Overview of Key Device Parameters
Generic Three Terminal Device

Output current is dependent on input voltage:

\[ I_o = f(V_i, V_o) \approx f(V_i) \]

Examples:

nPN BJT  n-channel JFET  NMOS  GaAs MESFET  vacuum tube
Bipolar: \( I_o \approx I_s e^{-\frac{qV_{BE}}{kT}} \)

("forward active")

JFET/MESFET: \( I_D = I_{DSG} \left( 1 - \frac{V_{GS}}{V_p} \right)^2 \)

("pinch-off" regime)

MOSFET: \( I_D = \frac{\mu_n C_{ox} W}{2 L} \left( V_{GS} - V_T \right)^2 \)

("saturation")

Vacuum Tube: \( I_o = G \left( V_i + \frac{V_o}{\mu} \right)^{3/2} \)
Generic Device Behavior

- slope is output resistance of device
  - resistor (triode) region (very small output voltage)
  - non-linear resistor region (small output voltage)
  - “constant” current region (large output voltage, current nearly independent of output)
Resistors and capacitors are non-linear

- $R_\pi$ and $R_o$ depend on bias point
- $R_g$ (intrinsic) depends on channel inversion level
- $R_b$ can change due to current spreading effects
- $C_{gs}$ varies from accumulation to depletion to inversion
- Junction capacitors vary with bias
In small signal regime, $R$ and $C$ linear about a bias point:

- For BJT: $r_x = r_b$
- For a FET input:

$$R_g = R_{nqs} + \frac{1}{d}R_{poly}$$
$$d = 3 \leftrightarrow 12$$
$$R_{nqs} \sim \frac{1}{5g_m}$$
Various Figures of Merit

- Intrinsic
- Voltage
- Gain \((a_0)\)
- Power Gain
- Unilateral Gain
- Noise
  - Noise figure \((NF\) and Noise Measure \(M)\)
  - Flicker noise corner frequency
- Unity Gain Frequency
- \(f_T\)
- Maximum Osc Freq \(f_{max}\)
- Gain (normalized to current): \(g_m/I\)
- Gain Bandwidth
Other Important Metrics

- Complementary devices
  - Device with same order of magnitude of $f_T / f_{max}$
  - Lateral pnp a “dog” compared to vertical npn

- Availability of Logic
  - Low power/high density
  - Useful with S/H (sample hold) and SC (switch capacitor) circuits
  - Important for calibration

- Breakdown voltage
  - Power amplifiers, dynamic range of analog circuitry

- Thermal conductivity
  - Power amplifiers

- Quality and precision of passives
  - Inductors, capacitors, resistors, and transmission lines
\( i_o = g_m v_{gs} = \frac{g_m v_{gs}}{j \omega (C_{gs} + C_{gd})} \)

Since \( i_d = g_m v_{gs} \), and \( v_{gs} j \omega (C_{gs} + C_{gd}) = i_s \), taking the ratio we have

\[ A_i = \frac{i_o}{i_i} = \frac{g_m}{j \omega (C_{gs} + C_{gd})} \]

Solving for \( |A_i = 1| \), we arrive at the unity gain frequency

\[ \omega_T = 2 \pi f_T = \frac{g_m}{C_{gs} + C_{gd}} \]
It is not at first obvious why $f_T$ plays such an important role in analog (and digital) circuits. To see this, consider the simple cascade amplifier, where a single stage amplifier drives an identical copy. The gain of this first stage is given by the intrinsic gain of the amplifier, and the 3-dB bandwidth is limited by the pole at the high impedance node

$$\omega_0 = \frac{1}{r_{o,1}((1 + |A_2|)C_{gs,2} + C_{d1,tot})}$$
Here $C_{d1,tot}$ is the total drain capacitance $(C_{gd} + C_{ds} + C_{wire} + \cdots)$ and the effect of Miller multiplication is captured by boosting the second stage input capacitance by the gain $A_2$. If the second stage is a cascode with a small voltage gain between the gate/drain, then $A_2$ can be made smaller than unity to minimize its impact. So if we neglect the Miller effect, we have

$$\omega_0 = \frac{1}{r_o(C_{gs} + C_{d,tot})}$$
In most applications, we intend to embed this amplifier in a feedback loop, so the gain-bandwidth product is of interest.

For a feedback system, we know that we can approximately trade gain for bandwidth. So if we place the amplifiers in a feedback loop, we have

\[ Gain = A_0 \]

\[ BW = \omega_0 \]

\[ Gain \times BW = A_0 \omega_0 < g_m r_o \frac{1}{r_o (C_{gs} + C_{d,tot})} \]

\[ = \frac{g_m}{C_{gs} + C_{d,tot}} \approx \omega_T \]
Even in a digital circuit, we find that $f_T$ plays a key role. Consider the time constant of a simple gate, such as an inverter. If the fan out of an inverter is unity, in other words the inverter drives an identical copy of itself, then the load of the inverter is approximately $C_{gs} + C_{d,tot}$. 
The discharge current takes a complicated form, but to first order the transistor acts like a switch with on-conductance $g_{ds} = g_m$ (in triode region), so the discharge time is given by the product

$$\tau_T = \frac{C_{gs}}{g_m} = \frac{1}{\omega_T}$$

For larger fan-out (FO) circuits (and/or gates), we just scale $f_T$ by the FO.
Bipolar Junction Transistors
Most transistor “action” occurs in the small npn sandwich under the emitter. The base width should be made as small as possible in order to minimize recombination. The emitter doping should be much larger than the base doping to maximize electron injection into the base.

A SiGe HBT transistor behaves very similarly to a normal BJT, but has lower base resistance $r_b$ since the doping in the base can be increased without compromising performance of the structure.
The resistor \( R_{\pi} \) dominates the input impedance at low frequency. At high frequency, though, \( C_{\pi} \) dominates.

\( C_{\pi} \) is due to the collector-base reverse biased diode capacitance.

\( C_{cs} \) is the collector to substrate parasitic capacitance. In some processes, this is reduced with an oxide layer.

\( C_{\pi} \) has two components, due to the junction capacitance (forward-biased) and a diffusion capacitance.
Due to Boltzmann statistics, the collector current is described very accurately with an exponential relationship

\[ I_C \approx I_S e^{\frac{qV_{BE}}{kT}} \]  \hspace{1cm} (1)

The device transconductance is therefore proportional to current

\[ g_m = \frac{dI_C}{dV_{BE}} = I_S \frac{q}{kT} e^{\frac{qV_{BE}}{kT}} = \frac{qI_C}{kT} \] \hspace{1cm} (2)

where \( kT/q = 26\text{mV} \) at room temperature. Compare this to the equation for the FET. Since we usually have \( kT/q < V_{gs} - V_T \), the bipolar has a much larger transconductance for the same current. This is the biggest advantage of a bipolar over a FET.
Control Terminal Sensitivity

BJT:

\[
10 = \exp \left( \frac{q \Delta V_{BE}}{kT} \right)
\]

\[
\Delta V_{BE} = \frac{kT}{q} \times \ln 10 \approx 60 \text{mV}
\]

MOSFET:

\[
10 = \left( \frac{V_{GS,1} - V_T}{V_{GS,2} - V_T} \right)^2
\]

\[
\sqrt{10} = \left( \frac{V_{GS,1} - V_T}{V_{GS,1} + \Delta V_{GS} - V_T} \right)
\]

\[
\Delta V_{GS} = \frac{1 - \sqrt{10}}{\sqrt{10}} (V_{GS,1} + V_T) \approx 1 \text{V}
\]
The unity gain frequency of the BJT device is given by

$$\omega_T = \frac{g_m}{C_\pi + C_\mu} = \frac{g_m}{C_{\text{diff}} + 2C_{je0} + C_\mu}$$  \hspace{1cm} (3)$$

where we assumed the forward bias junction has $C_{je} \approx 2C_{je0}$.

Since the base-collector junction capacitance $C_\mu$ is a function of reverse bias, we should bias the collector voltage as high as possible for best performance.

The diffusion capacitance is a function of collector current, $C_{\text{diff}} = g_m\tau_F$

$$\omega_T = \frac{g_m}{g_m\tau_F + 2C_{je0} + C_\mu} = \frac{1}{\tau_F + \frac{2C_{je0} + C_\mu}{g_m}}$$  \hspace{1cm} (4)$$
We can clearly see that if we continue to increase $I_C$, then $g_m \propto I_C$ increases and the limiting value of $f_T$ is given by the forward transit time $\omega_T \approx 1/\tau_F$.

In practice, though, we find that there is an optimum collector current. Beyond this current the transit time increases. This optimum point occurs due to the Kirk Effect. It’s related to the “base widening” due to high level injection. (Not Star Trek!)
Base Transmit Time

\[ C_i = C_j + C_{\text{diff}} \approx C_{\text{diff}} = \tau_F \cdot g_m \]

- Base transmit time
- Current gain unity freq.

\[ f_T = \frac{g_m}{2\pi C_i} \approx \frac{1}{2\pi \tau_F} \]

\[ f_T \approx \frac{2\mu_n}{2\pi W_B^2} \frac{kT}{q} \]

\[ f_T \propto \frac{1}{W_B^2} \]
CMOS FET Transistors
Modern CMOS process has very short channel lengths \((L < 100\text{nm})\). To ensure gate control of channel, as opposed to drain control (DIBL), we employ thin junctions and thin oxide \((T_{ox} < 5\text{nm})\).

Due to lithographic limitations, there is an overlap between the gate and the source/drain junctions. This leads to overlap capacitance. In a modern FET this is a substantial fraction of the gate capacitance (up to half).
In a standard CMOS process, only the PMOS is isolated by the n-well. All NMOS devices share a common body (psub). This means that the body-source cannot be tied together (unless grounded) and moreover, digital substrate noise couples to sensitive analog and RF circuits.

In many processes, a deep n-well (“DNW”) can be used to isolate the p-well. This is like a triple-well process, allowing isolated NMOS transistors. The DNW is formed from a ring of NW and an overlapping DNW region.
The junctions of a FET form reverse-biased pn junctions with the substrate (well), or the body node. This is another form of parasitic capacitance in the structure, $C_{db}$ and $C_{sb}$.

At DC, input is an open circuit. The input impedance has a small real part due to the gate resistance $R_g$ (polysilicon gate and NQS) and $R_{s,d}$ account for junction and contact resistance.
FET Simplified Models

- In the forward active (saturation) region, the input capacitance is given by $C_{gs}$

$$C_{gs} = \frac{2}{3} W \cdot L \cdot C_{ox} + C_{par}$$

- Don’t forget that layout parasitics increase the capacitance in the model, sometimes substantially (esp in deep submicron technologies). $R_o$ is due to channel length modulation and other short channel effects (such as DIBL).

- For low frequencies, the resistors are ignored. But these resistors play an important role at high frequencies.

- If the source is tied to the bulk, then the model simplifies.
Substrate parasitics, gate resistance, source/drain resistances, extra metal and contact resistance and capacitance. At very high frequencies, the distributed inductance of the leads.
Intrinsic Voltage Gain

- Important metric for analog circuits

\[ A_{v,mos} = g_m r_o = \frac{2I_{DS}}{V_{dsat}} \frac{V_A}{I_{DS}} = 2 \frac{V_A}{V_{dsat}} \]

\[ A_{v,bjt} = g_m r_o = \frac{qI_C}{kT} \frac{V_A}{I_C} = 2 \frac{V_A}{kT q} \]

- Communication circuits often work with low impedances in order to achieve high bandwidth, linearity, and matching.
- To achieve high \( f_T \), the \( V_{dsat} \) is relatively large so the current is increased to obtain sufficient gain.
Inductive loads are also common to tune out the load capacitance and form a resonant circuit. The gain is thus given by

\[ A_v = g_m R_p = g_m \omega_0 QL \]

In a given process, there’s a maximum \( Q \) that we can obtain, usually \( Q \sim 10 \) at 1 GHz in a typical 90nm process with thick metal options. The inductance cannot be increased without bound due to area limitations and ultimately due to the tuning requirements

\[ \omega_0 L = \frac{1}{\omega_0 (C_{gs} + C_{db} + C_{par})} \]

\[ \omega_0 L \leq \frac{1}{\omega_0 C_{gs}} \]

\[ A_v = g_m \omega_0 QL = g_m Q \frac{1}{\omega_0 C_{gs}} = \frac{g_m}{C_{gs}} \cdot Q \cdot \frac{1}{\omega_0} = Q \cdot \frac{f_T}{f_0} \]
CMOS is running out of steam because the DC gain per device is dropping with smaller $L$. This makes design very difficult, especially since the DC gain drops with supply voltage.

Most fast transistor come with the penalty of lower speed of operation.
For a bipolar device, the exponential current relationship results in a high constant normalized gain

\[
\frac{g_m}{I_C} = \frac{q}{kT} \approx \frac{1}{26\text{mV}}
\]

For a square law MOSFET, in saturation we have

\[
\frac{g_m}{I_{DS}} = \frac{2}{V_{GS} - V_T}
\]

In weak inversion, the MOSFET is also exponential

\[
\frac{g_m}{I_{DS}} = \frac{q}{nkT} \approx \frac{1}{26\text{mV}} \frac{1}{n}
\]

The factor \( n \) is set by the ratio of oxide to depletion capacitance
MOSFET in Subthreshold

- In sub-threshold, the surface potential varies linearity with $V_G$.
- The surface charge, and hence current, is thus exponentially related to $V_G$.

\[ \psi_s \propto V_G \]

\[ \frac{g_m}{I_D} \propto e^{\frac{qV_G}{nkT}} \]
The I-V curve of a given transistor with fixed dimension ($W$ and $L$) reveals most of the salient features. For example, a plot of $I_{ds}$ versus $V_{gs}$ for a family of $V_{ds}$ quickly reveals current drive capability. If a device is biased in weak or moderate inversion, then the logarithmic plot is more useful as it expands this regime of operation.
The leakage currents (sub-threshold slope) is important in analog applications that use the transistor as a switch. If the device cannot be turned off, then it’s very difficult to build high precision discrete time circuits.
Plotting the current versus drain-source voltage, or a plot of $I_{ds}$ versus $V_{ds}$ for a family of $V_{gs}$ shows the current saturation behavior of a device. The output conductance is more easily observed using small-signal parameters as discussed shortly, but certain trends can be observed even from the raw I-V curves.
Since the power dissipation is determined by and large by the DC current, we'd like to get the most “bang for the buck”.

From this perspective, the weak and moderate inversion region is the optimal place to operate.

The price we pay is the speed of the device which decreases with decreasing $V_{GS}$.

Current drive is also very small.
The value of $g_m$ increases with $V_{gs}$. For the diffusion component of current, at low overdrive, the increase in $g_m$ is exponential due to the exponential dependence of current on gate voltage.

For the drift component of current, $I_{ds}$ is proportional to the amount of charge in the channel and the carrier velocity. The channel charge scales in proportion to the gate bias, so for a fixed mobility, we expect the $g_m$ to increase linearly.
In fact, for low fields, due to Coulomb scattering, the mobility will experience an enhancement due to screening provided by the inversion layer, and the increase in $g_m$ is faster than linear.

On the other hand, due to high field effects, we know the mobility will eventually degrade as carriers are pushed closer to the silicon-insulator interface, where surface scattering causes the mobility to drop.
The trend for $g_m$ as a function of $V_{ds}$ can be divided into two regions. In the triode region of operation, increasing $V_{ds}$ increases the current, so the overall $g_m$ increases.

As the device nears saturation, one would expect the $g_m$ to saturate.
\( g_{ds} \) in triode region is very large since the drain voltage has a direct impact on the channel charge. In fact, to first order, \( g_{ds} = g_m \) since varying the gate has the same impact as varying the gate due to the presence of the inversion layer, making the drain terminal just as effective as the gate-source.

In a long channel device, in the so-called pinch-off region, the device terminal no longer controls the channel charge directly, and the output conductance drops dramatically.
The drain still modulates the depletion region width, which indirectly impacts the device through channel length modulation (CLM).

In short channel devices, both because of the increase in the relative magnitude of the change in channel length $\delta L$ relative to $L$, but also due to Drain Induced Barrier Lowering (DIBL).
Output Resistance Broken Down

\[
R_{\text{out}} \quad \text{k}\Omega
\]

\[
V_{ds} \ (\text{V})
\]

Niknejad

Advanced IC's for Comm
FET Unity Gain Frequency

- **Long channel FET:**
  - Note that there is a peak $f_T$ since eventually the mobility of the transistor drops due to high vertical fields.

- **Short channel limit:**
  \[
  f_T = \frac{1}{2\pi} \frac{g_m}{C_{gs} + C_{gb} + C_{gd}}
  \]

  Assuming $C_{gs} \ll C_{gb} + C_{gd}$

  \[
  f_T = \frac{1}{2\pi} \frac{g_m}{C_{gs}} = \frac{1}{2\pi} \frac{3}{2} \frac{\mu_n}{L^2} (V_{GS} - V_T)
  \]

  \[
  I_{DS} \approx v_{sat} Q_{inv} W = W C_{ox} (V_{GS} - V_T) v_{sat} \rightarrow g_m = W C_{ox} v_{sat}
  \]

  \[
  f_T = \frac{1}{2\pi} \frac{g_m}{C_{gs}} = \frac{3}{2} \frac{W C_{ox} v_{sat}}{W L C_{ox}} \propto \frac{1}{L}
  \]
CMOS transistors have steadily improved in performance just as predicted by theory. In the short channel regime the improvements are linear with scaling.

At the same time, the decreasing supply voltage has led to a reduced dynamic range. Also the maximum gain has not improved as much...
The well known improvement in $f_T$ with channel scaling is shown here. Since both $g_m$ and $C_{gs}$ depend linearity on the width, only the channel length $L$ matters.

Shorter channel lengths improve both the $g_m$ and lower the capacitance of the device. In the velocity saturated limit, only the drop in capacitance plays a role.
The MOS capacitor of modern devices does not follow classical equations due to polysilicon depletion and quantum effects ...
To go fast, you must burn power ...

It is important to note that $f_T$ is usually measured using scattering parameters and so using the “DC” values of $g_m$ and $C$ will fail to capture $f_T$ at high frequency. We will return to this point in the RF section of this chapter.
High Frequency
Measure or calculate two-port parameters at a particular frequency. To obtain maximum gain $G_{\text{max}}$, design an input and output matching network to satisfy the following conditions

$$Y_S = Y_{in}^*$$
$$Y_L = Y_{out}^*$$

where $Y_S$ and $Y_L$ are the source and load admittance seen by the two-port and $Y_{in}$ and $Y_{out}$ are the input and output admittance seen looking into the two-port when loaded by these matched admittances.

This leads to two equations and two unknowns (4 real).
Maximum Power Gain \( f_{\text{max}} \)

- \( f_{\text{max}} = \max \) freq. of activity = \( \max \) freq. of oscillation = freq. when power gain = 1

\[
G_p \approx \frac{\left( \frac{f_T}{f} \right)^2}{4r_x \left( g_o + g_m \frac{C_\mu}{C_\pi} \right) + 4r_x g_o}
\]

\[
G_{\text{max}} \approx \frac{f_T}{8\pi r_x C_\mu f_{\text{max}}^2} = 1
\]

\[
f_{\text{max}} = \sqrt{\frac{f_T}{8\pi r_x C_\mu}}
\]
FET $f_{\text{max}}$

$$f_T \approx \frac{g_m}{2\pi C_{gg}}$$

$$f_{\text{max}} \approx \frac{f_T}{2\sqrt{R_g (g_m C_{gd} / C_{gg}) + (R_g + r_{ch} + R_s)g_{ds}}}$$

- Minimize all resistances
  - $R_g$ – use many small parallel gate fingers, < 1$\mu$m each
  - $R_{sb}$, $R_{db}$ and $R_{bb}$ – substrate contacts < 1 – 2$\mu$m from device
  - $R_s$, $R_d$ – don’t use source/drain extensions to reduce $L$
SiGe HBT and CMOS FinFETs
SiGe Technology

- Higher Performance: Demonstrations of $f_T > 500\text{GHz}$ and approach 1THz.

- Problem:
  - As $W_B$ decreases $\rightarrow r_b$ increases

- Solution:
  - SiGe base allows for higher $f_T$ without reducing $W_B$

\[ \beta = \frac{I_C}{I_B} \]
A SiGe BJT is often called an HBT (heterojunction bipolar transistor)

- Ge epitaxially grown in base
  - Causes strain in crystal
  - Causes extra potential barrier for holes (majority carrier) in the base from flowing into emitter

Beneficial effects

- $W_B$ decreases, $N_B$ increases, $r_b$ low
- $N_E$ decreases, $C_j$ decreases
One of the primary advantages of the III-V based transistors is the higher peak mobility compared to Si. With short channel devices, most of the current flow is due to velocity saturation, which greatly limits the advantages of III-V over silicon.

The insulating substrate also allows higher Q passives.

The extra cost of these technologies limits it to niche applications such as very high frequencies, high performance, and power amplifiers.
To combat the problems with scaling of MOSFETs below 45nm, Berkeley researchers introduced the “FinFET”, a double gate device. (Intel uses this technology at the 22nm node)

Due to thin body and double gates, there is better “gate control” as opposed to drain control, leading to enhanced output resistance and lower leakage in subthreshold.
Gate straddles thin silicon fin, forming two conducting channels on sidewall

Multi-finger layout very similar to RF layout.
An Aside on Thermal Conductivity

- **GaAs**
  - Semi-insulating substrate
  - Not very good conductor of heat
  - High quality passive elements (next topic)

- **Si**
  - Semi-conducting substrate
  - Good conductor of heat
  - Lossy substrate leads to lower quality passives
An Aside on Thermal conductivity (2)

- Also depends on packaging
  - Example: in flip-chip bonding, thermal conductivity function of number of bumps rather than substrate
  - Back-side of die can lose heat through radiation or convection through air but thermal contact is much more effective

- Flip chip bonding

- Wire bonding
High BJT Transconductance

$$g_m = \frac{dI_C}{dV_{BE}} = I_S \frac{q}{kT} e^{\frac{qV_{BE}}{kT}} = \frac{qI_C}{kT}$$

- For fixed current, BJT gives more gain
- Precision
  - Important in multiplication, log, and exponential functions
  - More difficult in FETs due to process/temp. dependence
  - $I_S$ process dependent in BJT ... use circuit tricks
Advantages of BJT

\[\frac{g_m}{I_C} = \frac{q}{kT} = \frac{1}{25\text{mV}}\]
\[\frac{g_m}{I_D} = \frac{2}{V_{GS} - V_T} \approx \frac{1}{250\text{mV}}\]

\[I_C \propto e^{-E/kt} = e^{qV_{BE}/kT}\]

For high-speed applications, need to bias in strong inversion...Results in \(\sim 10\times\) lower efficiency

- For a BJT, this relationship is fundamental and related to the Boltzman statistics (approximation of Fermi-Dirac statistics)
- For a MOSFET, this relationship is actually only valid for a square-law device and varies with \(V_T\) (body bias) and temperature
Advantage of BJT over FET (2)

- Better precision
  - About 4 decades (420mV) of linearity
- Example:
  \[ V_{BE1} + V_{BE2} = V_{BE3} \]
  \[ V_{BE} = V_T \ln \left( \frac{I_C}{I_S} \right) \]
  \[ \frac{I_{C1} \cdot I_{C2}}{I_{S1} \cdot I_{S2}} = \frac{I_{C3}}{I_{S3}} \]

- Can build exp, log, roots, vector mag
- Lower 1/f noise corner
- Lower offset voltage

\[ V_{OS} \sim 1\text{mV} \]
Disadvantage of BJT

- $r_b$ hurts gain (power), NF
  - SiGe allows fast transistors with low $r_b$
- Exponential transfer function (advantage and disadvantage)
  - Exponential $\rightarrow$ non-linear
- Expensive in high volumes, cheaper in low volumes
- Absence of a switch
  - Old CMOS gets cheaper!
  - 45nm $\sim$ $1M$ mask $\rightarrow$ 0.25$\mu$m $50k$ mask
Advantage of FET over BJT

- Cheaper and more widely available (many fabs in US, Asia, and Europe)
- Square law $\rightarrow$ less distortion
- PMOS P-FET widely available, only $\sim 2\times$ less performance.
- Triode region $\rightarrow$ variable resistor
- Widely available digital logic
- Low leakage in gates
  - Sample and hold (S/H) and switch cap filters (SCF)
- Dense digital circuitry / DSP for calibration
  - Offset voltages and mismatches can be compensated digitally
- Dense metal layers allows MIM (“MOM”) capacitors for free
References and Further Reading

- UCB EECS 142/242 Class Notes (Niknejad/Meyer)
- UCB EECS 240 Class Notes (Niknejad/Boser)