$\beta_{DC} = \frac{N_{dE}}{N_{aB}} \cdot \frac{D_{nB} W_E}{D_{pE} W_B} \ . \tag{G.17}$$

Primarily, the DC gain is determined by the ratio of doping concentrations in the emitter and the base. This simple result reflects the distribution of current in the forward biased diode between electrons and holes. Transistors with high current gains always have much higher doping levels in the emitter than in the base.

The result of this idealized analysis implies that for a given device the DC gain should be independent of current. In reality this is not the case. Although modern transistors show a roughly constant current gain over decades of current, the current gain falls off both at very low and at high currents.

The decrease at low currents is due to recombination in the base–emitter depletion region. Since the most efficient recombination centers are near the middle of the bandgap, we set $E_t = E_i$ to obtain the recombination rate

$$\frac{dN_R}{dt} = \sigma v_{th} N_t \frac{pn - n_i^2}{n + p + 2n_i} \ . \tag{G.18}$$

Due to quasi-equilibrium in the space charge region

$$pn = n_i^2 e^{q_e V_{BE}/kT} \ . \tag{G.19}$$

This gives us the product of electron concentrations, but we also need the individual concentrations.

For a given forward bias the maximum recombination rate will coincide with the minimum concentration, *i.e.*

$$d(p + n) = 0 \tag{G.20}$$

and since $pn = \text{const}$,

$$dp = -dn = \frac{pn}{p^2} dp \tag{G.21}$$

or

$$p = n \tag{G.22}$$

at the point of maximum recombination. Hence, the carrier concentrations are

$$pn = n^2 = p^2 = n_i^2 e^{q_e V_{BE}/kT} \tag{G.23}$$

or

$$n = p = n_i e^{q_e V_{BE}/2kT} \ . \tag{G.24}$$

Inserting these concentrations into the expression for the recombination rate eqn G.18 yields

$$\frac{dN_R}{dt} = \sigma v_{th} N_t \frac{n_i^2 (e^{q_e V_{BE}/kT} - 1)}{2n_i (e^{q_e V_{BE}/2kT} + 1)} \ . \tag{G.25}$$

This maximum recombination rate will not prevail throughout the depletion region, but only over a region where the potential changes no more than $\sim kT/q_e$.

If we assume that the average field is V_{BE}/W_{BE}, a suitable averaging distance is

$$w = \frac{kT}{q_e}\frac{W_{BE}}{V_{BE}} \ . \tag{G.26}$$

Nevertheless, for simplicity assume that recombination is uniform throughout the depletion region. For $V_{BE} \gg kT/q_e$ this yields the recombination current to be made up by holes from the base current

$$J_{pr} = q_e \frac{dN_R}{dt} \approx \frac{1}{2} q_e \sigma v_{th} N_t n_i W_{BE} e^{q_e V_{BE}/2kT} \ . \tag{G.27}$$

Now the base current becomes

$$I_B = (J_{pE} + J_R) A_{JE} \ , \tag{G.28}$$

so the DC gain with recombination is

$$\beta_{DC} = \frac{q_e D_{nB} \dfrac{n_i^2}{N_{aB} W_B} e^{q_e V_{BE}/kT}}{q_e D_{pE} \dfrac{n_i^2}{N_{dE} W_E} e^{q_e V_{BE}/kT} + \dfrac{1}{2} q_e \sigma v_{th} N_t n_i W e^{q_e V_{BE}/2k_B T}} \ . \tag{G.29}$$

Because of the factor 1/2 in the exponent of the recombination term

$$\beta_{DC} = \frac{D_{nB} \dfrac{n_i^2}{N_{aB} W_B}}{D_{pE} \dfrac{n_i^2}{N_{dE} W_E} + \dfrac{n_i}{2\tau_0} W_{BE} e^{-q_e V_{BE}/2kT}} \ , \tag{G.30}$$

which can be rewritten in a more informative form as

$$\frac{1}{\beta_{DC}} = \frac{1}{\beta_0} + \frac{N_{aB} W_{BE}}{D_{nB}} W_B \frac{N_t \sigma v_{th}}{n_i e^{q_e V_{BE}/2kT}} \ , \tag{G.31}$$

where β_0 is the DC gain without recombination. Increasing the concentration of traps N_t decreases the current gain β_{DC}, whereas decreasing the base width W_B reduces the effect of traps. Since a smaller base width translates to increased speed (reduced transit time through the base), fast transistors tend to be less sensitive to trapping.

The current dependence enters through the exponential, which relates the injected carrier concentrations to the base–emitter voltage V_{BE}. With decreasing values of base–emitter voltage, *i.e.* decreasing current, recombination becomes more important. This means that the DC gain is essentially independent of current above some current level, but will decrease at lower currents.

Furthermore note that – although not explicit in the above expression – the "ideal" DC gain depends only on device and material constants, whereas the recombination depends on the local density of injected electrons and holes with

respect to the concentration of recombination centers. Consider two transistors with different emitter areas, but operating at the same current, so the larger transistor operates at a lower current density. The larger area transistor will achieve a given current at a lower base–emitter voltage V_{BE} than the smaller device, which increases the effect of trapping. Thus, the relative degradation of DC gain due to recombination depends on the current density.

This means that within a given fabrication process, a large transistor will exhibit more recombination than a small transistor at the same current. Stated differently, for a given current, the large transistor will offer more recombination centers for the same number of carriers.

Measured current gains *vs.* current are shown in Chapter 6. Data in Chapter 7 show the degradation of current gain after irradiation. Radiation damage increases the concentration of trapping sites N_t proportional to fluence Φ, so the above equation can be rewritten to express the degradation of current gain with fluence for a given current density (fixed V_{BE}):

$$\frac{1}{\beta_{DC}} = \frac{1}{\beta_0} + K\Phi \ , \tag{G.32}$$

where the constant K encompasses the device constants and operating point. In a radiation-damaged transistor the reduction in current gain for a given DC current will be less for smaller devices and furthermore faster transistors (small base width) will be less sensitive to radiation damage.

Why does the DC gain drop off at high currents?

1. With increasing current the high field region shifts towards the collector, effectively increasing the base width ("Kirk effect").
2. At high current levels the injected carrier concentration becomes comparable with the bulk doping. This reduces the injection efficiency and at very high current densities leads to bandgap narrowing.
3. Voltage drops in base and emitter resistance reduce the effective base–emitter and collector voltages.
4. Auger recombination.

For a discussion of these effects see Sze (1981).

References

Grove, A.S. (1967). *Physics and Technology of Semiconductor Devices*. Wiley, New York. ISBN 0-471-32998-3

Sze, S.M. (1981). *Physics of Semiconductor Devices* (2nd edn). Wiley, New York. ISBN 0-471-05661-8, TK7871.85.S988 1981

INDEX

^{60}Co irradiation, 303
3D detector, 305

ABCD readout IC, 350–353
absorption, photon, 23
acceptor, 58
accumulation layer, 428
activation, 420
active edge, 306
ADC, 5, 196
  channel profile, 197
  charge sensing, 5
  conversion time, 202, 204, 205, 207, 208
  dead time, 202
  differential nonlinearity, 199–201, 205, 207
    correction DAC, 206
  flash, 5, 203
  integral nonlinearity, 201
    vs. pulse shape, 202
  odd-even effect, 201
  parameters, 196
  pipelined, 208
  ramp discharge, 206
  rate effects, 202
  resolution, 197
  sigma-delta, 209
  sliding scale, 205
  stability, 203
  subranging, 208
  successive approximation, 204
  Wilkinson ADC, 206
aliasing, 213–214
amorphous silicon, 84
amplifier, 31–35, 91–101, 434–446
  bandwidth, 439–440
  bipolar transistor, 222–229
  capacitive source, 92
  cascode, 259–264, 412–413
  charge-sensitive, 93–101, 261
    input impedance, 100
    noise, 123–125
  common-base, 227
  common-collector, 228
    output resistance, 229
  common-emitter, 222
  composite, 256
  current, 434
  current sensitive, 91
  differential, 224–225, 257–258
  equivalent input noise, 145
  feedback, 98, 438–446
  folded cascode, 264, 412–413
  frequency response, 95–100, 439–440
  gain–bandwidth product, 97, 260, 262, 440
  linearity, 439
  loop gain, 263
  mode *vs.* input time constant, 91
  MOSFET, 242–243
  multi-stage, 98
  noise, 117–125, 129–133
    capacitive source, 123
    quantum limit, 132–133
  noise model, 117
  output impedance
    vs. feedback, 442–443
  phase response, 96–99
  pole, 263
  rise time, 36, 180
  time response, 98
  transconductance, 434
  transresistance, 434
  types, 434
  voltage, 434
  voltage sensitive, 91
amplifier stability criteria, 444–446
analog-to-digital conversion – see ADC, 5, 196–210
AND, 191
annealing
  anti-annealing, 286–288
  beneficial, 287, 288
  flash, 420
anti-aliasing filter, 213, 350
anti-annealing, 286–288
APD, 18, 88–90
APV25 readout IC, 348–350
astronomical imaging, 366
ATLAS pixel detector, 357–363
  layout, 357
  noise, 361
  readout, 360
  readout IC, 359–363
  sensors, 358
ATLAS SCT, 345–346, 352–356
  detector module, 353–355
    temperature distribution, 354
  layout, 345
  readout IC, 352–353

AToM readout IC, 323–326
avalanche
  breakdown, 18
  gain, 87
  noise, 88
  photodiode, 88–90

BaBar SVT, 320–327
  cabling, 326
  layout, 322
  readout IC, 323–326
  support, 322
backside readout, 131
ballistic deficit, 137–138
band bending, 239
bandgap, 12, 48–52, 84, 85
bandgap engineering, 86
bandwidth, 439–440
  noise *vs.* signal, 120
baseline restorer, 175
batch size, 425
beam pipe, 8, 319, 321, 323, 326, 338
Belle, 320
beneficial annealing, 287
bias current, temperature dependence, 284
bias resistors
  integrated, 427
binary readout, 38, 169, 350–353
  required S/N, 351–352
bipolar transistor, 217–222
  amplifier, 222–229
  current gain, 218, 220, 474
    vs. current density, 477
    vs. fluence, 477
    low current, 476
    vs. trapping, 476
  dimensions, 219
  doping levels, 219
  frequency response, 222
  input resistance, 227
  noise, 248–252
  output resistance, 226
  parasitic in CMOS, 429
  radiation effects, 292–295
  transconductance, 223
  transit frequency f_T, 222
Bode plot, 446
Boltzmann approximation, 448, 450, 451, 454
breakdown, 18, 88
built-in potential, 14, 15, 59
bulk material, 418–419
bump bonding, 357
bypass capacitor, 401–404, 414–415

C-V plot, 66
cable
  impedance, 386
  polyimide ribbon, 397
  propagation delay, 387
  reflections, 386
  self shielding, 397
cabling, SVT, 326
calibration circuitry
  ATLAS pixel detector, 360
  ATLAS SCT, 353
  BaBar SVT, 325
capacitance
  backplane, 16
  diode depletion region, 15, 65–66
  fringing, 15
  interstrip, 16, 428
capacitive matching, 246, 252–255, 269–270
capacitors, 414–415
  ceramic dielectrics, 414
  impedance *vs.* frequency, 414
  microphonics, 414
  series inductance, 414
  series resistance, 414
carrier concentrations, 447–451
  doped crystals, 450–451
carrier generation, 465–467
carrier lifetime, 82, 290–292
carrier velocity, 12–13, 67–69, 231–232
cascode amplifier, 259–264
CCD, 11, 24, 330–337, 366–368
  astronomical imaging, 366–368
  fully depleted, 367
  ILC, 333
  multiple readouts, 332
  output circuit, 335
  peak supply current, 335
  power dissipation, 335
  radiation resistance, 337
  readout speed, 334, 335
  signal charge, 330
  thinned, 331, 366
  tiling, 368
CDF SVX detector, 337–339
  readout IC, 339–341
channel profile, 197
channeling, 419
charge calibration, 94
charge collection, 9, 16–17, 67–83
charge coupled device, 24
charge, induced, 72–82
  parallel plate geometry, 75–77
  strip detector, 78–82
charge-sensitive amplifier, 93–101, 261

charge trapping, 82–83
chemical vapor deposition, 420
circuits
  digital, 191–196
  equivalent, 22, 32, 117–118, 142–145, 226, 243–245, 248–249, 434–437
clock interpolation, 209
closed loop gain, 439
CMOS, 428–429
  device structure, 428–429
  inverter, 193–194
  parasitic bipolar transistor, 429
  power dissipation, 195
CMOS imager, 27, 363–364
  radiation effects, 364
  tiling, 363
CMS pixel detector, 363
CMS tracker, 356
  layout, 346
^{60}Co irradiation, 303
collection time, 17–18, 68–71
  overbias, 70–71
  partial depletion, 17, 69–70
common mode choke, 408
common mode currents, 408
common mode noise, 326
common mode rejection, 224–225, 258
common-base amplifier, 227
common-collector amplifier, 228
compensated material, 288
complex phase notation, 432–433
compound semiconductors, 85
concentration
  equilibrium, 449
  free carrier, 447–451
    doped crystals, 450–451
  intrinsic, 449
conditioning, sensor, 376
conflicts and compromises, 315–316
contact, ohmic (non-rectifying), 420
conversion time, 202, 204, 205, 207
convolution, 135, 159, 160
corner parameters, 373
correlated double sampling, 153, 160–166, 340
count rate effects, 202
CR-nRC shaper, 136–137
CR-RC shaper, 134, 135
CR-RC shaper, 134–138
cross-coupled noise, 129–132
crosstalk, detector signal, 101
crystal
  indices, 44
  orientation, 44, 346, 348, 419
current
  CCD peak supply, 335
  diffusion, 451, 467
  diode
    reverse saturation, 61
  drift, 451
  emitter current density, 293
  generation, 465–468, 470
  reverse bias, 16, 468
    annealing, 284
    temperature dependence, 16
  shared paths, 398–405
  signal, 71–82
    pad detector, 82
    strip detector, 81
current gain, 218, 474
  differential, 222
  direct, 222
  vs. fluence, 477
current injection, 399
current mirror, 262
current *vs.* voltage mode, 125
Czochralski-grown Si, 26, 419
  detectors, 286

DØ silicon detector, 339–341
  readout IC, 339–341
damage constant, bias current, 284
dangling bonds, 237
dark current, 16
data acquisition, 40, 333
dead time, 202
deconvolution pulse processor, 348–350
defect engineering, 286
deformation, VXD3 ladder, 332
delay line pulse shaper, 164
DEPFET, 28, 364–365
depletion
  capacitance, 65–66
  electric field, 62
  full, 15, 68, 70–71
  partial, 15, 69–70
  region, 13, 59–66
  voltage, 16, 62–63
  width, 15, 62, 64
depletion region, 59–66
depletion voltage, 16, 62–63, 68
  vs. strip pitch, 426
deposition, 420
detection limits, 29
detector
  3D, 305
  active adge, 306
  ATLAS SCT, 345
  BaBar SVT, 321
  back-to-back, 305

Belle, 321
breakdown, 88
capacitance *vs.* voltage, 66
CCD, 330–337
signal charge (VXD3), 330
CDF SVX, 337–341
CMS silicon tracker, 345
CMS tracker, 346–350
Czochralski silicon, 419
diamond, 306
diode structure, 14
double-sided, 66
fabrication, 422–425
forward biased operation, 305
GLAST tracker, 368–369
hadron colliders, 337
hybrid, 39
integrated capacitors, 427
internal gain, 87
ladder, 319
materials, 83–86
amorphous silicon, 84
CdZnTe, 86
compound semiconductors, 85
diamond, 86, 306
GaAs, 84
polycrystalline, 86
silicon carbide, 306
table, 85
module, 39
n-on-n, 305
noise model, 127
noise summary, 166
pixel, 11–12, 24–29, 330–337, 357–368
signal cross-coupling, 101
silicon carbide, 306
strip
double-sided, 9, 66
single-sided, 9, 66
structure, 66, 151, 426–428
strip or pixel structure, 66
structure, diode, 14
VXD3, 330–337
detector ladder
CCD VXD3, 331
deformation, VXD3, 332
detector module
ATLAS pixel detector, 357–358
material, 363
ATLAS SCT, 353–355
material, 355
BaBar, 322–323
material, 323
temperature distribution, 354
VXD3 ladder, 332
detector process flow, 423–425
device fabrication, 418–425
DI water, 425
diamond, 86
charge collection length, 83, 86
lattice, 44
radiation effects, 306
dielectric materials, 414, 427
dielectric relaxation time, 70
dielectric, multilayer, 427
differential amplifier, 224–225, 257–258
differential nonlinearity, 199–201, 205, 207
correction DAC, 206
differential signal transmission, 355, 369, 406–408
differentiator, 134
diffusion, 419
constant, 452
transverse, 20
diffusion current, 467, 468
digital circuits, 191–196
digital signal processing, 210–216
digitization, on-chip
time over threshold, 325, 359
Wilkinson ADC (SVX ICs), 339
digitizer, 5, 196–210
diode
breakdown, 88
equation, 61
forward bias, 61
reverse bias, 61–66
reverse bias current, 16, 61, 83–85, 468–470
reverse bias voltage, 13, 61–66
reverse current, 470
"turn-on" voltage, 456
diode equation, 447–457, 468–470
direct transitions, 462
displacement damage, 278
comparison, 280
donor, 56
donor removal, 285
doping, 13, 56, 419–420, 450–451
acceptor, 58
energy level, 58, 469
activation, 420
diffusion, 419
donor, 56
energy level, 57, 469
wavefunction, 57
ion implantation, 419
n-type, 56
net concentration, 64
p-type, 57
profile, 90

doping distribution
  APD, 90
  bipolar transistor, 219
dose rate, 295

Early voltage, 225
edge-on sensors, 372
effective doping level, 285
Einstein relation, 20, 67, 452
electron–hole pairs, 12, 47–52
electronic noise, 29–35, 107–126
electronics
  ATLAS pixel readout, 358–363
  ATLAS SCT readout, 352–353, 355
  BaBar readout, 323–327
  CDF SVX readout, 339–341
  CMS tracker readout, 348–350
  DØ silicon detector, 339–341
  readout, 29
  VXD3 readout, 332–333
emission and capture processes, 281, 460
emitter follower, 228
  output resistance, 229
  self oscillation, 229
enclosed geometry MOSFETs, 302
energy
  acceptor level, 58, 469
  bands, 44–47, 60–61
    vs. bias, 61
  donor level, 57, 469
  excitation, 43
  intrinsic level, 451
  ionization, 48–52
    vs. bandgap, 51
    high energy quanta, 51
    low energy photons, 50
    SiO_2, 282
    table, 85
  phonon, 52
    average in Si, 54
energy bands, 44–47, 60–61
epitaxial (epi) layer, 27
equilibrium shield, 303
equivalent circuit, 434–437
  input, 436
  noise, 32, 142
  output, 437
  signal, 43
equivalent input noise voltage, 244
equivalent noise charge, 31–35, 116
  bipolar transistor, current dependence, 249
  calculation, 141, 146–148
  evaluation, 138–142
  vs. shaping time, 34, 148
etching, 421
evaporator, 420
event rate, LHC, 342

fabrication
  detector, 422–425
  device, 418–425
  semiconductor device, 418–429
failure modes, single-point, 374
Fano factor, 19, 52–55
Faraday shield, 395
feedback, 438–446
  linearity improvement, 439
  self-oscillation, 444–446
  series, 440–441
  shunt, 440–442
Fermi level, 48, 60
  doped crystals, 451
  quasi-, 454–455
Fermi–Dirac distribution, 447
FET, 229–241
  input impedance, 437
  noise, 243–248, 251–252
field effect transistor, 229–241
field line pinning, 394
field oxide, radiation effects, 283, 302
field, electric
  depletion region, 16–17, 62, 68
  overbias, 17, 70–71
  partial depletion, 16
filter
  anti-aliasing, 213, 350
  FIR, 213
  high-pass, 134
  low-pass, 134
finite impulse response (FIR) filter, 213
flash ADC, 203
flash annealing, 420
flex hybrid, 353
flip flop, 191
float zone silicon, 27, 286, 418
floating strips, 21
fluctuations
  number, 31, 107
  signal *vs* baseline, 106
  velocity, 31, 107
folded cascode, 264, 412–413
forward biased detector, 305
FOXFET biasing, 428
FPGA, 196, 213
frequency response, 439–440, 443–446
  peaking, 446
frequency, unity gain, 97, 260, 440
fully depleted detector, 15, 70–71

GaAs, 84
  FETs, 295, 296

radiation resistance, 277, 281, 295, 296
gain, open and closed loop, 439
gain–bandwidth product, 97, 260, 262, 440
Gaussian noise distribution, 31
generation current, 465–468, 470
gettering, 419, 423
effect on bias current, 423
GLAST, 368–369
"ground loops", 398–405
ground plane
patterned, 410
potential distribution, 403
ground, safety, 327
grounding, 326, 398–405
guard ring, detector, 14, 64, 427
guard ring, shield, 395

HDI, 323
high-pass filter, 134
hole, 13, 58
formation, 47
trapping, 282
hybrid, 39

ideality factor, diode, 470
imaging
microdischarge, 376
x-ray, 369–372
impact ionization, 87
improvements, device technology, 271
impurities, energy levels, 468
indirect transitions, 463
induced charge, 72–82
input impedance, 100
FET, 437
series feedback, 441
shunt feedback, 441, 442
input resistance
bipolar transistor, 227
integral nonlinearity, 201
integrated bias resistors, 427
integrated capacitors, 427
integration
electro–mechanical, 6–8
monolithic of detectors and electronics, 24–29
integrator, 134
interface traps, 296–298, 419
International Linear Collider, 333
interpolation
capacitive, 21, 321–322
charge, 21, 321–322
interstrip capacitance, 428
interstrip isolation, 428
intrinsic carrier concentration, 449
intrinsic energy level, 451
inversion
moderate, 267–270
strong, 240, 241, 267, 268, 272
weak, 239–241, 267–270
inverter
CMOS, 193–194
NMOS, 192–193
PMOS, 192–193
ion implantation, 419
ionization coefficient, 87, 88
ionization damage, 278, 282
ionization energy, 12, 13, 48–52
SiO_2, 282
table, 85
irradiation, ^{60}Co, 303
isolation
bias supplies, 409
common mode currents, 408–409
interstrip, 66
junction, 302
oxide, 302
trench, 429

JFET, 230–236
gate noise current, 244
noise, 243, 244
output curves, 232, 234
transconductance, 234
Johnson noise, 109
junction field effect transistor (JFET), 230–236

kTC noise, 144, 336

ladder, detector, 331
Landau distribution, 19
latch, 192
latchup, 429
lattice
constant, 44
diamond, 44
law of mass action, 450
layout
ATLAS pixel detector, 357
ATLAS SCT, 345
BaBar SVT, 322
barrel and disk, 337
CDF SVXII, 338
CMS tracker, 346
DØ silicon detector, 339
MarkII, 319
LHC, 342–344
event rate, 342
lifetime
carrier, 82, 290–292

detector, 289
light leaks, 389
linear collider, 333
linearity, amplifier, 439
local potential referencing, 416
local series feedback, 443
logic
arrays, 195
CMOS, 193–194
FPGA, 196
functions, 191
inverter, 192–194
power dissipation, 195
symbols, 192
synchronous, 195
timing, 192
logic symbols, 193
"long tailed pair", 224–225
loop gain, 263, 443
Lorentz deflection, 346
low dose enhancement, 295
low power, optimization, 266–275
low-pass filter, 134
LVDS, 408

Mark II vertex detector, 319
mask, 421
mass action, 450
material
ATLAS pixel detector, 363
BaBar detector module, 323
bulk semiconductor, 419
DØ silicon detector, 339
SCT detector module, 355
VXD3 ladder, 332
materials
bulk semiconductor, 418
detector, 83–86
amorphous silicon, 84
CdZnTe, 86
compound semiconductors, 85
diamond, 86
GaAs, 84
polycrystalline, 86
table, 85
materials, detector, 83–86
medical imaging, 148
MESFET, 296
metal oxide field effect transistor (MOSFET), 236–242
metallization, 420
microdischarge, 376, 426
effect of strip geometry, 427
microphonics, 390
mid-gap states, 468
minimum noise
bipolar transistor, 249–251
vs. power, 271
FET, 247
vs. power, 269
minority carrier
injection, 464
lifetime, 465, 470
mobility, 12–13, 67
vs. field, 69
mobility–lifetime product, 82
moderate inversion, 267–270
module, detector, 39
monolithic active pixel sensor, 26
monolithic integration of detectors and electronics, 24–29
MOS accumulation, 239
MOS capacitor, 237
MOS depletion, 239
MOSFET, 236–242
amplifier, 242–243
enclosed geometry, 302
gate noise current, 245
noise, 243
saturation voltage, 240
subthreshold regime, 241
types – NMOS and PMOS, 241
$\mu\tau$ product, 82
multilayer dielectric, 427
mythology, grounding, 398–405

n-on-n detectors, 289, 346, 358
n-well, 429
NAND, 191
nanotechnology, 272
N_{eff}, 285
components, 286
temperature dependence, 286
time dependence, 289
NIEL, 279
noise
vs. capacitance, 33
vs. shaping time, 34
amplifier, 117–125, 129–133
capacitive source, 123
charge-sensitive, 123–125
current mode, 125
input noise, 118, 145
noise model, 117
quantum limit, 132–133
ATLAS pixel detector, 361
avalanche, 88
BaBar SVT readout, 324
backside readout, 131
vs. bandwidth, 110, 119
bias current, 143
bipolar transistor, 248–252

vs. current, 249, 265
detector module, 265
optimum, 249
post radiation, 295
vs. power, 271
vs. capacitance, 146–148, 150, 152, 166–169
capacitive matching, 246, 252–255, 269–270
CCD, 331
combined probability function, 171
common mode, 326
comparison bipolar *vs.* FET, 251
complex sensors, 127
contributions in detector module, 265
corner frequency, 248, 251
correlated, 115
correlated double sampling, 160–166
$1/f$ noise, 164–166
cross-coupled, 129–132
DEPFET readout, 365
detector, summary, 166
electronic, 29–35
equivalent circuit, 32
equivalent input noise, 118
equivalent input voltage, 244
equivalent noise charge, 31–35, 116
calculation, 141, 146–148
CR-RC shaper, 147
evaluation, 138–142
estimate, eqn for, 35, 250–251
FET, 243–248, 251–252
$1/f$, 248
frequency domain analysis, 135–136, 138–147
gated integrator, 158
Gaussian distribution, 31
JFET, 243, 244
gate noise current, 244
kTC, 144, 336
long strip detectors, 327–329
low frequency, $1/f$, 109–110, 113–114
low-frequency, $1/f$
MOSFET and JFET, 248
matching, 121
transformer, 122
measurement
oscilloscope, 140
rms voltmeter, 139
spectrum analyzer, 139
minimum
bipolar transistor, 249–251
FET, 247
MOSFET, 243
gate noise current, 245
post radiation, 299–300
vs. power, 269
"noiseless" resistances, 114
occupancy, 173
noise measurement, 174
occupancy *vs.* efficiency, 174
parallel resistance, 144
power supply, 389
quantization, 214
rate, 171
"series" and "parallel", 32
shape factors, 33, 158
vs. shaping time, 146–148, 150, 152, 166–169
shot, 31, 109
AC coupling, 148
slope, 168
spectral density
low frequency, $1/f$, 113–114
shot noise, 111
thermal (Johnson), 110
SVX3 readout IC, 341
SVX4 readout IC, 341
thermal (Johnson), 109
vs. time constant, 136
time domain analysis, 153–166
transistor, 35
VXD3, 331
weighting function, 156
white, 31
noise analysis
frequency domain, 135–136, 138–147
time domain, 153–166
noise corner frequency, 248, 251
non-rectifying contact, 420
nonionizing energy loss, 279
NOR, 191
normalized transconductance, 235
vs. channel length, 268, 272
vs. current density, 268
np product, 450
Nyquist criterion, 213
Nyquist plot, 446

occupancy, 173
odd-even effect, 201, 207
ohmic contact, 420
open loop gain, 439
OR, 191
OR, exclusive, 191
orientation, crystal, 419
output impedance
vs. feedback, 442–443
series feedback, 443
shunt feedback, 442
output resistance, 256
bipolar transistor, 226

cascode amplifier, 260
NMOS and PMOS, 257
PMOS
vs. current and channel length, 257
overbias, 17, 68, 70–71
"overdepletion", 17, 70
oversampling, 214
oxide, 422
oxide charge, 282, 283, 292
oxide isolation
bipolar transistor, 293
MOSFET, 283
oxide passivation, 66
oxygenated silicon, 287, 291, 292
oxygenation, 286

p-spray, 66, 428
parasitic bipolar transistor, 429
partial depletion, 15, 69–70
passivation
oxide, 422
PSG, 422, 429
surface, 422
patterning, 421–422
peak detector, 206
peaking time, 3
peaking, frequency response, 446
phase margin, 444
phasors, 432–433
phonon excitation, 52, 67
phosphosilicate glass, 422
photodiode, 26
avalanche, 18
readout, 148
photodiodes, 86–91
avalanche
reach-through, 88–90
photoelectric
absorption, 23
effect, 12
photoelectron, 23
photolithography, 421–422
photon absorption, 23
photoresist, 421
negative, 421
positive, 421
pickup
current injection, 399
inductive, 396
light, 389
microphonics, 390
power supply noise, 389
RF, 391
pile-up, 3
pinch off
JFET, 231
MOSFET, 239
voltage, 231
pinholes, 427
pipeline
analog, 340
digital, 324–325, 351
pipelined ADC, 208
pitch, strip, 9, 321–322, 337, 345, 348, 352, 369
pixel detector, 330–337, 357–368
astronomical imaging, 366–368
ATLAS, 357–363
CMOS imager, 363–364
CMS, 363
DEPFET, 364–365
VXD3, 330–337
pixel device, 11–12, 24–29
active, 26–29
hybrid, 11
monolithic, 24–29
random access, 11
planar process, 422
pn product, 450
pn-junction, 13, 59, 451–457
pole, 263
pole–zero cancellation, 177
polysilicon
doped, 420
gate, 429
position resolution
CCDs, 366
potential
built-in, 59
weighting
strip detector, 78
weighting (induced charge), 75
potential referencing, 416
power dissipation
ATLAS ABCD IC, 353
ATLAS pixel IC, 360
BaBar AToM IC, 324
CCD, 335
GLAST readout ICs, 369
GLAST tracker, 369
SVX4 readout IC, 341
power minimization *vs.* noise, 269–271
power supply noise, 389
power supply rejection ratio, 261
power, strip *vs.* pixel detector, 274
preamplifier, 3
prefilter, 158
process flow, detector, 423–425
propagation delay, 194
pseudo-rapidity, 342
PSG, 422
pulse height, 4

pulse shaper, 34
pulse shaping, 3, 32
  ballistic deficit, 137–138
  bipolar *vs.* unipolar, 178
  correlated double sampling, 153, 160–166
  CR-nRC, 136–137
  CR-RC, 4, 134–138
    noise, 147
  delay line, 164
  gated integrator, 158
  noise *vs.* shaping time, 148
  noise *vs.* time constants, 135–136
  sequence, 159
  time constants, 135
  time variant, 158–166
pulse stretcher, 206
punch-through biasing, 428

quantization noise, 214
quantum noise limit, 132–133
quasi-Fermi level, 454–455

radiation effects
  ^{60}Co irradiation, 303
  annealing, 288
  anti-annealing, 286–288
  ATLAS pixel system, 359
  BaBar ATom IC, 324
  bias current
    annealing, 284
    temperature dependence, 284
  bipolar transistor
    current density, 293
    current gain, 292–295
    f_T, 293
    low dose enhancement, 295
    oxide charge, 293
  CMOS imager, 364
  cryogenic operation, 286, 306
  crystal orientation, 348
  damage constant, bias current, 284
  deep submicron CMOS, 298
  defect engineering, 286
  diamond, 306
  diodes, 283
  displacement damage, 278
    threshold, 280
  donor removal, 285
  doping, effective, 285
  dose units, 279
  electronic circuits, 306
  emission and capture processes, 281
  GaAs, 277, 281
    FETs, 277, 295, 296
  IC isolation structures, 302
  ionization, 278, 292
  ionization damage, 282
  JFETs, 296
  mitigation in detectors, 304–306
  mobility, 291
  MOSFET
    enclosed geometry, 302
    noise, 299–300
    oxide charge, 296–298
    threshold shift, 298
    threshold shift *vs.* oxide thickness, 298
    transconductance, 299
  n-on-n detectors, 289, 346
  N_{eff}, 285
    components, 286
    temperature dependence, 286
    time dependence, 289
  noise
    bipolar transistor, 295, 296
    MOSFET, 299–300
  nonionizing energy loss, 279
  oxide charge, 282, 292
  oxygenated silicon, 287, 291, 292
  oxygenation, 286
  reverse bias current, 284
  Si JFETs, 295
  SVX4 readout IC, 341
  trapping, 282, 289
  type inversion, 285
radiation field at LHC, 344
radiation length, 8
Ramo's theorem, 73
rapidity, 342
rate effects, 202
rate of noise pulses, 171
reach-through photodiode, 88–90
readout
  ATLAS pixel IC, 360
  bussing, 36–38
  detector backside, 131
  VXD3, 333
readout IC
  ABCD, 350–351
  ATLAS pixel, 359–363
    threshold distribution, 360
  ATLAS SCT, 352–353
  BaBar SVT, 323–326
  CMS silicon tracker, 348–350
  size, 325, 339, 340, 350
  SVX2, 339–341
  SVX3, 341
  SVX4, 341
recombination, 462, 465
recombination rate, 464
reflections, 386

resistivity, 18, 65
resolution
  double-pulse, 353
  double-track, 330, 345
  energy, 19, 29, 318, 365, 368, 370, 371
    intrinsic in Ge and Si, 55
    scintillator *vs.* Ge detector, 105
  impact parameter, 7
  position, 7, 19
    ATLAS pixels, 360
    ATLAS SCT, 346
    BaBar, 320
    CCD, 331
    CMS tracker, 348
    interpolation, 21
    Lorentz deflection, 346
  time, 324, 348
  z, *vs.* dip angle, 321
reverse bias current, 61, 83–85, 468
RF pickup, 391
ringing, 446
rise time
  vs. bandwidth, 36, 180
  cascaded amplifiers, 36, 180

safety ground, 327
sampling, 211, 213
saturation velocity, 69
scattering, multiple, 8
segmentation, 318
self oscillation, 386, 411, 446
semiconductor device technology, 418–429
sensor, 2
  conditioning, 376
  noise model, 127
  position sensitive, 9
  principle, 8
sensors
  ATLAS pixel, 358
  ATLAS SCT, 345
  BaBar, 321
  CDF SVX, 339
  CMS tracker, 348
  GLAST, 369
  long strip, 327–329
"series" and "parallel" noise, 32
series feedback, 440–441
series resonance, 414
shape factors, 33, 158
shaper
  frequency response, 415
  interference suppression, 415
shared current paths, 398–405
shielding, 392
  apertures, 393
  dielectric, 395–396
  field line pinning, 394
  guard ring, 395
Shockley equation, 61, 456
shot noise, 109
shunt feedback, 440–442
signal
  current
    pad detector, 82
    strip detector, 81
  detector
    cross-coupling, 101
  equivalent circuit, 43
  formation, 55
  processing, 134–188, 210–216
signal charge, 12
signal path
  control of, 401–405, 410–413
  detector, 412
  folded cascode, 412
  local loops, 401
  output driver, 401
signal polarity, 341
signal scan, 174
signal transmission
  differential, 355, 369, 406–408
signal-to-noise ratio, 3, 29
  vs. bandwidth, 108
  vs. input time constant, 125
silicide, 429
silicon drift chamber, 11, 25
silicon nitride, 427
single pole response, 263
single-point failure modes, 374
skin depth, 393
sliding scale, 205
small angle stereo, 339, 345
small-angle stereo, 10–11
smart pixels, 28
SNAP, 367
space charge, 13
space points, 331
sputtering, 420
stability criteria, 444–446
stacked sensors, 371
star ground, 405
stereo angle, 339, 345
strip detector
  long, 327–329
  structure, 427
strip detector structures, 426–428
strip pitch, 9, 321–322, 337, 345, 348, 352, 369
strong inversion, 240, 241, 267, 268, 272
successive approximation ADC, 204
support

BaBar SVT, 322
surface
accumulation, 239
depletion, 239
inversion, 239
surface passivation, 422
SVX
detector (CDF)
readout ICs, 338, 341
SVX detector (CDF), 337–339
readout IC, 339–341
SVX4 readout IC, 341
symbols, logic, 193

tail cancellation, 177
TDC, 209–210
analog ramp, 209
clock interpolation, 209
counter, 209
technology
semiconductor device, 418–429
technology, device improvements, 271
temperature
operating, 289
termination, cable, 388
test input, 94
ATLAS pixel detector, 360
ATLAS SCT, 353
BaBar SVT, 325
thermal noise, 109
thermal runaway, 304
threshold, 38
threshold discriminator systems, 169–175
minimum threshold, 173
noise rate, 172
threshold distribution, 360
threshold scan, 174
threshold shift, 298
tiling, 316, 378
ATLAS pixels, 358
CCDs, 368
CMOS imager, 363
time
dielectric relaxation, 70
jitter, 35, 180
stamp, 37
walk, 36, 182, 183
time over threshold
ATLAS pixel readout, 359
BaBar SVT, 325
time-to-digital conversion, 209–210
timing measurements, 35–36, 179–188
constant fraction timing, 185, 186
pulse shaping, 180
resolution *vs.* S/N, 187
results, 187
threshold, 183
zero crossing timing, 184
token passing, 37
ToT, 325
transconductance
bipolar transistor, 223
JFET, 234
NMOS, 241
normalized, 235
vs. channel length, 268, 272
vs. current density, 268
transistor
bipolar, 217–222
field effect, 229–241
JFET, 230–236
MOSFET, 236–242
transit frequency f_T, 222
transmission capacity, 191
trapping, 82–83, 282, 289
trench isolation, 429
trigger, 37
trim DAC, 353, 360
"turn-on" voltage, 456
type inversion, 285

undersampling, 213
unity gain frequency, 97, 260, 440

velocity
carrier, 12–13, 67
saturation, 18, 69
thermal, 68
vertex detection, 6
vertex detector
BaBar SVT, 320–327
Mark II, 319
VXD3, 330–337
virtual ground, 441
VXD3, 330–337
electronics, 332–333

wafer size, 286
water, deionized, 425
waveguide below cutoff, 394
weak inversion, 239–241, 255, 267–270
weighting potential, 75
Wilkinson ADC, 206

x-ray
detection, 26
detection efficiency, 370
imaging, 23, 369–372
spectroscopy, 151–153, 369–372

zone refining, 418