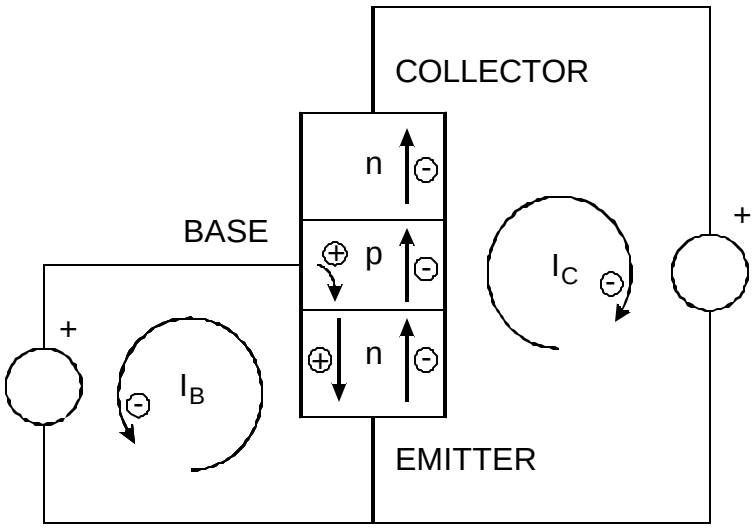
V. Amplifying Devices and Microelectronics

1. Bipolar Transistors

Consider the *npn* structure shown below.



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. In an ideal diode

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

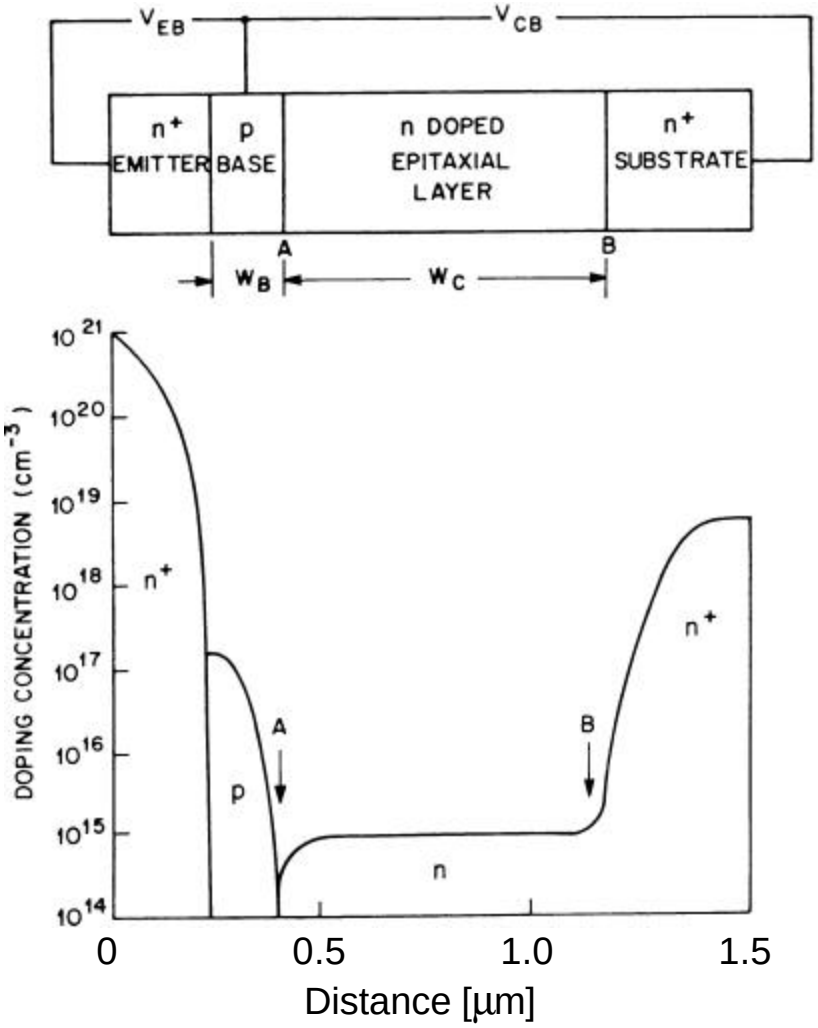
If the ratio of doping concentrations in the emitter and base regions N_D/N_A is sufficiently large, the collector current will be greater than the base current.

- 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.

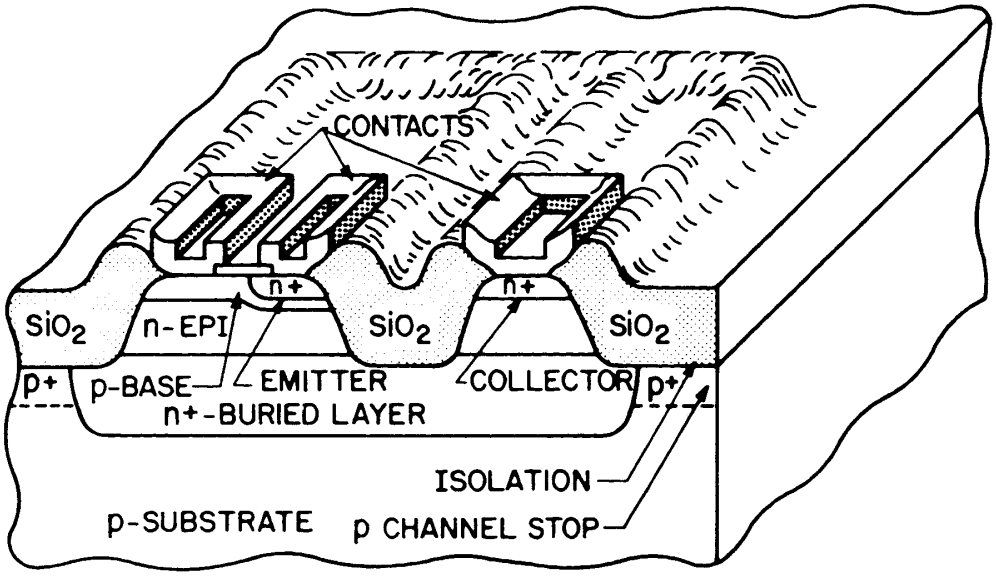
Since the current inside the transistor comprises both electrons and holes, the device is called a bipolar transistor.

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



(adapted from Sze)

High-speed bipolar transistors are implemented as vertical structures.



(from Sze)

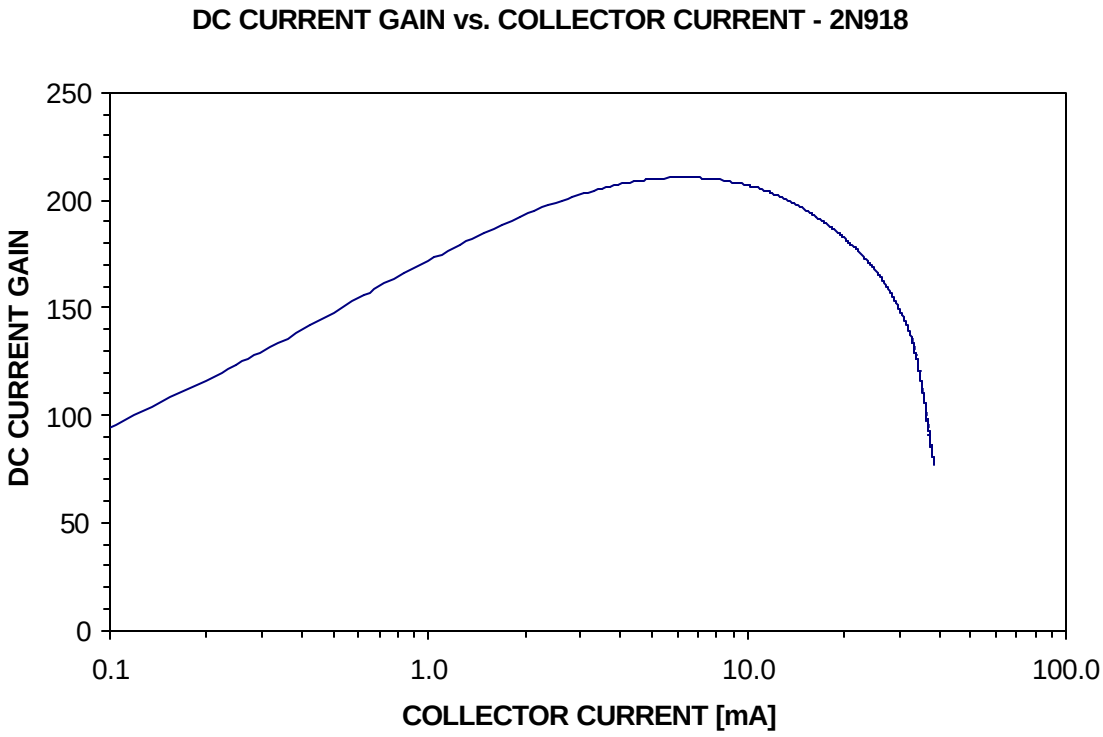
The base width, typically 0.2 μm or less in modern high-speed transistors, is determined by the difference in diffusion depths of the emitter and base regions.

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.

The result of this simple analysis implies that for a given device the DC current gain should be independent of current. In reality this is not the case.



The decrease at low currents is due to recombination in the base-emitter depletion region.

At high currents the current gain drops because of resistive voltage losses, shift of the high field region at the collector (increased base width), loss of injection efficiency as the carrier density approaches the doping concentration, etc.

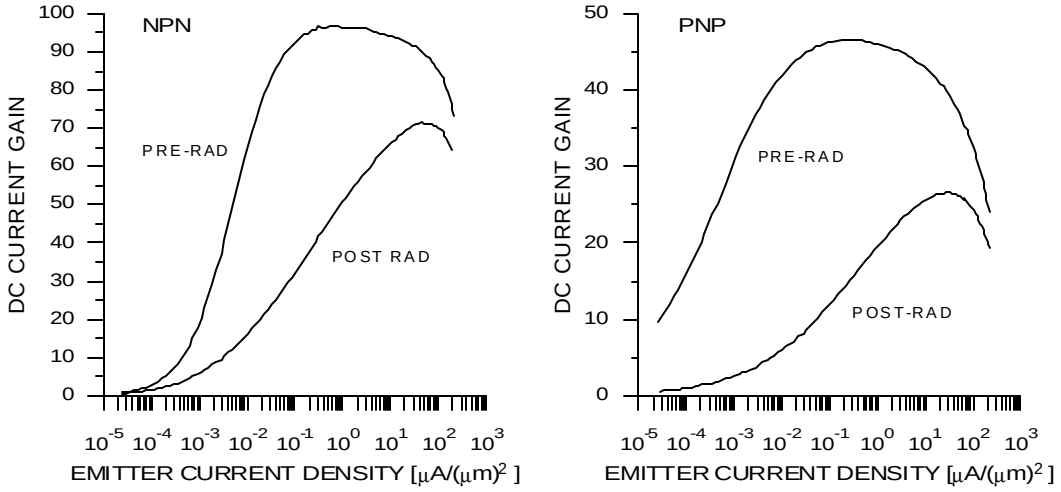
The “ideal” DC current gain depends only on device and material constants, whereas the recombination depends on the local density of injected electrons and holes with respect to the concentration of recombination centers.

Thus, the relative degradation of DC current gain due to recombination depends on the current density.

> 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.

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

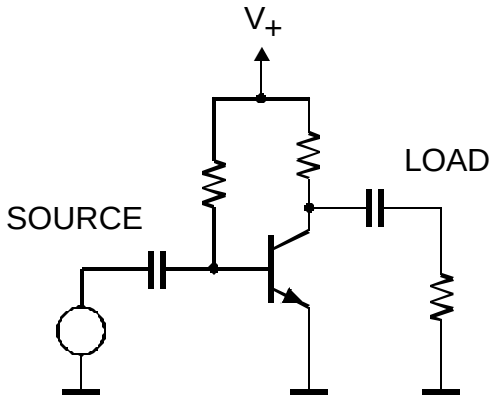


In a radiation-damaged transistor the reduction in DC current gain for a given DC current will be less for smaller devices.

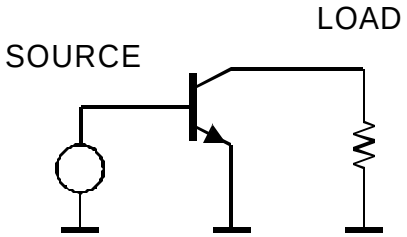
Bipolar Transistors in Amplifiers

Three different circuit configurations are possible:

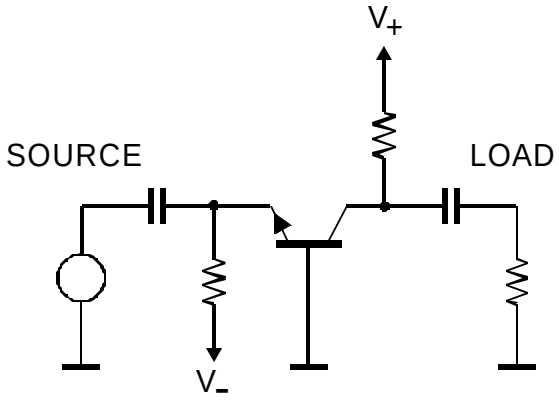
1. Common Emitter

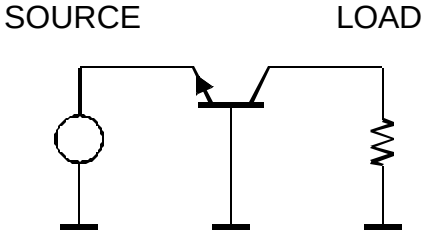


Equivalent Circuit

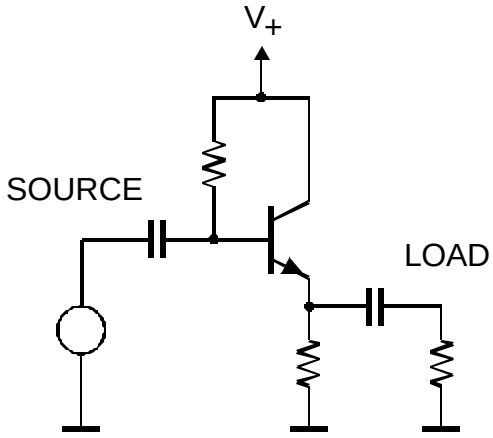


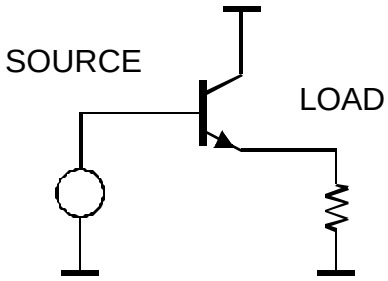
2. Common Base



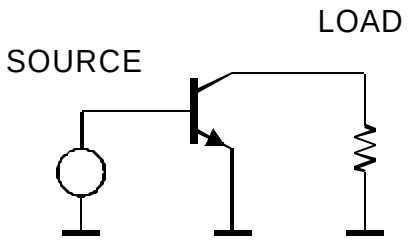


3. Common Collector (Emitter Follower)





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.

As discussed in a previous lecture, 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

$$I_C = b_{DC} I_B = b_{DC} I_R e^{q_e V_{BE}/k_B T}$$

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

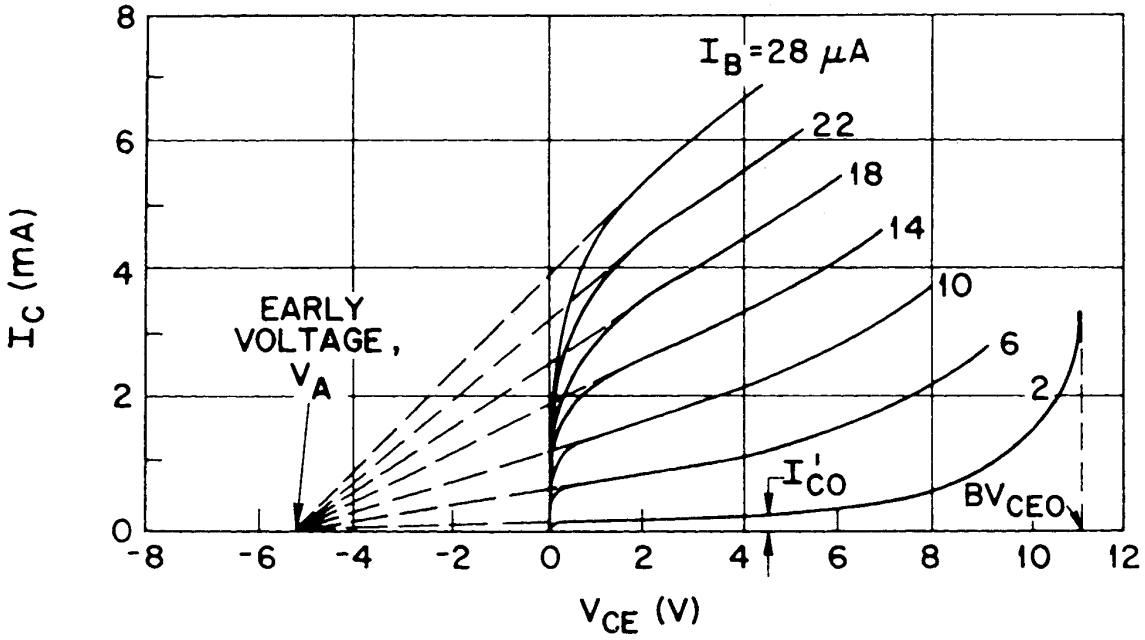
$$g_m \equiv \frac{dI_C}{dV_{BE}} = b_{DC} I_R \frac{q_e}{k_B T} e^{q_e V_{BE}/k_B T} = \frac{q_e}{k_B T} I_C$$

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

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

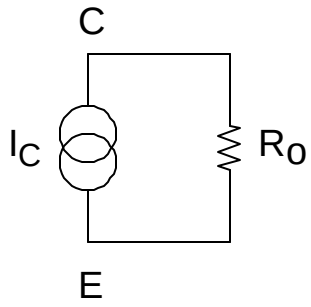
The obtainable voltage gain of an amplifier depends on the output characteristics of the transistor.



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

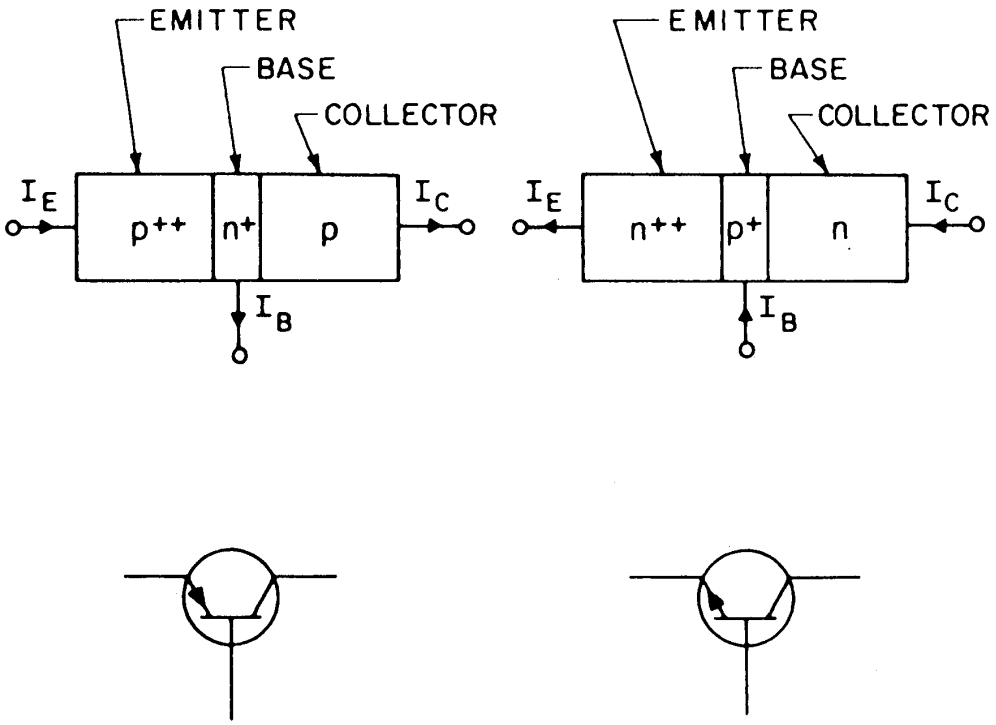
$$A_{v,\max} = \frac{dI_C}{dV_{BE}} R_o = g_m R_o \approx \frac{I_C}{k_B T / q_e}\frac{V_A}{I_C} = \frac{V_A}{k_B T / q_e}$$

which at room temperature is about ($40V_A$).

- 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 *pnp* or *npn* structures.



The polarities of the applied voltages are the opposite:

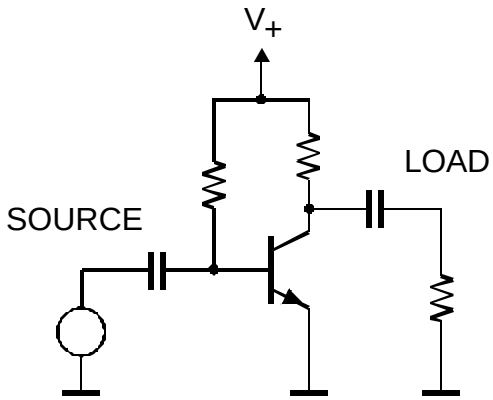
npn: positive collector-emitter and base emitter voltages

pnp: negative collector-emitter and base emitter voltages

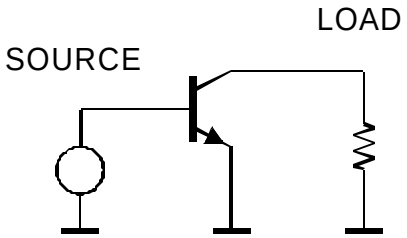
The basic amplifier equations are the same for both transistor types.

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

a) Common Emitter configuration



Equivalent Circuit



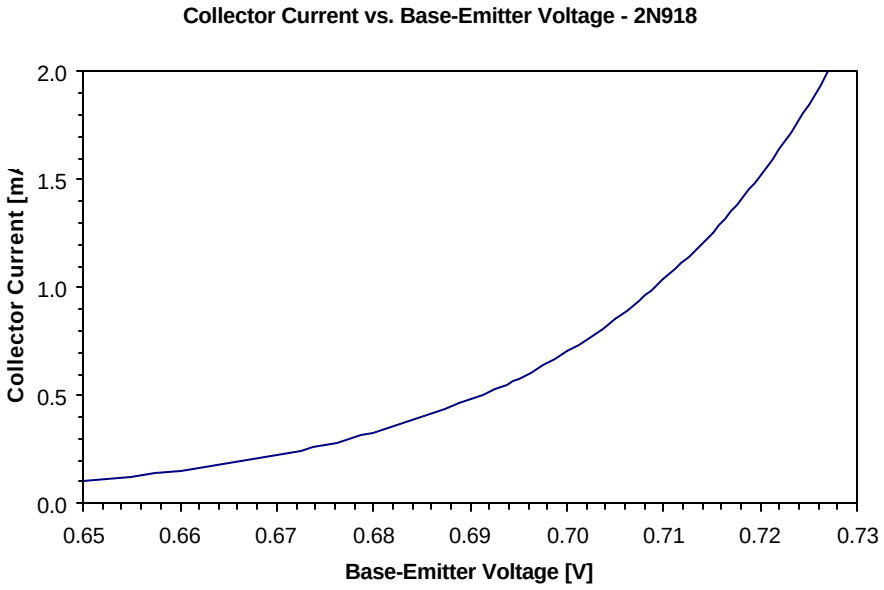
The input signal is applied to the base, the output taken from the collector.

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

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

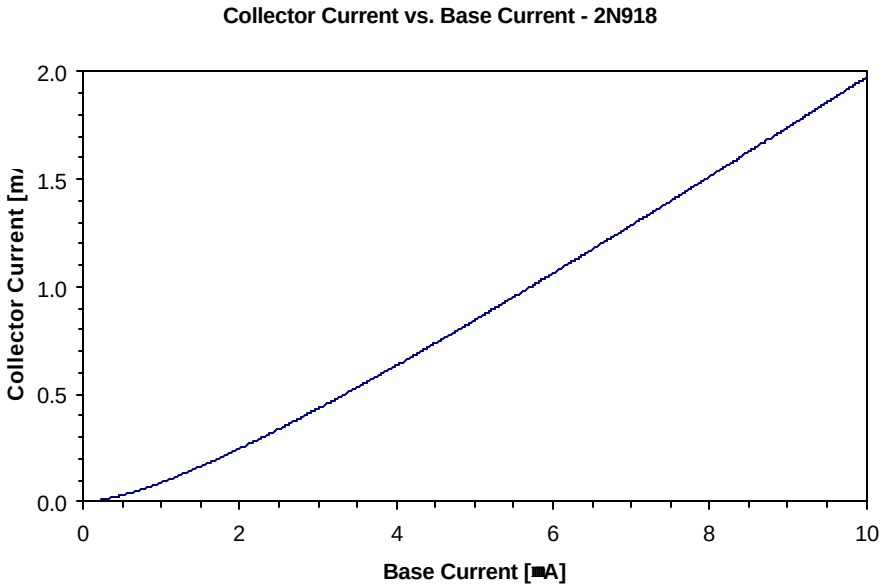
For b_{DC} = 100 and I_c= 1 mA, R_i= 2600 Ω.

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.

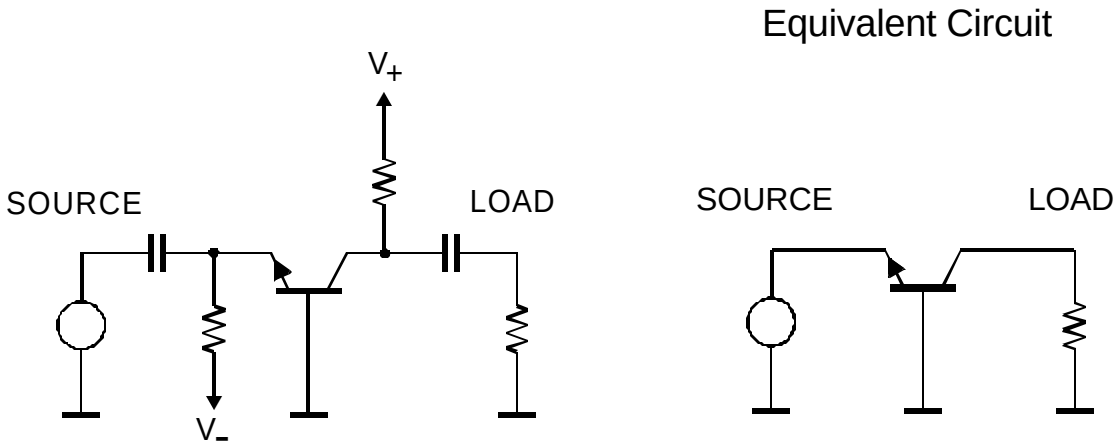


In audio amplifiers, for example, this causes distortion.
Distortion may be limited by restricting the voltage swing, which to some degree is feasible because of the high transconductance.

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

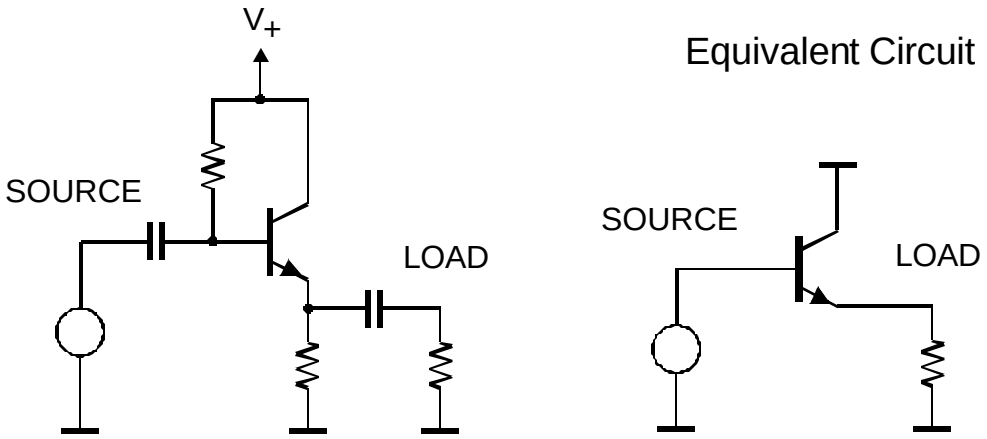
$$R_i = \frac{0.026}{I_C}$$

i.e. R_i= 26 Ω at I_c= 1 mA.

The input resistance is about $1/b$ 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,

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

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

$$\frac{dV_i}{dI_i} = \frac{dV_i}{dI_B} \approx b\, R_L$$

Thus, the input resistance depends on the load: $R_i \approx bR_L$

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

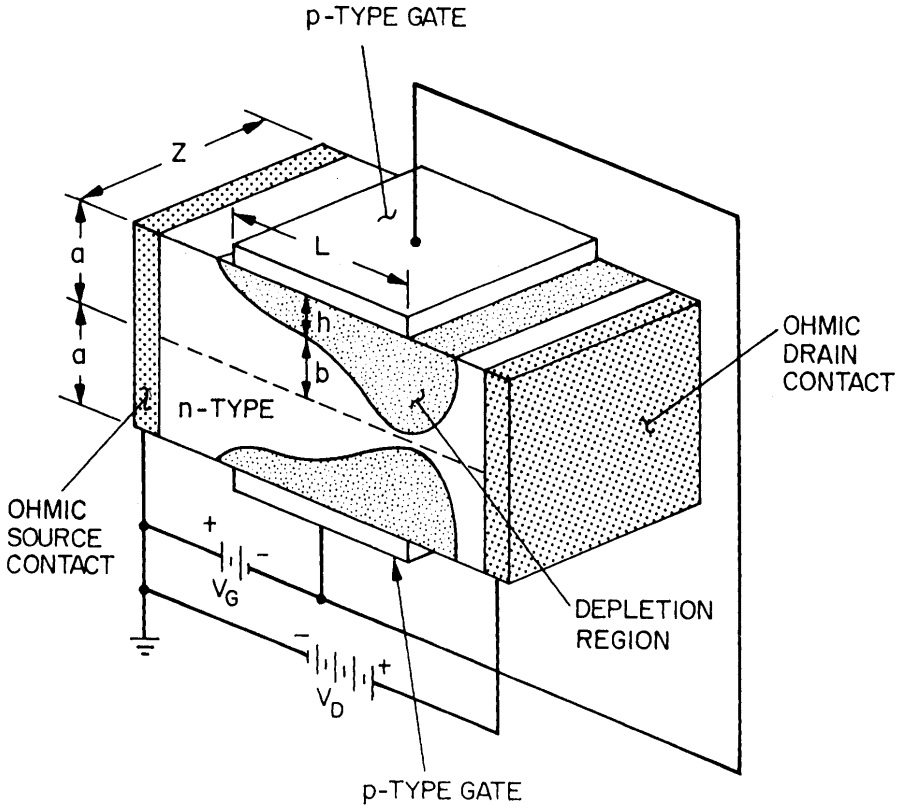
2. Field Effect Transistors

Field Effect Transistors (FETs) utilize a conductive channel whose resistance 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.

In the implementation shown below a pair of gate electrodes of opposite doping with respect to the channel are placed at opposite sides of the channel. Applying a reverse bias forms a depletion region that reduces the cross section of the conducting channel.



(from Sze)

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

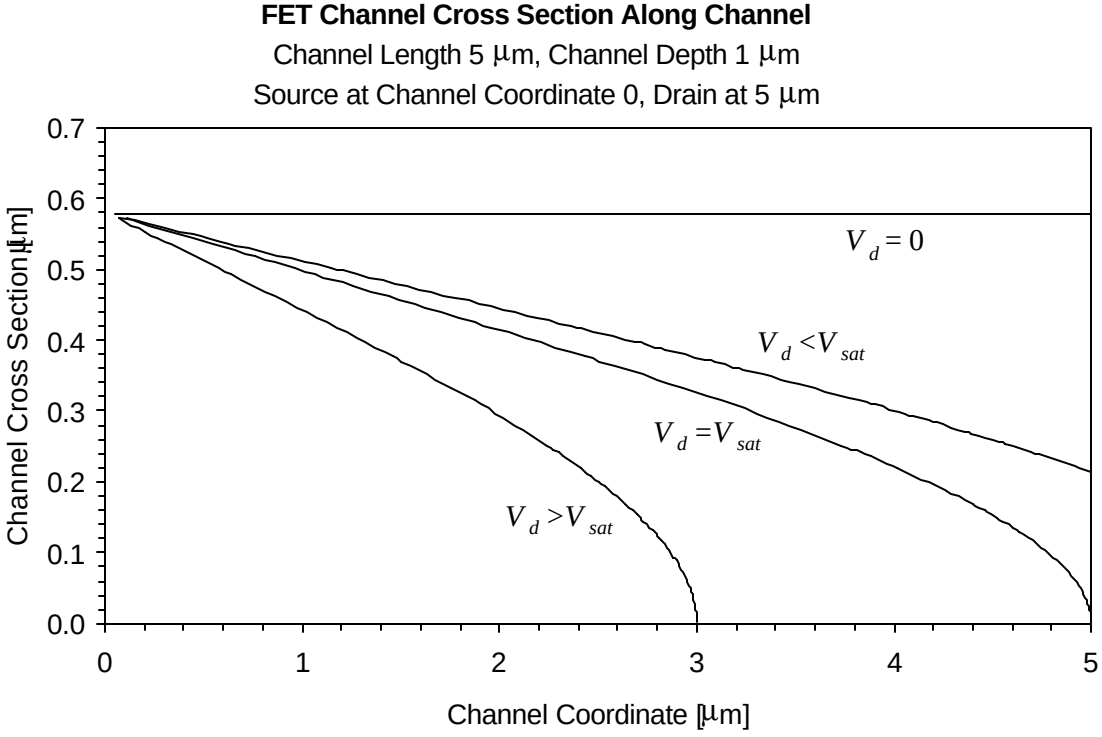
First assume that the drain voltage is 0.
Increasing the reverse gate potential will increase the depletion width, i.e. reduce the cross section of the conducting channel, until the channel is completely depleted. The gate voltage where this obtains is the “pinch-off voltage” V_P.

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”.

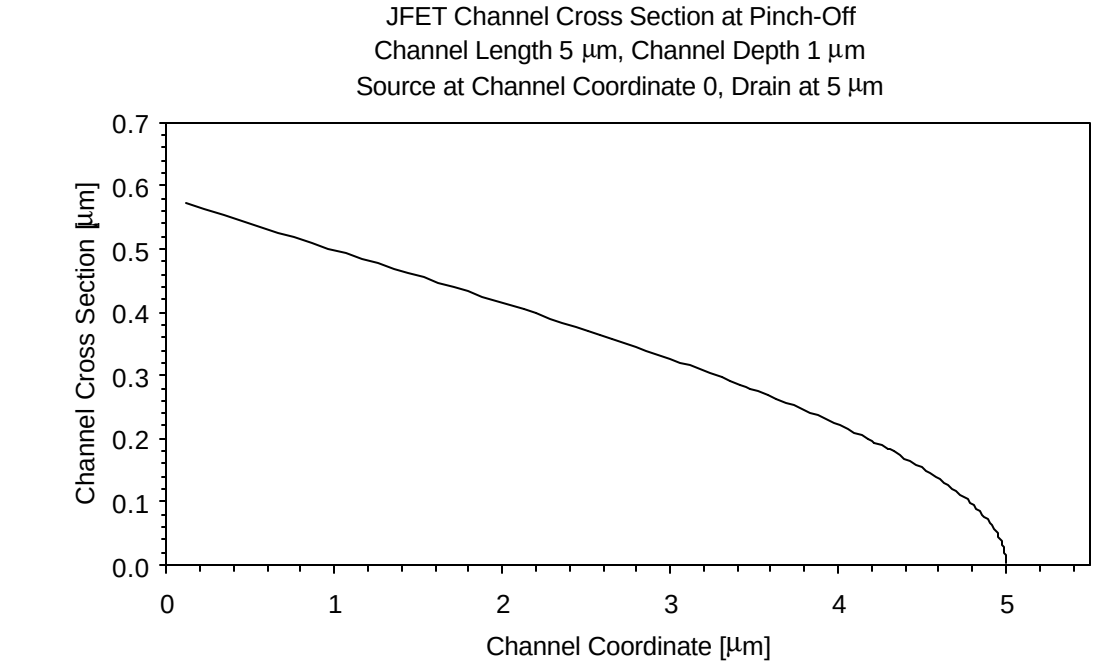
Increasing the drain voltage beyond this point moves the pinch-off point towards to the source.

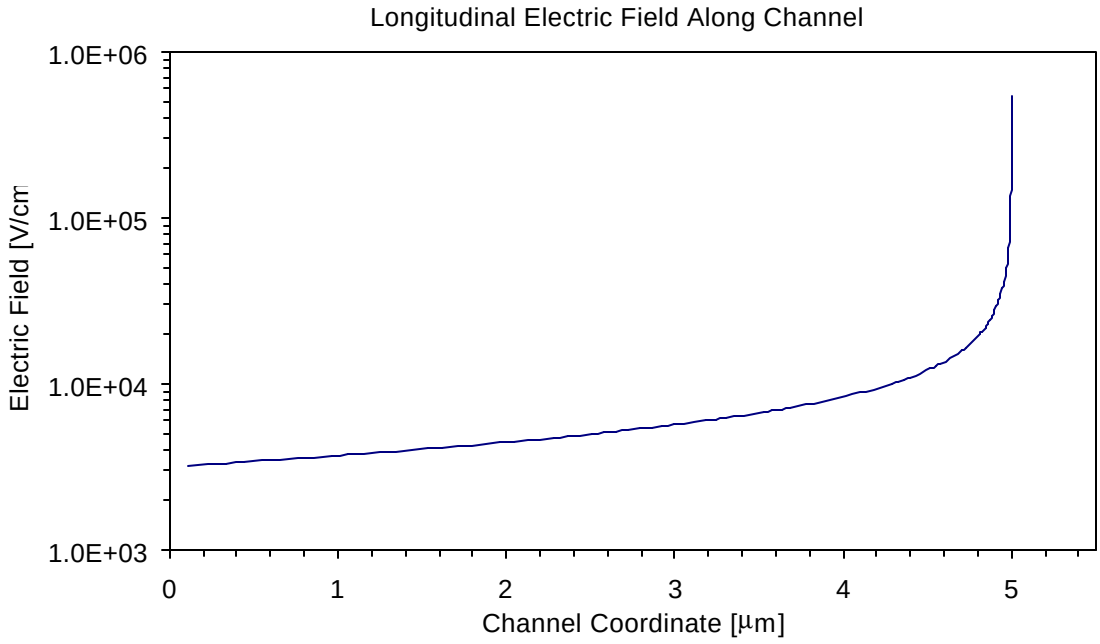


Pinching off the channel does not interrupt current flow. All thermally excited carriers have been removed from the depleted region, but carriers from the channel can still move through the potential drop to the drain.

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

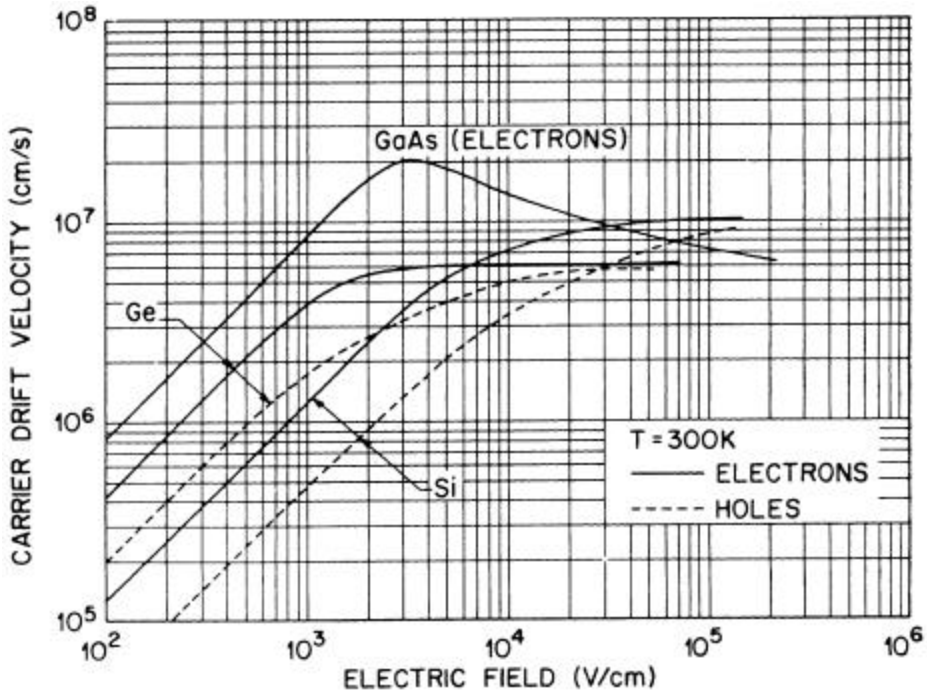
Current flow along the channel changes the local potential.
As the channel cross section decreases, the incremental voltage drop increases, i.e. the longitudinal drift field that determines the carrier velocity increases.





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.



(from Sze)

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, i.e. at high fields silicon acts as an incremental insulator ($dI/dV = 0$).

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.

- 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.

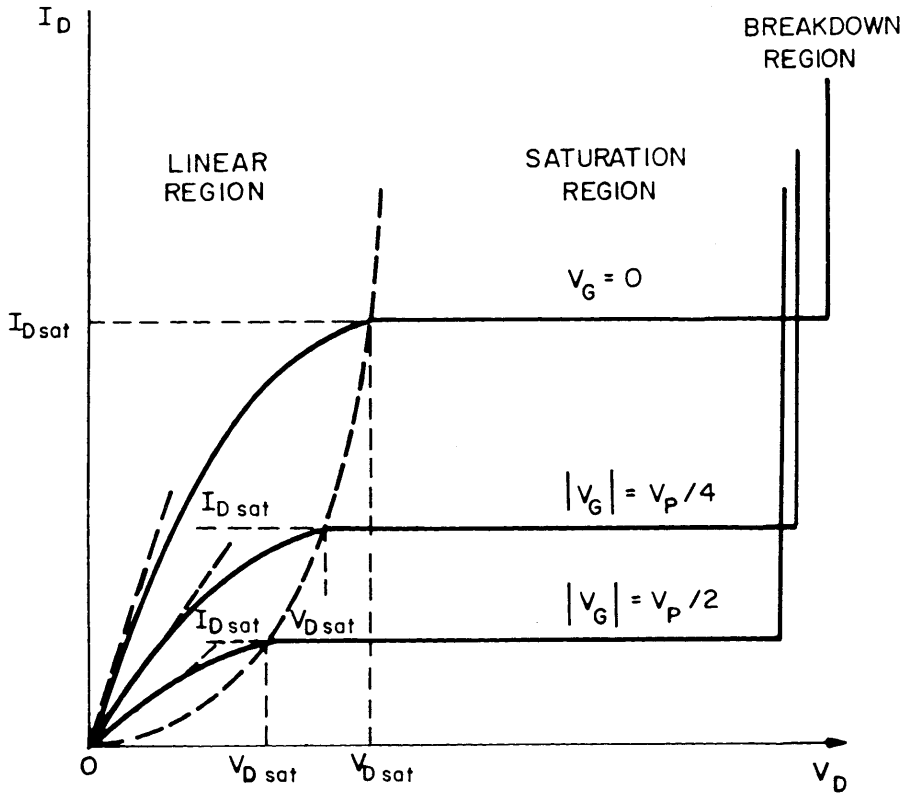
The drain current increases linearly with drain voltage.

"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"



(from Sze)

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

$$I_{DSS} = \frac{1}{6e} m (q_e N_{ch})^2 d^3 \frac{W}{L}$$

is determined by the carrier mobility m, the doping level in the channel N_{ch} and the channel depth d, width W and length L. e 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

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

Obviously, a high carrier mobility will increase the transconductance, since for a given carrier concentration this will increase the magnitude of the current.

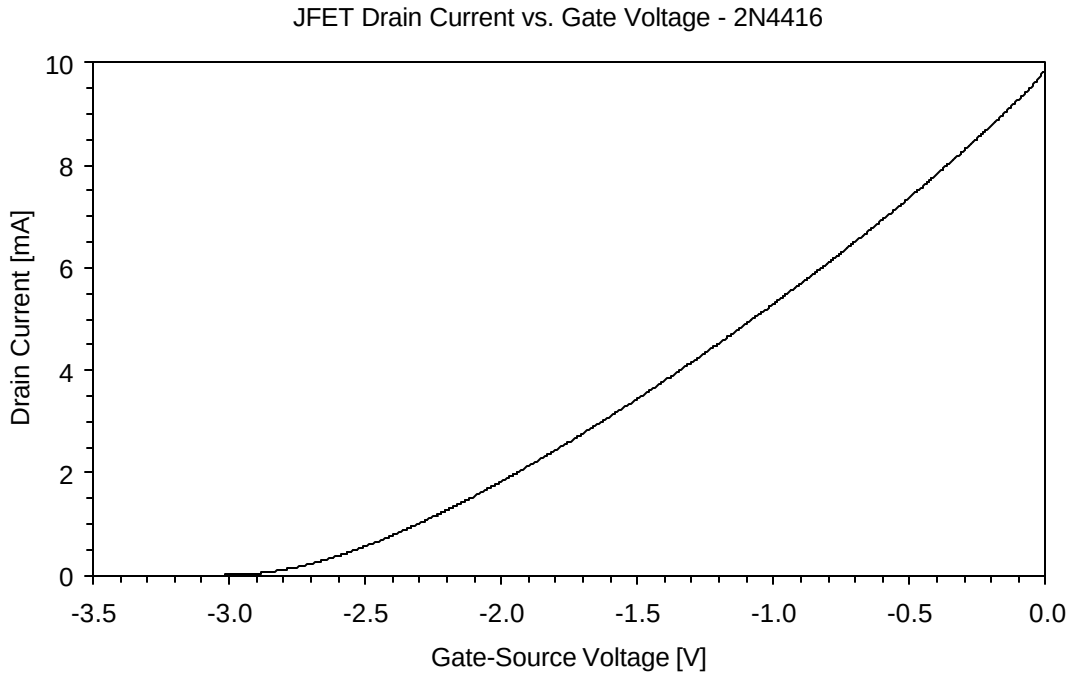
1. The proportionality of transconductance to width W is trivial, since it is equivalent to merely connecting device in parallel, so the normalized transconductance g_m/W is used to compare technologies.
2. The normalized transconductance increases with the number of carriers per unit length $N_{ch}d$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

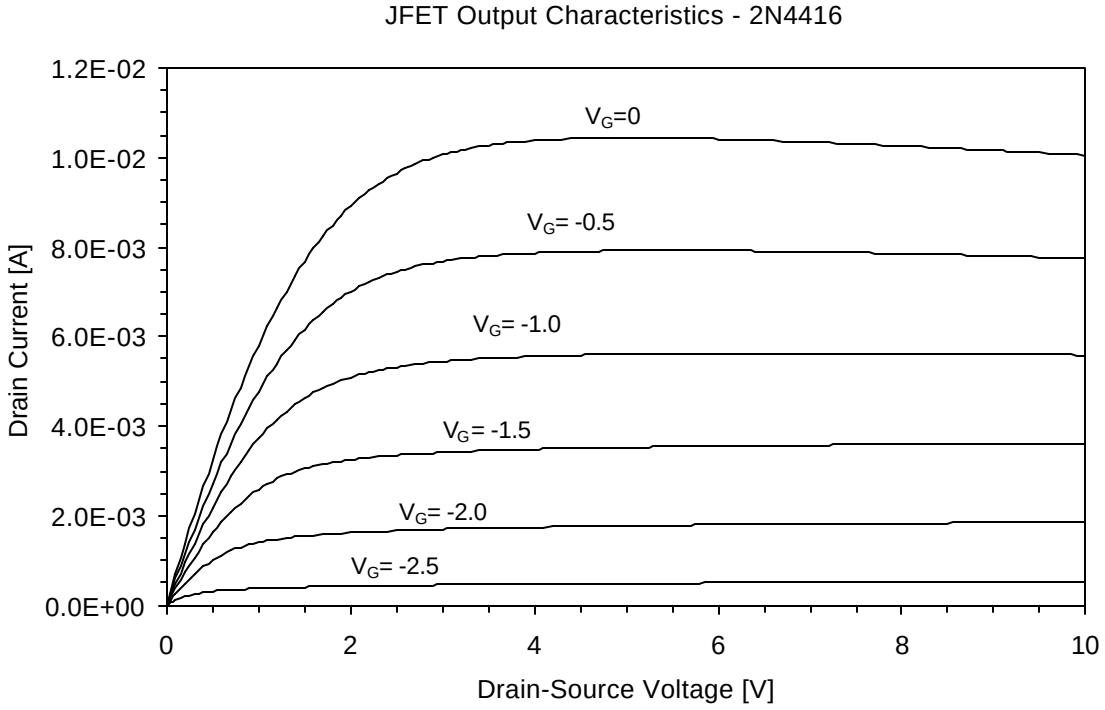
$$g_m = \left|\frac{dI_D}{dV_G}\right| = \frac{2I_{DSS}}{V_P}\left(1 - \frac{V_G + V_{bi}}{V_P}\right) = \frac{2\sqrt{I_{DSS}}}{V_P}\sqrt{I_D}$$

   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



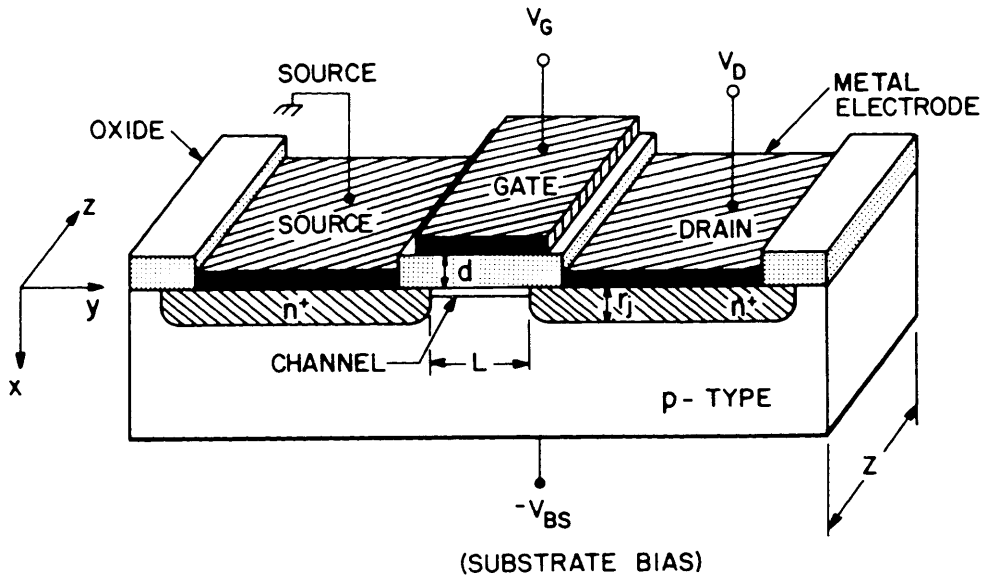


The slight drop in drain current for V_G=0 is due to self-heating.

Metal Oxide Field Effect Transistors (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.

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.



(from Sze)

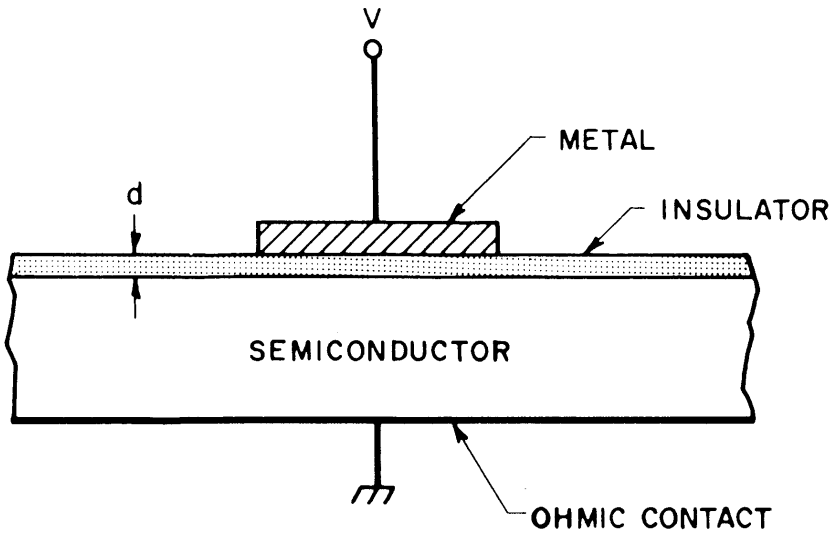
The source and drain are n^+ regions in a p-substrate.

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.

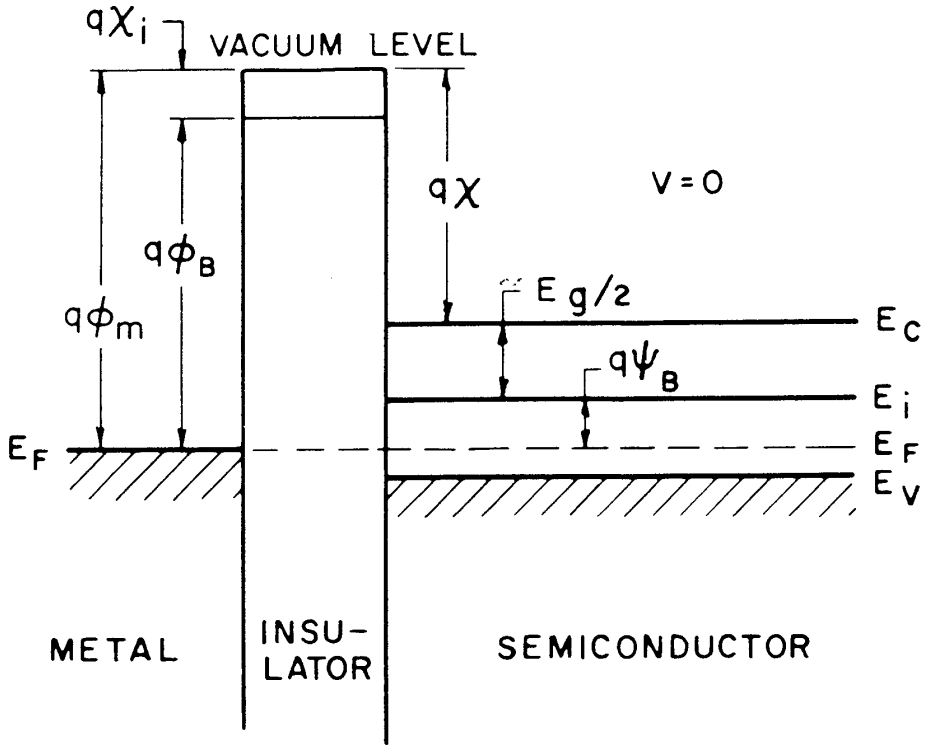
As in a JFET the combination of gate and drain voltages control the conductivity of the channel.

Formation of the Channel - The MOS Capacitor



(from Sze)

Band structure in an ideal MOS capacitor on a *p*-Substrate

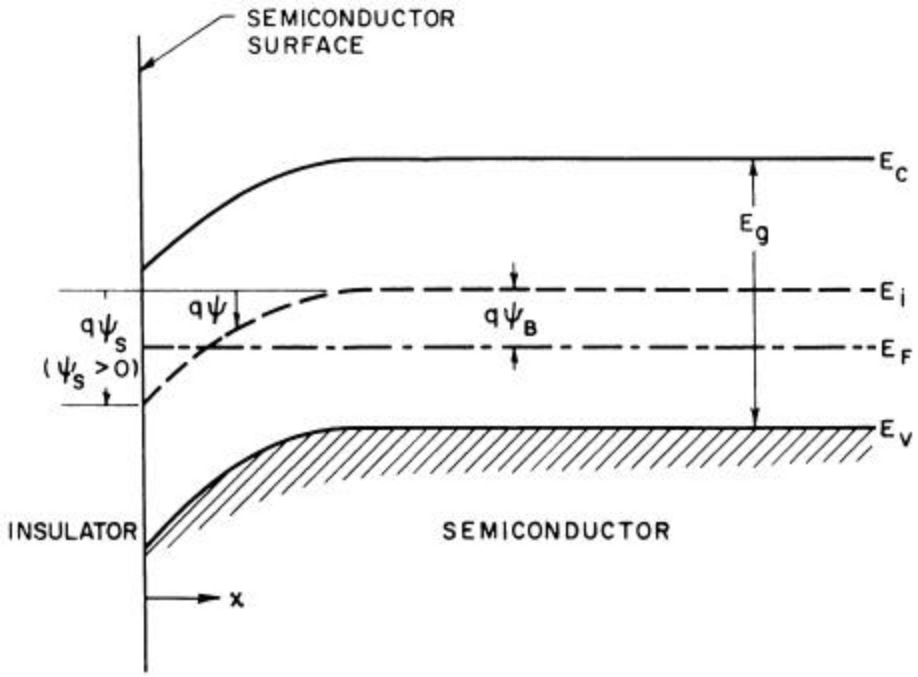


(from Sze)

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.

Energy band diagram of a p-type semiconductor



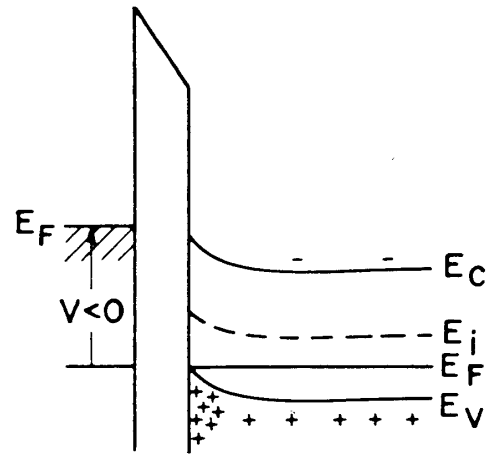
(from Sze)

As shown above the surface potential Y_s is positive.

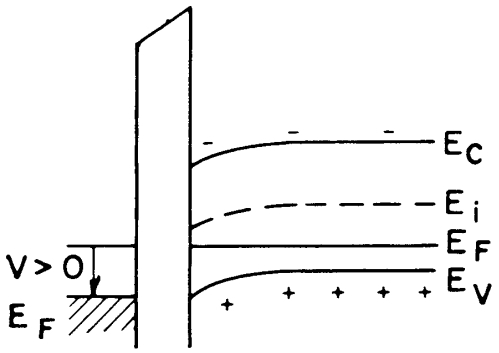
$Y_S < 0$	the bands bend upwards, increasing the hole concentration at the surface.
$Y_B > Y_S > 0$	the bands bend slightly downwards, reducing the concentration of holes at the surface (depletion)
$Y_S > Y_B$	the conduction band edge dips below the Fermi level, leading to an accumulation of electrons 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.

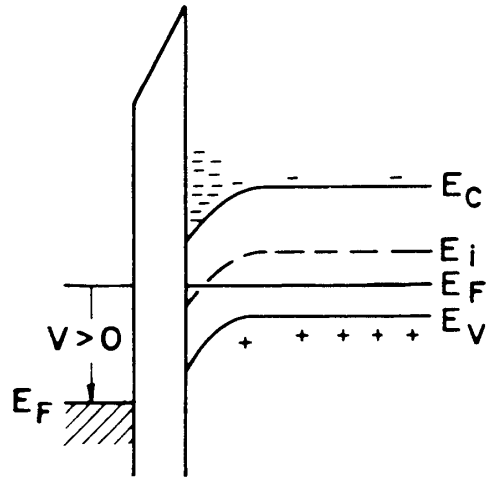
Surface concentration vs. band-bending in *p*-type material



Accumulation



Depletion



Inversion

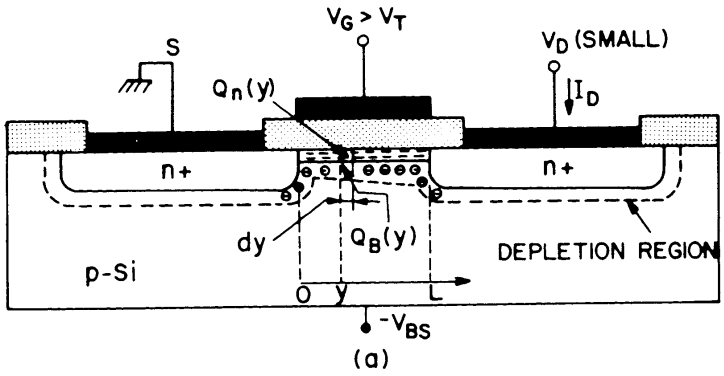
(from Sze)

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”.

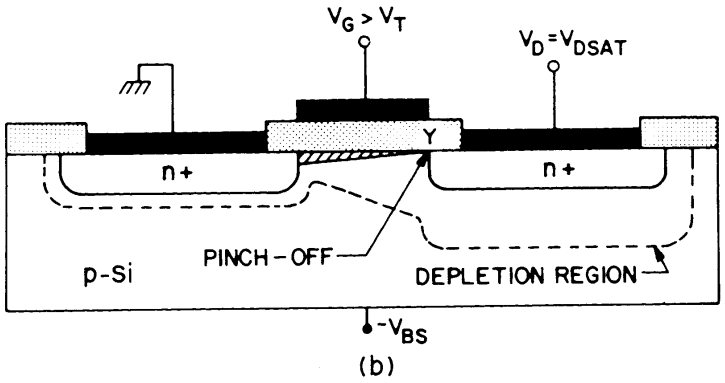
Low drain voltage
$V_D < V_{sat}$

Resistive channel



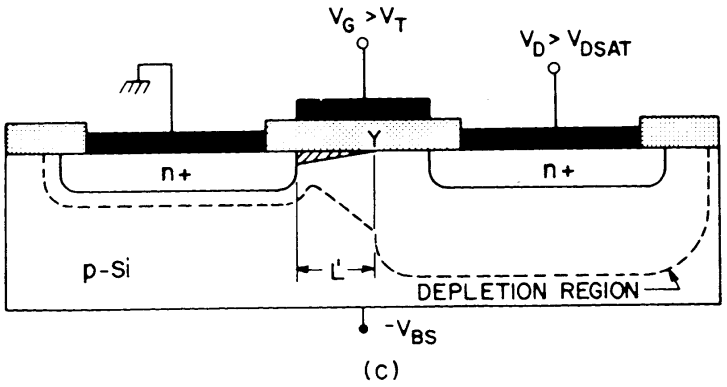
$V_D = V_{sat}$

Onset of current saturation



$V_D > V_{sat}$

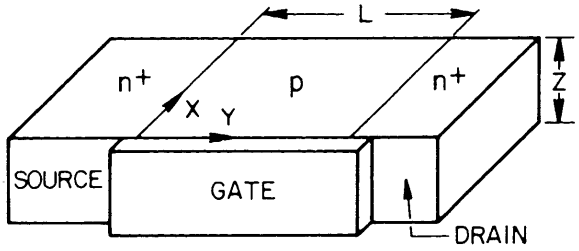
Output current saturated



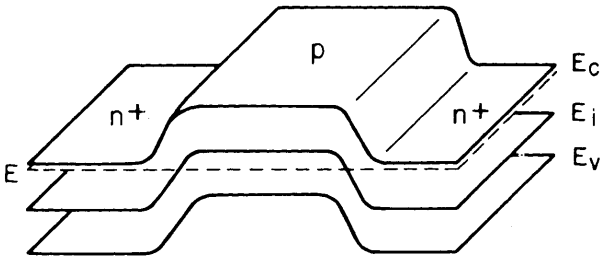
(from Sze)

Three-Dimensional Potential Diagram of an *n*-Channel MOSFET

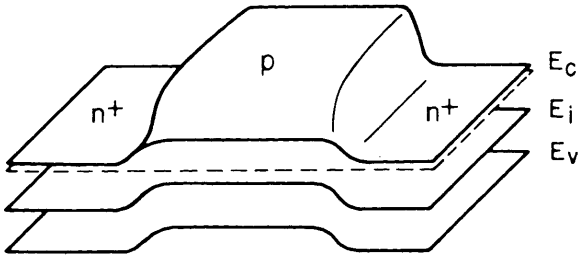
Device orientation



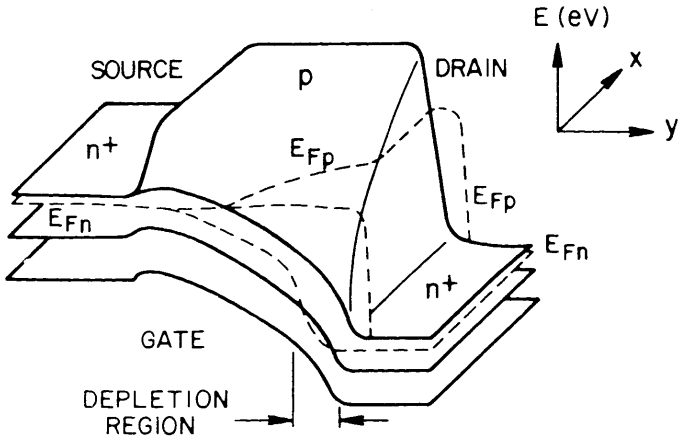
Flat-band condition



Formation of inversion layer



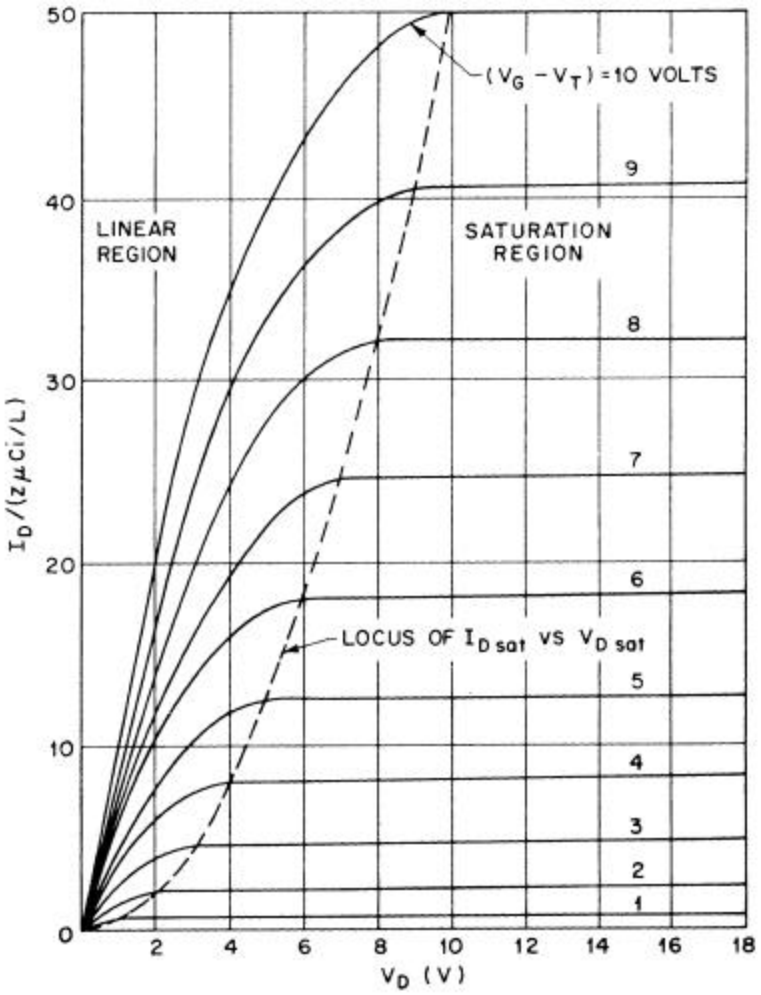
Drain voltage beyond saturation



(from Sze)

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

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



(from Sze)

In saturation

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

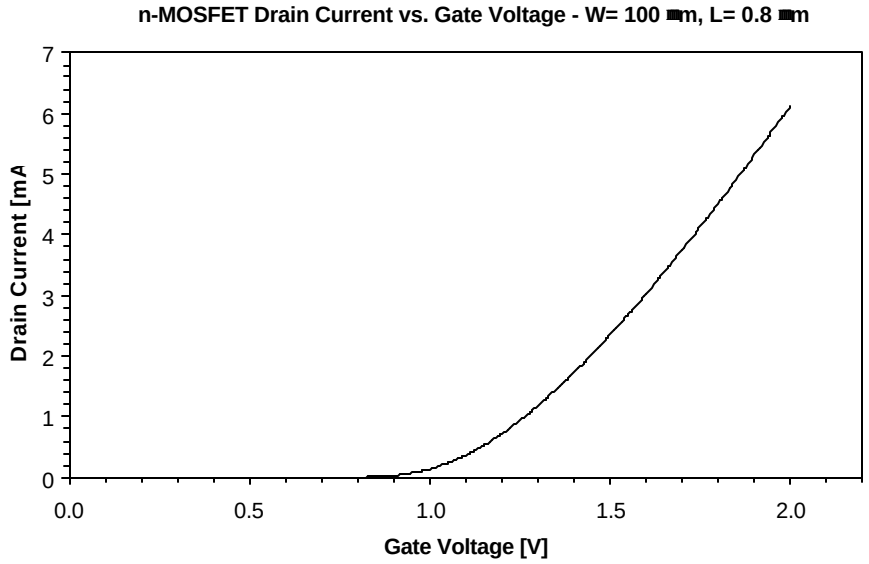
where C_i is the gate capacitance per unit area e_{ox}/d_{ox} and V_T is the gate voltage corresponding to the onset of strong inversion ("threshold voltage")

From this the transconductance is

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

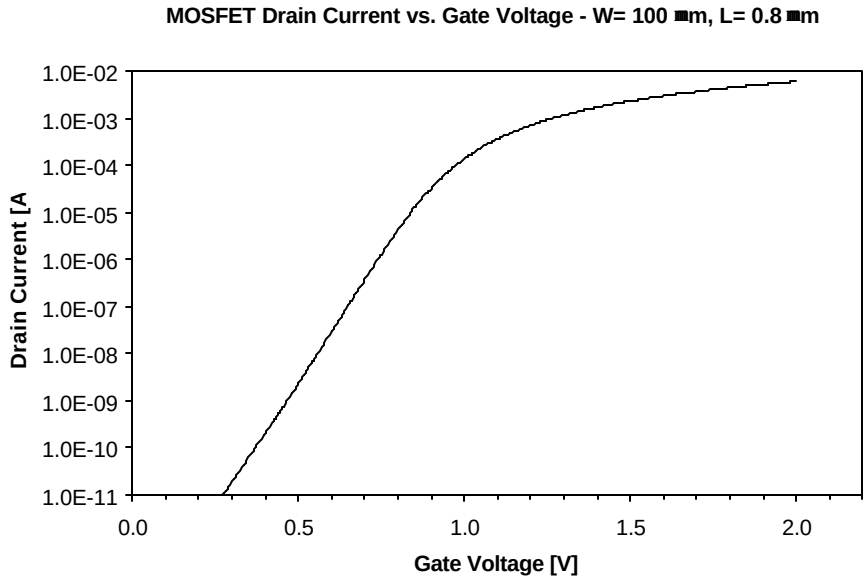
For a given width W and drain current I_D the transconductance is increased by decreasing the channel length L and the thickness of the gate oxide d_{ox}. As for the JFET, the transconductance depends primarily on current.

Measured characteristics of an *n*-channel MOSFET with 0.8 μm channel length and 20 nm gate oxide thickness



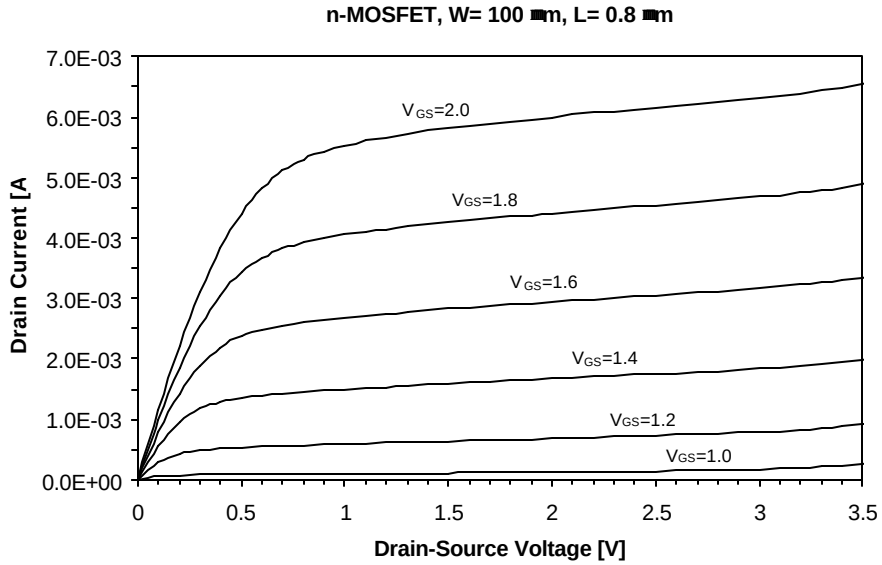
For this device the threshold voltage V_T is about 1.2 V.

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



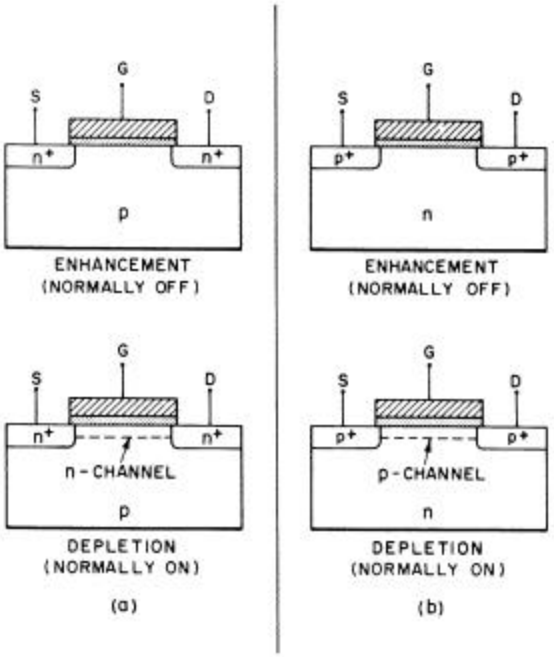
In the subthreshold regime the drain current is proportional to the inversion carrier concentration, i.e. it increases exponentially with gate voltage.

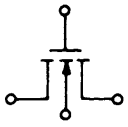
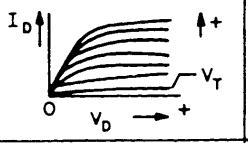
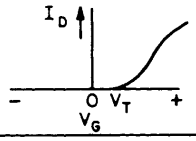
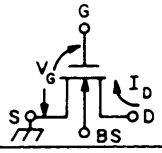
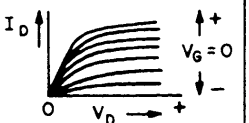
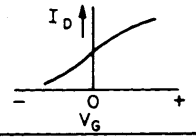
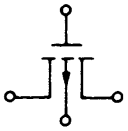
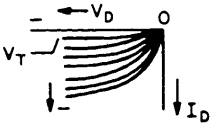
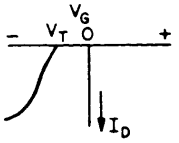
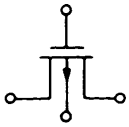
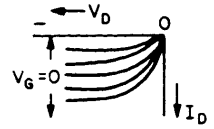
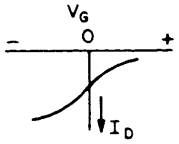
The increase in output saturation voltage with gate voltage can be seen clearly in the output curves.



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), as illustrated below

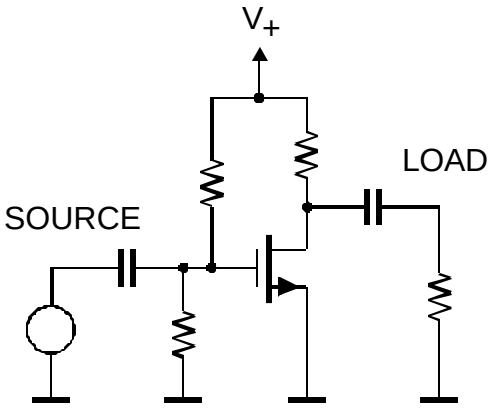


TYPE	ELECTRICAL SYMBOL	OUTPUT CHARACTERISTIC	TRANSFER CHARACTERISTIC
N-CHANNEL ENHANCEMENT (NORMALLY OFF)			
N-CHANNEL DEPLETION (NORMALLY ON)			
P-CHANNEL ENHANCEMENT (NORMALLY OFF)			
P-CHANNEL DEPLETION (NORMALLY ON)			

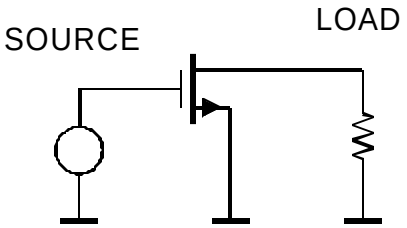
MOS Transistors in Amplifiers

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

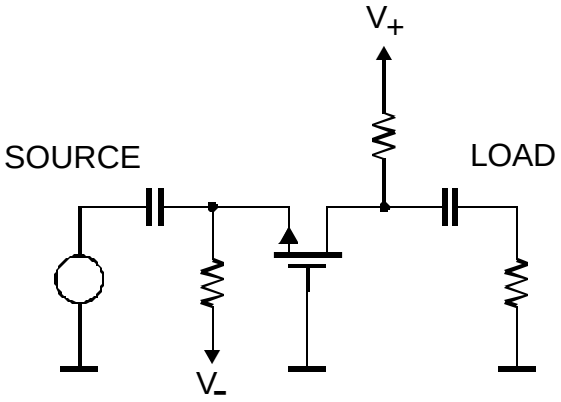
1. Common Source

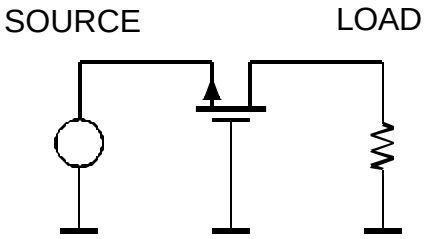


Equivalent Circuit

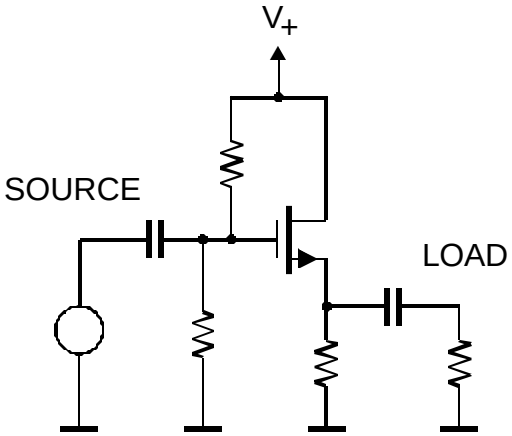


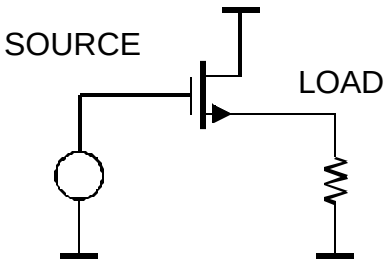
2. Common Gate





3. Common Drain (Source Follower)

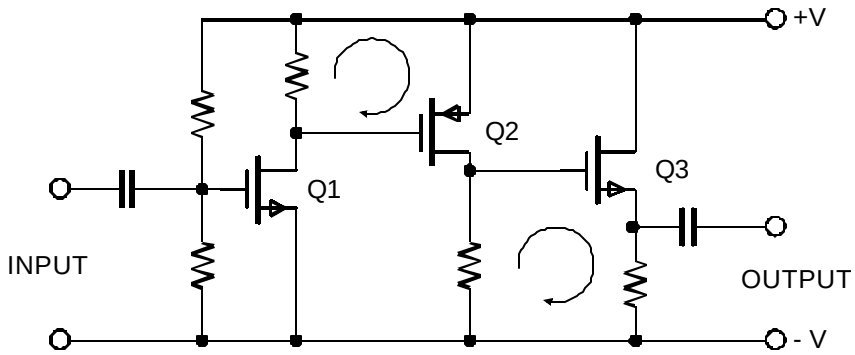




Advantages of MOS/CMOS

- high input resistance (capacitive gate)
- complementary devices (NMOS, PMOS)
- magnitude of gate voltages allows simple direct coupling
- tailoring of device characteristicschoice by choice of geometry (W, L)
- low-power, high-density logic circuitry (CMOS)

Amplifier cascade using NMOS and PMOS devices



Drawbacks

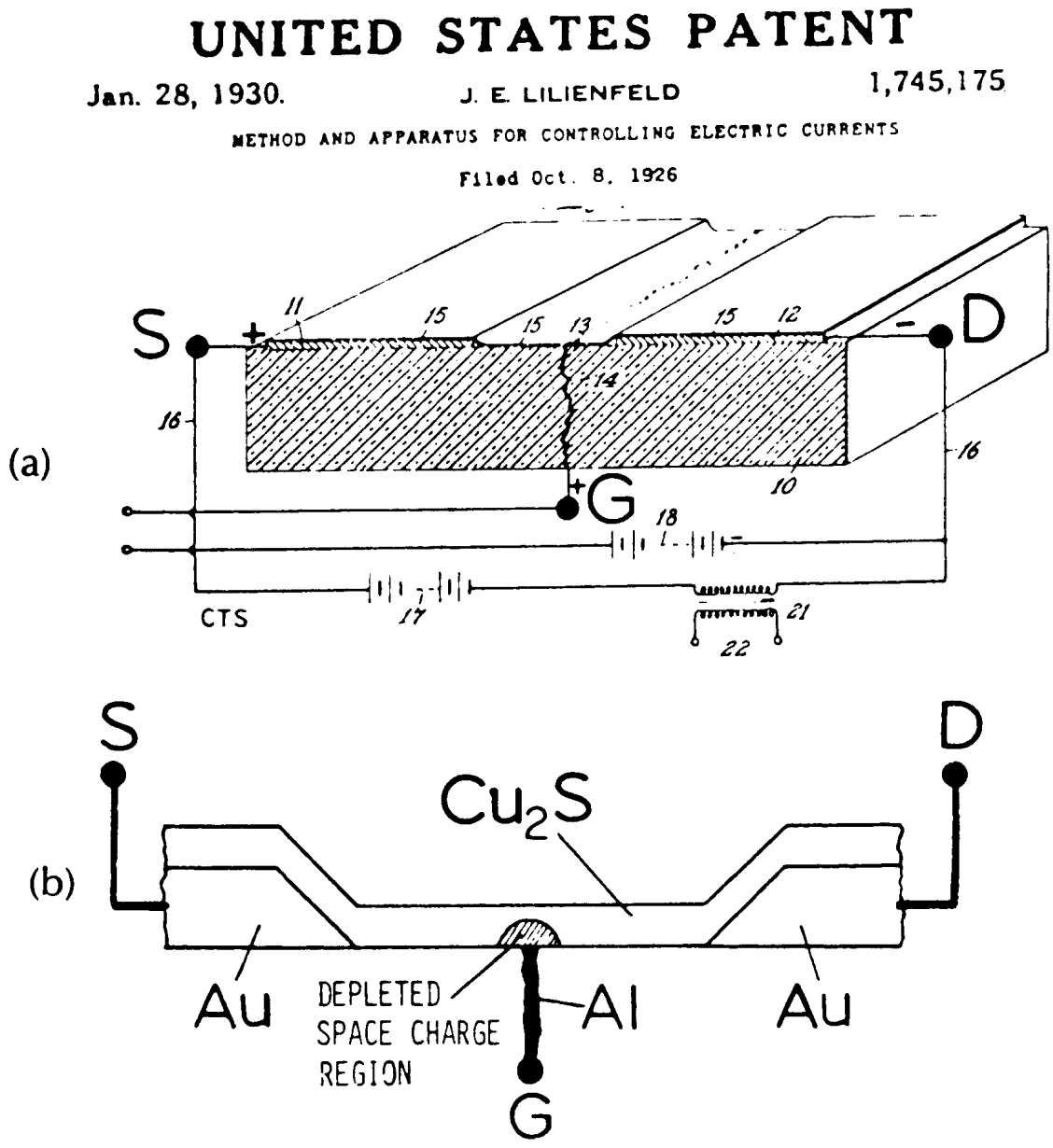
relative to bipolar transistors …

- 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)

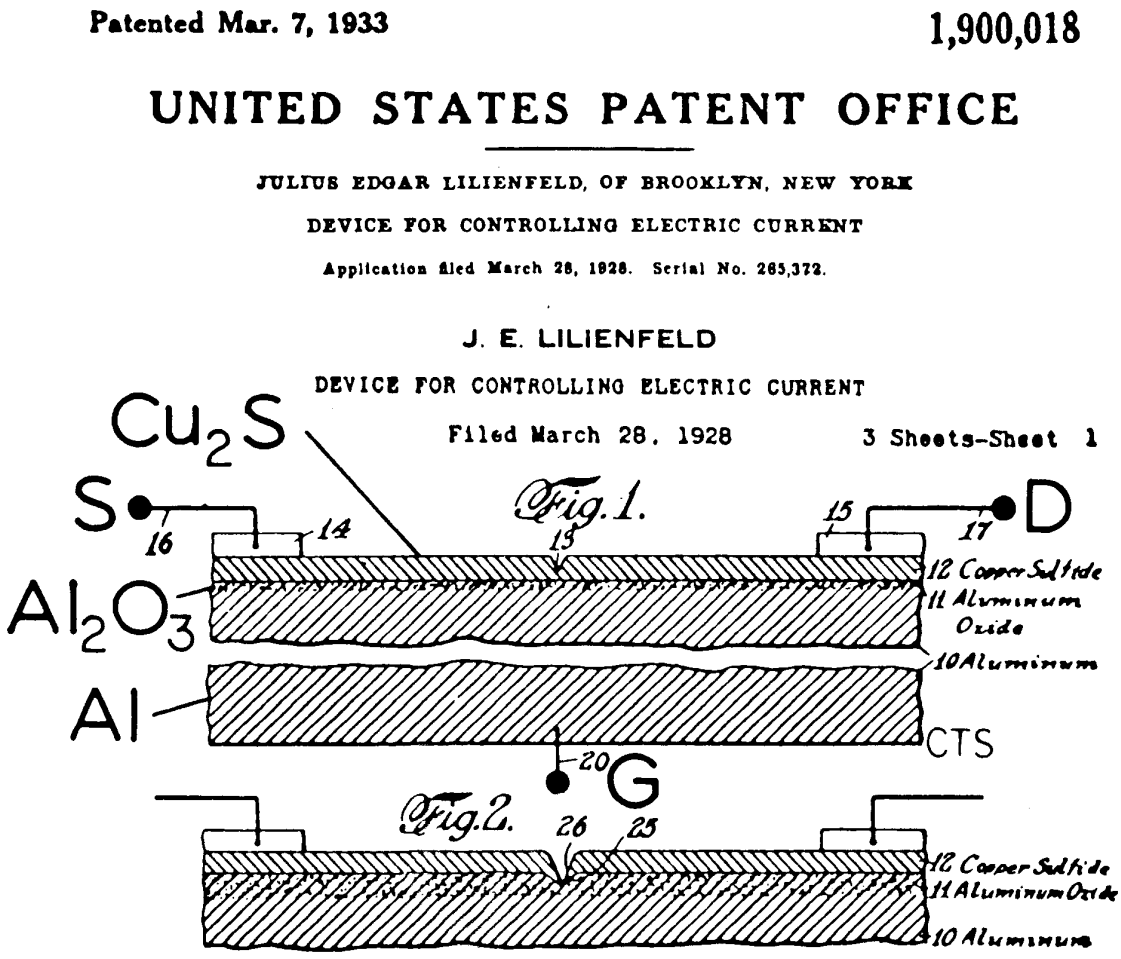
Modern fabrication processes combine bipolar transistor and MOS technology to exploit best features of both (BiCMOS).

… 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



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

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.

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.

3. Noise in Transistors

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

$$i_{nd}^2 = \frac{N_{Ctot} q_e}{L^2} m_0 \; 4 k_B T_e$$

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 m_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

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

one can express the noise current as

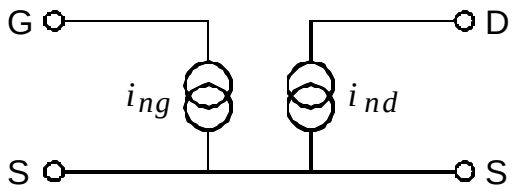
$$i_{nd}^2 = g_n g_m \; 4 k_B T_0$$

where T_0= 300K and g_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

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

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

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

The noise coefficient g_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.

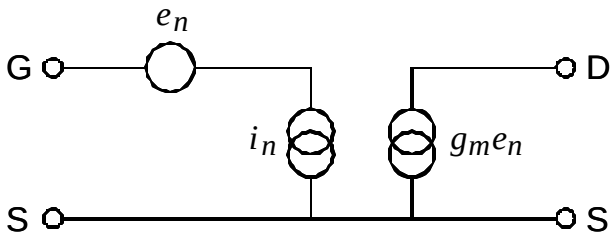
In MOSFETs the DC gate current is very low (at oxide thicknesses <10 nm it is dominated by electron tunneling through the oxide). Here the capacitive coupling from the channel to the gate is significant. To a good approximation

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

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. A 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.
If $F_i = F_v = 1$

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

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 3250 el, so the current contribution is negligible, except in very low frequency applications.

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

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

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

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

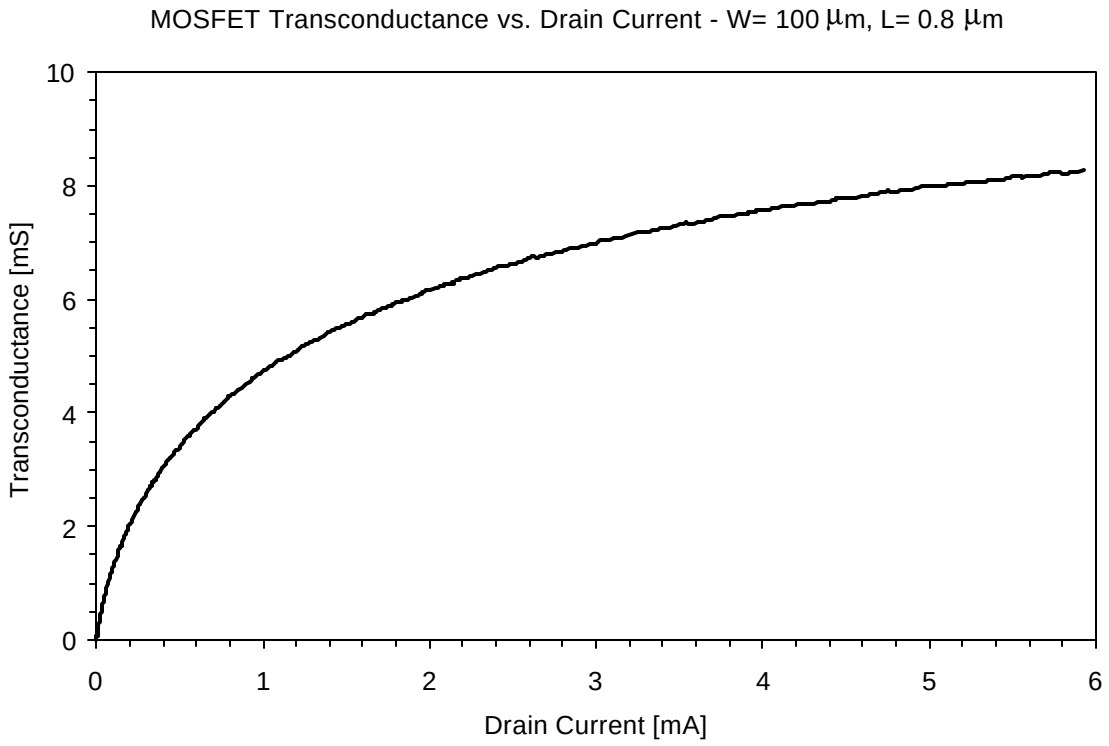
$C_i << C_{det}$: The detector capacitance dominates, so the effect of increased transistor capacitance is negligible.
As the device width is increased the transconductance increases and the equivalent noise voltage decreases, so S/N improves.

$C_i > C_{det}$: 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 as shown below.



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

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

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

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

where

$$k_n \equiv \frac{g_m}{g_n C_i}$$

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 $\quad g_m/C_i = 3 \cdot 10^{-9}$ s^{-1} and $g_n = 1$.

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

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

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

Measured values of the noise coefficient g_n for various n- and p-MOSFETs of various geometries for three normalized drain currents I_d/W.

Type	**NMOS**	**PMOS**	**NMOS**	**PMOS**	**NMOS**	**PMOS**	**NMOS**	**PMOS**
Width	75	75	1332	1332	888	888	1332	1332
Length	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	0.50
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	0.60
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

Short channel n-MOSFETs tend to have higher noise coefficients at short channel lengths, probably due to increased electron temperature at high fields.

The table also shows the noise degradation after irradiation.

Since they are majority carrier devices, MOSFETs are insensitive to displacement damage, but they are affected by ionization damage, which leads to charge buildup in the oxide and the formation of interface states.

See H. Spieler, Introduction to Radiation-Resistant Semiconductor Devices and Circuits, tutorial, in A.H. Lumpkin, C.E. Eyberger (eds.) *Beam Instrumentation, Proceedings of the Seventh Workshop*, AIP Conference Proceedings 390, American Institute of Physics, Woodbury, NY, 1997, ISBN 1-56396-612-3

and http://www-atlas.lbl.gov/strips/strips.html

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

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

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 for various device types:

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

Specific implementations can improve on these values.

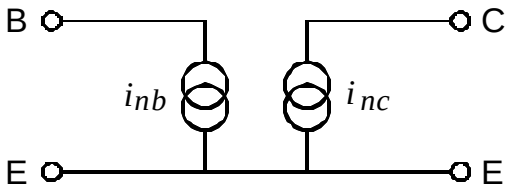
One should note that this model is not universally applicable, since excess noise usually does not exhibit a pure $1/f$ dependence; especially in PMOS devices one often finds several slopes. In practice, one must test the applicability of this parameterization by comparing it with data before applying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

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.

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

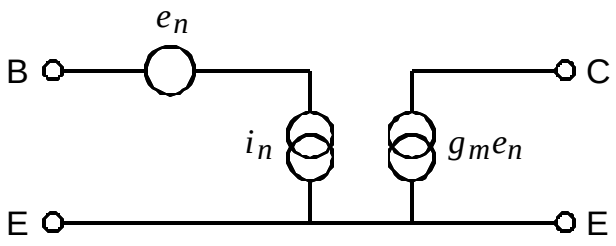
The noise current in the collector is the shot noise originating in the base-emitter junction associated with the collector component of the emitter current.

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

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 $I_B = I_C / b_{DC}$

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

The current noise term increases with I_C, whereas the second (voltage) noise term decreases with I_C.

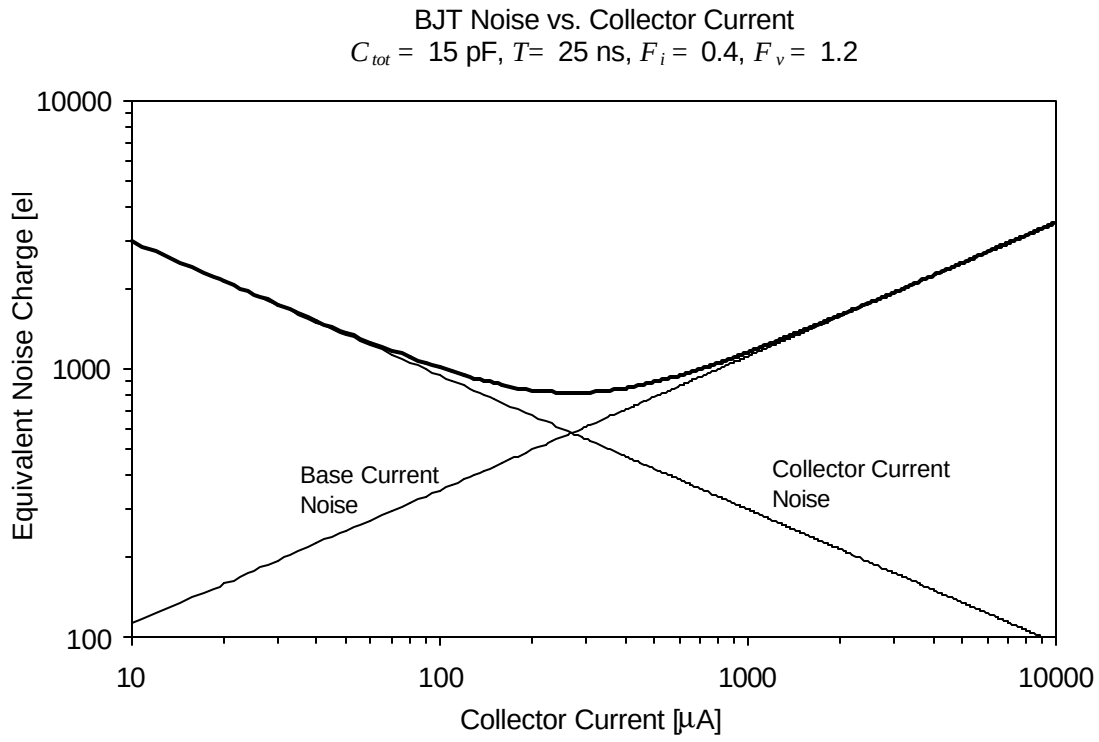
Thus the noise attains a minimum

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

at a collector current

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

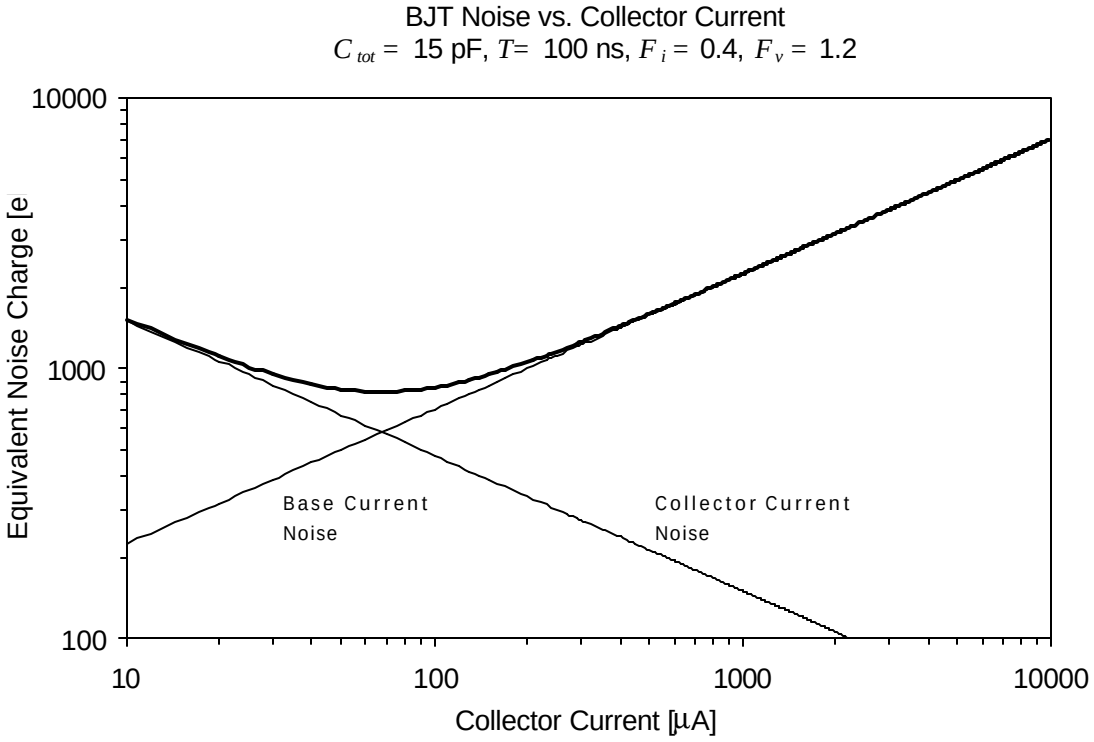
Note: This corresponds to the general criterion of noise matching derived in Section III, i.e. the detector impedance at the peaking frequency of the shaper equals e_n / i_n .



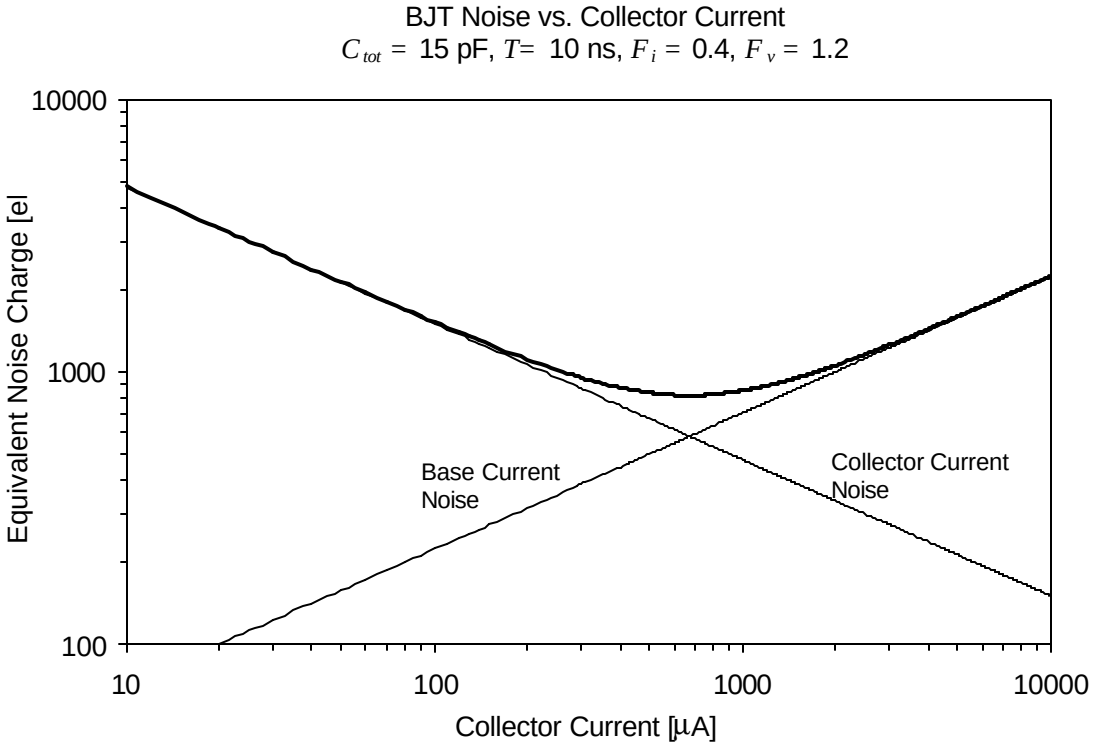
For a given shaper, the minimum obtainable noise is determined only by the total capacitance at the input and the DC current gain of the transistor, *not by the shaping time*.

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

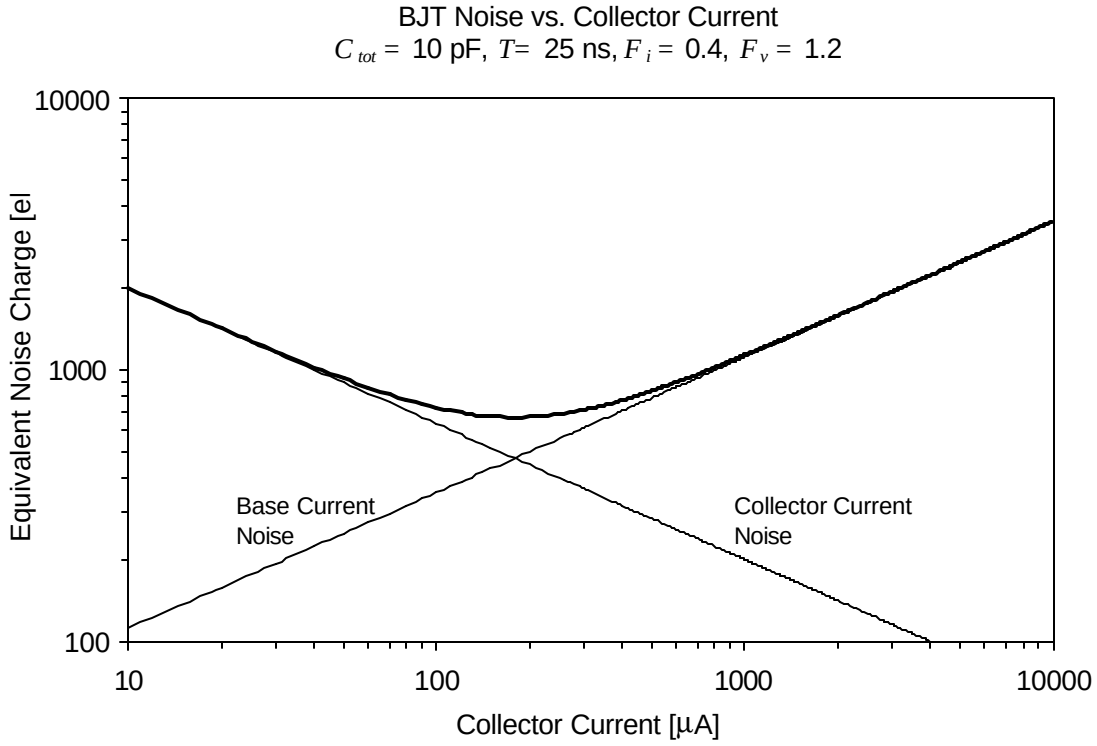
T= 100 ns

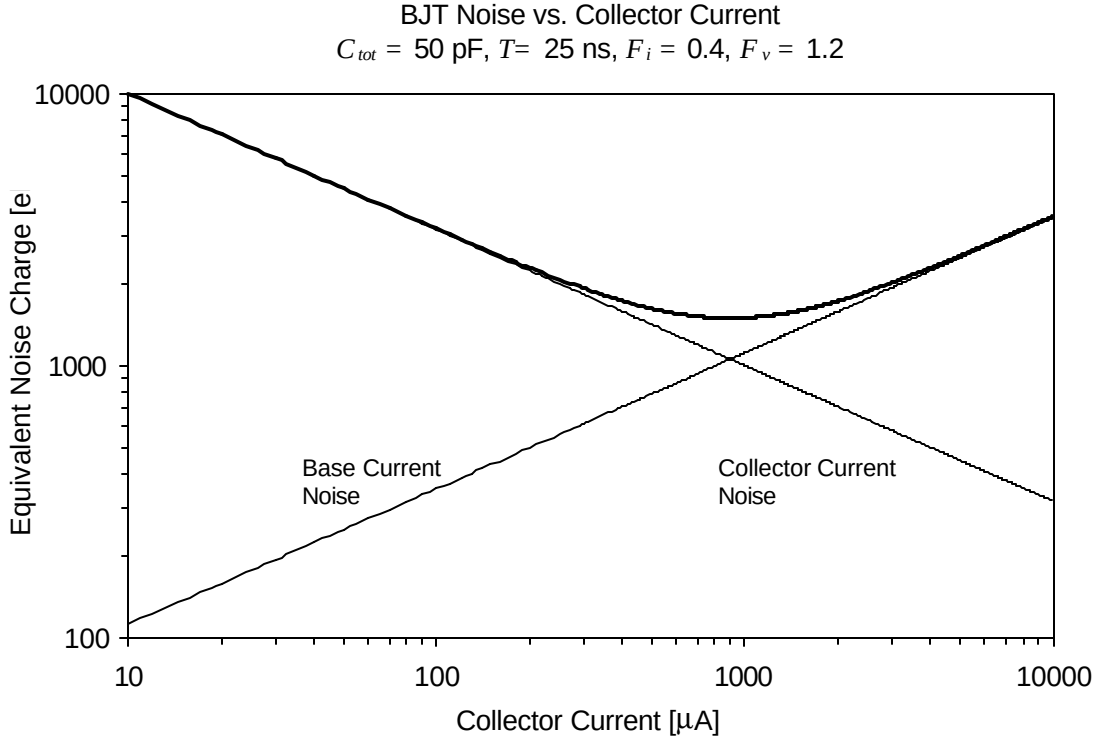


T= 10 ns



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





Simple Estimate of obtainable BJT noise

For a CR-RC shaper

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

obtained at $$I_c = 26 \left[\frac{m\text{A} \cdot \text{ns}}{\text{pF}} \right] \cdot \frac{C_{tot}}{t} \sqrt{b_{DC}}$$

Since typically $b_{DC} \approx 100$, 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 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.

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

where D_B is the diffusion constant in the base and w_{Ti} is a frequency that characterizes carrier transport in the base. w_{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 w_{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

$$Q^2_{n,\min} = 4k_BT\frac{C_{tot}}{\sqrt{b_{DC}}}\sqrt{F_iF_v} \quad ,$$

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?

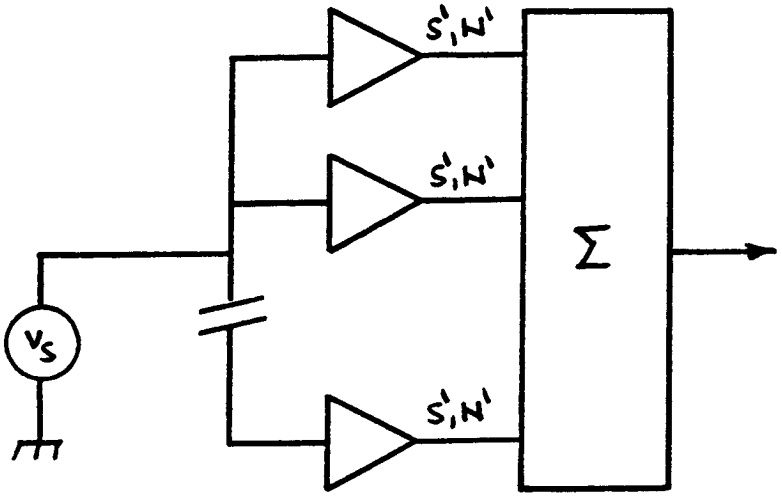
Since the base current noise increases with shaping time, bipolar transistors are only advantageous at short shaping times.

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.

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 amplifiers with both current and voltage noise. As in previous derivations of the equivalent noise charge, the amplifiers are assumed to have voltage-sensitive inputs. Furthermore, to simplify the analysis, the amplifiers do not utilize feedback.



Of course, in considering the current and voltage noise contributions, one can follow a formal argument based on the noise charge

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

Since the current noise contribution does not depend on capacitance, matching the amplifier input capacitance to the detector capacitance should be irrelevant. On the other hand, since the voltage contribution does depend on capacitance, a correlation between e_n and C_i can yield an optimization condition.

Nevertheless, reviewing the formation of signal and noise in detail is useful to clarify the limits of capacitive matching.

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_n' is the equivalent noise current of a single amplifier, for n amplifiers the total input noise current flowing through the signal source impedance is

$$i_n = \sqrt{n} \cdot i_n'$$

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

$$v_n = i_n Z_i$$

This voltage is applied in parallel to all amplifier inputs, so at the output of an individual amplifier (assuming gain A) the noise level is

$$N'(n) = Av_n = A\sqrt{n}\ i_n' Z_i$$

At the output of the summing circuit, the cumulative noise from all amplifier outputs becomes

$$N(n) = \sqrt{n} N'(n) = An\ i_n' Z_i$$

Since the amplifiers respond to voltage, the magnitude of the signal applied to all amplifiers is the same, which for a signal current i_s is

$$v_s = i_s Z_i$$

In the summed output the signals add coherently, so that

$$S(n) = nAi_s Z_i$$

and the signal-to-noise ratio

$$\frac{S(n)}{N(n)} = \frac{nAi_s Z_i}{Ani_n' Z_i} = \frac{i_s}{i_n'} = \frac{S(1)}{N(1)} \quad ,$$

the same as for a single amplifier

Paralleling amplifiers does not affect the signal-to-noise ratio if only current noise is present.

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, so the signal-to-noise ratio is unaffected.

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 infinite, 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_n' and an input capacitance C_i'.

Then the input signal voltage

$$v_s = \frac{Q_s}{C}$$

where C is the total input capacitance including the detector

$$C = C_{\det} + nC_i'$$

The signal at each amplifier output is

$$S' = Av_s = A\frac{Q_s}{C_{\det} + nC_i'}$$

The noise at each amplifier output is

$$N' = Av_n'$$

After summing the n outputs the signal-to-noise ratio

$$\frac{S(n)}{N(n)} = \frac{nS'}{\sqrt{n}N'} = \sqrt{n}\frac{S'}{N'} = \sqrt{n}\frac{A\dfrac{Q_s}{C_{\det} + nC_i'}}{Av_n'} = \frac{Q_s}{v_n'/\sqrt{n}} \cdot \frac{1}{C_{\det} + nC_i'}$$

which assumes a maximum when $C_{\det} = nC_i'$.

Under this "capacitive matching" condition $\sum C_i' = C_{\det}$ the signal-to-noise ratio

$$\frac{S}{N} = \frac{Q_s}{v_n'}\sqrt{\frac{C_{\det}}{C_i'}}\frac{1}{C_{\det} + nC_i'} = \frac{Q_s}{v_n'}\sqrt{\frac{C_{\det}}{C_i'}}\frac{1}{2C_{\det}}$$

or

$$\frac{S}{N} = \frac{1}{2}\frac{Q_s}{v_n'\sqrt{C_i'}} \cdot \frac{1}{\sqrt{C_{\det}}}$$

Since v_n' and C_i' 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.

Example: A MOSFET operated in weak inversion has a transconductance that 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.

For both BJTs and FETs, the minimum obtainable noise increases with the square root of detector capacitance, although the physical origins for this behavior are quite different in the two types of devices.

Optimization for Low-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.

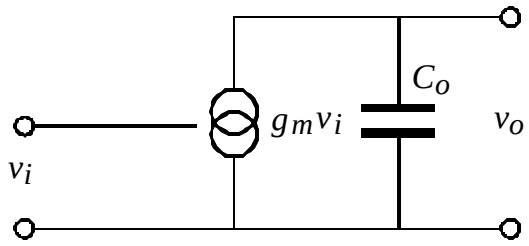
What is the optimum operating current for a given device?

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 g_m 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

$$t_o = R_L C_o = \frac{1}{w_o}$$

Hence the gain-bandwidth product

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

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

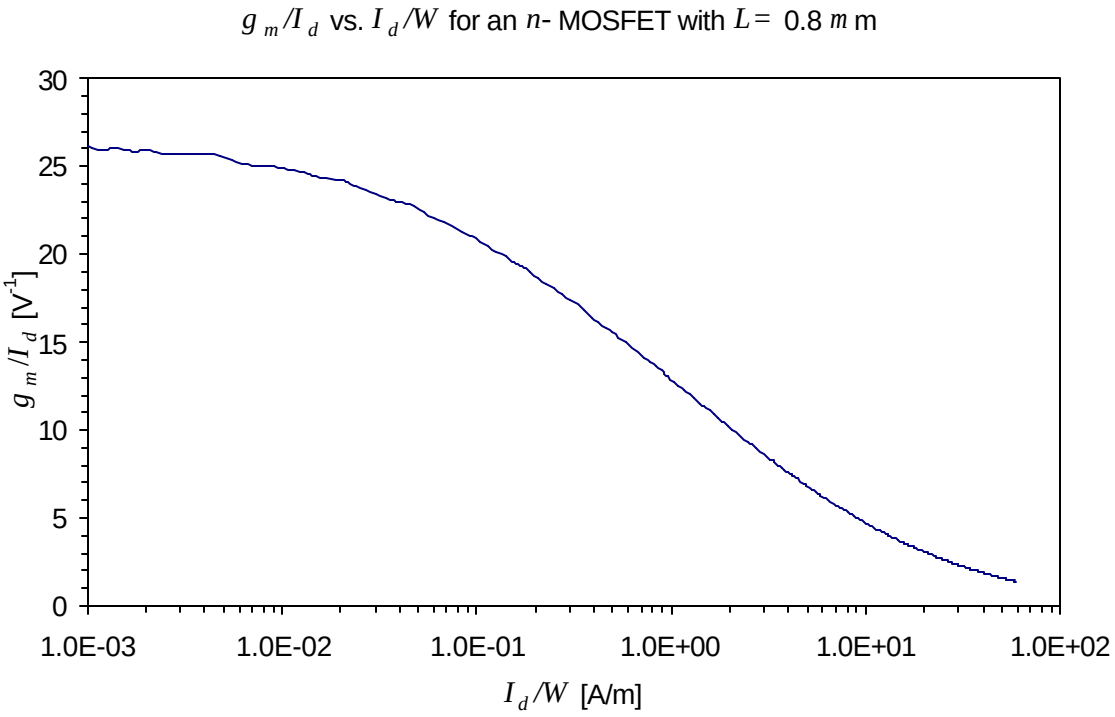
$$g_m = \frac{q_e}{k_B T} I_C$$

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.

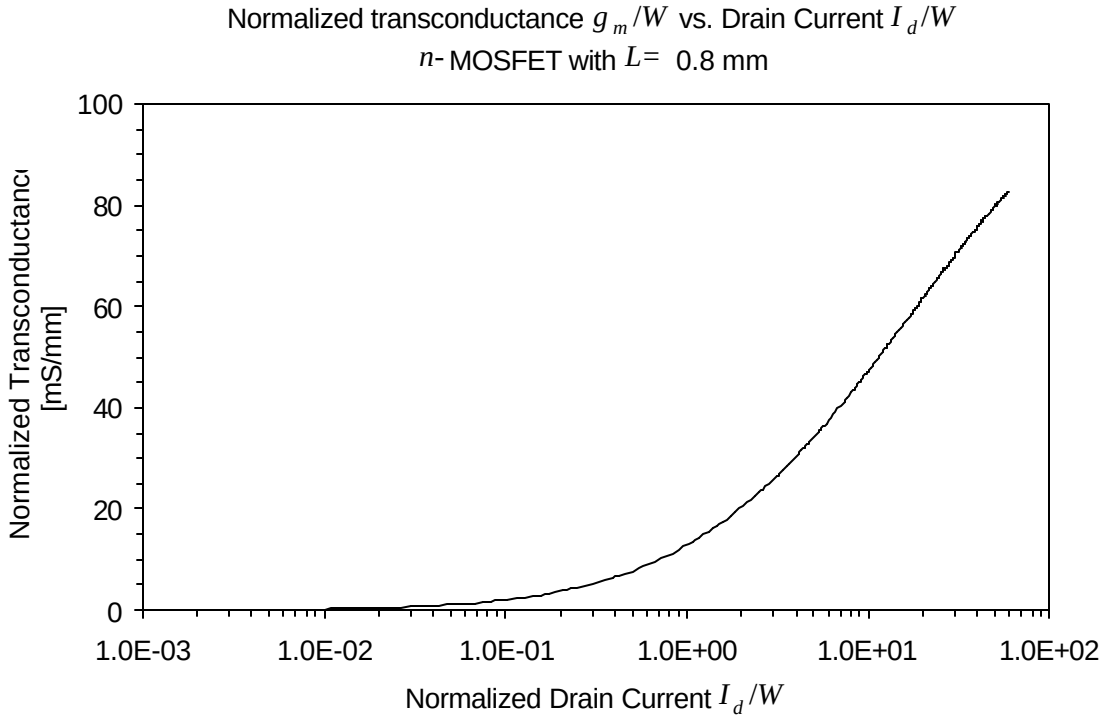
The figure below shows g_m/I_d as a function of normalized drain current I_d/W for a MOSFET with 0.8 μm channel length. This is a universal curve for all transistors using the same technology and channel length.



At low currents the MOSFET starts out with constant g_m/I_d, equal to a bipolar transistor. This is the weak inversion regime.

One then sees a rapid decrease in the regime 0.1 < I_d < 10 A/m (moderate inversion) and finally a gradual decrease at I_d > 10 A/m (strong inversion).

Note that although g_m/I_d is decreasing with current, the transconductance itself is increasing, but at a substantial penalty in current.



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 is logarithmic, and that the high transconductance in strong inversion comes at the expense of substantial power.

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.

Having chosen a value of normalized drain current I_d/W that provides an adequate gain-bandwidth product, the required noise can be achieved adjusting the width of the FET, limited by capacitive matching.

Since in this scaling the current density remains constant, both the current and the transconductance increase proportionally with width. Thus the noise

$$v_n^2 \propto \frac{1}{g_m} \propto \frac{1}{I_D}$$

$$Q_n^2 \propto v_n^2 C_{tot}^2 \propto \frac{C_{tot}^2}{I_D}$$

so for a given noise level the required power

$$P_D \propto I_D \propto C_{tot}^2$$

A similar result obtains for bipolar transistors. The most efficient operating regime with respect to power is below the current for minimum noise

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

In this regime the noise is dominated by voltage noise, so

$$Q_n^2 \approx \frac{2(k_B T)^2}{q_e I_C} C_{tot}^2 \frac{F_v}{T}$$

and as above for a given noise level

$$P \propto I_C \propto C_{tot}^2$$

The capacitance of the detector element strongly affects the required current in the input transistor, so reducing the capacitance not only improves noise performance, but also reduces power consumption.

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.

- 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 (discussed in VIII.5)

$$P \propto C^2$$

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

n times as many pixels as strips

- $$P_{pixel,tot} \approx \frac{4}{n} P_{strip}$$

- 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.

4.SQUIDs

Superconducting Quantum Interference Devices

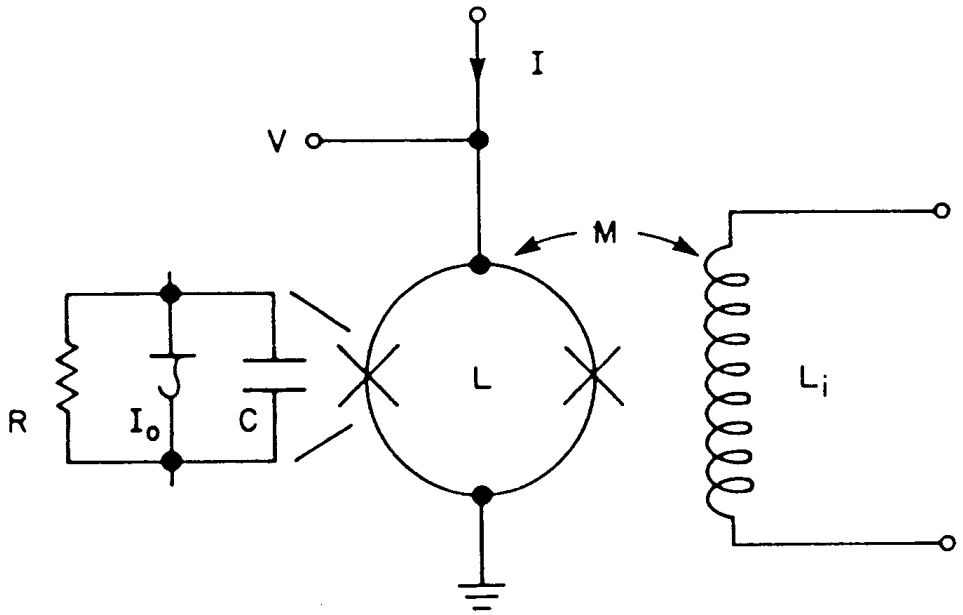
Two Josephson junctions connected in parallel to form superconducting ring:

Two key ingredients:

1. Phase between two tunneling currents in Josephson junction is determined by current.

2. Magnetic flux in superconducting loop is quantized:

$$\Delta\Phi_0 = \frac{p\hbar c}{e} = 2.0678 \cdot 10^{-7}\ \text{gauss cm}^2 = 2.0678 \cdot 10^{-15}\ \text{Vs}$$

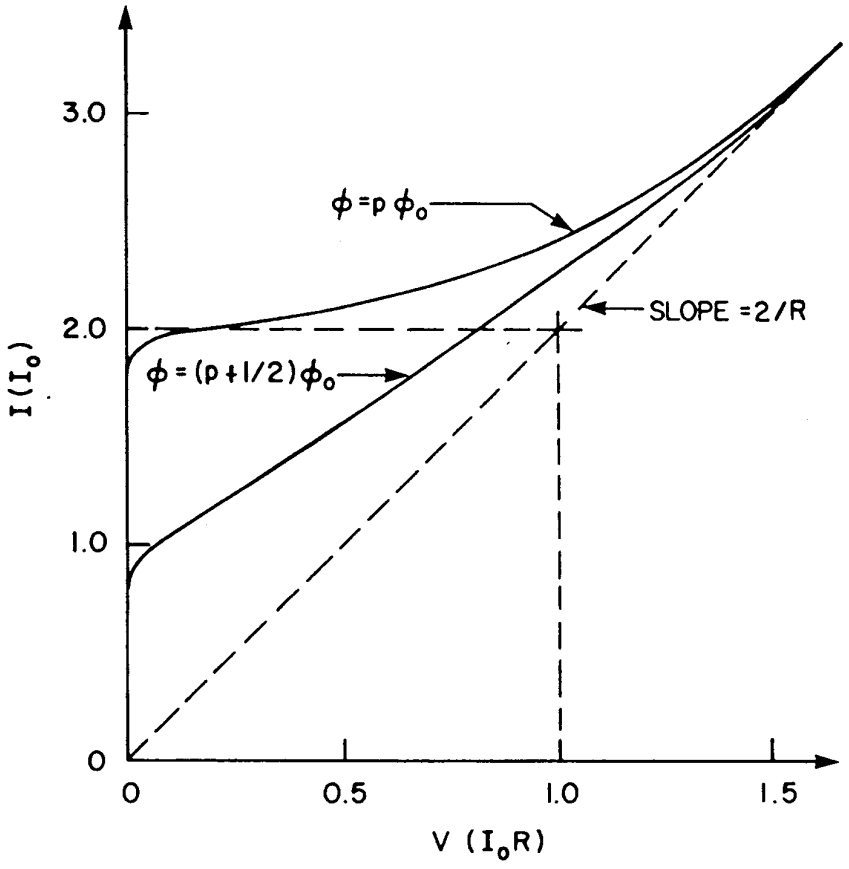


SQUID is biased by current I.

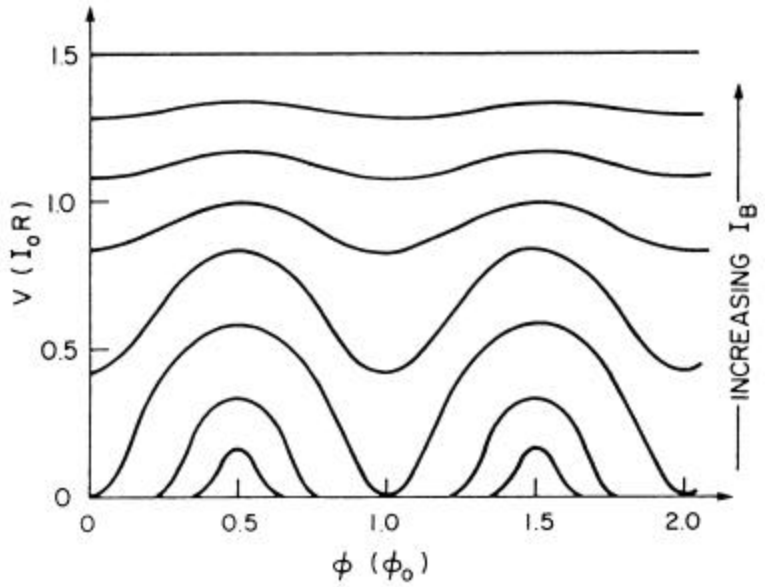
Input signal is magnetic flux due to current through coupling coil L_i.

Output is voltage V .

Current-Voltage Characteristics

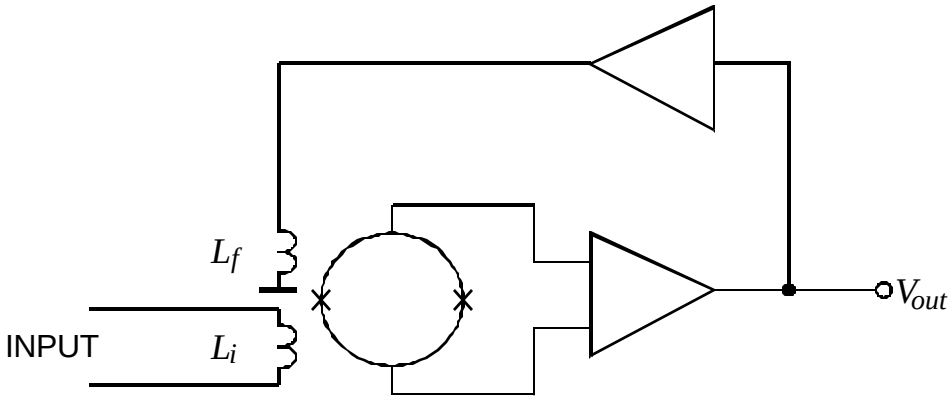


Output voltage V vs. flux Φ/Φ_0 as bias current I_B is increased



However,

- Input signal may not exceed ¼ flux quantum
  (output periodic in Φ_0)
- Feedback loop required to lock flux at proper operating point
  (flux locked loop)



Feedback circuit limits frequency response.

Typical Parameters

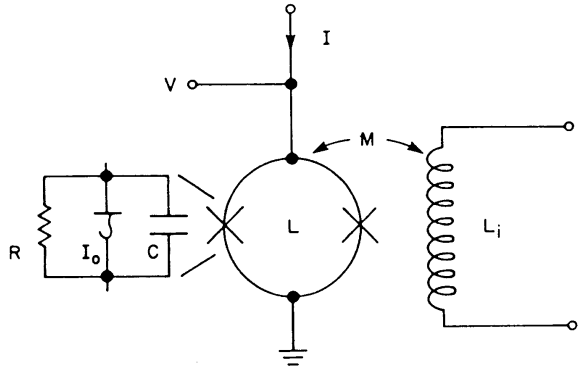
Operating Temprature:	0 – 5 K (also high T_C SQUIDs)
Flux Sensitivity:	V_Φ=150 μV/Φ_0
Flux Noise:	1 to 10 μΦ_0
SQUID Inductance:	100 – 500 pH
Input Inductance:	100 nH to 1 μH

Depending on the choice of input inductance, a typical input sensitivity is 0.1 μA/Φ_0.

SQUID Noise

For a detailed discussion see C.D. Tesche and J. Clarke, J. Low Temp. Phys **29** (1977) 301 and **37** (1979) 397

Noise is dominated by resistors shunting the Josephson junctions.



The resistors introduce

a noise voltage across the SQUID loop

and

a circulating current in the SQUID loop

with spectral power densities

$$S_V(f) = 16kTR \quad \text{and} \quad S_I(f) = 11kT / R$$

Strictly speaking, the two are correlated, but we'll neglect this here.

The noise voltage corresponds to a flux in the SQUID

$$\Phi_{vn}^2 = \frac{S_V}{V_\Phi^2}$$

The circulating current builds up a real flux in the SQUID loop

$$\Phi_{in}^2 = L^2 S_I$$

The input signal is coupled to the SQUID loop through an input coil with mutual inductance M.

For a given input current I_i the flux coupled into the SQUID loop

$$\Phi_i = MI_i$$

so the equivalent flux due to the resistor noise voltage translates to an equivalent input noise current

$$i_n = \frac{\sqrt{S_V}}{MV_\Phi}$$

The fluctuating flux due to the circulating noise current induces a noise voltage across the input coil

$$e_n = -iwM\sqrt{S_I}$$

Thus, the optimum source resistance

$$R_{opt} = \left|\frac{e_n}{i_n}\right| = \left|\frac{-iwM\sqrt{S_I}}{\sqrt{S_V}/MV_\Phi}\right| = k_L L L_i w V_\Phi \sqrt{\frac{S_I}{S_V}}$$

where k_L is the coefficient of coupling. Let's assume perfect coupling $k_L = 1$ for simplicity. Since $V_\Phi \approx \frac{R}{L}$

$$R_{opt} = \sqrt{\frac{11}{16}} k^2 w L_i \approx w L_i$$

and noise matching is achieved by adjusting the input inductance. Note that the optimum source impedance is frequency dependent.

For an inductive source (i.e. an external coil used as a magnetic sensor), one obtains a wideband noise match.

However, even with wideband noise matching, the noise temperature depends on frequency

$$T_N = \frac{\sqrt{e_n i_n}}{2k} = \frac{7wT}{V_\Phi} \ ,$$

so the noise level increases with frequency.

The noise temperature of optimized SQUIDs in the He temperature range is typically

$$T_N \approx \frac{f}{10^8 \ [\mathrm{Hz/K}]} \ ,$$

so at 1 MHz $T_N = 10$ mK.

With respect to their input noise, SQUIDs and transistors are duals.

Recall that in bipolar and field effect transistors the input noise current is physically present at the input, whereas the input noise voltage is a transformed quantity, i.e. the output noise current translated to the input. This is why one speaks of an *equivalent* noise voltage.

In the SQUID we have the opposite situation:

> The input noise voltage is the induced voltage due to flux fluctuations in the SQUID loop, so it is physically present at the input.
>
> The input noise current, on the other hand, derives from the output noise voltage, transformed to the input, so it is an *equivalent* noise current.

5. Microelectronics

Fabrication of Semiconductor Devices

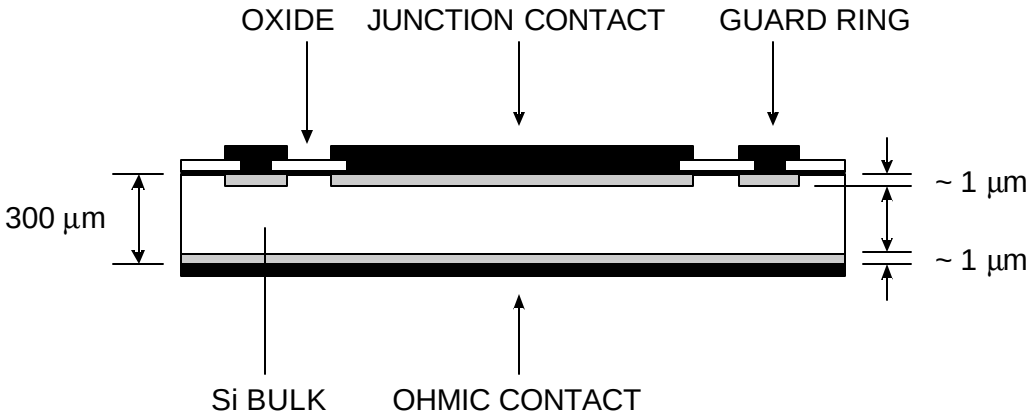
Ingredients of a semiconductor device fabrication process

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 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”.

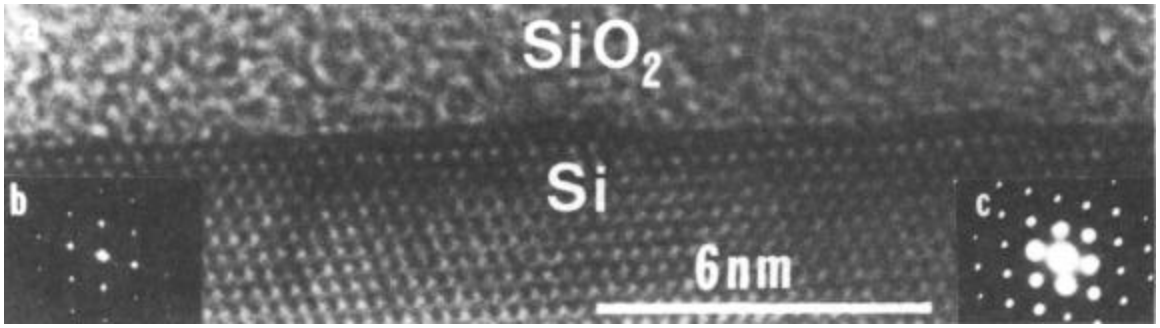
   This is accomplished by heating (annealing).

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)

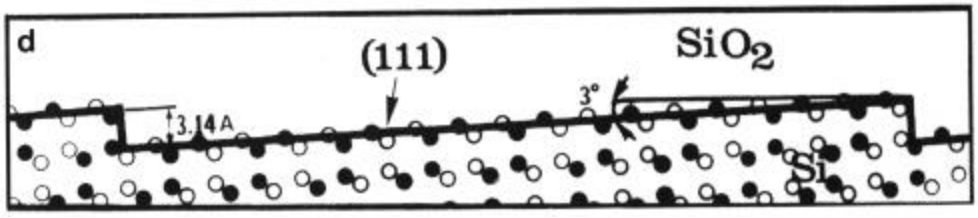
In these respects SiO_2 on Si is unequalled – indeed this is probably the single key ingredient that allows Si technology to achieve a circuit density that is at least an order of magnitude greater than in any other semiconductor.

Atomic resolution electron microscope image of SiO_2-silicon interface



|¬ ¾ ¾ ¾ ¾ ¾ ¾ ¾ ¾ ¾ ®|

ß



(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.

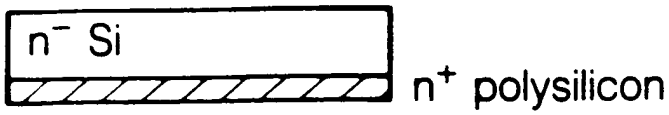
A photoresist is applied to the surface.

Exposure to light through a “mask” to cross-links the polymer in the desired areas.

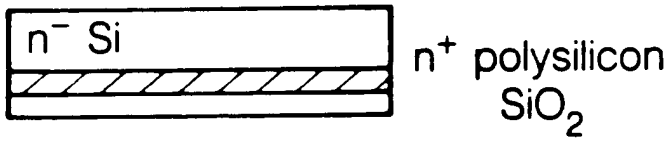
(or the inverse – once can use positive or negative resist)

The exposed portions are removed by an appropriate solvent.

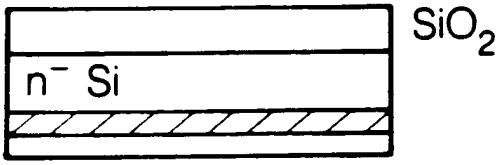
1. Backside contact (typically P diffusion or implant) – here P-doped polysilicon (discussed later)



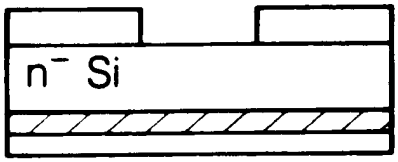
2. Deposition of silicon dioxide on the backside



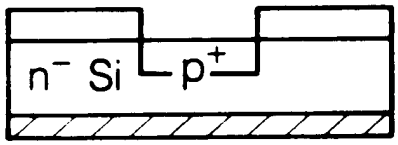
3. Thermal oxidation of the top surface



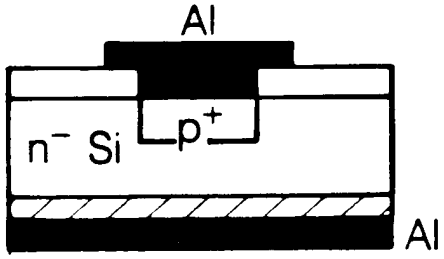
4. Photolithography and etching of the silicon dioxide



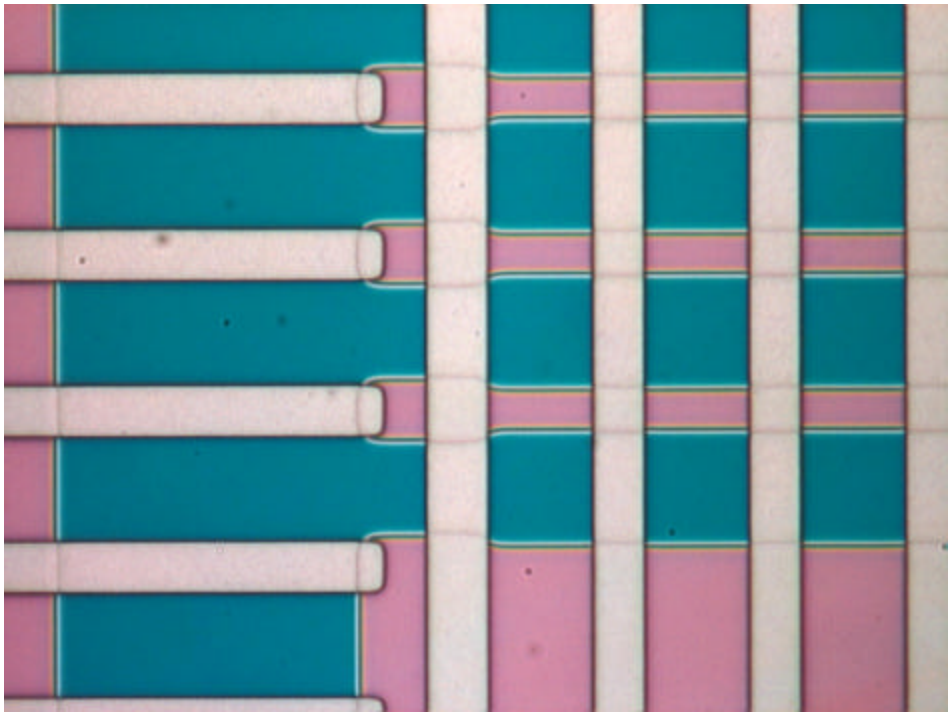
5. Boron doping to form p-type electrode and subsequent removal of backside oxide

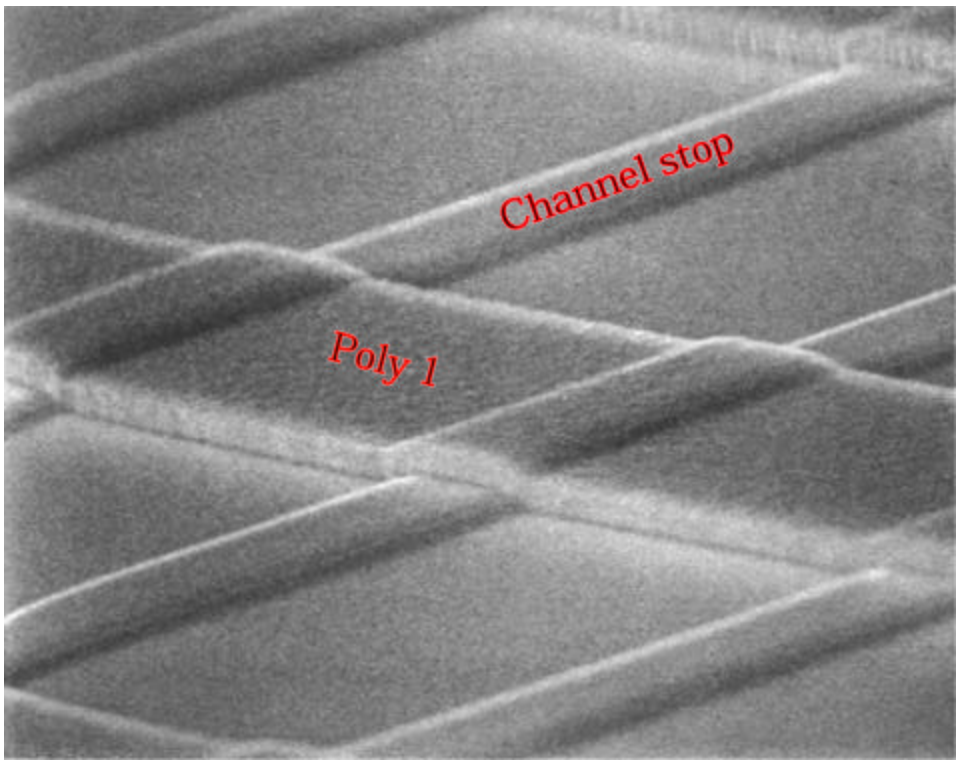


6. Aluminum metallization and patterning to form contacts



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

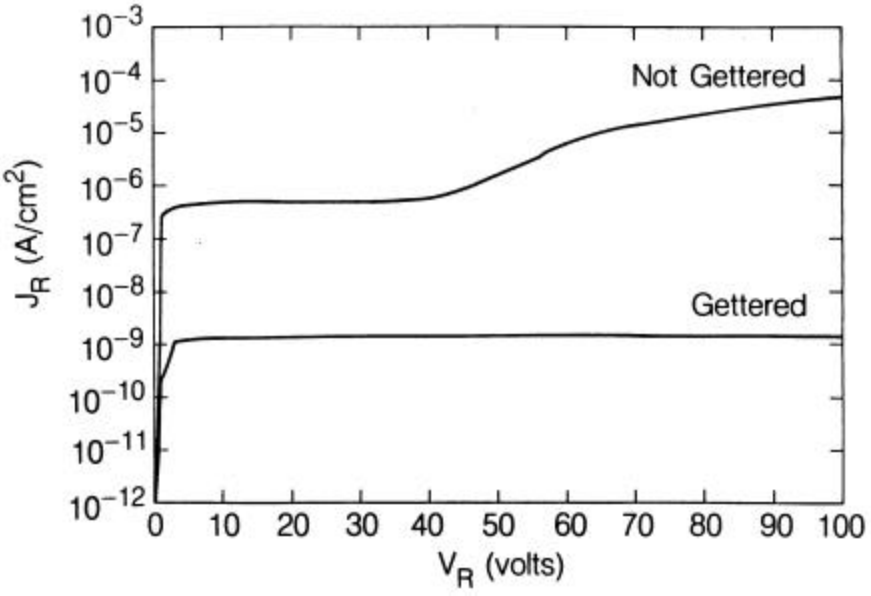
Gettering can be used to remove contaminants from the sensitive regions by providing capture sites for contaminants. This requires that the critical contaminants are sufficiently mobile so that they will diffuse to the gettering sites and be captured.

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)

Next – just to get a flavor of the required attention to detail – listings of process steps in fabricating a detector diode.
First, a simplified sequence and then a more detailed sequence.

Detector Fabrication Process

(S. Holland, LBNL)

High-temperature process with very effective gettering to actively remove electrically active impurities from detector volume.

Fully compatible with conventional IC fabrication. Detectors and electronics have been integrated monolithically

S. Holland, H. Spieler: IEEE Trans. Nucl. Sci. **NS-37** (1990) 463
W. Snoeys et al.: IEEE Trans Nucl. Sci. **NS-39** (1992) 1263

1. Gettering Layer
   (also forms ohmic electrode of detector diode)
   a) Deposition of ~0.5 μm low-temperature oxide (LTO)
      ~1 hr at 400 °C
   b) Spin photoresist on wafer front side
   c) Etch backside oxide
   d) Deposit ~1.2 μm phosphorus-doped polysilicon
   e) Spin photoresist on wafer backside
   f) Remove front-side polysilicon
   g) Deposit capping LTO
   h) Spin photoresist on wafer back side to protect capping layer
   i) Remove LTO from wafer front side
2. Grow thermal oxide (steam, 4 hrs at 900 °C)
   forms 400 nm thick oxide
3. Deposit photoresist
4. Expose photoresist through mask with strip pattern, develop
5. Etch exposed oxide
6. Introduce *p*-dopant, typically boron, to form strip electrodes
   Two methods possible:
   a) thermal diffusion: expose to B_20_3 source for 30 min at 900 °C
   b) ion implantation: 30 keV B ions at dose of $2 \cdot 10^{15}$ 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

9. Spin on photoresist
10. Expose through contact mask to form strip metallization and bonding pads, develop resist
11. Etch metal
12. Coat front side with photoresist
13. Etch away backside oxide
14. Deposit 100 nm aluminum on back side to form ohmic contact
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:

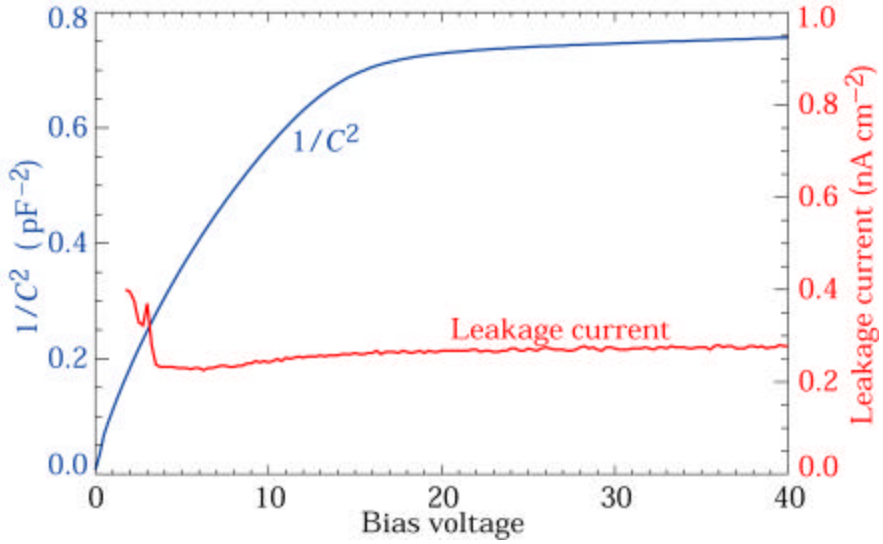
Wafer cleaning prior to each furnace step (except post-metallization anneal):

1. immerse in 5:1 H_2S0_4:H_20_2 at 120 °C
2. HF etch
3. rinse in Dl water
4. immerse in 5:1:1 solution of H_20:NH_4OH:H_20_2 at 65 °C
5. dilute HF etch
6. rinse in Dl water
7. immerse in 5:1:1 solution of H_20:HCI:H_20_2 at 65 °C
9. rinse in Dl water

This is just an outline!

Many more detailed steps are involved, so that a detector fabrication run requires 4 to 5 weeks.

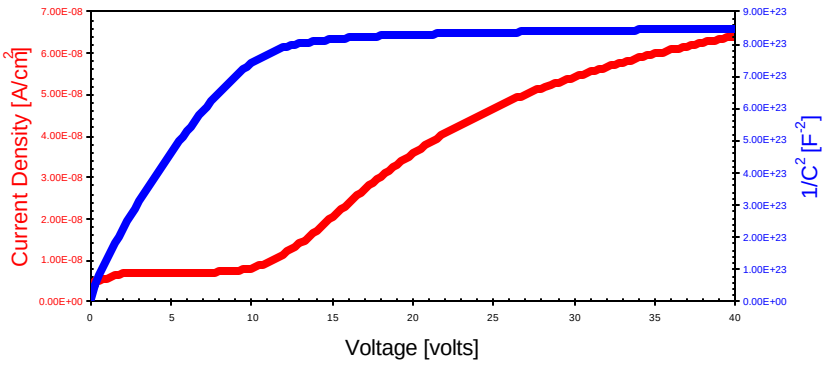
Measured Diode Characteristics
(S. Holland)



Presence of defects or impurities changes I-V curve:

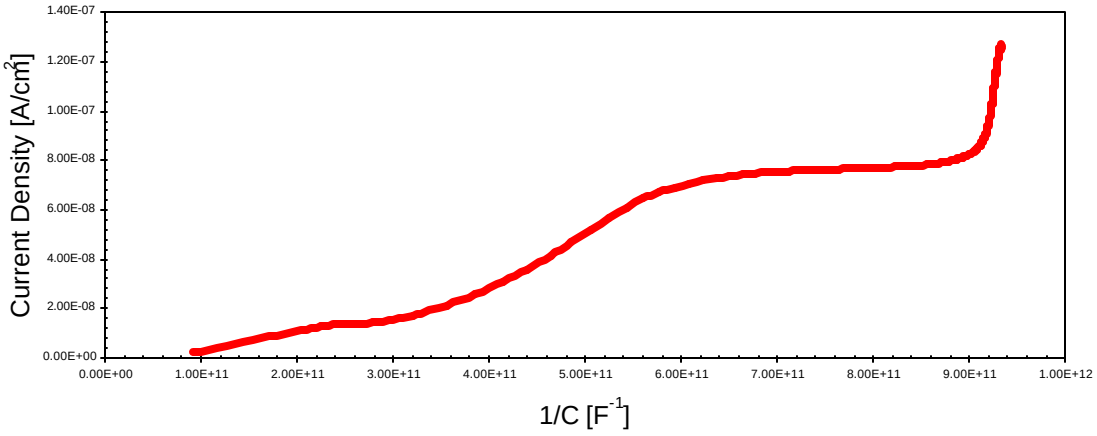
damage at back surface (polishing?)

2k x 4k 1-3 RB 2mm^2 pin diode 2/26/2001



bulk contamination

2k x 4k 1-3 UR 2mm^2 pin diode 2/26/2001

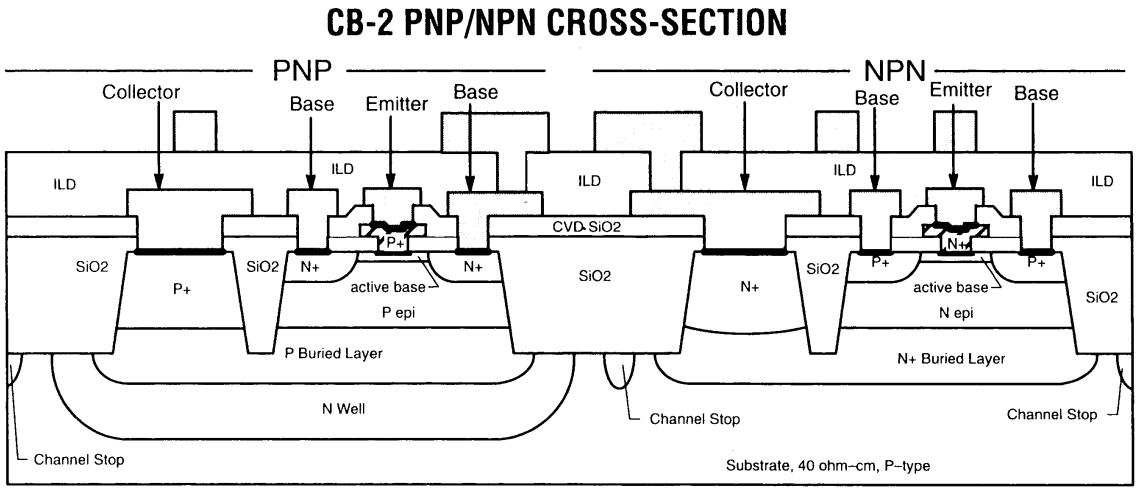


Integrated Circuits

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 Section VII).

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%.

  ⇒ 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 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

LBNL Microsystems Laboratory

Modern fabrication facility for specialized silicon devices. Operated by Physics Division.

Ultra-clean facility with process capabilities not available in conventional detector laboratories

- class 1 clean room
- wide range of deposition processes
- sub-micron lithography
- water system with sub-ppb impurity levels

Emphasis on

- test of device concepts and technology beyond industrial mainstream
- rapid turn-around

Although primary emphasis on R&D and proof-of-principle, limited production runs (~1 - 3 m^2) feasible.

Complementary to campus microfabrication facility (Cory Hall)

Current Topics

Fully depleted CCDs for faint light imaging

primary focus

Successful fabrication of 200 x 200, 2048 x 2048, and 2048 x 4096 $(15\mu m)^2$ CCD's

Photodiode arrays for medical imaging

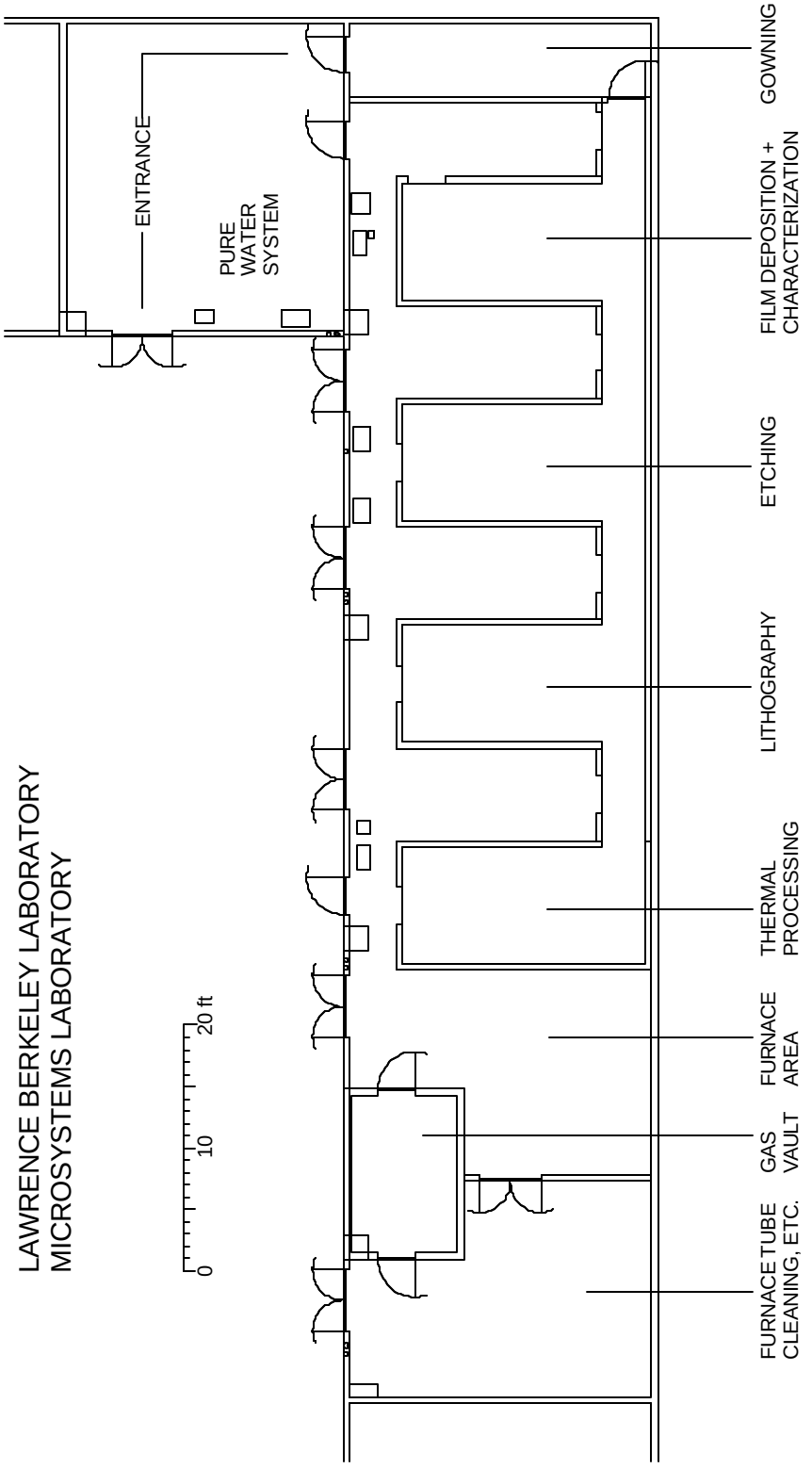
Test structures for device R&D

Laboratory Facilities

- ~700 ft^2 of class 1 - 10 fabrication area
- Floor plan chosen to allow maintenance with minimal disruption of clean room activities
- Tight temperature (±1C) and humidity (±2%) control
- Dedicated high-purity water system

Process equipment includes

- Semi-automated wet process stations for wafer cleaning and etching
- Film deposition: multi-station sputtering system for aluminum, silicon dioxide and indium tin oxide
- Sub-micron lithography
  GCA wafer stepper with programmable apertures
  MTI wafer track for highly uniform photoresist deposition
  Perkin Elmer full wafer mask aligner
- Plasma etchers for Polysilicon and silicon nitride
- Thermal processing
  5 atmospheric furnace tubes
  for oxidation, doping by diffusion, drive-in of ion implants, sintering, and annealing
  3 low-pressure furnace tubes
  for a variety of film depositions, *e.g.*
  doped and undoped polysilicon
  (for detector bias resistors, MOSFETs)
  Si nitride
  High-temperature oxide
  (for combined oxide-nitride films, *e.g.* for detector blocking capacitors)
  Low-temperature oxide
  (thick passivation layers)
  Furnaces are computer-controlled with extensive safety systems


LAWRENCE BERKELEY LABORATORY
MICROSYSTEMS LABORATORY
HELMUTH SPIELER
7-FEB-89
ENTRANCE
PURE WATER SYSTEM
GOWNING
FILM DEPOSITION + CHARACTERIZATION
ETCHING
LITHOGRAPHY
THERMAL PROCESSING
FURNACE AREA
GAS VAULT
FURNACE TUBE CLEANING, ETC.
0
10
20 ft

Wet Process Stations

Furnaces

Sputtering System

Lithography – full wafer mask aligner

Plasma Etchers

Service Chase

All routine maintenance can be performed from the back-side, i.e. without entering the clean room.