VI. Semiconductor Devices and Microelectronics

1. Bipolar Transistors

   Bipolar transistors in amplifiers

2. Field Effect Transistors

   Junction Field Effect Transistors (JFETs)

   Metal Oxide Semiconductor FETs (MOSFETs)

   MOSFETs in amplifiers

3. Noise in Transistors

   Noise in Field Effect Transistors

   Optimization of Device Geometry

   Noise in Bipolar Transistors

   Noise Optimization – Capacitive Matching

   Optimization for Low Power

4. Microelectronics

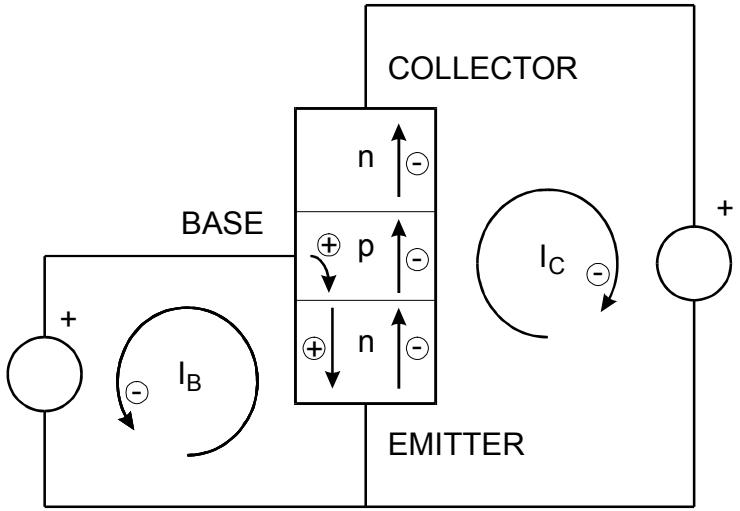
   Fabrication of Semiconductor Devices

   Integrated Circuits

For more details and an excellent physics explanation, see
S.M. Sze, *Physics of Semiconductor Devices*, Wiley-Interscience publication

1. Bipolar Transistors

Consider an npn structure



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. The magnitude of this current depends on the n-p doping ratio.

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.

In an ideal diode

$$\frac{I_C}{I_B} = \frac{I_{nBE}}{I_{pBE}} = \frac{D_n / N_A L_n}{D_p / N_D L_p} = \frac{N_D}{N_A} \frac{D_n L_p}{D_p L_n}$$

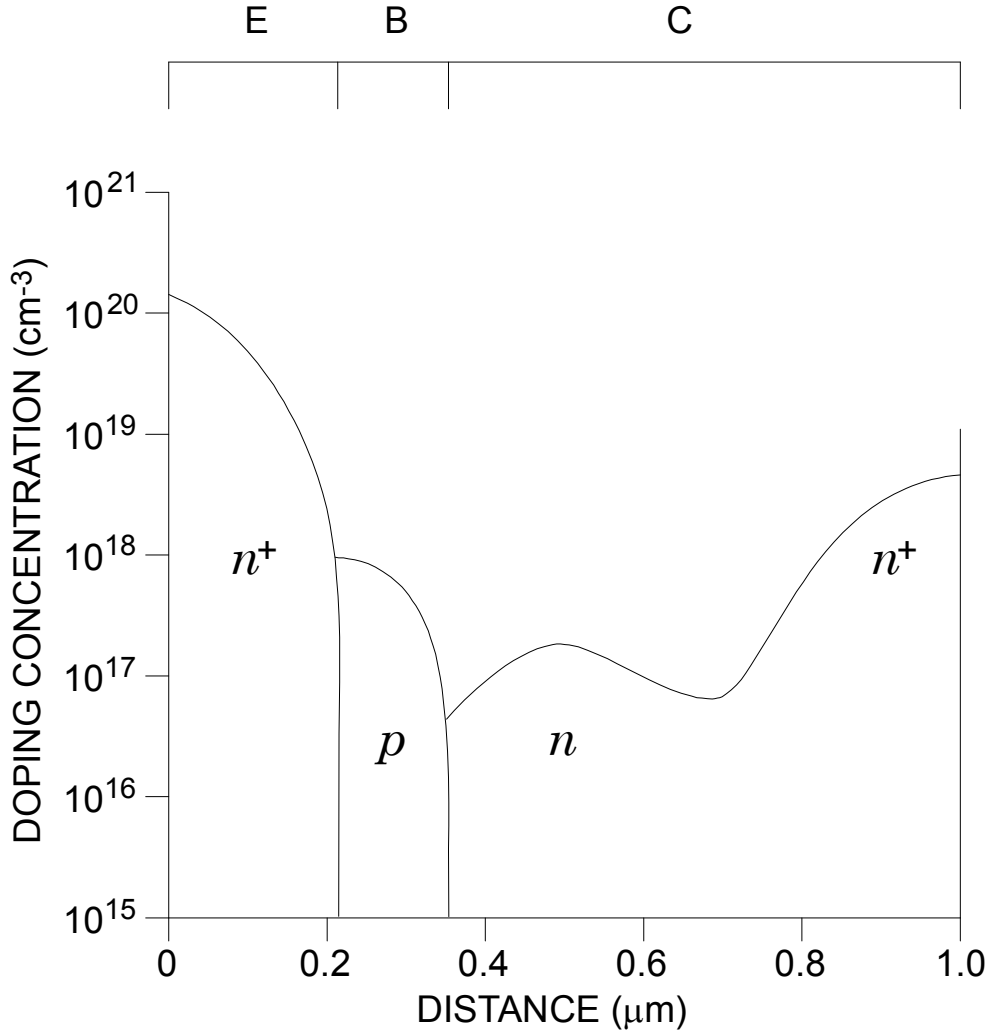
If the ratio of doping concentrations in the emitter and base regions N_D / N_A is sufficiently large, the collector current will be greater than the base current.

$\Rightarrow$ DC current gain

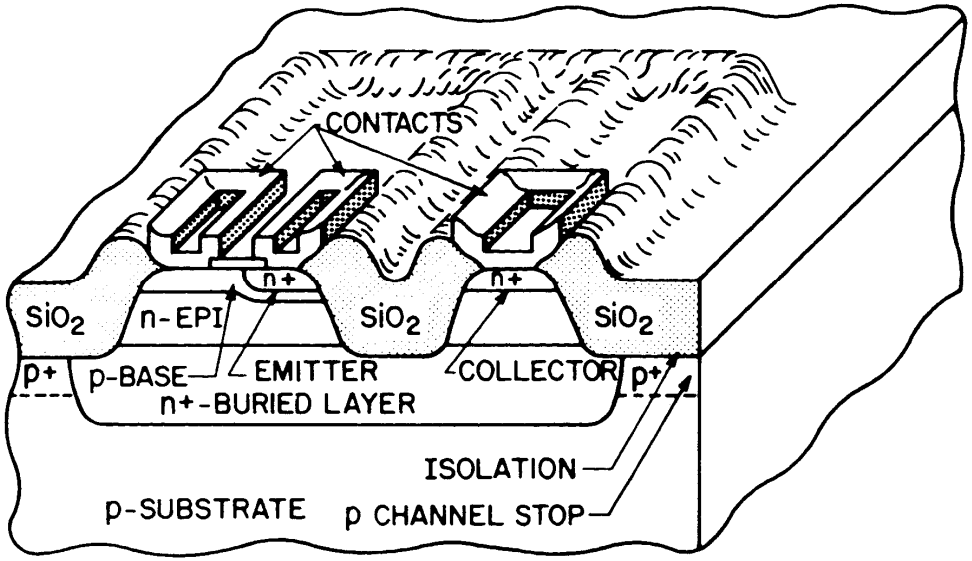
Furthermore, we expect the collector current to saturate when the collector voltage becomes large enough to capture all of the minority carrier electrons injected into the base.

Since the current inside the transistor comprises both electrons and holes, the device is called a bipolar transistor.

Dimensions and doping levels of a modern high-frequency transistor (5 – 10 GHz bandwidth)



High-speed bipolar transistors are implemented as vertical structures.



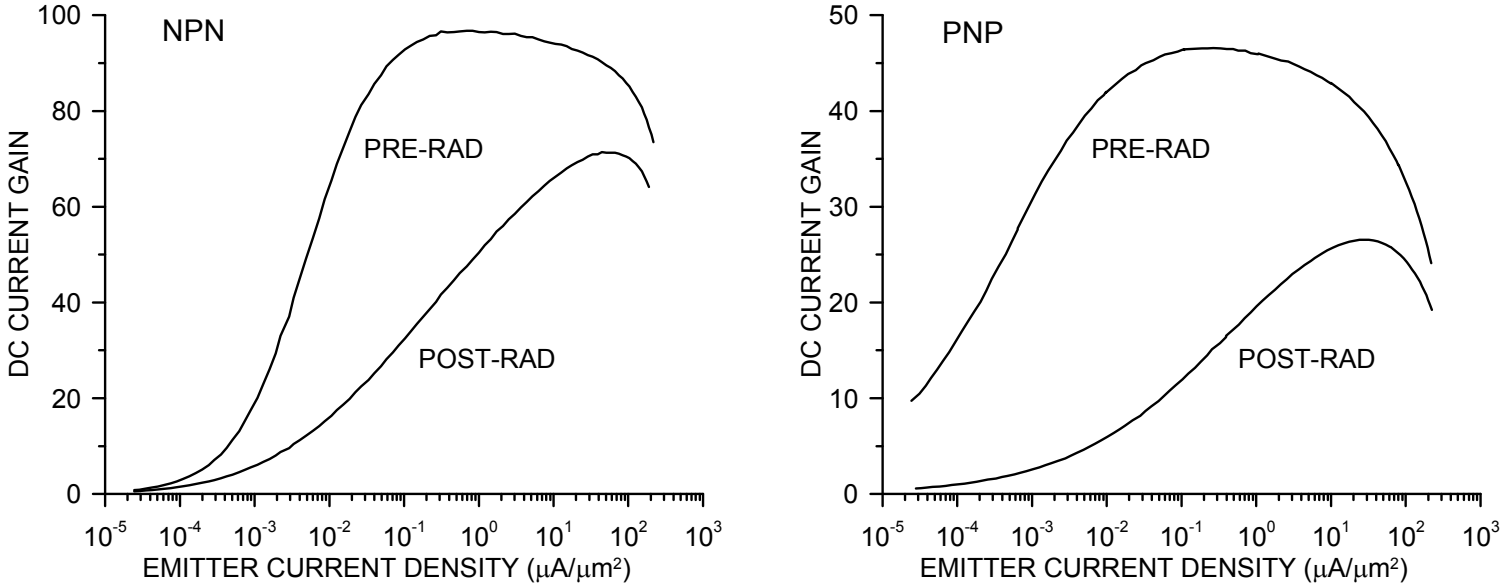
(From Sze 1981, ©Wiley and Sons, reproduced with permission)

The base width, typically 0.2 μm or less in modern high-speed transistors, is determined by the difference in diffusion depths of the emitter and base regions.

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.

As shown in the preceding figure, the collector region is usually implemented as two regions: one with low doping (denoted “epitaxial layer in the figure) and the other closest to the collector contact with a high doping level. This structure improves the collector voltage breakdown characteristics.

As shown in the plots below, modern devices exhibit DC current gain that is quite uniform over orders of magnitude of emitter current.



The figures also show the degradation of current gain after irradiation, here after exposure to the equivalent of 10^{14} minimum ionizing protons/cm^2 (discussed in next Section VII)

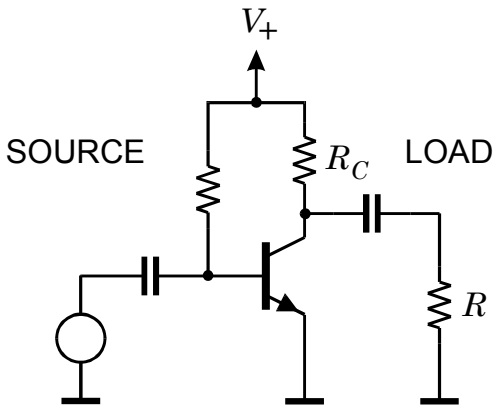
The decrease of DC current gain at low current densities due to increased recombination is apparent.

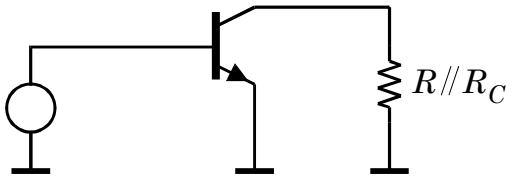
In a radiation-damaged transistor the reduction in DC current gain for a given DC current will be less for smaller devices.

Bipolar Transistors in Amplifiers

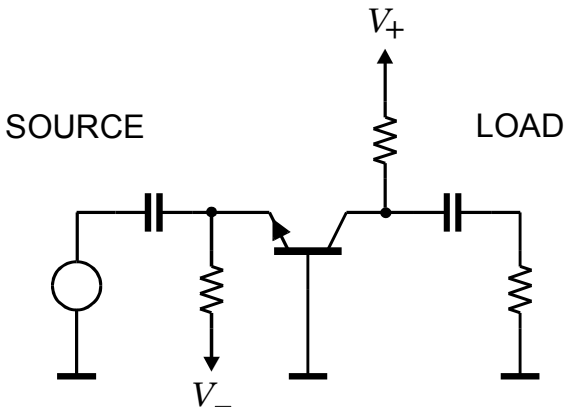
Three different circuit configurations are possible:

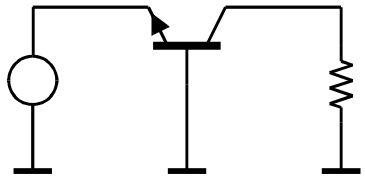
1. Common Emitter



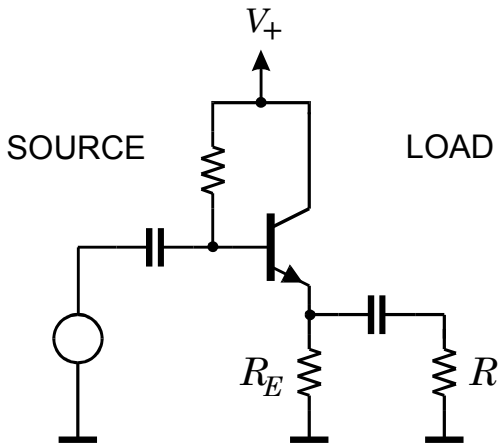


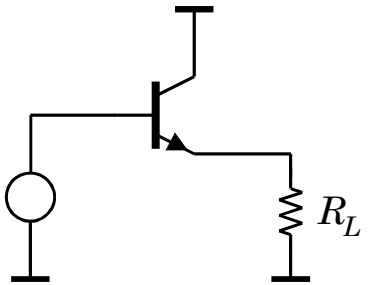
2. Common Base





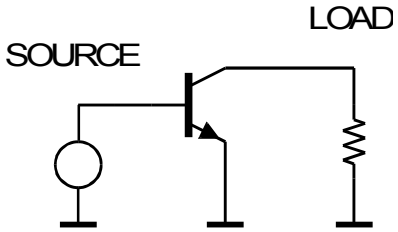
3. Common Collector (Emitter Follower)





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 (CE) configuration.



The voltage gain $$A_V = \frac{dV_{out}}{dV_{in}} = \frac{dI_C}{dV_{BE}} R_L = g_m R_L$$

Since the dependence of base current on base-emitter voltage is given by the diode equation

$$I_B = I_R(e^{q_e V_{BE}/k_B T} - 1) \approx I_R e^{q_e V_{BE}/k_B T}$$

the resulting collector current is $$I_C = \beta_{DC} I_B = \beta_{DC} I_R e^{q_e V_{BE}/k_B T}$$

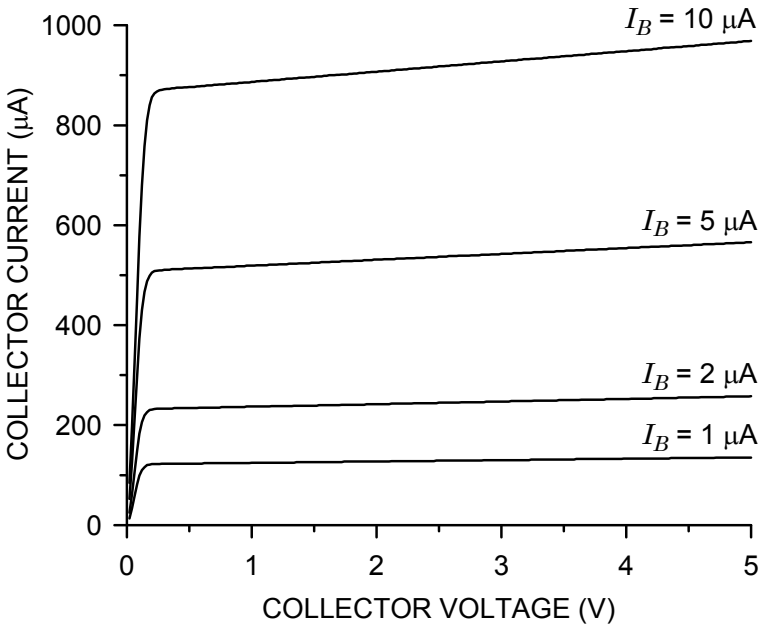
and the transconductance, i.e. the change in collector current vs. base-emitter voltage

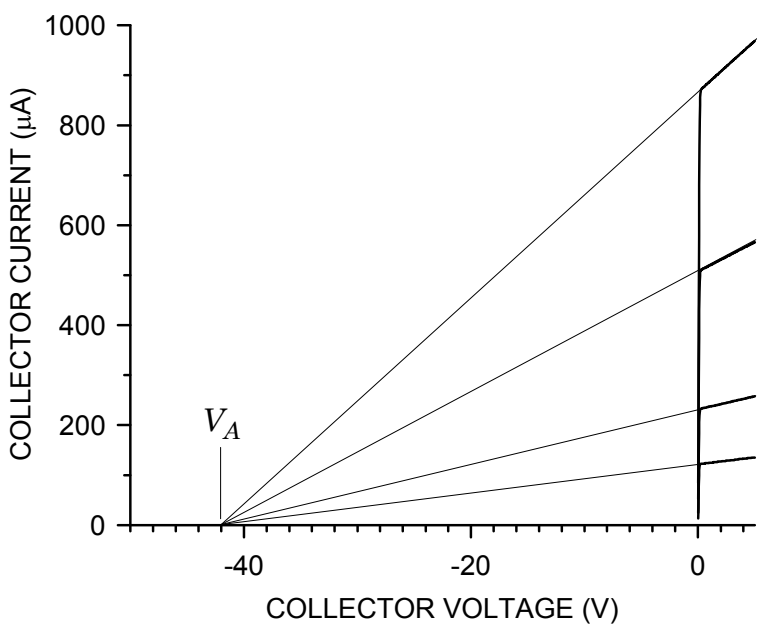
$$g_m \equiv \frac{dI_C}{dV_{BE}} = \beta_{DC} I_R \frac{q_e}{k_B T} e^{q_e V_{BE}/k_B T} = \frac{q_e}{k_B T} I_C$$

The transconductance depends only on collector current, so for any bipolar transistor – regardless of its internal design – setting the collector current determines the transconductance.

Since at room temperature $k_B T / q_e$ = 26 mV, $$g_m = \frac{I_C}{0.026} \approx 40 I_C$$

The obtainable voltage gain of an amplifier depends on the output characteristics of the transistor.

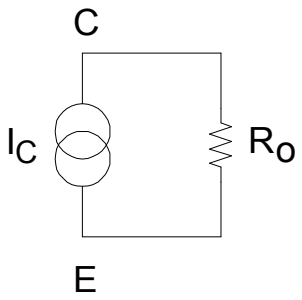




At low collector voltages the field in the collector-base region is not sufficient to transport all injected carriers to the collector without recombination. At higher voltages the output current increases gradually with voltage (saturation region), due to the change in effective base width.

An interesting feature is that the extrapolated slopes in the saturation region intersect at the same voltage V_A for $I_c = 0$ ("Early voltage").

The finite slope of the output curves is equivalent to a current generator with a shunt resistance



where $R_o = K\frac{V_A}{I_C}$

K is a device-specific constant of order 1, so usually it's neglected.

The total load resistance is the parallel combination of the external load resistance and the output resistance of the transistor. In the limit where the external load resistance is infinite, the load resistance is the output resistance of the amplifier.

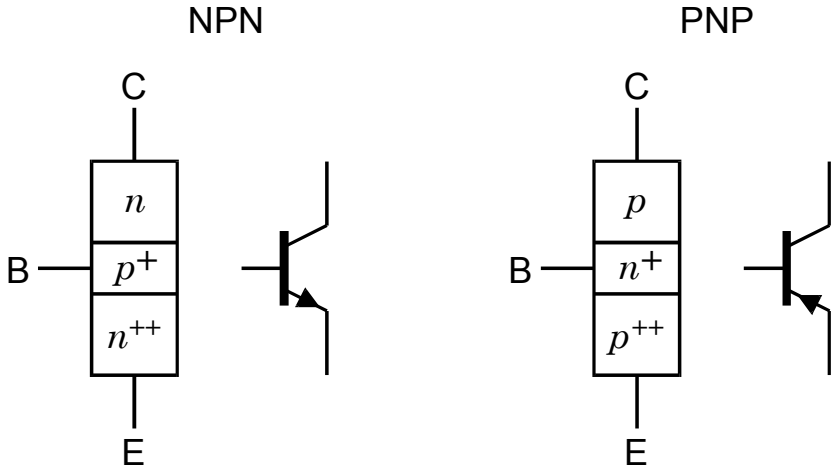
The maximum obtainable voltage gain is

$$A_{v,\max} = \frac{dI_C}{dV_{BE}} R_o = g_m R_o \approx \frac{I_C}{k_B T / q_e} \frac{V_A}{I_C} = \frac{V_A}{k_B T / q_e}$$

which at room temperature is about $40 V_A$.

- Note that to first order the maximum obtainable voltage gain is independent of current.
- Transistors with large Early voltages allow higher voltage gain.

Bipolar transistors can be implemented as npn or pnp structures.



The polarities of the applied voltages are the opposite:

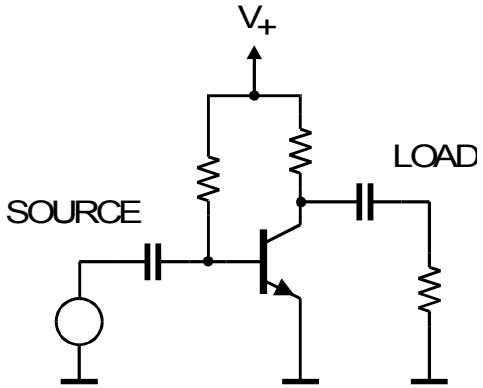
npn: positive collector-emitter and base emitter voltages

pnp: negative collector-emitter and base emitter voltages

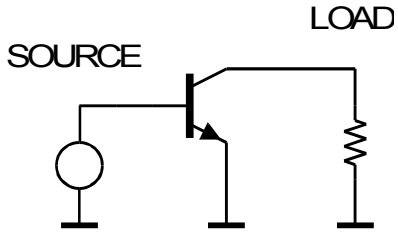
The basic amplifier equations are the same for both transistor types.

The availablility of complementary transistors offers great flexibility in circuit design.

a) Common Emitter Configuration



Equivalent Circuit



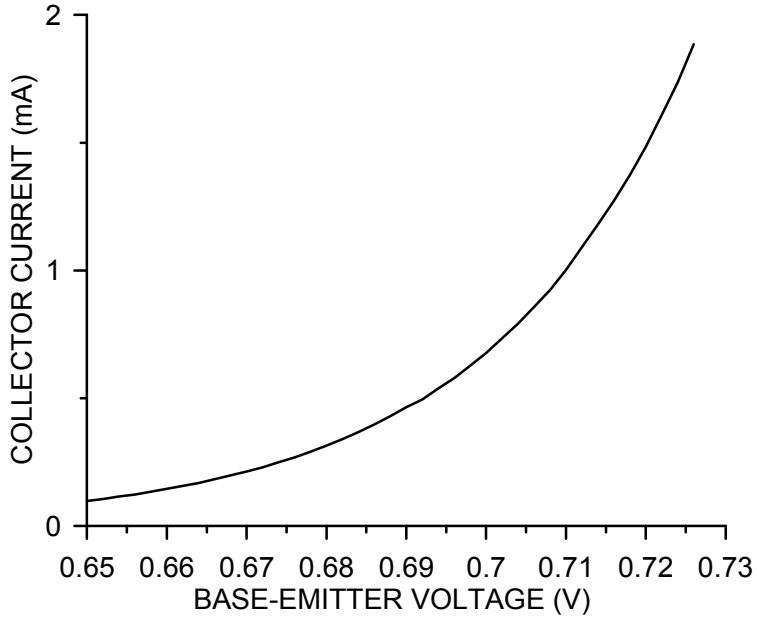
The input signal is applied to the base, the output taken from the collector.

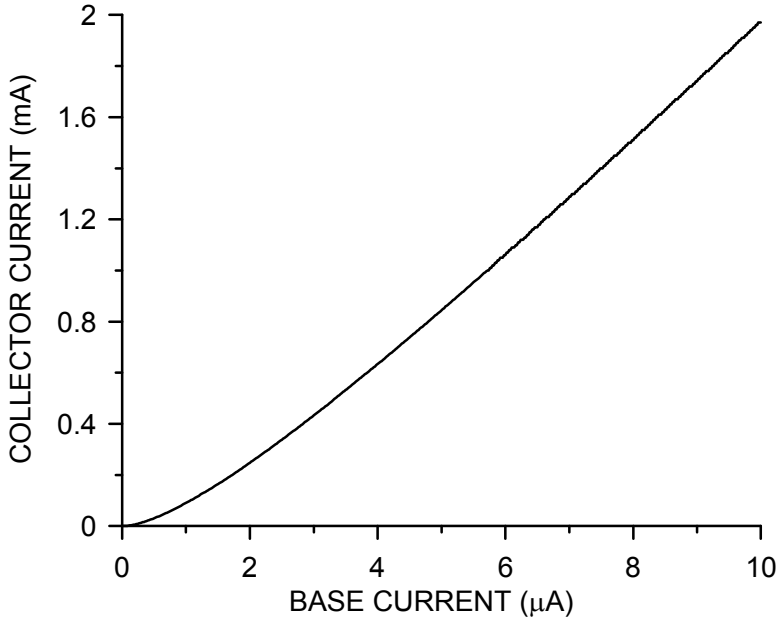
The input resistance is proportional to the DC current gain and inversely proportional to the collector current.

$$R_i = \frac{dV_{BE}}{dI_B} \approx \beta_{DC} \frac{dV_{BE}}{dI_C} = \frac{\beta_{DC}}{g_m} = \frac{k_B T}{q_e} \frac{\beta_{DC}}{I_C}$$

For β_{DC} = 100 and I_c= 1 mA, R_i= 2600 Ω.

Although the bipolar transistor is often treated as a voltage-driven device, the exponential dependence of base current on input voltage means that as an amplifier the response is very non-linear.



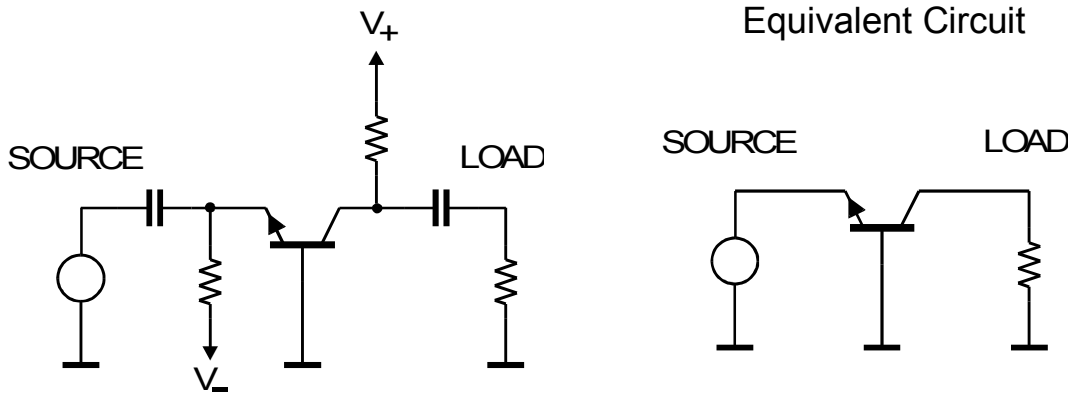


In audio amplifiers, for example, this causes distortion.

Distortion may be limited by restricting the voltage swing, which to some degree is feasible because of the high transconductance.

With current drive the linearity is much better.

b) Common Base Configuration



The input signal is applied to the emitter, the output taken from the collector.

This configuration is used where a low input impedance is required.

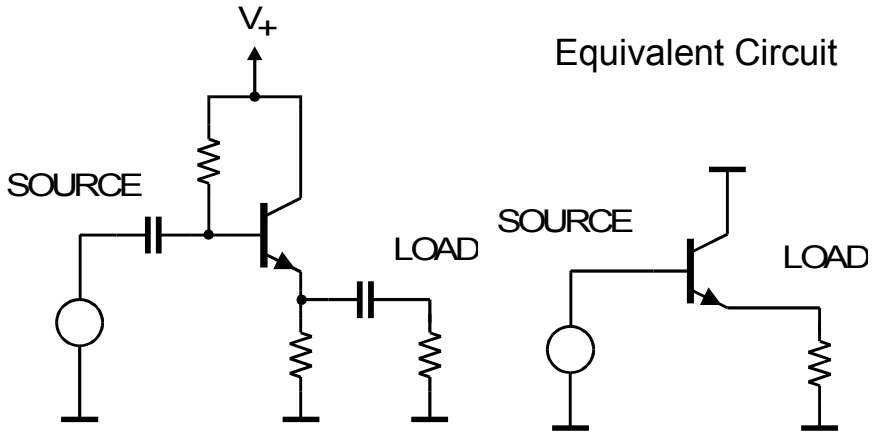
$$R_i = \frac{dV_{EB}}{dI_E} \approx \frac{dV_{EB}}{dI_C} = \frac{1}{g_m} = \frac{k_B T}{q_e}\frac{1}{I_C}$$

Since at room temperature $k_B T / q_e$ = 26 mV: $R_i = \frac{0.026}{I_C}$ i.e. R_i=26 Ω at I_c= 1 mA.

The input resistance is about $1/\beta$ times smaller than in the common emitter configuration.

c) Common Collector Configuration

The signal is applied to the base and the output taken from the emitter ("emitter follower").



The load resistance R_L introduces local negative feedback,

$$V_i = V_{BE} + I_E R_L \approx V_{BE} + \beta I_B R_L$$

Since V_{BE} varies only logarithmically with I_B, it can be considered to be constant (≈ 0.6 V for small signal transistors), so

$$\frac{dV_i}{dI_i} = \frac{dV_i}{dI_B} \approx \beta R_L$$

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 1.

The output resistance of the emitter follower is

$$R_o = -\frac{dV_{out}}{dI_{out}} = -\frac{d(V_{in} - V_{BE})}{dI_E} \approx \frac{dV_{BE}}{dI_E} \approx \frac{1}{g_m}$$

(since the applied input voltage is independent of emitter current, $dV_{in} / dI_E = 0$).

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.

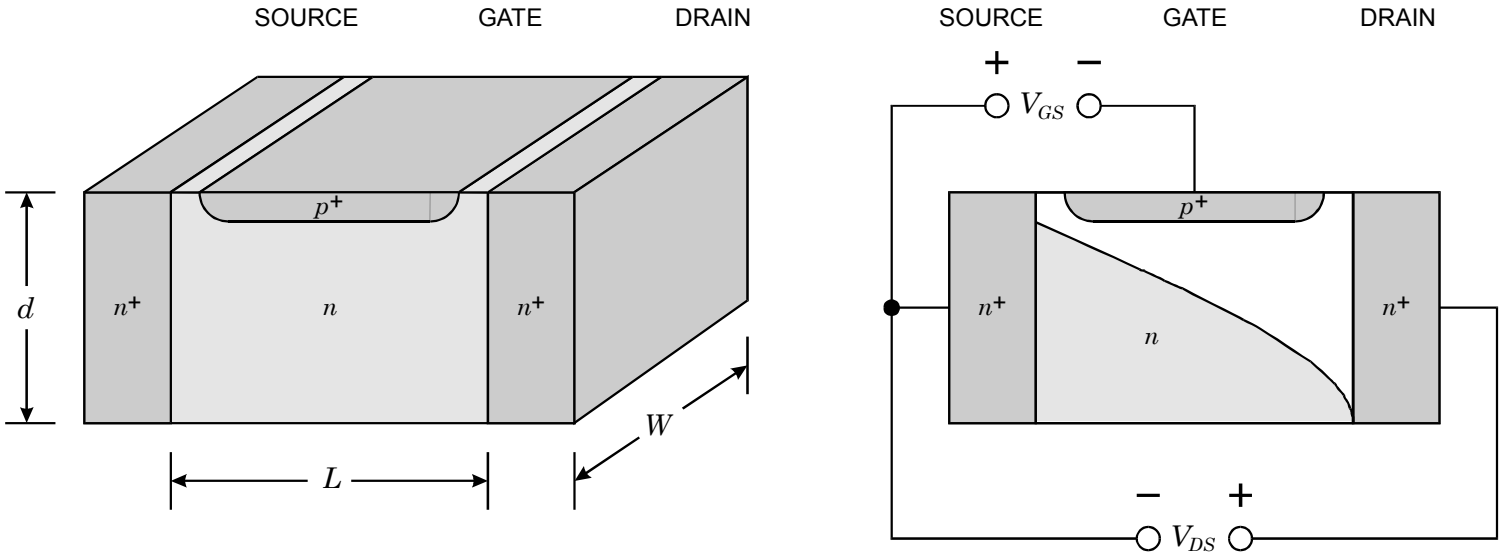
The high input resistance allows the preceding stage to have high gain and the low output impedance provides an output voltage source.

2. Field Effect Transistors

Field Effect Transistors (FETs) utilize a conductive channel that is controlled by an applied potential.

1. Junction Field Effect Transistor (JFET)

In JFETs a conducting channel is formed of n or p-type semiconductor (GaAs, Ge or Si). Connections are made to each end of the channel, the Drain and Source.

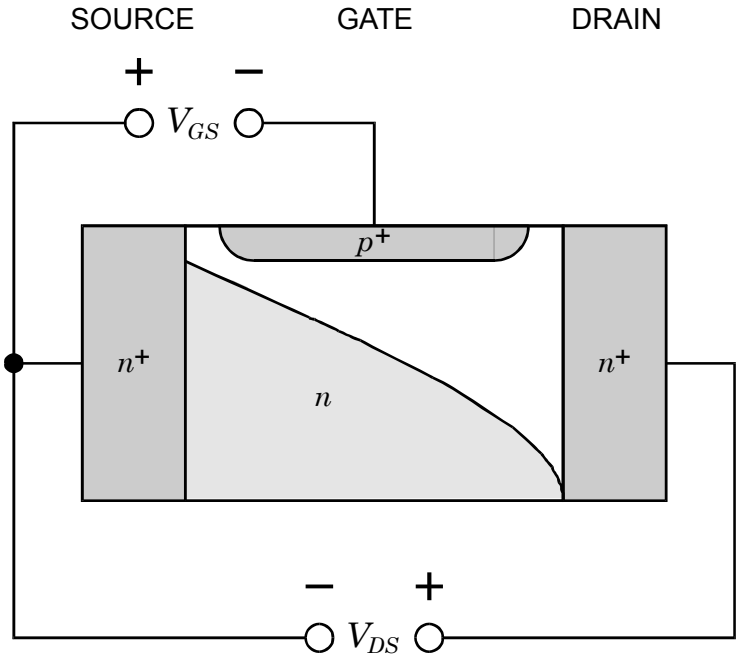


Applying a reverse bias to the gate electrode forms a depletion region that reduces the cross section of the conducting channel.

Changing the magnitude of the reverse bias on the gate modulates the cross section of the channel.

First assume that the drain voltage is 0.

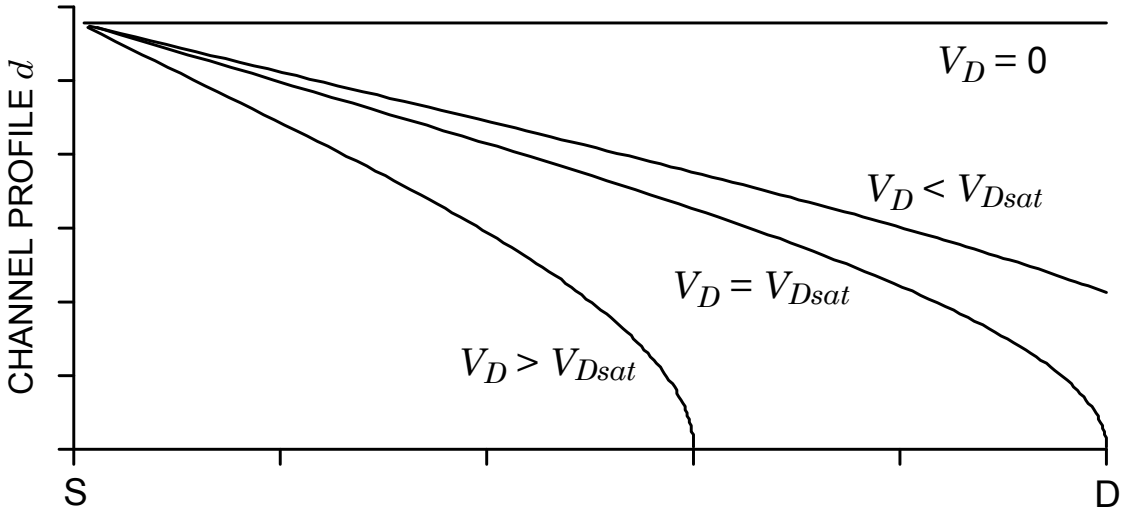
Increasing the reverse gate potential will increase the depletion width, i.e. reduce the cross section of the conducting channel, until the channel is completely depleted. The gate voltage where this obtains is the “pinch-off voltage” V_P.



Now set both the gate and drain voltages to 0. The channel will be partially depleted due to the “built-in” junction voltage.

Now apply a 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 voltage is sufficient to fully deplete the channel, the device is said to be “pinched off”.

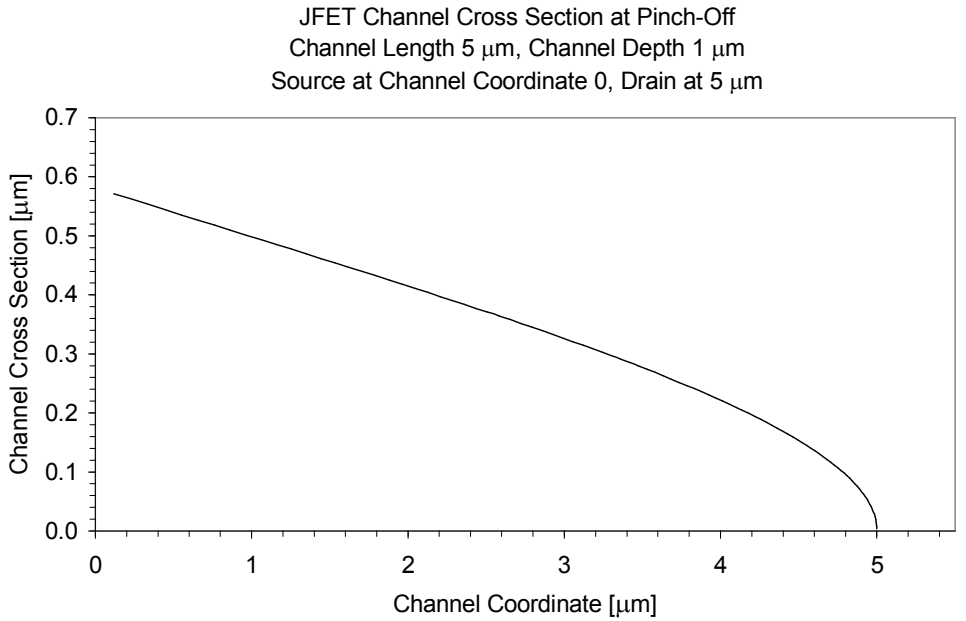


Increasing the drain voltage beyond pinch-off moves the pinch-off point towards to the source.

Pinching off the channel does not interrupt current flow.

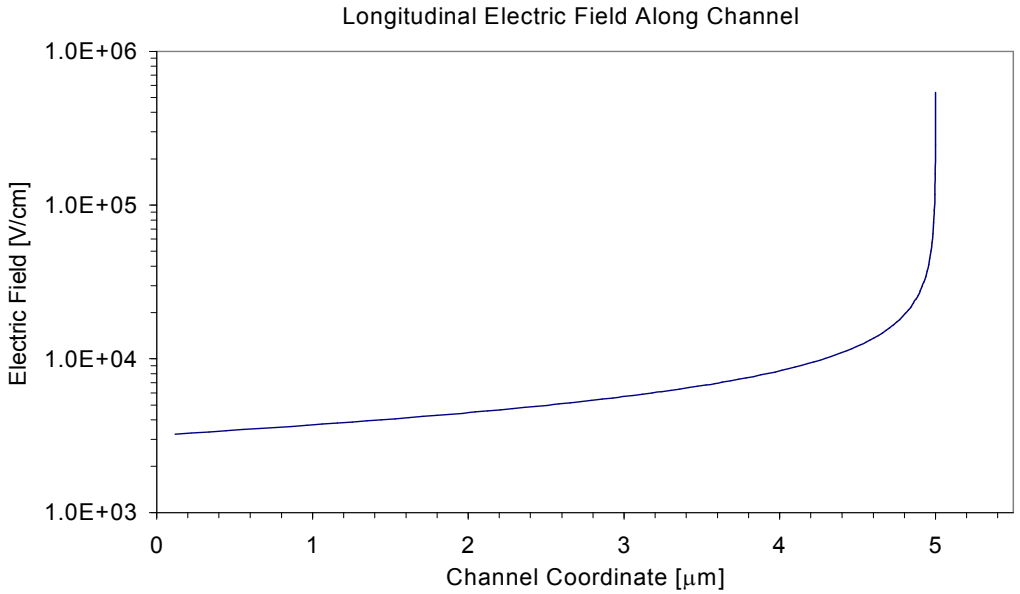
All thermally excited carriers have been removed from the depleted region, but carriers from the channel can still move through the potential drop to the drain.

The profile of the depletion region is not determined by the static potentials alone.



Current flow along the channel changes the local potential.

As the channel cross section decreases, the incremental voltage drop increases, i.e. the longitudinal drift field that determines the carrier velocity increases.



At high electric fields the mobility decreases.

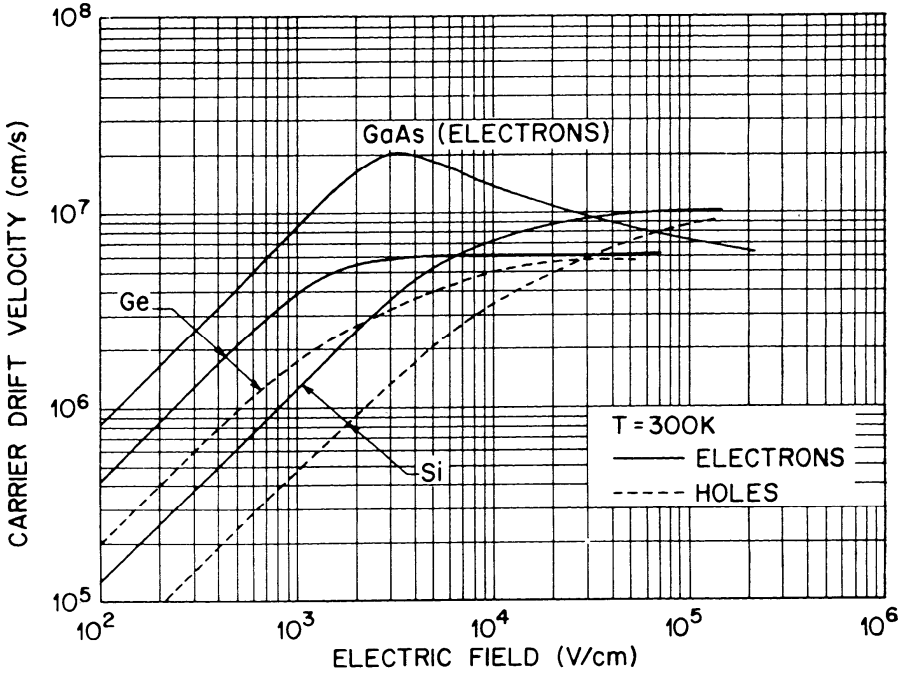
This comes about because as the carriers' velocity increases they begin to excite optical phonons. At fields above 10^5 V/cm practically all of the energy imparted by an increased field goes into phonon emission.

Since the velocity saturates at high fields, the current

$$I = N_C q_e v$$

also saturates, since the number of carriers N_C remains constant.

$\Rightarrow$ At high fields silicon acts as an incremental insulator ($dI/dV = 0$).



(From Sze 1981, ©Wiley and Sons, reproduced with permission)

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.

$\Rightarrow$ The current increases only gradually with drain voltage.

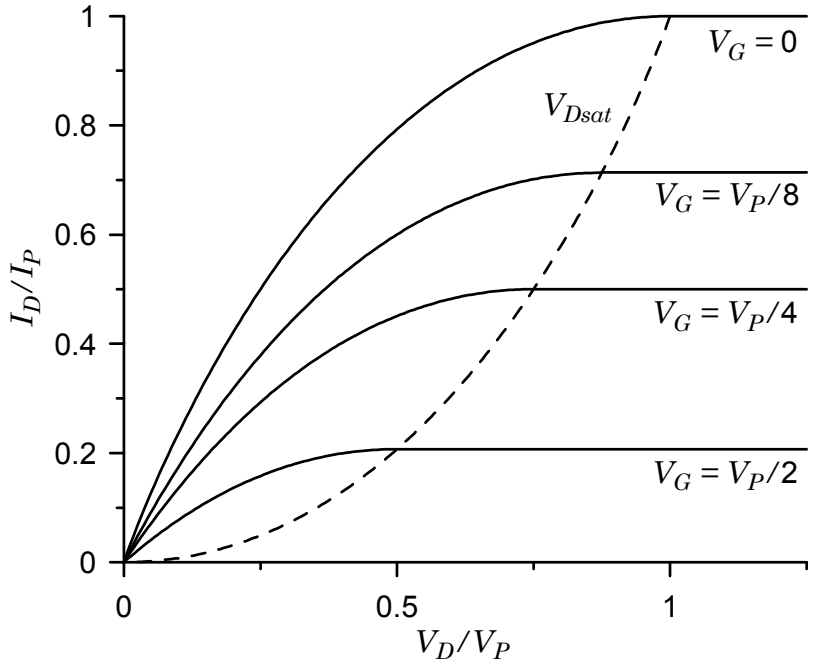
Current Voltage Characteristics

Low drain and gate voltages:

The resistive channel extends from the source to the drain.

The drain current increases linearly with drain voltage.

“Linear Region”



Gate and drain voltages sufficiently high to pinch off the resistive channel:

The drain current remains constant with increasing drain voltage.

“Saturation Region”

The drain saturation voltage V_{Dsat} increases as the gate voltage is changed from the static pinch off voltage V_P towards 0.

For use in amplifiers the characteristics in the saturation region are of the most interest.

To a good approximation $$I_D = I_{DSS}\left(1 - \left(\frac{V_G + V_{bi}}{V_P}\right)\right)^2$$

where V_p is the "pinch-off" voltage, i.e. the gate voltage at which the channel is fully depleted.

The drain saturation current $$I_{DSS} = \frac{1}{6\varepsilon}\mu(q_e N_{ch})^2 d^3 \frac{W}{L}$$

is determined by the carrier mobility μ, the doping level in the channel N_{ch} and the channel depth d, width W and length L. ε is the dielectric constant.

The transconductance $$g_m = \left|\frac{dI_D}{dV_G}\right| = \frac{2I_{DSS}}{V_P}\left(1 - \frac{V_G + V_{bi}}{V_P}\right)$$

is maximum for $V_G = 0$, i.e. maximum drain current, and for small pinch off voltages.

Then $$g_m\big|_{V_G=0} \approx \frac{2I_{DSS}}{V_P}$$

The transconductance depends primarily on current:

$$g_m = \frac{2I_{DSS}}{V_P}\left(1 - \frac{V_G + V_{bi}}{V_P}\right) = \frac{2\sqrt{I_{DSS}}}{V_P}\sqrt{I_D}$$

The applied voltages only provide the boundary conditions to set up the required current. To maintain performance it is important to control the current (rather than the voltages).

To see how device parameters affect the transconductance, we'll ignore the built-in voltage since it varies only weakly with doping

$$V_{bi} = (k_B T / q_e) \log(N_{ch} / n_i).$$

With this approximation $g_m\big|_{V_G=0} \approx \frac{2I_{DSS}}{V_P} \approx \frac{W}{L}\frac{\mu(q_e N_{ch})^2 d^3}{3q_e N_{ch} d^2} \propto \frac{W}{L}\mu N_{ch} d$

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 device in parallel, so the normalized transconductance g_m / W is used to compare technologies.
2. The normalized transconductance increases with the number of carriers per unit length $N_{ch} d$ and decreasing channel length L.
3. Transconductance increases with drain current

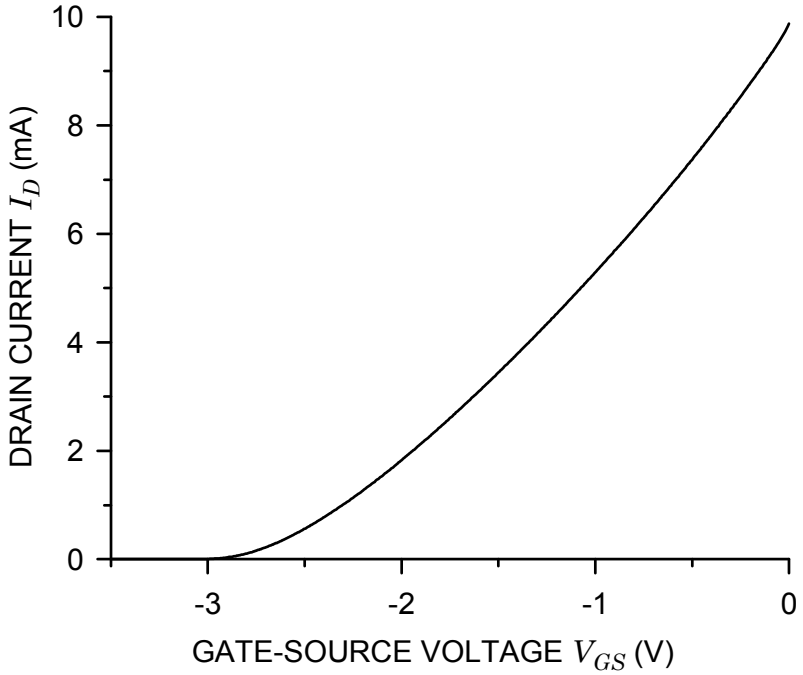
$$g_m = \left|\frac{dI_D}{dV_G}\right| = \frac{2I_{DSS}}{V_P}\left(1 - \frac{V_G + V_{bi}}{V_P}\right) = \frac{2\sqrt{I_{DSS}}}{V_P}\sqrt{I_D}$$

   i.e. drain current is the primary parameter; the applied voltages are only the means to establish I_D.

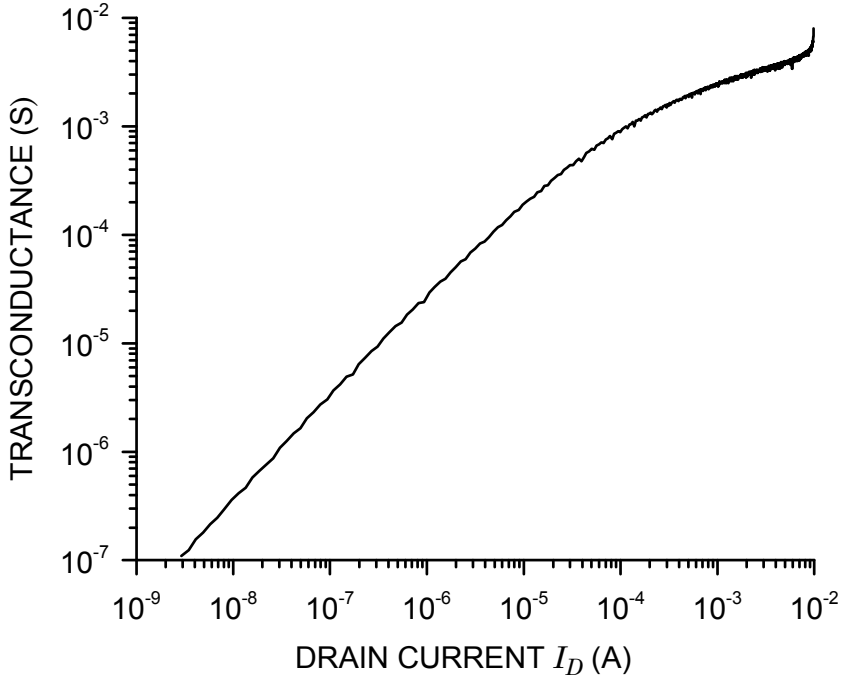
All of these optimizations also increase the power dissipation.
For low power systems optimization is more involved.

Measured JFET Characteristics

Drain current vs. gate voltage

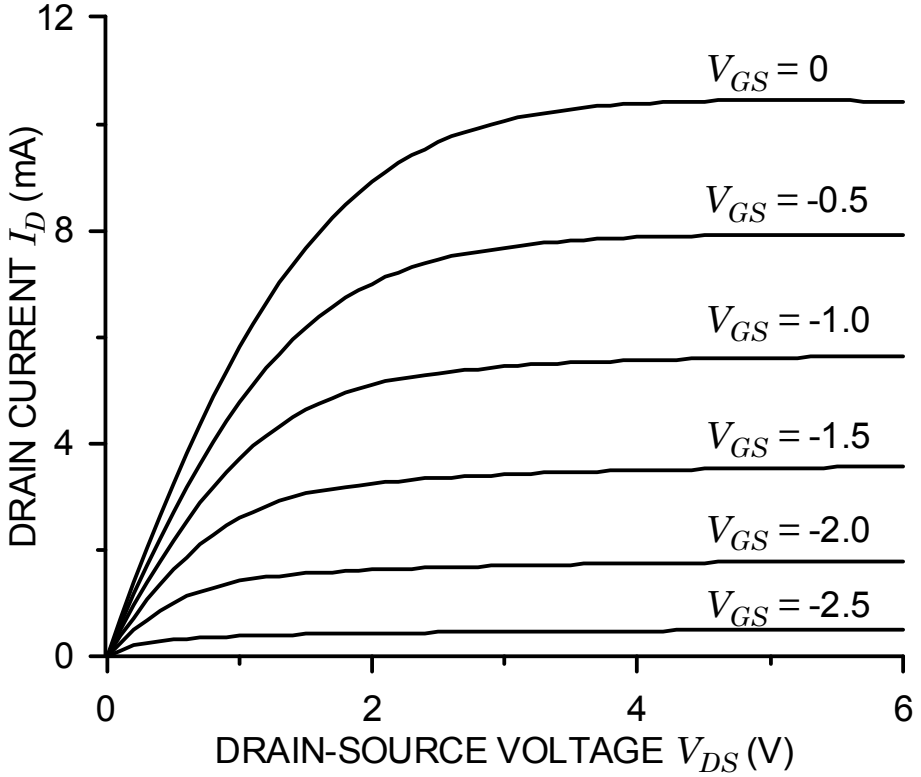


Transconductance



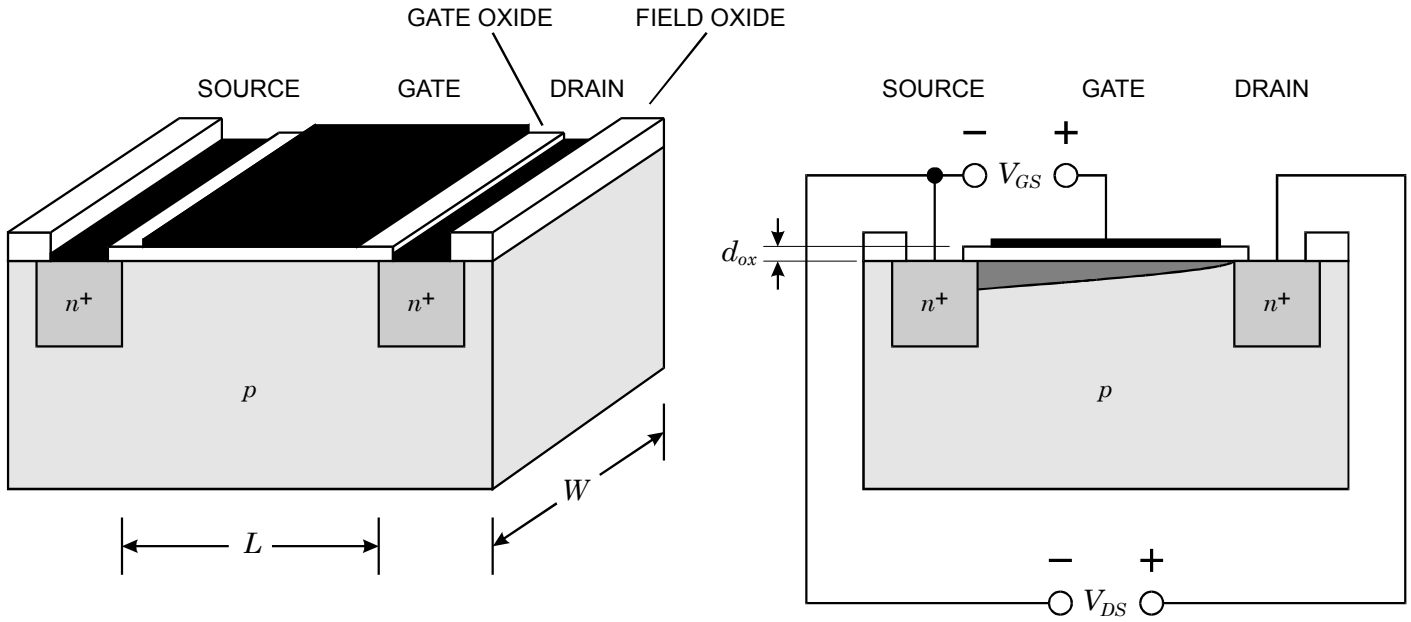
Measured JFET Characteristics

Output curves



Metal Oxide Semiconductor Field Effect Transistors (MOSFETs)

Unlike a JFET, where a conducting channel is formed by doping and its geometry modulated by the applied voltages, the MOSFET changes the carrier concentration in the channel.



The source and drain are n^+ regions in a p-substrate.

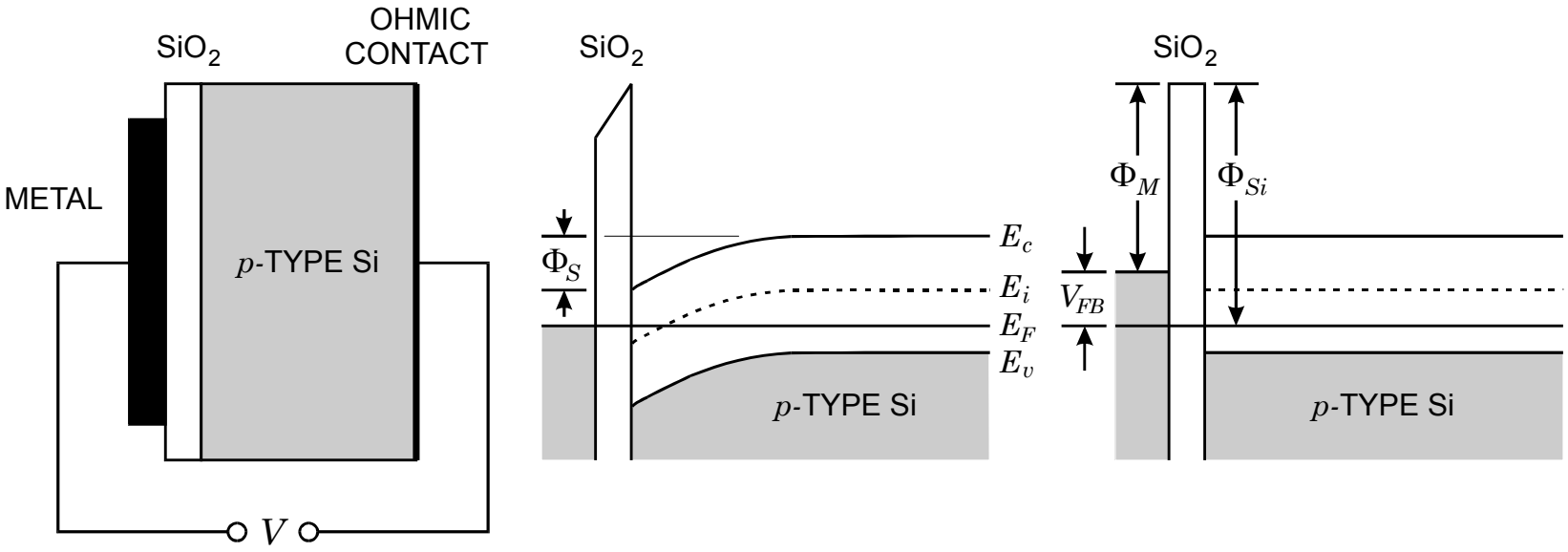
The gate is capacitively coupled to the channel region through an insulating layer, typically SiO_2.

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.

Formation of the Channel - The MOS Capacitor

Band structure in an ideal MOS capacitor on a p-substrate

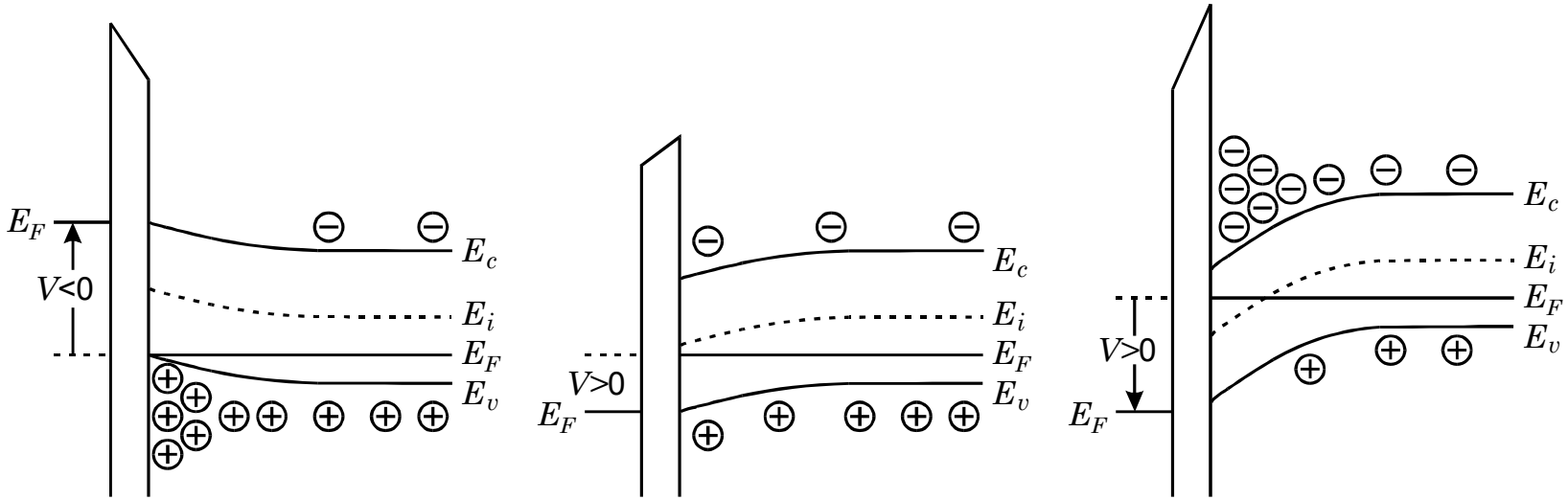


In its natural state, however, the band structure is not flat.

The discontinuity in the crystal structure and charge trapped at the surface change the potential at the surface, so the bands bend.

Application of the appropriate potential to the metal electrode can adjust the surface potential so that the bands are flat.

Surface concentration vs. band-bending in p-type material



ACCUMULATION DEPLETION INVERSION

If the surface potential Ψ_S is positive, electrons are attracted to the silicon surface.

Ψ_B is the difference between the external and intrinsic Fermi levels $E_F - E_i$. Flat for V = 0.

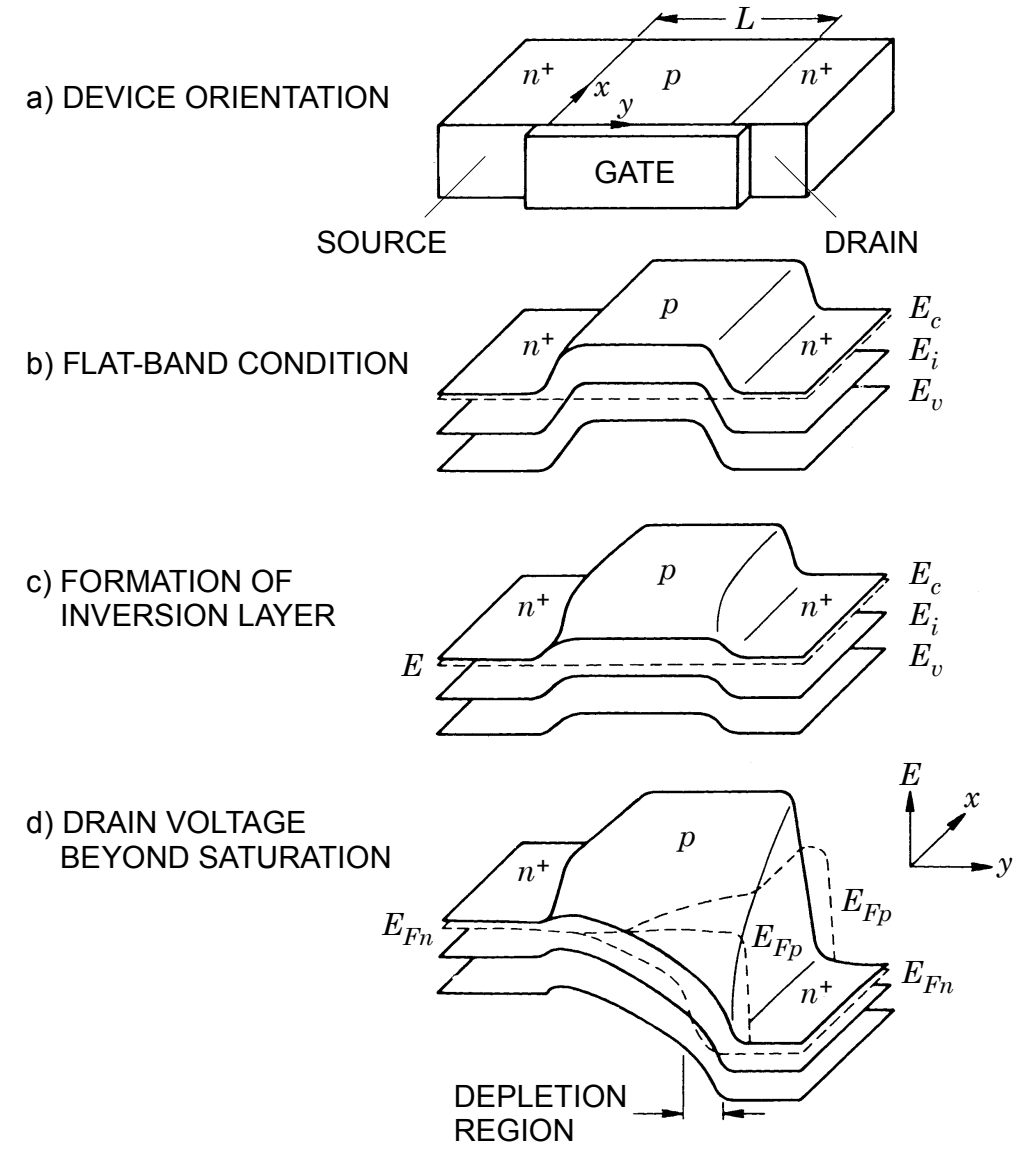
Ψ_S < 0 the bands bend upwards, increasing the hole concentration at the surface.

$\Psi_B > \Psi_S > 0$ the bands bend slightly downwards, reducing the concentration of holes at the surface (depletion)

$\Psi_S > \Psi_B$ the conduction band edge dips below the Fermi level, leading to an accumulation of electrons at the surface (inversion)

In the absence of any special surface preparation the surface of silicon is n-type, i.e. p-type silicon inverts at the surface.

Three-Dimensional Potential Diagram of an n-Channel MOSFET



(Adapted from Sze 1981, ©Wiley and Sons, reproduced with permission)

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.

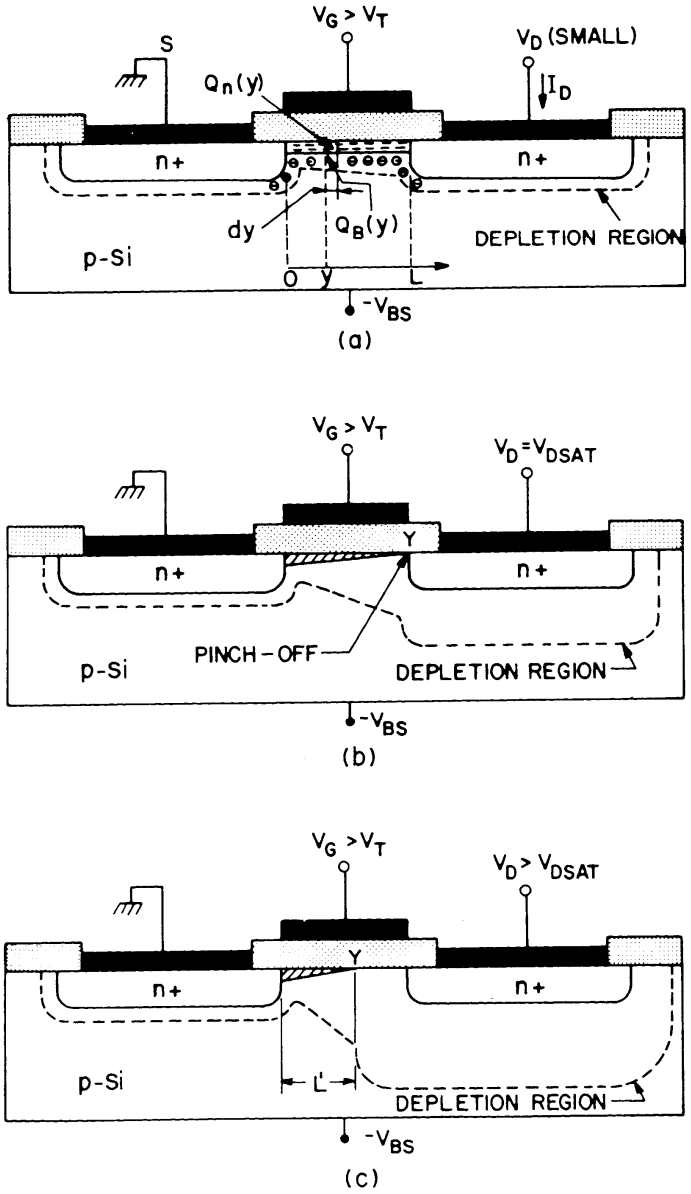
As in the JFET, the combination of current flow in the channel and the applied potentials forms a depletion region that is greatest near the drain.

At a sufficiently large drain potential the channel "pinches off".

(a) Low drain voltage $V_D < V_{sat}$:
Resistive channel

(b) $V_D = V_{sat}$: Onset of current saturation

(c) $V_D > V_{sat}$: Output current saturated

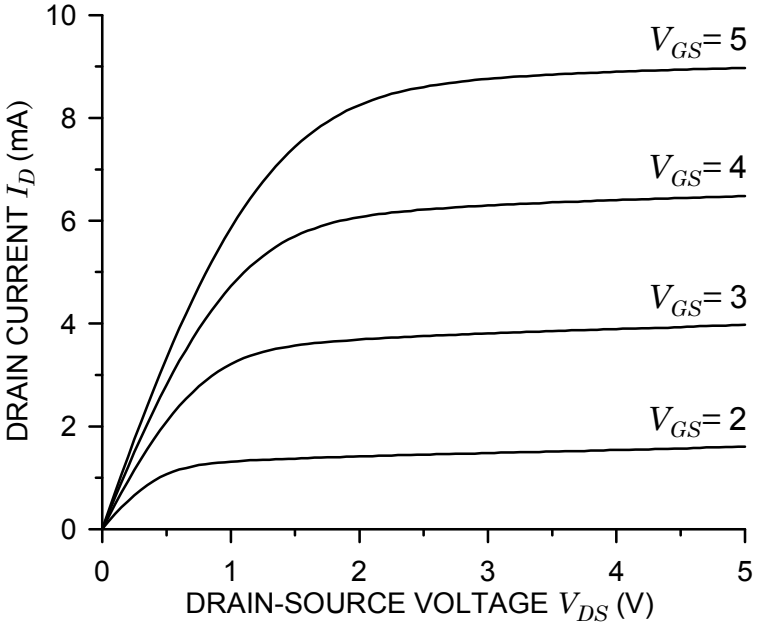


(From Sze 1981, ©Wiley and Sons, reproduced with permission)

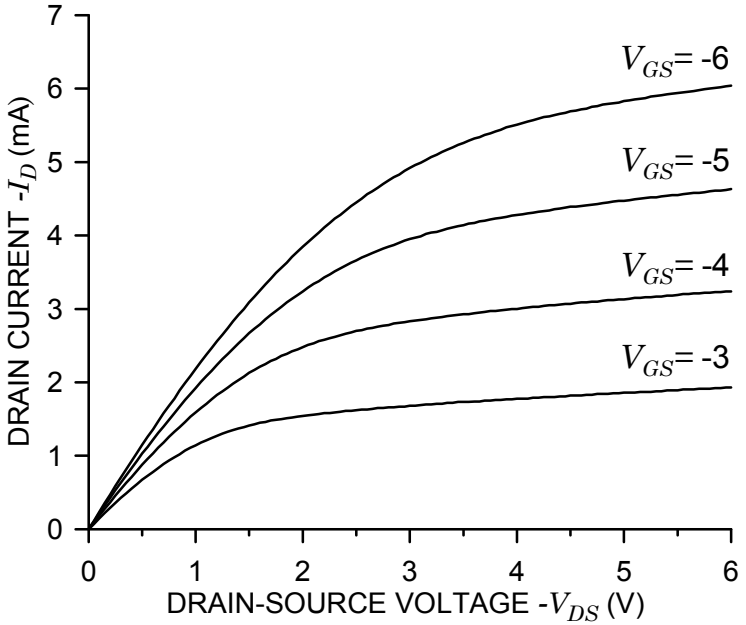
The output curves of a MOSFET are similar to a JFET.

The drain voltage required to attain saturation increases with the operating current.

NMOS



PMOS



In strong inversion and for $V_D > V_{D,sat}$

$$I_D = \frac{W}{L}\frac{\mu\ C_i}{2}(V_G - V_T)^2$$

where C_i is the gate capacitance per unit area $\varepsilon_{ox} / d_{ox}$ and V_T is the gate voltage corresponding to the onset of strong inversion ("threshold voltage")

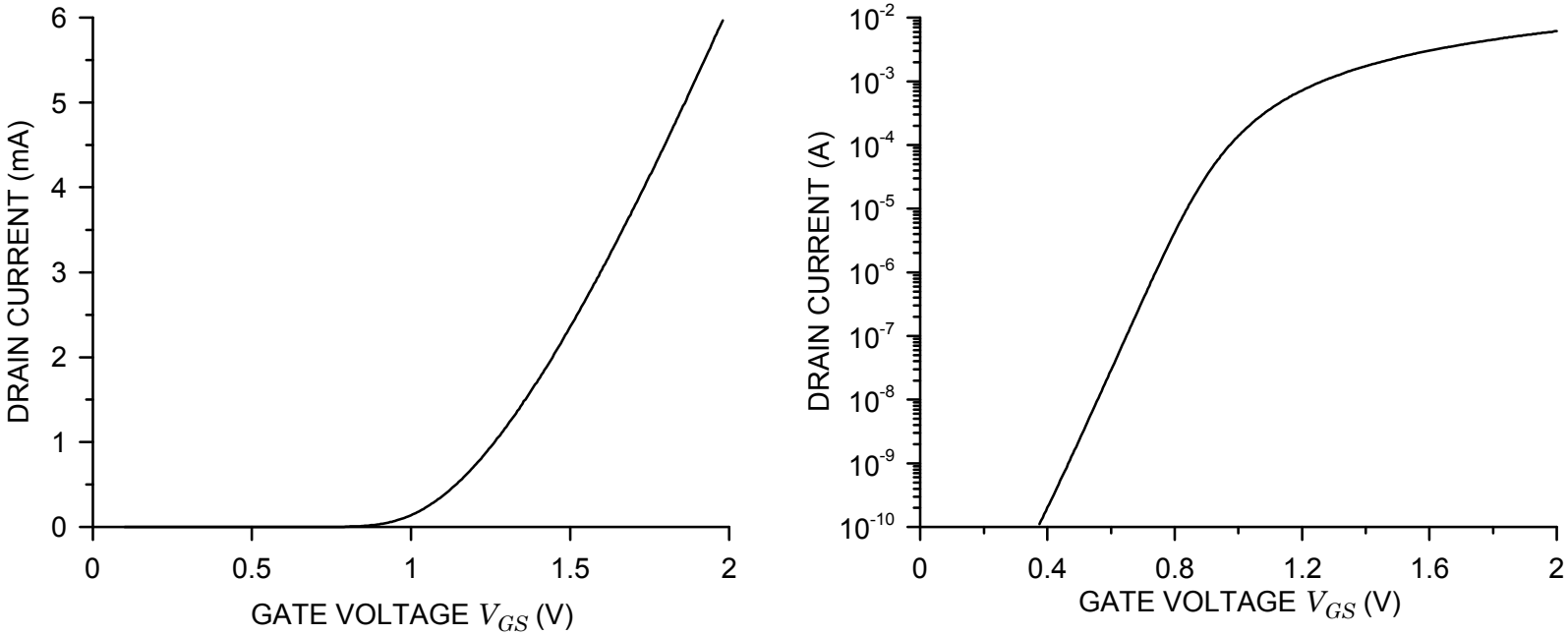
From this the transconductance is

$$g_m = \frac{W}{L} C_i \mu\ (V_G - V_T) = \frac{W}{L}\frac{\varepsilon_{ox}}{d_{ox}}\mu\ (V_G - V_T) = \sqrt{\frac{W}{L}\cdot\frac{\varepsilon_{ox}}{d_{ox}}\mu \cdot I_D}$$

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}.

As for the JFET, the transconductance depends primarily on current.

Measured characteristics of an n-channel MOSFET with 0.8 μm channel length and 20 nm gate oxide thickness (W = 100 μm)

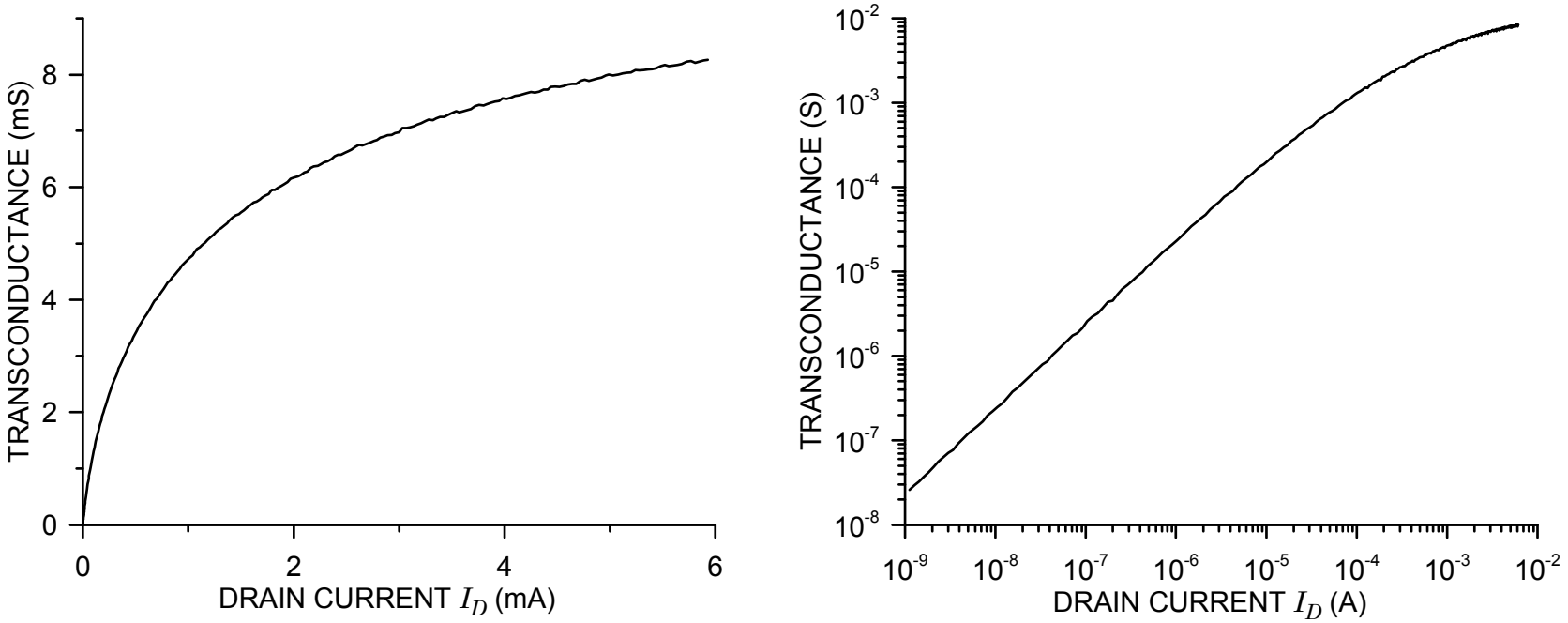


For this device the threshold voltage V_T is about 1.2 V.

The transition from weak to strong inversion becomes more apparent in a logarithmic plot.

In the subthreshold regime the drain current is proportional to the inversion carrier concentration, i.e. it increases exponentially with gate voltage.

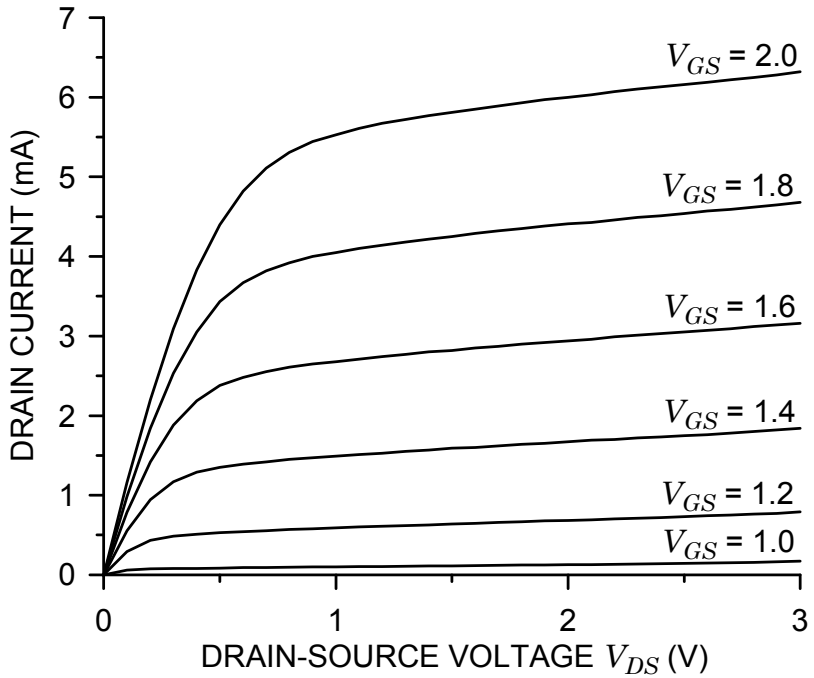
The transconductance increases with drain current:



In **strong inversion** ($I_D > 10^{-3}$ A) the transconductance increases with the square root of current following the expression given above.

At low currents (**weak inversion**) the transconductance increases linearly with current.

The increase in output saturation voltage with gate voltage can be seen clearly in the output curves.

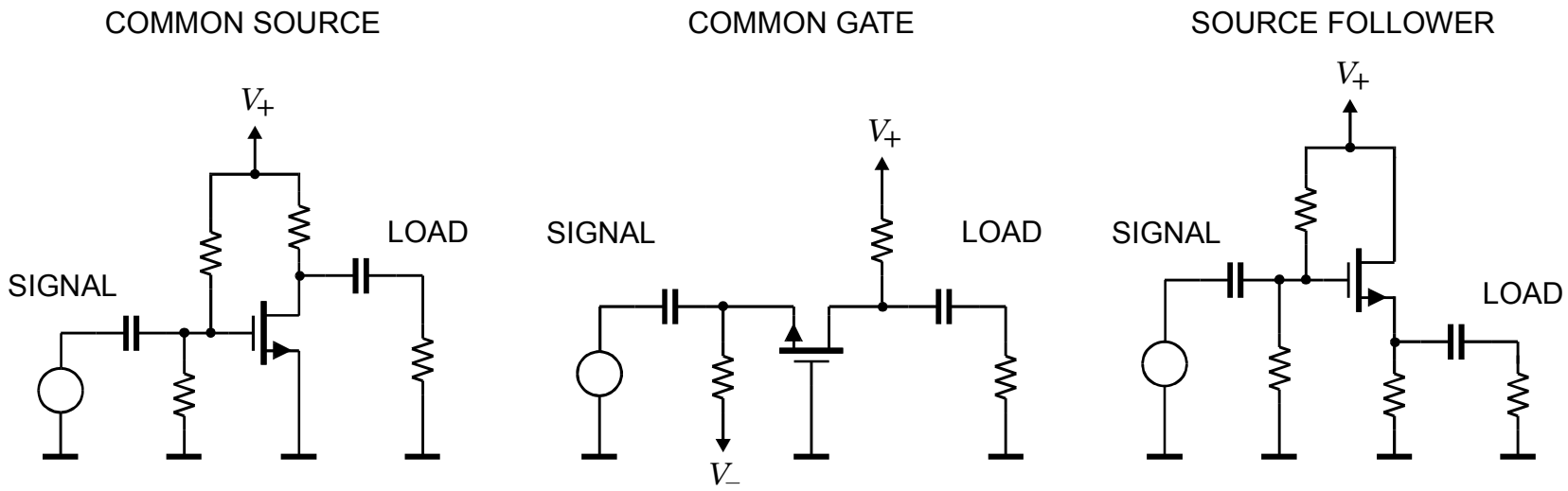


Depending on the substrate doping MOSFETs can be implemented with either n or p-channels.

A thin surface layer can be implanted to adjust the threshold voltage. With this devices can be normally on at zero gate voltage (depletion mode) or normally off, i.e. require additional voltage to form the inversion layer (enhancement mode).

MOS Transistors in Amplifiers

As shown for BJTs, three different circuit configurations are possible:

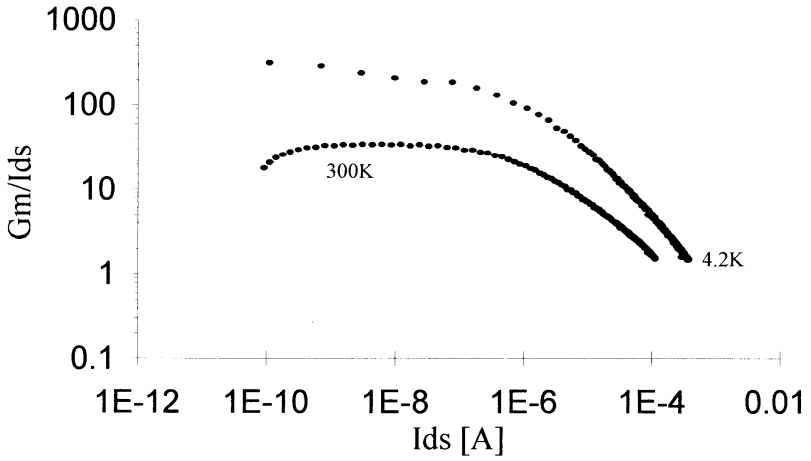


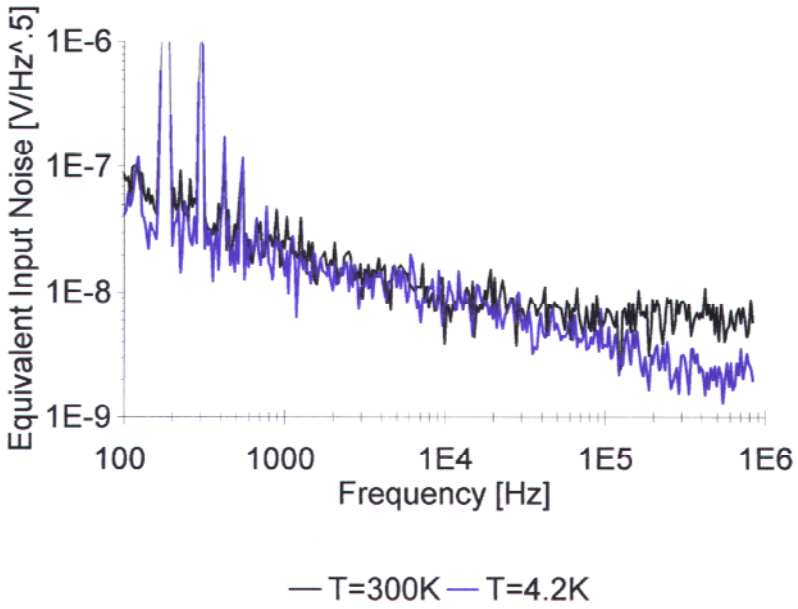
Low Temperature Operation

JFETs and BJTs “freeze out” at low temperatures, as thermal activation is insufficient to maintain the required free carrier density.

MOSFETs do not rely on thermal activation to set the charge concentration in the channel. External potentials bend the bands to establish the inversion layer. However, source and drain must be “degenerately doped” to prevent freeze-out and substrate freeze out can introduce additional noise.

Data of MOSFETs on high-resistivity substrates: Noise for W= 50 μm, L= 5 μm, I_D=100 μA.



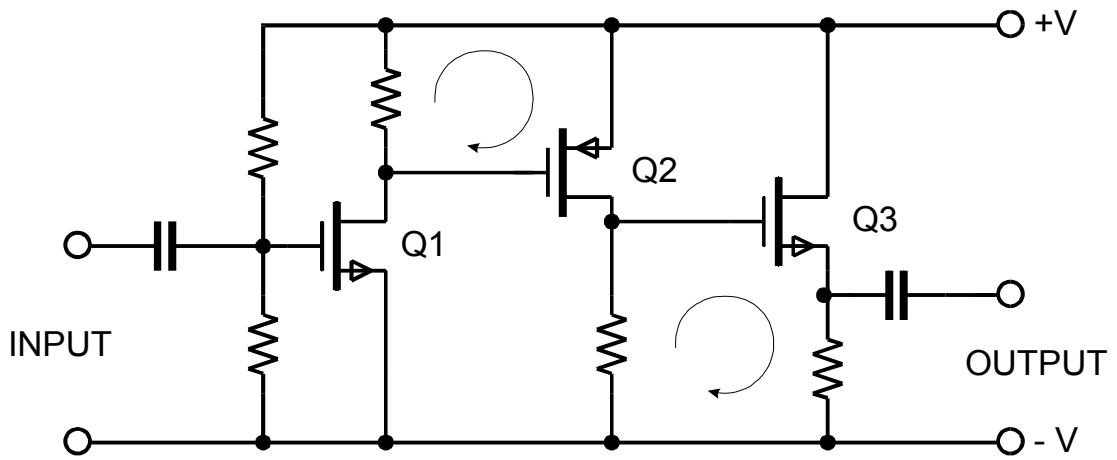


Transconductance increases at 4K.

Low-freq. noise unaffected, but at >100 kHz 3-fold improvement.

Advantages of MOS/CMOS

- high input resistance (capacitive gate)
- complementary devices (NMOS, PMOS)
- magnitude of gate voltages allows simple direct coupling
- tailoring of device characteristics by choice of geometry (W, L)



Amplifier cascade using NMOS and PMOS devices

- low-power, high-density logic circuitry (CMOS)

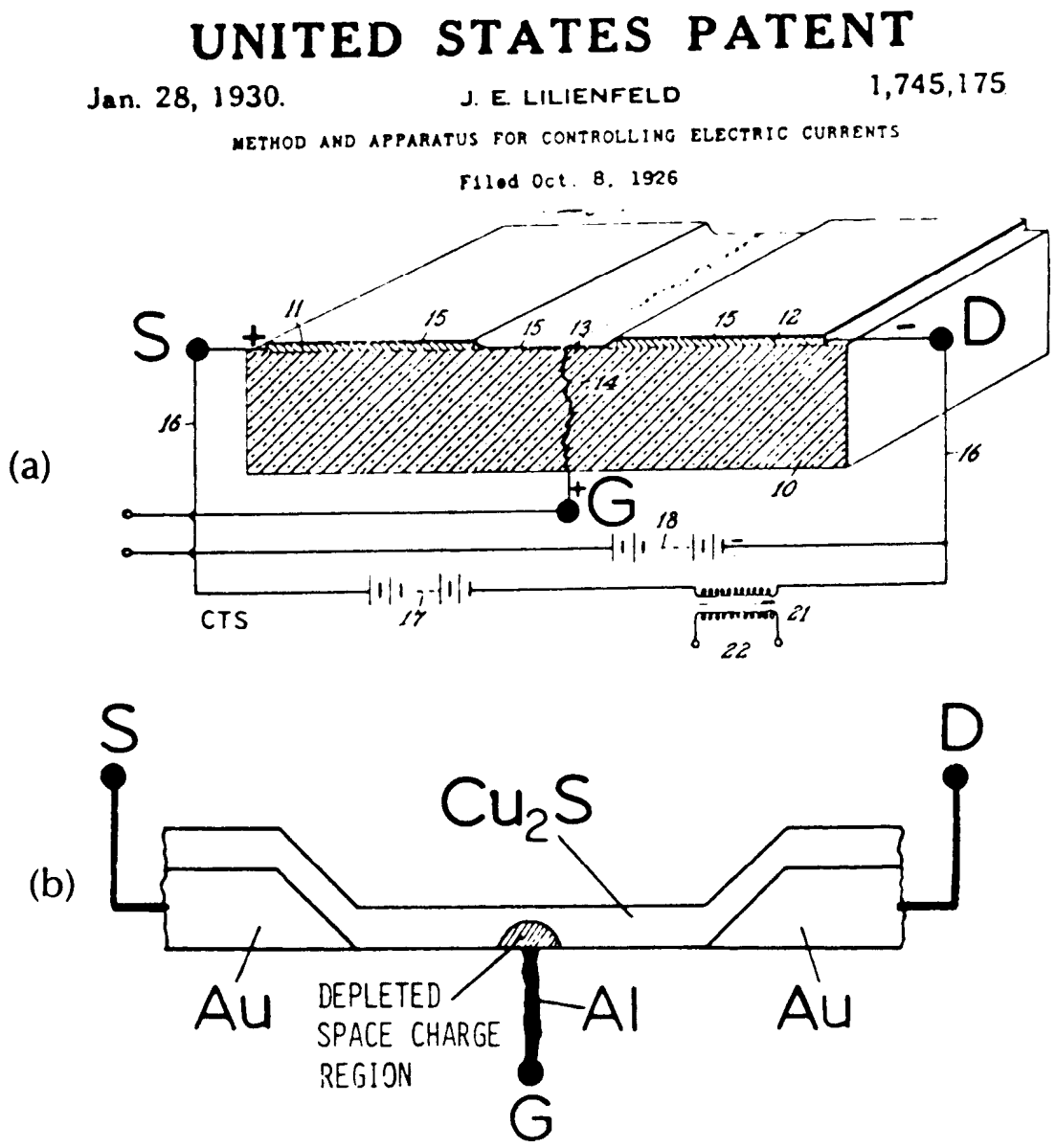
Drawbacks relative to bipolar transistors …

- more current for given transconductance (speed)
- more current for given noise level in detector circuits using short shaping times
- inferior device matching
- analog characteristics less predictable (more difficult to model)

Modern IC fabrication processes combine bipolar transistor and MOS technology to exploit best features of both (BiCMOS).

… So what else is new?

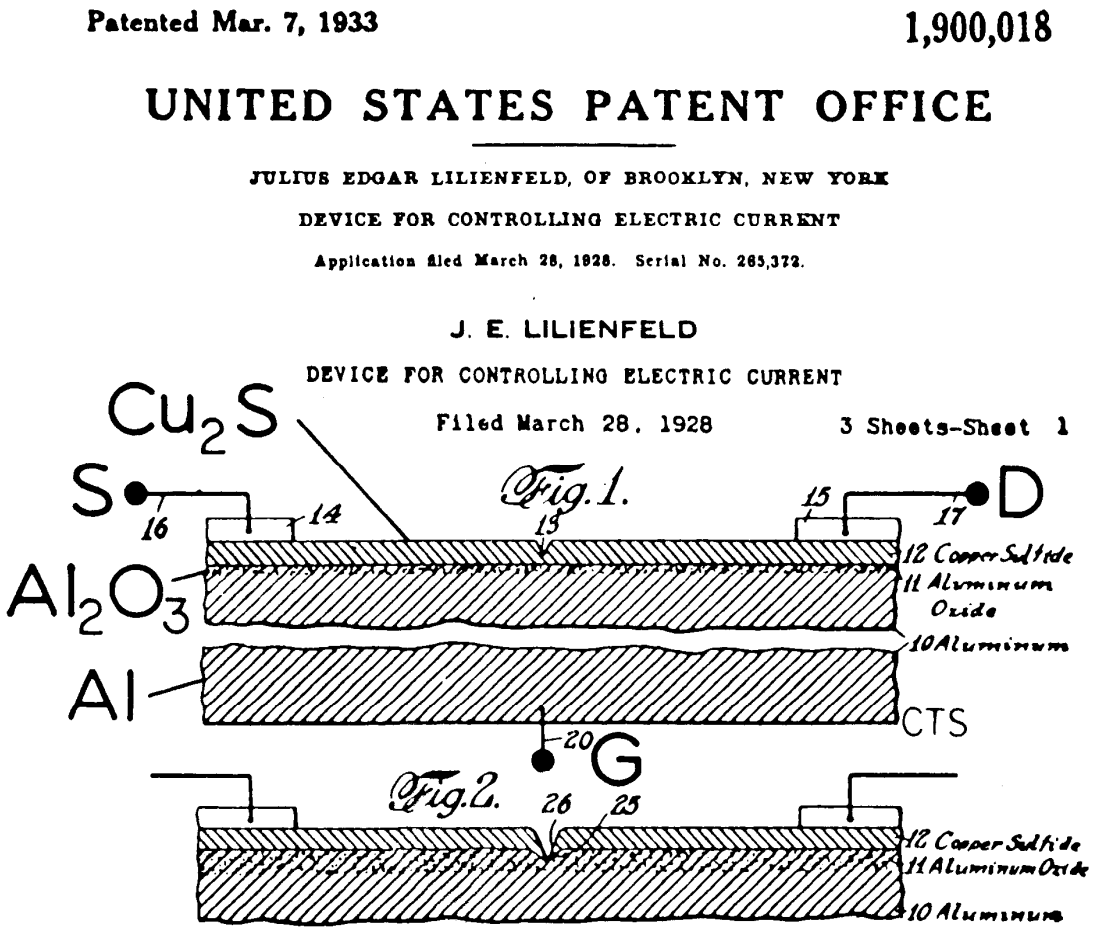
The first patent awarded for a junction field effect transistor was submitted in 1926



In 1928 Lilienfeld submitted a patent application for a MOSFET

Although Lilienfeld appears to have fabricated prototypes, the results were not reproducible, because surface states and impurity levels could not be controlled.

Furthermore, unknown to everyone at the time, the dynamics of electrons and holes and practically all of semiconductor physics had yet to be understood.



Patented Mar. 7, 1933 1,900,018

UNITED STATES PATENT OFFICE

JULIUS EDGAR LILIENFELD, OF BROOKLYN, NEW YORK

DEVICE FOR CONTROLLING ELECTRIC CURRENT

Application filed March 28, 1928. Serial No. 265,372.

J. E. LILIENFELD

DEVICE FOR CONTROLLING ELECTRIC CURRENT

Filed March 28, 1928 3 Sheets-Sheet 1

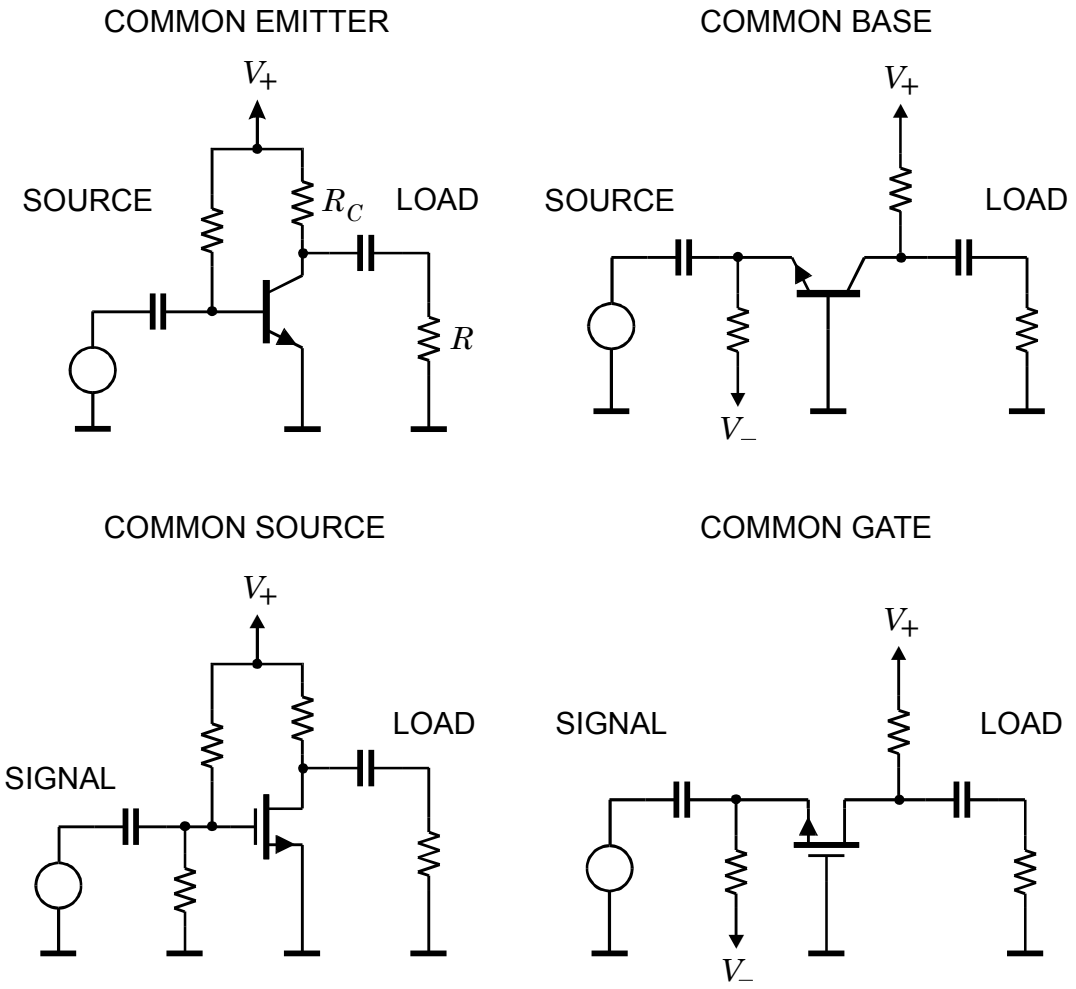
Nevertheless, these concepts provided the impetus for the research that lead to the first bipolar transistor in 1947, JFET in 1953 and practical Si MOSFETs in 1960.

Transistor Gain – Basic Principles

Both bipolar and FET/CMOS transistors in common emitter/source amplifiers and common base/gate amplifiers have current source outputs.

For the common emitter and common source amplifiers the voltage gain is determined by

- The ratio of output current to input voltage, i.e. the transconductance g_m
- The output load resistance: Parallel connection of the resistance to the voltage source and the load resistance.



In bipolar transistors the transconductance is easily described, driven primarily by the collector current I_C :

$$g_m \equiv \frac{dI_C}{dV_{BE}} = \beta_{DC} I_R \frac{q_e}{k_B T} e^{q_e V_{BE}/k_B T} = \frac{q_e}{k_B T} I_C$$

The transconductance depends only on collector current,

so for any bipolar transistor – regardless of its internal design –

setting the collector current determines the transconductance.

Since at room temperature $k_B T / q_e =$ 26 mV,

$$g_m = \frac{I_C}{0.026} \approx 40 I_C$$

In JFETs and MOSFETs the transconductance also depends on geometry, e.g. in MOSFETs

$$g_m = \frac{W}{L} C_i \mu \, (V_G - V_T) = \frac{W}{L} \frac{\varepsilon_{ox}}{d_{ox}} \mu \, (V_G - V_T) = \sqrt{\frac{W}{L} \cdot \frac{\varepsilon_{ox}}{d_{ox}} \mu \cdot I_D}$$

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 JFETs

$$g_m = \frac{2 I_{DSS}}{V_P} \left(1 - \frac{V_G + V_{bi}}{V_P} \right) = \frac{2\sqrt{I_{DSS}}}{V_P} \sqrt{I_D} = \frac{2}{V_P} \sqrt{\frac{1}{6\varepsilon} \mu (q_e N_{ch})^2 d^3 \frac{W}{L}} \sqrt{I_D}$$

As for the for both types of FET, for a given geometry the transconductance depends primarily on current.

However, unlike bipolar transistors, the transconductance for a given current also depends on geometry and current density.

The above expressions are only for high current density!

At low current densities, i.e. larger area devices at the same current, a given transconductance can be achieved at lower current.

This is an important consideration in large-scale systems.

3. Noise in Transistors

This section will discuss the details of noise formation, but here are some basic principles.

In both FET and bipolar transistors the major noise is in the output, formed in the internal signal channel and not directly present at the input.

- In the bipolar transistor the output noise is the fluctuation of charge components in the output current, so it is an output current noise:

$$i_{nc}^2 = 2q_e I_C$$

- In FETs the noise source is the thermal noise of the channel resistance. Since the output signal is a current, this resistor noise is expressed as a noise current. The resistor noise depends on the channel parameters and current density.

$\Rightarrow$ In both bipolar transistors and FETs the major noise is output current noise.

To permit direct analysis of the signal-to-noise ratio for a given signal level, the output noise current is referred to the input by dividing the output current by the transconductance

$$v_{ni} = \frac{i_{no}}{g_m}$$

In bipolar transistors the base current contributes direct input noise.

a) Noise in Field Effect Transistors

The primary noise sources in field effect transistors are

a) thermal noise in the channel

b) gate current in JFETs

Since the area of the gate is small, this contribution to the noise is very small and usually can be neglected.

Thermal velocity fluctuations of the charge carriers in the channel superimpose a noise current on the output current.

The spectral density of the noise current at the drain is

$$i_{nd}^2 = \frac{N_{C,tot} q_e}{L^2} \mu_0 \; 4 k_B T_e$$

The current fluctuations depend on the number of charge carriers in the channel $N_{C,tot}$ 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 measureable device parameters.

Since the transconductance in the saturation region

$$g_m \propto \frac{W}{L}\mu N_{ch} d$$

one can express the noise current as

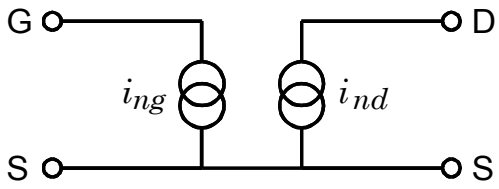
$$i_{nd}^2 = \gamma_n g_m \ 4k_B T_0 \ ,$$

where T_0 = 300K and γ_n is a semi-empirical constant that depends on the carrier concentration in the channel and the device geometry.

In a JFET the gate noise current is the shot noise associated with the reverse bias current of the gate-channel diode

$$i_{ng} = 2q_e I_G$$

The noise model of the FET



The gate and drain noise currents are independent of one another.

However, if 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

$$e_{ng} = Z\ i_{ng}$$

leading to an additional noise current at the output $g_m e_{ng}$, so that the total noise current at the output becomes

$$i_{no}^2 = i_{nd}^2 + (g_m Z i_{ng})^2$$

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

$$e_{ni}^2 = \frac{i_{no}^2}{g_m^2} = \frac{i_{nd}^2}{g_m^2} + Z i_{ng}^2 \equiv e_n^2 + Z i_n^2$$

i.e. referred to the input, the drain noise current i_{nd} translates into a noise voltage source

$$e_n^2 = 4 k_B T_0 \frac{\gamma_n}{g_m}$$

The noise coefficient γ_n is usually given as 2/3, but is typically in the range 0.5 to 1 (exp. data will shown later).

This expression describes the noise of both JFETs and MOSFETs.

At low frequencies the gate current of a JFET dominates the input noise current. At high frequencies capacitive coupling to the channel resistance introduces an additional noise current.

In MOSFETs the DC gate current is very low (at oxide thicknesses <10 nm it is dominated by electron tunneling through the oxide).

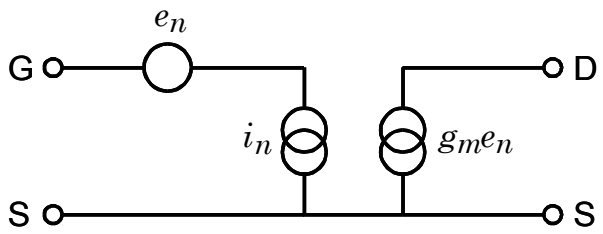
Here the capacitive coupling from the channel to the gate is significant. To a good approximation

$$i_n^2 \approx 20kT_0 f^2 \frac{C_{ox}^2}{g_{m,sat}} ,$$

i.e. the input noise current increases with frequency.

For an optimized device geometry g_m / C_{ox} is a constant for a given fabrication process.

In this parameterization the noise model becomes



where e_n and i_n are the input voltage and current noise.

As shown above these together contribute to the input noise voltage e_{ni}, which in turn translates to the output through the transconductance g_m to yield a noise current at the output $g_m e_{ni}$.

The equivalent noise charge

$$Q_n^2 = i_n^2 F_i T + e_n^2 C_i^2 \frac{F_v}{T}$$

For a typical JFET $g_m = 0.02$, $C_i = 10$ pF and $I_G < 150$ pA.

For $F_i = F_v = 1$

$$Q_n^2 = 1.9 \cdot 10^9 T + \frac{3.25 \cdot 10^{-3}}{T}$$

As the shaping time T decreases, the current noise contribution decreases and the voltage noise contribution increases.

For $T = 1$ μs the

current contribution is 43 el

and the

voltage contribution 57 el,

so the two contributions are about equal.

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

$$g_m \propto W \quad \text{and} \quad C_i \propto W$$

so that the ratio

$$\frac{g_m}{C_i} = const$$

Then the signal-to-noise ratio can be written as

$$\left(\frac{S}{N}\right)^2 = \frac{(Q_s / C)^2}{e_n^2} = \frac{Q_s^2}{(C_{det} + C_i)^2} \frac{g_m}{4k_B T_0 \Delta f}$$

S/N is maximized for $C_i = C_{det}$ (capacitive matching):

$$\left(\frac{S}{N}\right)^2 = \frac{Q_s^2}{\Delta f} \frac{1}{4k_B T_0} \left(\frac{g_m}{C_i}\right) \frac{1}{C_i \left(1 + \frac{C_{det}}{C_i}\right)^2}$$

$C_i \ll C_{det}$: The detector capacitance dominates, so the effect of increased transistor capacitance is negligible.

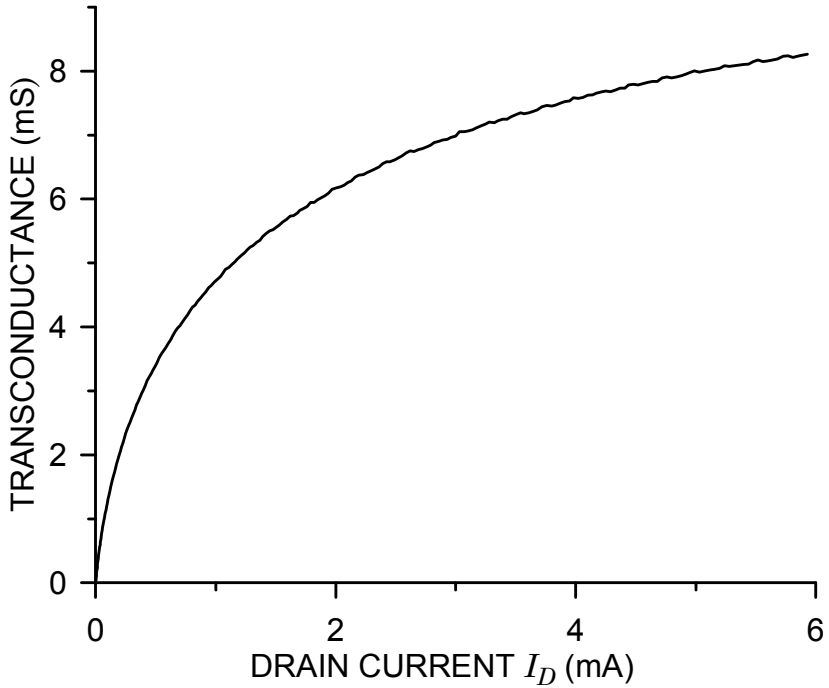
As the device width is increased the transconductance increases and the equivalent noise voltage decreases, so S/N improves.

$C_i > C_{det}$: The equivalent input noise voltage decreases as the device width is increased, but only with $1/\sqrt{W}$, so the increase in capacitance overrides, decreasing S/N.

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
(e.g. MOSFET, W= 100 μm, L= 0.8 μm):



However, noise only decreases up to a certain current. The reason is that the noise from parasitic source and gate resistances becomes significant.

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

$$g_m' = \frac{g_m}{W} \quad \text{and} \quad C_i' = \frac{C_i}{W}$$

and use these quantities to scale to any other device width. Since the equivalent input noise voltage

$$e_n^2 \propto \frac{1}{g_m}$$

the normalized input noise voltage is $e_n' = e_n\sqrt{W}$

Using these quantities the equivalent noise charge can be written as

$$Q_n^2 = 4k_B T_0 \frac{\gamma_n}{Wg_m'} \frac{F_v}{T}(C_d + C_s + WC_i')^2$$

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

$$Q_{n,min} = \sqrt{\frac{16 k_B T_0}{\kappa_n} \frac{F_v}{T} (C_d + C_s)}$$

where $\kappa_n \equiv \frac{g_m}{\gamma_n C_i}$ is a figure of merit for the noise performance of the FET.

Example: CMOS transistor with 1.2 μm channel length

At I_d/W = 0.3 A/m g_m/C_i = $3 \cdot 10^{-9}$ s^{-1} and

γ_n = 1.

For a *CR-RC* shaper with a 20 ns shaping time and an external capacitance

$C_d + C_s$ = 7.5 pF

$Q_{n,min}$ = 88 aC = 546 electrons,

achieved at a device width W = 5 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 = 1 μs

$Q_{n,min}$ = 1.8 aC = 11 electrons,

although in practice additional noise contributions will increase the obtainable noise beyond this value.

Low Frequency Noise (‘’1/f noise”)

The preceding discussion has neglected $1/f$ noise, which adds a constant contribution independent of shaping time

$$Q_{nf}^2 \propto A_f C_{tot}^2$$

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

$$A_f = \frac{K_f}{WL\ C_g^2}$$

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:

p-MOSFET	$K_f \approx 10^{-32}$ C^2/cm^2
n-MOSFET	$K_f \approx 4{\cdot}10^{-31}$ C^2/cm^2
JFET	$K_f \approx 10^{-33}$ C^2/cm^2

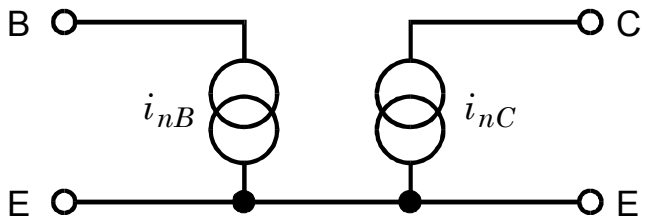
One should note that this model is not universally applicable, since excess noise usually does not exhibit a pure $1/f$ dependence; especially in PMOS devices one often finds several slopes. In practice, one must test the applicability of this parameterization by comparing it with data before applying 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.

b) Noise in Bipolar Transistors

In bipolar transistors the shot noise from the base current is important.

The basic noise model is the same as shown before, but the magnitude of the input noise current is much greater, as the base current will be 1 – 100 μA rather than <100 pA.



The base current noise is shot noise associated with the component of the emitter current provided by the base.

$$i_{nb}^2 = 2q_e I_B$$

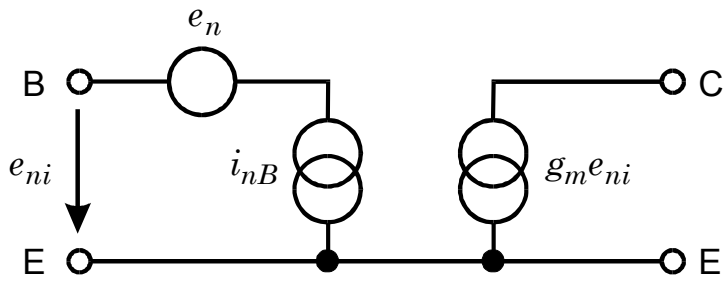
The noise current in the collector is the shot noise originating in the base-emitter junction associated with the collector component of the emitter current.

$$i_{nc}^2 = 2q_e I_C$$

Following the same argument as in the analysis of the FET, the output noise current is equivalent to an equivalent noise voltage

$$e_n^2 = \frac{i_{nc}^2}{g_m^2} = \frac{2q_e I_C}{(q_e I_C / k_B T)^2} = \frac{2(k_B T)^2}{q_e I_C}$$

yielding the noise equivalent circuit



where i_n is the base current shot noise i_{nb}.

The equivalent noise charge

$$Q_n^2 = i_n^2 F_i T + e_n^2 C_i^2 \frac{F_v}{T} = 2q_e I_B F_i T + \frac{2(k_B T)^2}{q_e I_C} C_{tot}^2 \frac{F_v}{T}$$

Since $I_B = I_C / \beta_{DC}$

$$Q_n^2 = 2q_e \frac{I_C}{\beta_{DC}} F_i T + \frac{2(k_B T)^2}{q_e I_C} C_{tot}^2 \frac{F_v}{T}$$

The current noise term increases with I_C,

whereas the second (voltage) noise term decreases with I_C.

Thus the noise attains a minimum

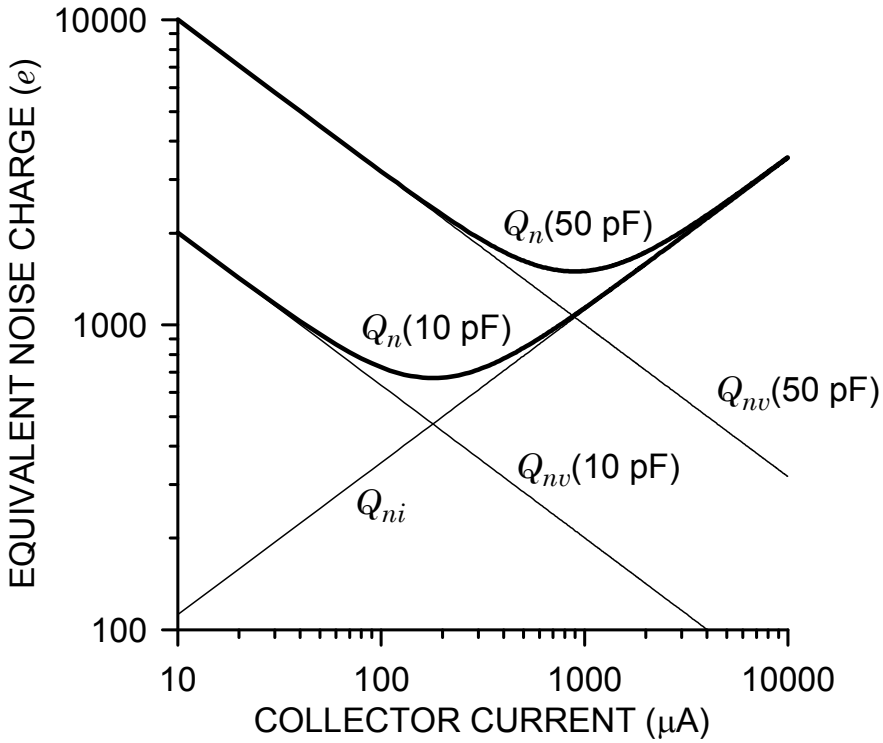
$$Q_{n,min}^2 = 4k_B T \frac{C_{tot}}{\sqrt{\beta_{DC}}} \sqrt{F_i F_v}$$

at a collector current

$$I_C = \frac{k_B T}{q_e} C_{tot} \sqrt{\beta_{DC}} \sqrt{\frac{F_v}{F_i}} \frac{1}{T}$$

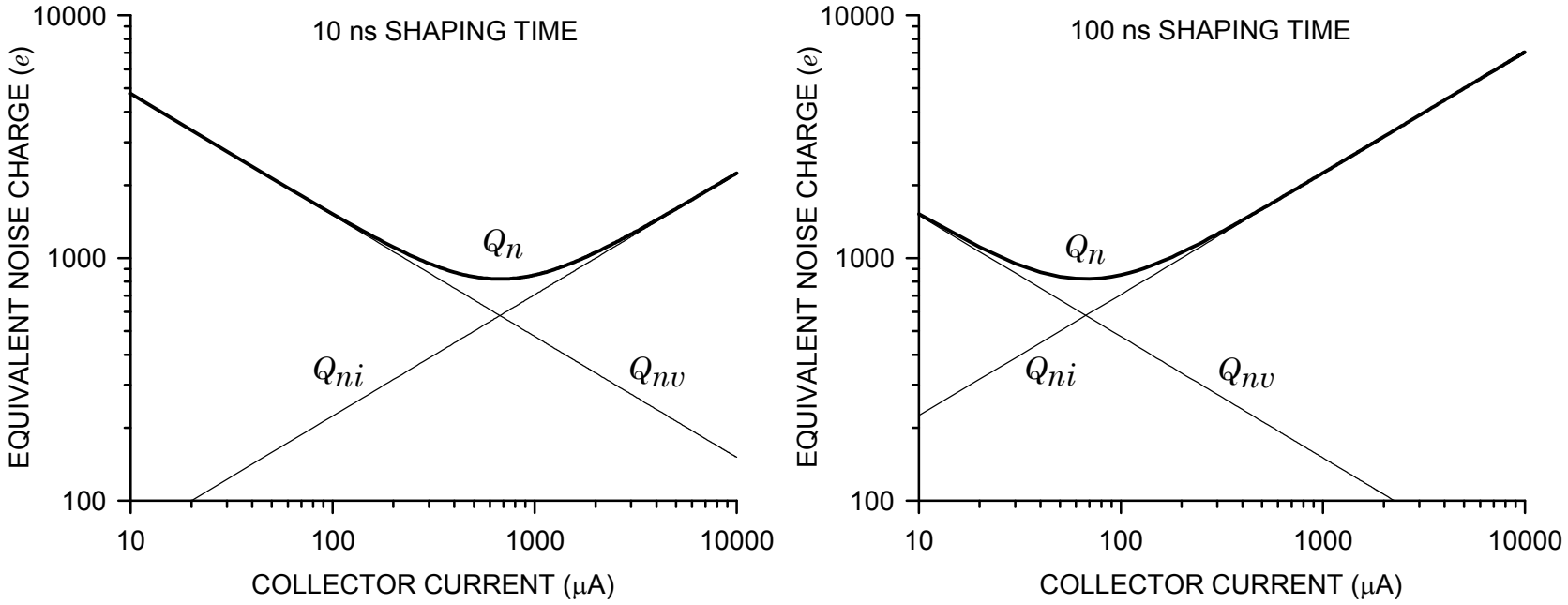
Note: This corresponds to the general criterion of noise matching derived in Section III, i.e. the detector impedance at the peaking frequency of the shaper equals e_n / i_n.

BJT equivalent noise charge vs. collector current for a CR-RC shaper with T_P = 25 ns.



Increasing the capacitance at the input shifts the collector current noise curve upwards, so the noise increases and the minimum shifts to higher current.

For a given shaper, the minimum obtainable noise of a BJT is determined only by the total capacitance at the input and the DC current gain of the transistor, *not by the shaping time*.



The shaping time only determines the current at which this minimum noise is obtained

Simple Estimate of obtainable BJT noise

For a *CR-RC* shaper

$$Q_{n,min} = 772 \left[\frac{\mathrm{el}}{\sqrt{\mathrm{pF}}}\right] \cdot \frac{\sqrt{C_{tot}}}{\sqrt[4]{\beta_{DC}}}$$

obtained at

$$I_c = 26 \left[\frac{\mu\mathrm{A}\cdot\mathrm{ns}}{\mathrm{pF}}\right] \cdot \frac{C_{tot}}{\tau}\sqrt{\beta_{DC}}$$

Since typically $\beta_{DC} \approx 100$, these expressions allow a quick and simple estimate of the noise obtainable with a bipolar transistor.

Note that specific shapers can be optimized to minimize either the current or the voltage noise contribution, so both the minimum obtainable noise and the optimum current will be change with respect to the above estimates.

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 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}, and
2. the diffusion capacitance C_{DE}.

The transition capacitance in a device with $e_n \approx 1\ \mathrm{nV}/\sqrt{\mathrm{Hz}}$ 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.

$$C_{DE} = \frac{\partial q_B}{\partial V_{be}} = \frac{q_e I_E}{k_B T}\left(\frac{W}{2D_B}\right) \equiv \frac{q_e I_E}{k_B T} \cdot \frac{1}{\omega_{Ti}}$$

where 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 where the current gain of the transistor is unity.

Inserting some typical values, I_E =100 μA and ω_{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 current gain but different input capacitances. Since the minimum obtainable noise

$$Q_{n,min}^2 = 4k_B T \frac{C_{tot}}{\sqrt{\beta_{DC}}} \sqrt{F_i F_v} \quad ,$$

increasing the transistor input capacitance merely increases the total input capacitance C_{tot} and the obtainable noise.

When to use FETs and when to use BJTs?

Since the base current noise increases with shaping time, bipolar transistors are only advantageous at short shaping times.

However, their lower noise and power may be important.

With current technologies FETs are best at shaping times greater than 50 to 100 ns, but decreasing feature size of MOSFETs will improve their performance.

Optimizing the operating point is crucial for MOSFETs, if dissipated power is important.

Optimization of Low Noise and Power

Optimizing the readout electronics in large vertex or tracking detector systems is not optimizing one characteristic, e.g. noise, alone, but finding an optimum compromise between noise, speed, and power consumption.

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 and input capacitance determines the voltage noise contribution, the device with the higher input capacitance must have a lower equivalent noise voltage v_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.

What is the optimum operating current for a given device?

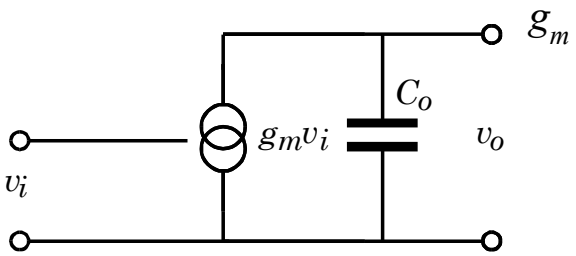
Both response time, i.e. bandwidth, and noise depend on a common parameter, transconductance.

The relationship between noise and transconductance was shown above.

The dependence of bandwidth on transconductance is easy to derive.

Consider an amplifying device with transconductance and a load resistance R_L.

The total capacitance at the output node is C_O.



The low frequency voltage gain is $$A_v = \frac{v_o}{v_i} = g_m R_L$$

The bandwidth of the amplifier is determined by the output time constant $\tau_o = R_L C_o = \frac{1}{\omega_o}$

Hence the gain-bandwidth product

$$A_v \omega_o = g_m R_L \cdot \frac{1}{R_L C_o} = \frac{g_m}{C_o}$$

is independent of the load resistance R_L, i.e. the voltage gain, but depends only on the device transconductance g_m and the capacitance at the output C_O.

The capacitance at the output node C_O depends on circuit topology and basic characteristics of the IC technology used.

Often, the bandwidth is determined less by the inherent device speed, than by the stray capacitance to the substrate.

Since increasing transconductance yields both improved bandwidth and noise, a useful figure of merit for low power operation is the

ratio of transconductance to device current g_m / I.

In a bipolar transistor

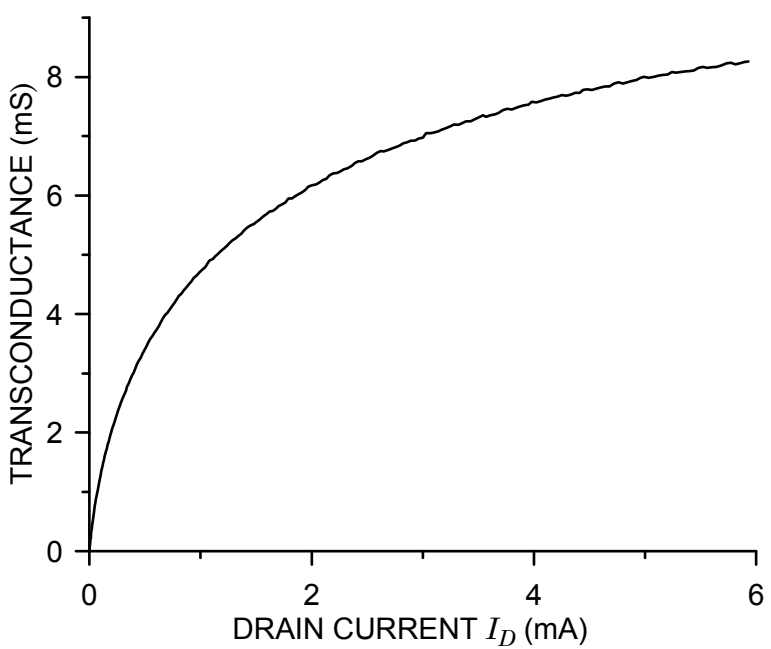
$$g_m = \frac{q_e}{k_B T} I_C$$

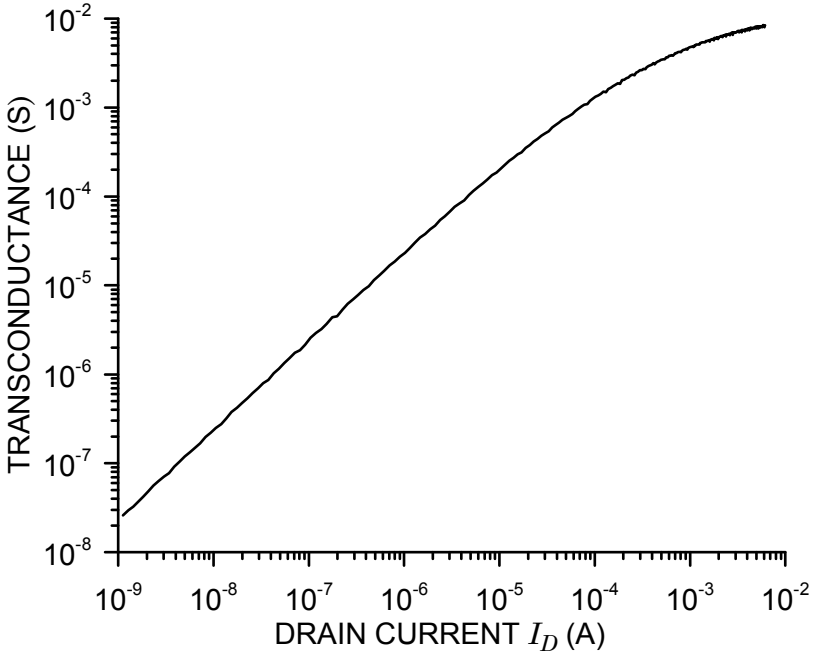
so g_m / I_C is constant

$$\frac{g_m}{I_C} = \frac{q_e}{k_B T}$$

In an FET the dependence of transconductance on drain current is more complicated.

Measured MOSFET Transconductance vs. Drain Current



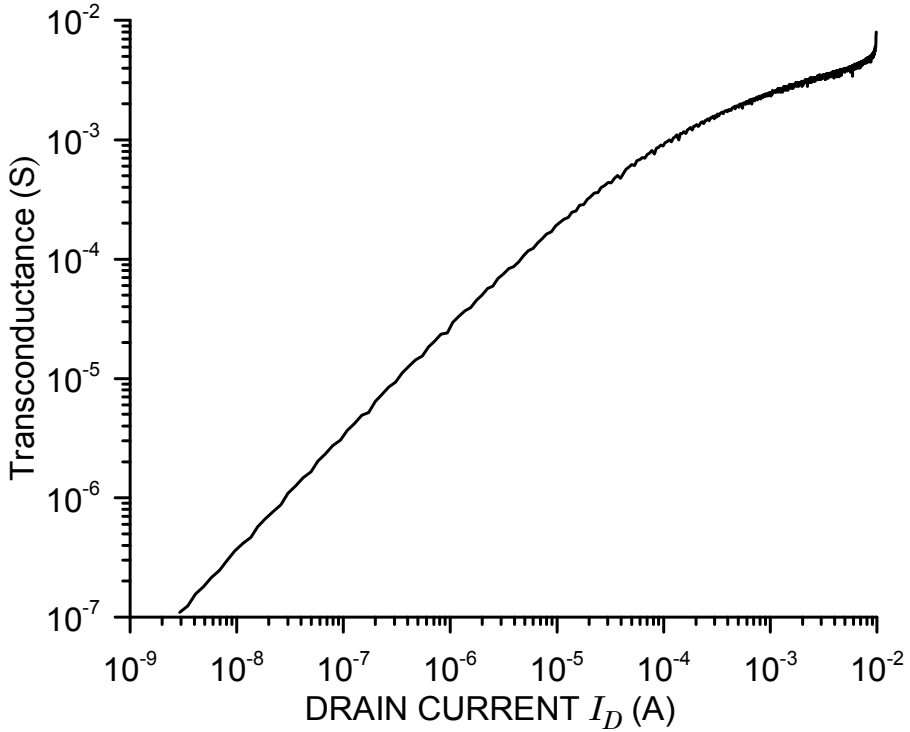


↑ weak inversion ↑ strong inversion

In weak inversion the transconductance increases linearly with current: $g_m \approx \frac{q_e}{k_B T} I_C$

In strong inversion (usual textbook equation): $g_m = \sqrt{\frac{W}{L} \cdot \frac{\varepsilon_{ox}}{d_{ox}} \mu \cdot I_D}$

JFETs show similar characteristics (although physics is different)



Capacitive Matching with FETs

Since in FETs $$e_n^2 \approx \frac{4kT}{g_m} ,$$

minimum noise is obtained at maximum transconductance.

For a given device, transconductance is maximized by operating at the highest allowable current.

The maximum current – and thus transconductance – can be increased further by increasing the device width (equivalent to connecting devices in parallel).

However, this also increases the input capacitance, which also enters into the equivalent noise charge.

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

$$g_m \propto W \quad \text{and} \quad C_i \propto W$$

so that the ratio

$$\frac{g_m}{C_i} = const$$

Then the signal-to-noise ratio can be written as

$$\left(\frac{S}{N}\right)^2 = \frac{(Q_s/C)^2}{v_n^2} = \frac{Q_s^2}{(C_d + C_i)^2}\frac{g_m}{4k_B T_0 \Delta f}$$

$$\left(\frac{S}{N}\right)^2 = \frac{Q_s^2}{\Delta f}\frac{1}{4k_B T_0}\left(\frac{g_m}{C_i}\right)\frac{1}{C_i\left(1+\frac{C_d}{C_i}\right)^2}$$

S/N is maximized for $C_i = C_d$ (capacitive matching).

$C_i \ll C_d$: 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.

$C_i > C_d$: The equivalent input noise voltage decreases as the device width is increased, but only with $1/\sqrt{W}$, so the increase in capacitance overrides and S/N gets worse.

Caution: Capacitive matching is not a generally valid principle

It only applies when there is a linkage between the input capacitance and the equivalent input noise voltage.

It does not apply to current noise

It does not apply to MOSFETs operating in weak inversion.

Transconductance depends only on drain current, not on geometry

Increasing the size (channel width, length) of a MOSFET
will increase the input capacitance,

but at constant current the transconductance – and thus the voltage noise – remain unchanged.

Increased capacitance + constant noise voltage $\Rightarrow$ higher noise charge

In bipolar transistors the diffusion capacitance of the base-emitter junction increases with current, but in modern high-frequency transistors this is negligible compared to the junction capacitance.

$\Rightarrow$ Capacitive matching is not a useful concept for bipolar transistors

In large scale systems power is a major concern,
so we must optimize noise vs. power.

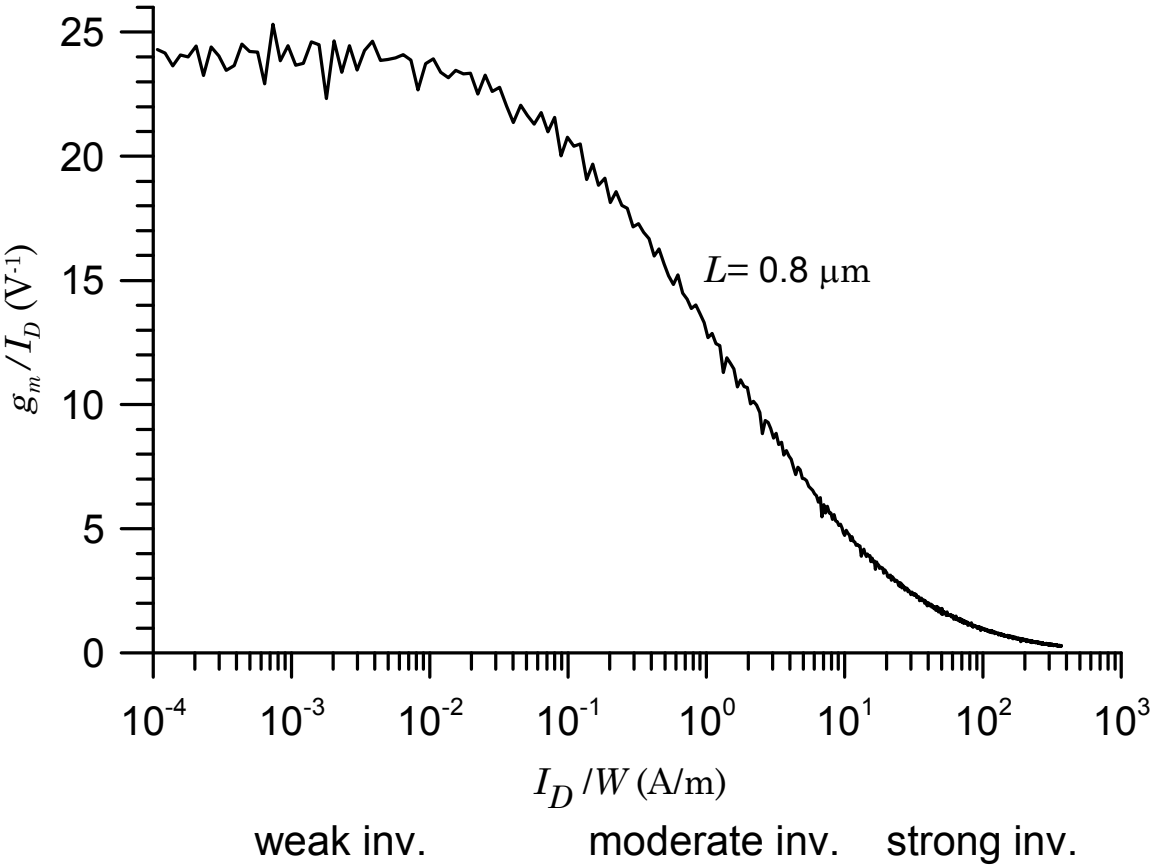
Since transconductance sets both the noise and speed, power efficiency improves when we increase the ratio of transconductance to operating current g_m / I

In a bipolar transistor

$$\frac{g_m}{I_C} = \frac{e}{kT} = \frac{1}{26\ \text{mV}} \approx 40$$

In a MOSFET 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 scalable relationship we plot g_m / I_D vs. I_D / W.



This is a universal curve for all transistors using the same technology and channel length.

- Weak inversion is very power efficient, but often does not yield required gain and noise.

- Increasing the FET width to reduce current density also increases the input capacitance, so for a given transconductance the noise charge increases.

- Strong inversion yields maximum gain and minimum noise, but at maximum current.

In large-scale systems where both sensitivity and total power can be critical, neither of these “standard” operating modes may be optimum.

Although not discussed in many courses and textbooks (and often ignored), the transition from weak to strong inversion often yields a practical compromise.

As shown in the previous plot, moderate inversion ranges over several decades of current density.

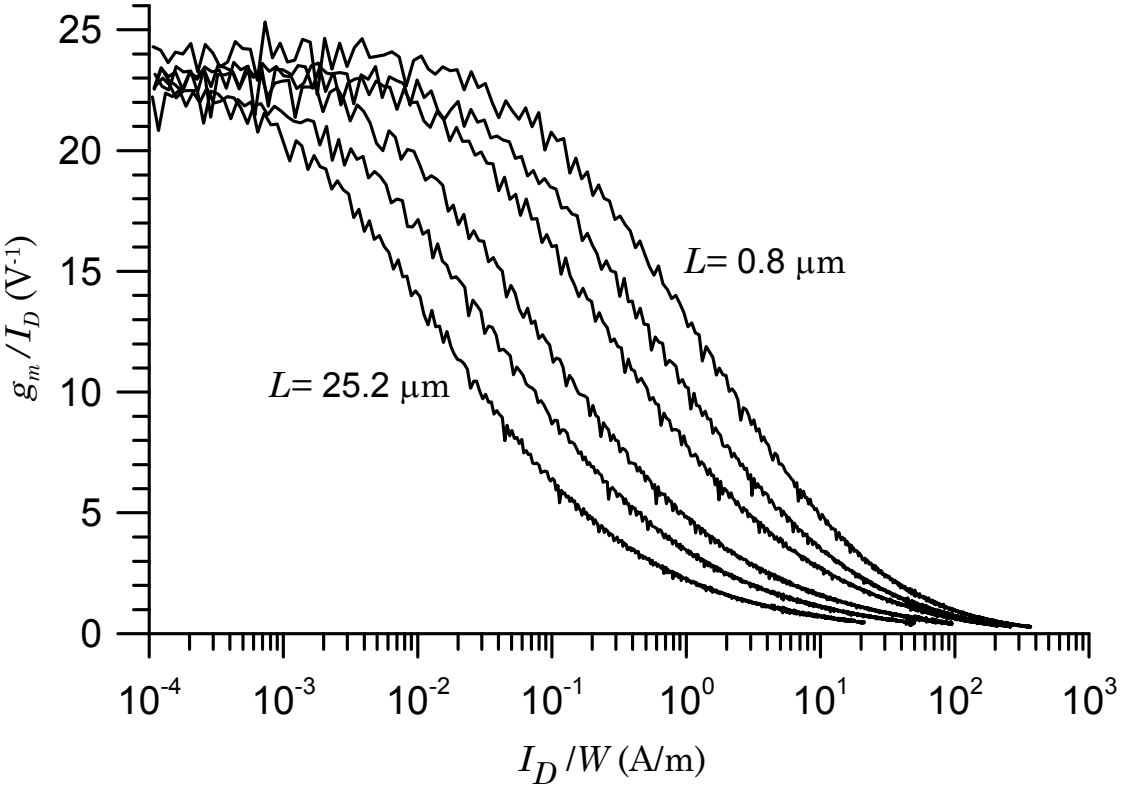
- Moderate inversion is often an optimum design criterion for large-scale systems.

Reducing the channel length improves power efficiency.

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.



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 1/10 the power.

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 $I_D / W \big|_{g_m / I_D}$,

which for a given technology depends on the channel length.

While keeping the current density constant, adjust the width to change the transconductance.

As the width is changed the drain current

$$I_D = W \cdot \left(I_D / W \right)_{g_m / I_D}$$

changes proportionally.

This value of drain current sets the transconductance

$$g_m = W \cdot \left(I_D / W \right)_{g_m / I_D} \cdot \left(g_m / I_D \right)_{selected}$$

Scaling the width changes the

- drain current
- transconductance
- input capacitance

proportionally.

Starting from a small device, 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.

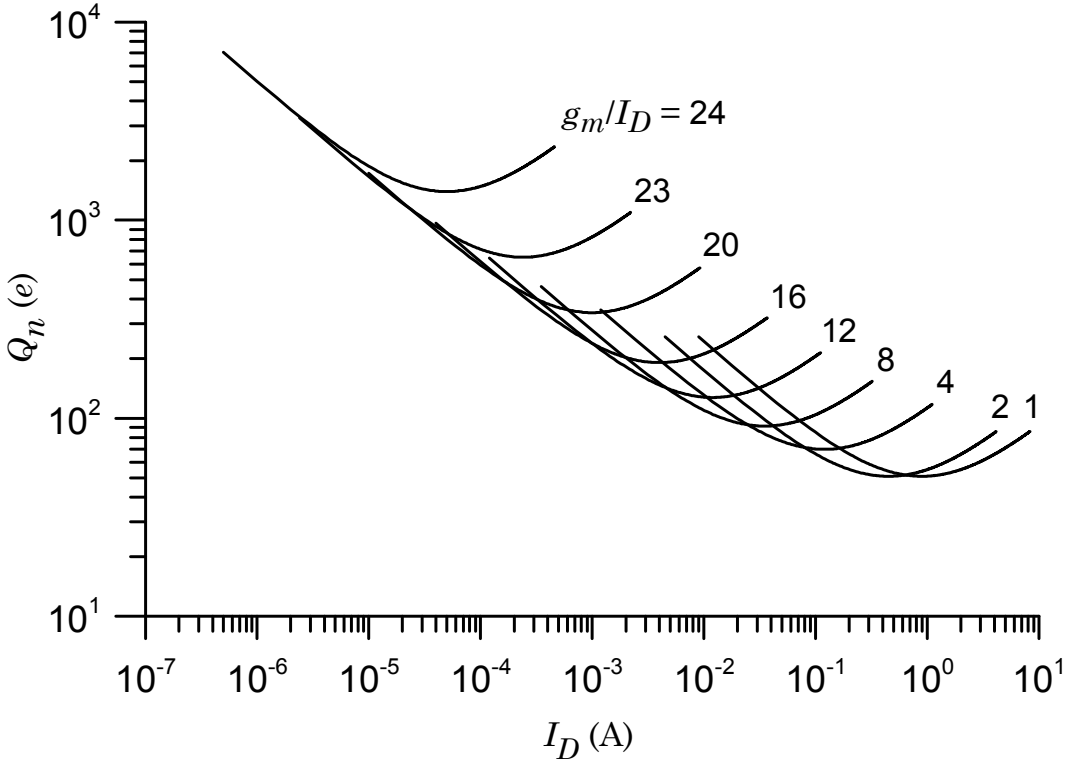
Example:

Desired noise level is 1000 el.

Input capacitance 1 fF/μm width.
Detector capacitance: 10 pF

A normalized transconductance g_m / I_D = 24 (weak inversion) allows a minimum noise of 1400 el at a drain current of 50 μA.

Increasing the current density to g_m / I_D = 20 shifts the operating mode towards moderate inversion and yields 340 el 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 el noise level at a drain current of 30 μA, less than the 50 μA needed for 1400 el noise at g_m / I_D = 24.

Much smaller values of g_m / I_D yield the desired 1000el noise at higher currents.

⇒ Capacitive matching is not a good criterion for low-power systems.

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

$$I_D \propto \frac{1}{Q_n^2} ,$$

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

$$e_n^2 \propto \frac{1}{g_m} \propto \frac{1}{I_D}.$$

Since the equivalent noise charge

$$Q_n^2 \propto e_n^2 C^2 \propto \frac{C^2}{I_D},$$

when operating well below capacitive matching, the required power for a given noise level increases with the square of sensor capacitance.

$$P_D \propto I_D \propto C_d^2.$$

A similar result obtains for bipolar transistors. The most efficient operating regime with respect to power is at a current lower than the current needed for minimum noise

$$I_C = \frac{kT}{e} C \sqrt{\beta_{DC}} \sqrt{\frac{F_v}{F_i}} \frac{1}{T_S},$$

since the minimum is rather shallow. In this regime the noise is dominated by voltage noise, so

$$Q_n^2 \approx \frac{2(kT)^2}{eI_C} C^2 \frac{F_v}{T_S}$$

and for a given noise level C / I_C is constant, so the required power

$$P \propto I_C \propto C^2$$

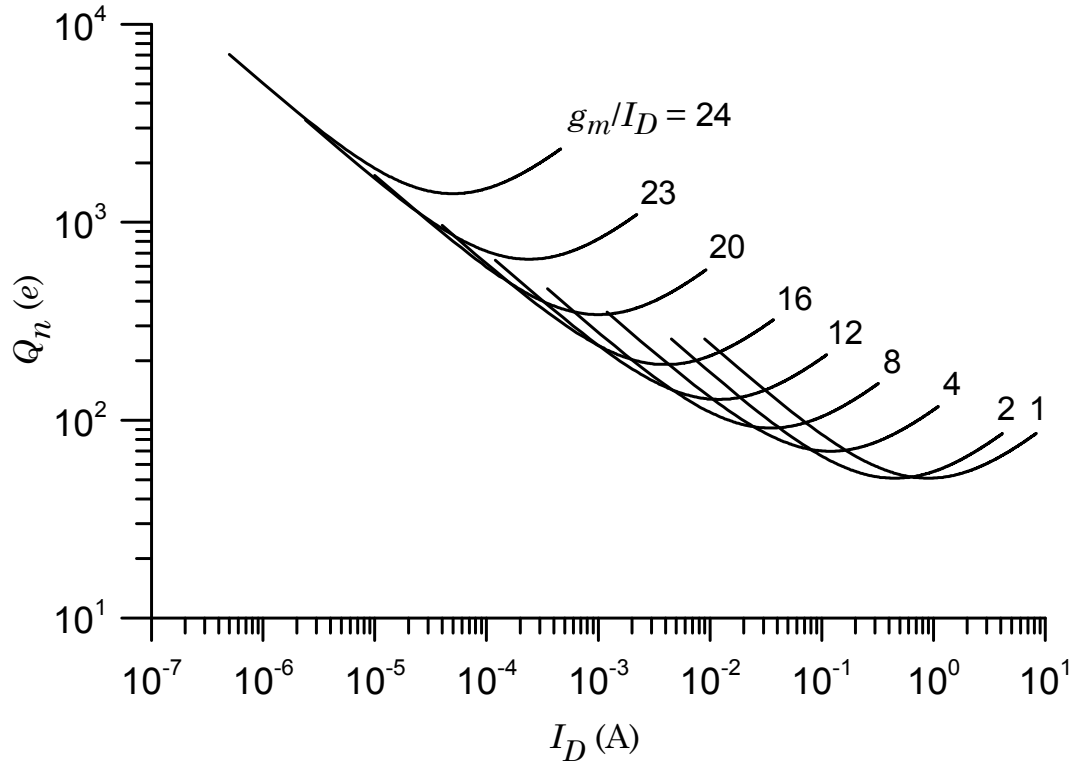
In this form of optimum scaling the required power in the input device scales with the square of capacitance at the input.

The required power also increases with the square of the desired signal-to-noise ratio.

These scaling rules represent optimum scaling, where the device input capacitance is negligible, i.e. well below capacitive matching.

For comparison, in this example a reduction in minimum noise at capacitive matching from 1400 to 50 el requires a 2000-fold increase in current.

Optimum scaling requires an 800-fold increase in current.



The difference is due to the penalty incurred by the device capacitance.

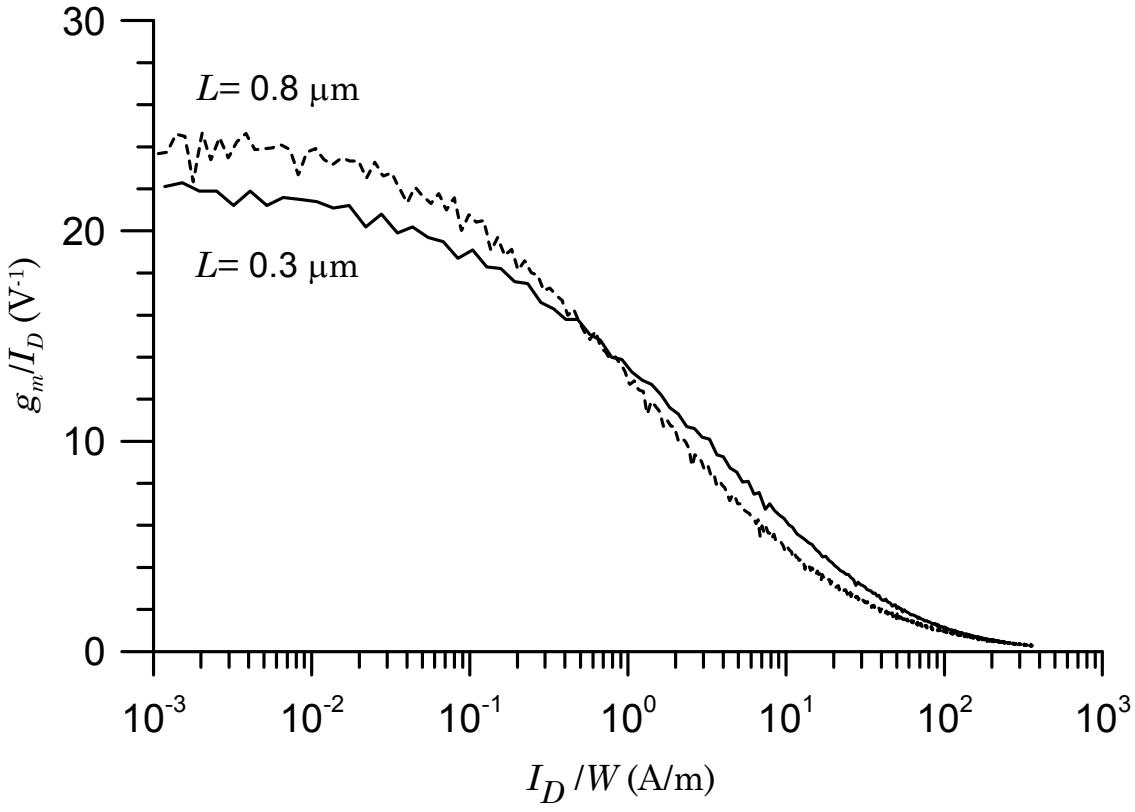
Technology Improvements

To what extent do improvements in device technology improve amplifier performance?

From data shown above we expect reductions in channel length to improve power efficiency.

Measured data for NMOS devices with L=0.8 and 0.3 μm, fabricated in 0.8 and 0.25 μm CMOS processes, respectively.

In the 0.25 μm process the transition from weak to moderate inversion occurs in the same current range as in the 0.8 μm MOSFET and the normalized transconductance in weak inversion is distinctly lower.



Why does the 0.3 μm channel length not show the expected improvement?

Scaling to smaller feature size involves more than lateral scaling, i.e. resolution 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.

Predicted improvements must be verified by measurements on practical processes!

Power-Efficient Systems – the Sensor

- Equivalent noise charge

$$Q_n^2 = i_n^2 F_i T + e_n^2 C^2 \frac{F_v}{T}$$

if noise current negligible $\quad Q_n^2 \approx e_n^2 C^2 \frac{F_v}{T}$

$$Q_n \propto e_n C = e_n \frac{A}{d} \varepsilon$$

ε is dielectric constant

- Energy resolution

$$\Delta E = E_i Q_n \propto E_g Q_n \propto e_n E_g \frac{A}{d} \varepsilon$$

E_g is bandgap

  - For constant ΔE and power (constant e_n), the product of band-gap and dielectric constant of detector material must remain constant.
  - Required power to maintain noise is proportional to $\left(E_g \varepsilon\right)^2$

Comparison: Power Dissipation of a Random Access Pixel Array vs. Strip Readout

If a strip readout for the LHC requires 2 mW per strip on an 80 μm pitch, i.e. 250 mW/cm width, is it practical to read out 15000 pixels per cm^2?

strip detector: n strips
pixel detector: n x n pixels

The capacitance is dominated by the strip-strip or pixel-pixel fringing capacitance.

$\Rightarrow$ capacitance proportional to periphery (pitch p and length l)

$$C \propto 2(l+p) \quad \Rightarrow \quad C_{pixel} \approx \frac{2}{n} C_{strip}$$

In the most efficient operating regime the power dissipation of the readout amplifier for a given noise level is proportional to the square of capacitance $P \propto C^2$

$$\Rightarrow \quad P_{pixel} \approx \frac{4}{n^2} P_{strip}$$

n times as many pixels as strips $\quad \Rightarrow \quad P_{pixel,tot} \approx \frac{4}{n} P_{strip}$

$\Rightarrow$ 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, random addressable pixel arrays can be implemented with overall power densities comparable to strips.

Summary of considerations in developing front-end electronics

- Minimizing capacitance at the input
- Assess the required noise level – not necessarily the minimum noise
- What is the required shaping time?

  At fast shaping times, bipolar transistor front-ends may be optimum
- High-density readout often constrains integrated circuitry to MOSFET alone.
- Is overall power dissipation critical?

  Moderate inversion in MOSFETs

  Note that many physicists and electronic designers are ignorant of moderate inversion in MOSFETs

  Simulation parameters are not always reliable in this regime, so check with device measurements.
- Consider potential reliability problems

  e.g. exponential effects of control parameters

  cost and inadequate development time to assess risks

4. Microelectronics

Fabrication of Semiconductor Devices

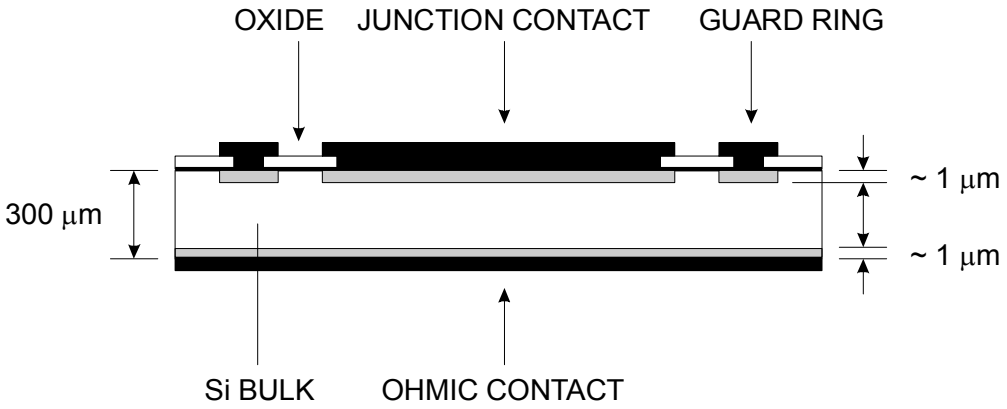
Ingredients of a semiconductor device fabrication process

1. bulk material, e.g. Si, Ge, GaAs
2. dopants to create p- and n-type regions
3. metallization to make contacts
4. passivation to protect the semiconductor surfaces
   from electrical and chemical contaminants

Practically all semiconductor devices are fabricated in a planar geometry (very few exceptions, e.g. large volume coaxial detectors)

1. the starting point is a semiconductor wafer
2. dopants are introduced from the surfaces

Typical planar detector diode structure



The p-n junction is formed by a high-conductivity to low-conductivity junction, so the depletion depth extends primarily into the bulk.

The guard ring is an additional junction that isolates the main junction from the edge of the wafer

Dopants are introduced by

1. Thermal diffusion in a gaseous ambient at ~1000 °C

or

2. Ion implantation

   Accelerate dopant ions to 20 – 100 keV, depending on desired penetration depth

   The implanted ions will initially be distributed interstitially, so to render them electrically active as donors or acceptors they must be introduced to substitutional lattice sites (“activation”.

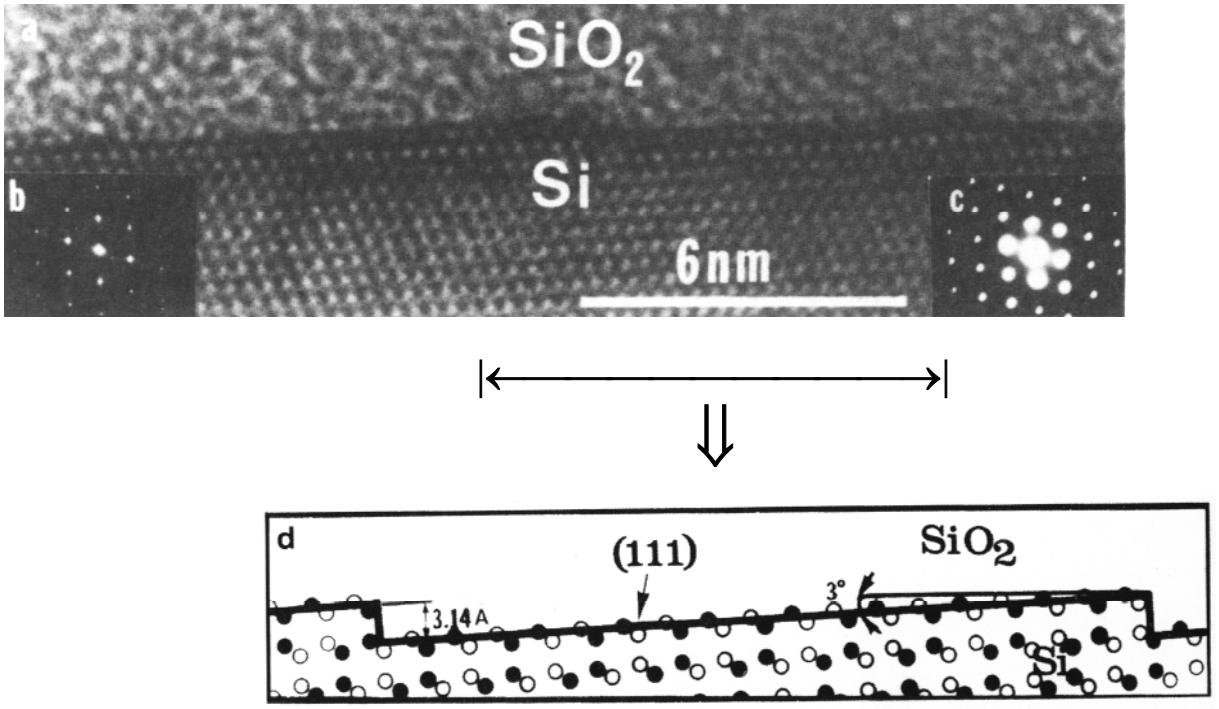
   This is accomplished by heating (annealing).

Protective layers immediately adjacent to the active semiconductor bulk must form a well controlled interface to the semiconductor lattice, to

1. minimize additional charge states (“dangling” bonds)
2. avoid mechanical stress (mismatch of thermal expansion coefficients)

In these respects SiO_2 on Si is unequalled – indeed this is probably the single key ingredient that allows Si technology to achieve a circuit density that is at least an order of magnitude greater than in any other semiconductor.

Atomic resolution electron microscope image of SiO_2-silicon interface



(Gronsky et al., LBNL National Center for Electron Spectroscopy)

The highest quality oxides are “grown”, i.e. the silicon is exposed to an oxidizing ambient, which diffuses into the silicon and forms SiO_2.

Oxide can also be deposited. The quality of the interface is much inferior to grown SiO_2, so deposited oxide is used primarily for protective layers on non-critical surfaces or after the silicon has already been protected by a grown oxide.

Growth of a high-quality oxide with minimum contamination is time consuming, so a common technique is to grow a thin oxide layer to provide a good electrical interface and then deposit an additional layer of lesser quality oxide.

Oxide can be deposited at relatively low temperatures (low temperature oxide – LTO), which is advantageous if the duration of high-temperature steps must be limited (to minimize diffusion and preserve shallow junctions)

Metallization is applied either by evaporation or sputtering.

Since all of these processes are only to be applied to specifically controlled areas, “masks” are used to expose only selected areas to difusion, ion implantation, or etchants.

The patterning is accomplished by photolithography.

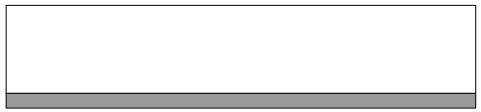
A photoresist is applied to the surface.

Exposure to light through a “mask” to cross-links the polymer in the desired areas.
(or the inverse – once can use positive or negative resist)

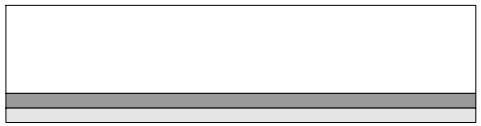
The exposed portions are removed by an appropriate solvent.

Key Process Steps

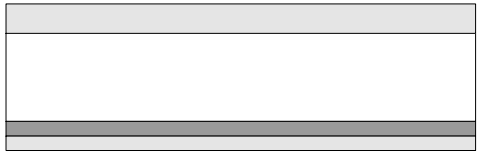
a) DEPOSIT P-DOPED POLY-Si BACKSIDE CONTACT



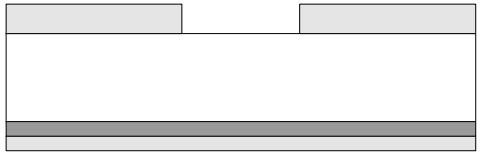
b) BACK CONTACT PROTECTED BY Si-NITRIDE CAPPING LAYER



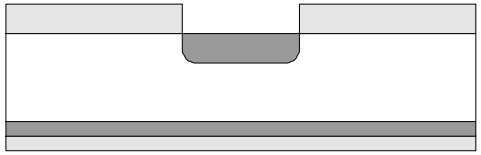
c) THERMAL OXIDATION OF TOP SURFACE



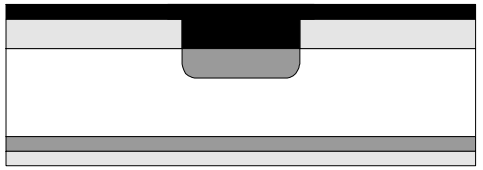
d) OPEN WINDOW FOR p^+ ELECTRODE



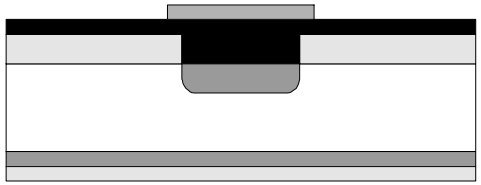
e) B-DOPING TO FORM p^+ ELECTRODE



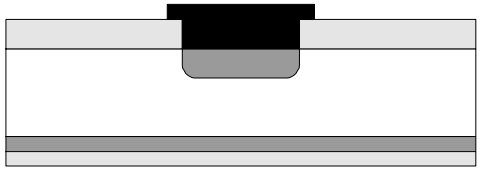
f) ALUMINUM METALLIZATION FOR FRONT CONTACT



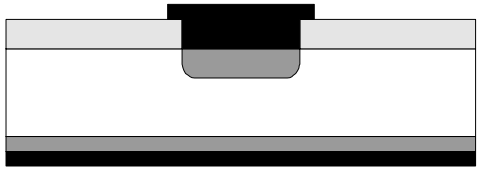
g) PHOTORESIST MASK FOR FRONT CONTACT



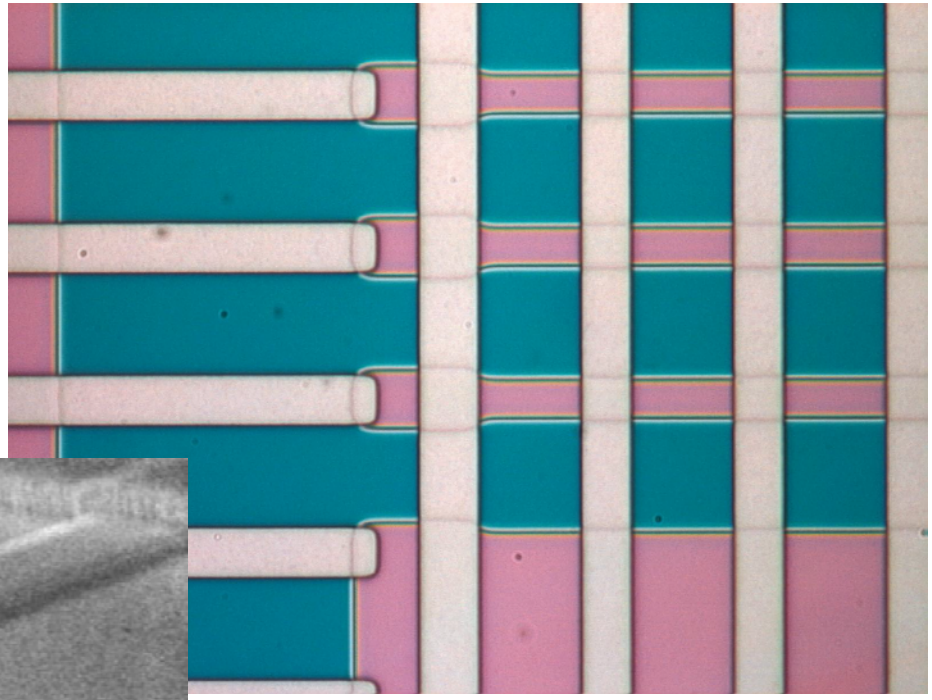
h) ETCH FRONT CONTACT

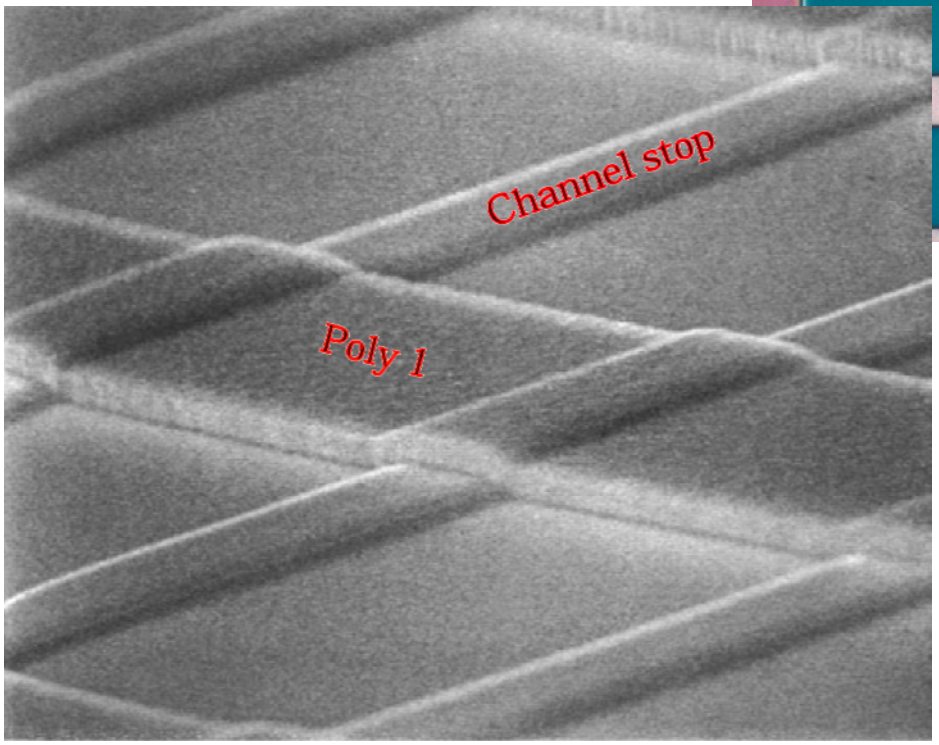


i) ALUMINUM METALLIZATION FOR BACK CONTACT



Optical and SEM Photographs of a Finished Device (CCD)





All of these process steps provide many opportunities for the introduction of deleterious contaminants.

Especially critical are

a) wet-process steps

Immersion in a liquid bath exposes the sample to many more molecules than in air, so liquid chemicals and the water used for dilution must be extremely pure (sub ppb contaminant levels).

b) thermal processing

High temperatures promote diffusion.

Two approaches have been taken in the fabrication of silicon detectors with low reverse bias currents.

a) Low temperature processing

(J. Kemmer, Nucl. Instr. and Meth. **226** (1984) 89)

pro: relatively simple and economical (no deposition systems required)
most commonly used for detectors

con: marginal activation of implants,
restricts use of most IC techniques
not compatible with monolithically integrated
electronics on same substrate

b) Gettering

(S. Holland, IEEE Trans. Nucl. Sci. **NS-36** (1989) 282, Nucl. Instr. and Meth. **A275** (1989) 537)

pro: very effective and removal of critical contaminants
reproducible
fully compatible with conventional IC processing

con: requires polysilicon deposition
some additional process complexity

Gettering can be used to remove contaminants from the sensitive regions by providing capture sites for contaminants.

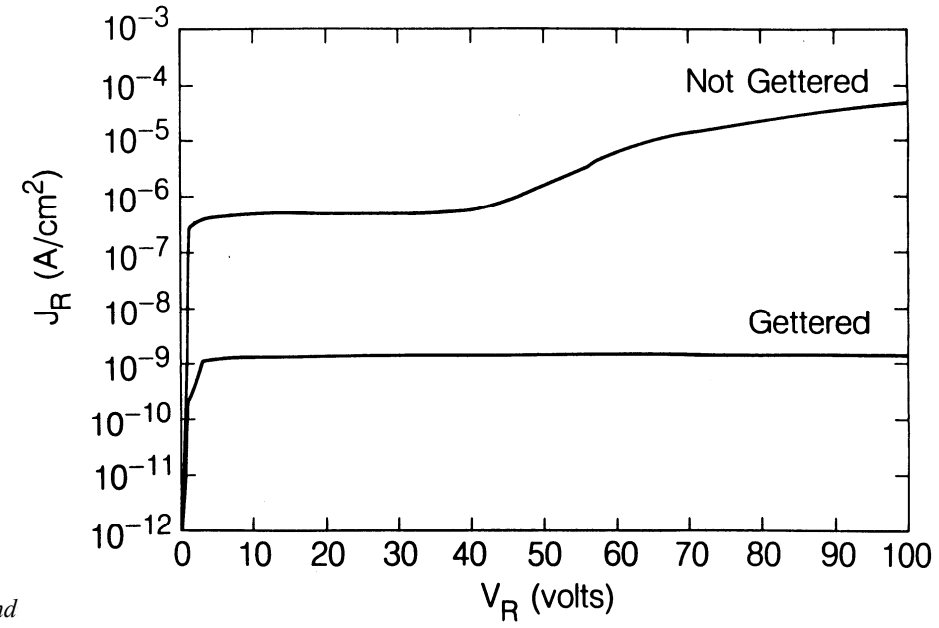
This requires that the critical contaminants are sufficiently mobile so that they will diffuse to the gettering sites and be captured.

Fortuitously, the most common contaminants that introduce mid-gap states are fast diffusers!

Disordered materials tend to be efficient getters (e.g. polysilicon).
Gettering can be promoted by chemical affinity (Phosporus)

Both can be combined, e.g. P-doped polysilicon

Reduction of diode reverse bias current by gettering (S. Holland, LBL):

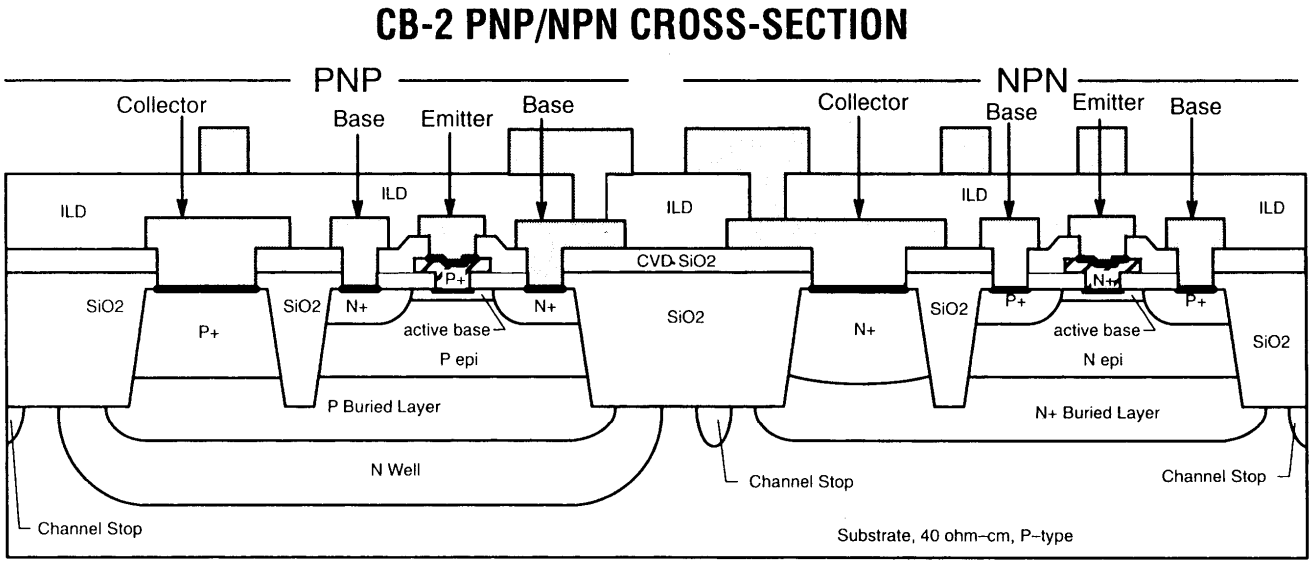


Integrated Circuits

A detector is a rather simple device.

The complexity of an integrated circuit is much greater, with a correspondingly larger number of masks and processing steps.

Complementary bipolar transistor process (npn and pnp transistors on same wafer)



(Maxim Integrated Products)

Modern IC processes typically require 15 to 18 masks, with more complicated processes using 24 or more.

Highly integrated front-end circuitry is a standard part of modern silicon vertex detectors in high-energy physics.

Apart from the essential reduction in size, custom designed ICs
(Application Specific Integrated Circuits – ASICs) offer significant electrical advantages

- The input device can be tailored to the detector and application

- Non-standard operating points can be used to obtain the optimum balance between

  noise

  speed

  power

- Can apply processes in ways not intended by the vendor

  Most low-noise vertex detector ICs have been designed using “digital” fabrication processes never intended for analog applications

  exploit high circuit density and

  provide mixed analog-digital circuitry

- Interference mitigation specific to experimental needs can be incorporated
  (see “Why Things Don’t Work”).

Circuit topology and design differs from discrete circuitry

- High-density CMOS processes do not provide resistors
  no problem for amplifiers (use capacitive feedback networks),
  but sometimes poses problems for DC biasing

- Absolute values of components not well-controlled, but relative matching very good.
  For example, tolerance of capacitors typ. 20%,
  but identical capacitors matched to 1%.

  $\Rightarrow$ use circuitry that depends on relative sizing, i.e.
  ratios of capacitances
  ratios of device sizes (FET widths)
  differential circuitry or balanced circuits

- High density circuitry invariably drives up complexity
  Circuit complexity can easily overwhelm designers
  Beware of “feature creep”.

  Complexity can increase reliability !
  – if architecture and circuitry are well-controlled

- Successful designs require great discipline

  Powerful simulation tools are available for
  - high-level simulations of architecture
  - circuit-level simulations
  - layout verification

  Although it is possible to perform “microsurgery” in some situations, in general this technology is very unforgiving.

  Once a circuit is “cast in silicon” and it doesn’t function, no tweaking or “cut and try” can change it.

  Extreme care is required in
  - the design of the devices and process,
  - the design of the circuit:
  - Are models used for simulation reliable?
  - Does the circuit provide latitude for process variations?
  - “Hooks” for external adjustments?
  - Conducting the fabrication process
  - Evaluation of prototypes

  Think hard about potential interference sources!

  Many research groups have produced successful ICs working reliably in experiments.

  Depends more on brain-power than money – it also helps to understand the physics.