The enhancement-type n-channel MOSFET

- **body**: p-type single-crystal silicon substrate
- **source/drain**: heavily doped n-type silicon (n⁺)
- **oxide**: thin insulating layer of SiO₂ (10-100 nm thick)
- **gate**: conducting layer (metal, polysilicon) on top of gate oxide
- **body diodes**: (S/B and D/B) kept reverse-biased and turned off by keeping the body voltage at the lowest potential
- **channel region**: area where current flows between the drain and source; length L (1-100 μm) and width W (1-500 μm)
- **symmetry**: definition of source and drain depends on the current for IC devices where body contact is ignored.

- Note that a real device typically has the body and source terminals connected together. If a positive bias is applied to the node relative to the drain, the body/drain diode will turn on and conduct current.
- Also note that d = 0.01W or 0.01L — these devices are much thinner in the q-axis than drawn here.

---

**No gate voltage (Vgs = 0)**

No current can flow between the drain and source due to the back-to-back diodes.

Creating a channel (Vgs > 0)

- positive charge from Vgs on the gate causes positive charge in the channel region to be repelled, creating a depletion region.
- negative charge from the source/drain regions is attracted to the positive gate potential, forming an induced channel of electrons between source and drain.
- when sufficient negative charge has gathered in the channel region to allow current to flow between the source and drain, the channel region is inverted. The value of Vgs required to invert the channel is called the threshold voltage Vth.
Applying a small $V_{gs}$ ($0 < V_{gs} < 0.2 V$)

- $V_{ds} = V_{t}$

### Negative Carriers (electrons) Flow from Source to Drain
- Current flows from drain to source.
- $i_0 = f(V_{ds})$ (density of electrons in channel)
- Electron density = $f(V_{ds})$
- Channel conductance $\frac{i}{V_{ds}} \propto V_{ds} - V_{t}$

Looking at just the channel:

- For small $V_{gs}$, the MOSFET is a linear resistance that is inversely proportional to $V_{ds} - V_{t}$.
- The voltage $V_{gs}$ enhances the channel conduction, hence the enhancement-type device.

Increasing $V_{gs}$ ($V_{gs} > V_{t}$, $V_{gs}$ constant)

- $V_{ds}$ appears as a voltage drop distributed along the length of the channel.
- At the source: channel voltage = 0
  gate-channel voltage = $V_{gs}$
- At the drain: channel voltage = $V_{ds}$
  gate-channel voltage = $V_{gs} - V_{ds}$

As $V_{gs}$ increases, the channel resistance increases due to the tapering of the channel.

When the voltage at the drain increases to the point that $V_{ds} - V_{gs} = V_{t}$ ($V_{gs} = V_{ds} - V_{t}$), the channel is pulled off. Increasing $V_{gs}$ further does not affect the channel shape.
In terms of our $I-V$ characteristic, we have three regions:

I) drain current increases linearly

$$i_d \propto V_{GS}-V_T$$

II) drain current curve bends because channel resistance ($\frac{\rho_{ch}}{W}$) increases as channel becomes doped.

III) drain current flattens out (saturates) when the channel becomes pinch off.

The drain current saturates for $V_{GS} \geq V_{th} - V_T$

$V_{GS} \geq V_{th} - V_T$ Saturation region

$V_{GS} \leq V_{th} - V_T$ Triode or linear region

Last time we looked at the NMOS $I-V$ characteristic

$V_T$ = threshold voltage - the voltage required on the gate to induce a conducting channel between the source and drain. $V_{GS} \leq V_T$ = cutoff $i_d = 0$

There are two circuit symbols used for the n-channel enhancement MOSFET:

- Arrow indicates direction of current flow and identifies device as n-channel FET.
Now let's put the FET in a circuit

\[
\begin{array}{c}
\text{\textbf{i}_b} \\
\text{\textbf{v}_{ds}} \\
\text{\textbf{i}_d} \\
\text{\textbf{v}_{gs}} \\
\end{array}
\]

\[
\text{i}_0 = 0
\]

We will use two equations to define the currents

\text{triode} \quad \text{i}_0 = k'_n \left( \frac{W}{L} \right) \left[ (v_{gs} - v_t) v_{ds} - \frac{1}{2} v_{ds}^2 \right]

\text{saturation} \quad \text{i}_0 = \frac{1}{2} k'_n \left( \frac{W}{L} \right) (v_{gs} - v_t)^2 \quad \text{for} \quad v_{gs} = v_{ds} - v_t

\text{k}'_n \equiv \text{transconductance parameter}

\text{typically} \quad 20 \leq k'_n \leq 100 \mu A/V^2

Consider three cases:

1) \quad v_{gs} > v_t, \quad v_{ds} \text{ small} \quad \Rightarrow \quad v_{ds} \rightarrow 0

\text{i}_0 \approx k'_n \left( \frac{W}{L} \right) (v_{gs} - v_t) v_{ds}

\text{i}_0 \propto v_{ds}

\text{current increases linearly with v_{ds}}

2) \quad v_{ds} > v_t, \quad v_{ds} < v_{gs} - v_t

\text{triode region} \quad \text{i}_0 \text{ no longer increasing linearly due to the } \left(-\frac{1}{2} v_{ds}^2\right) \text{ term.}

3) \quad v_{gs} > v_t, \quad v_{ds} \geq v_{gs} - v_t

\text{saturation} \quad \text{i}_0 \text{ independent of v_{ds}}

\text{The boundary between the triode region and saturation is given by } v_{ds} = v_{gs} - v_t

\text{triode} \quad \text{i}_0 = k'_n \left( \frac{W}{L} \right) \left[ (v_{gs} - v_t) v_{ds} - \frac{1}{2} v_{ds}^2 \right]

\quad = k'_n \left( \frac{W}{L} \right) \left( v_{gs} - v_t \right) (v_{ds} - v_t) - \frac{1}{2} \left( v_{gs} - v_t \right)^2

\text{i}_0 = \frac{1}{2} k'_n \left( \frac{W}{L} \right) (v_{gs} - v_t)^2 \quad \text{saturation}
\[ V_{bs} = V_{be} - V_T \]

\[ i_D = \frac{1}{2} k_n \left( \frac{W}{L} \right) (V_{bs} - V_T)^2 = \frac{1}{2} k_n \left( \frac{W}{L} \right) V_{bs}^2 \]

The boundary between triode and saturation is given by a square relationship on the I-V characteristic.

---

**Example 1**

\[ V_{BS} = 5V \]
\[ V_T = 0.8V \]
\[ W = 50\mu m \]
\[ L = 5\mu m \]

What is \( i_D \) for \( V_{bs} = 0.5V, 1.2V, 3V \)?

a) \( V_{bs} = 0.5V < V_T \) \( \Rightarrow \) cutoff

\[ i_D = 0 \]

b) \( V_{bs} = 1.2V \) \( \text{try saturation} \)

\[ i_D = \frac{1}{2} k_n \left( \frac{W}{L} \right) (V_{bs} - V_T)^2 = \frac{1}{2} \left( \frac{40\mu A}{V} \right) \left( \frac{50\mu m}{5\mu m} \right)(1.2V - 0.8V)^2 \]

\[ i_D = 32\mu A \]

\[ V_R = (13k\Omega)(32\mu A) = 0.44V \]

\[ V_{BS} = 5V - 0.44V = 4.56V \]

\[ 4.56V > 1.2V - 0.8V \Rightarrow \text{Saturation} \]

c) \( V_{bs} = 3V \) \( \text{try saturation} \)

\[ i_D = \frac{1}{2} \left( \frac{40\mu A}{V} \right) \left( \frac{50\mu m}{5\mu m} \right) (3V - 0.8V)^2 \]

\[ i_D = 0.97mA \]

\[ V_{BS} = 5V - (0.97mA)(3.3k\Omega) = 1.81V \]

\[ 1.81V < 3V - 0.8V = 2.2V \]

not saturated

\[ \text{try triode} \]

\[ i_D = k_n \left( \frac{W}{L} \right) \left[ (V_{bs} - V_T) V_{bs} - \frac{1}{2} V_{bs}^2 \right] \]

\[ V_{BS} = 5V - 3300 i_D \]

\[ i_D = (40\mu A) \left( \frac{50\mu m}{5\mu m} \right) \left[ (3V - 0.8V)(5 - 3300 i_D) - \frac{1}{2} (5 - 3300 i_D)^2 \right] \]

\[ 2500 i_D = 0.11 - 7260 i_D - 12.5 + 16500 i_D - 5945000 i_D^2 \]

\[ 5945000 i_D^2 - 6740 i_D + 1.5 = 0 \]

\[ i_D = \frac{6740 \pm \sqrt{(6740)^2 - 4(5945000)(1.5)}}{2(5945000)} \]

\[ i_D = \frac{6740 \pm 3872}{10890000} = 0.0747 \mu A \]

The only answer that makes sense is

\[ i_D = 0.0747 \mu A \text{ triode region} \]

\[ V_{BS} = 1.88V \]
Example 2.

What are \( i_0 \) and \( v_{GS} \)?

\[ v_{GS} = v_{GS} > v_{GS} - v_t \text{ always saturated} \]

\[ i_0 = \frac{1}{2} k' \left( \frac{W}{L} \right) (v_{GS} - v_t)^3 \]

\[ i_0 = \frac{10 - v_{GS}}{15k} \quad v_{GS} = 10 - 15000i_b \]

\[ i_0 = \frac{1}{2} k' \left( \frac{20 \mu A}{0.1 \mu m} \right) \left( 10 - 15000i_b - 2 \right)^2 \]

\[ 10000i_0 = 64 - 129000i_b - 225(0.9)^2 \]

\[ i_0 = 0.4 \text{ mA} \quad \Rightarrow \quad i_0 = 0.71 \text{ mA} \]

Example 3.

\[ k' = 1 \text{ mA/\mu m}^2 \]

\[ v_t = 1 \text{ V} \]

\[ v_g = \frac{10 \text{ mA}}{10 \text{ mA} + 10 \text{ M\Omega}} \]

\[ 10 \text{ V} = 5 \text{ V} \]

\[ v_{GS} = 5 - 6000i_b \]

Assume saturation.

\[ i_b = \frac{1}{2} k' \left( v_{GS} - v_t \right)^2 \]
\[ i_d = \frac{1}{2} (\frac{V_{GS}}{V_{TH}}) (5 \cdot 6000 i_p - 1)^3 \]

\[ 20000 i_d = 16 - 48000 i_0 + 36 \times 10^4 i_0^2 \]

\[ 36 \times 10^4 i_0^2 - 50 \times 10^3 i_0 + 16 = 0 \]

\[ i_0 = \begin{cases} +190 \, \mu A & \text{or} \\ +500 \, \mu A & \end{cases} \]

For \( \beta \): \( \gamma = 6000 i_0 = 5.3 \, V \geq \% \)

doesn't make sense

\[ \therefore \quad i_0 = 500 \, \mu A \]

\[ V_s = 6000 \, i_0 = 3 \, V \]

\[ V_{GS} = 2 \, V \]

\[ V_s = 10 - 6000 i_0 = 7 \, V \]

\[ V_{GS} = V_s - V_t = 9 \, V \]

\[ V_{DS} = 4 \, V \geq V_{GS} - V_t = 3 \, V - 1 \, V = 2 \, V \]

saturation - assumption correct

Recall for the n-channel MOSFET:

- **cutoff** \( V_{GS} < V_t \quad i_d = 0 \)
- **triode** \( V_{GS} < V_t < V_{DS} \quad i_d = \frac{1}{2} \frac{V_{GS} - V_t}{V_{th}} \left( V_{GS} - \frac{1}{2} V_t \right) \)
- **saturation** \( V_{DS} > V_{GS} - V_t \quad i_d = \frac{1}{2} \frac{V_{GS} - V_t}{V_{th}} \left( V_{DS} - V_t \right)^2 \)

\[ k' = \text{transconductance parameter} \]

\[ k_s = \mu V_{th} \]

\[ \mu = \text{mobility of electrons in channel} \]

\[ \left( \frac{V_{th}}{V_{th}} \right) = \frac{1}{\mu} \]

\[ C_{ox} = \text{oxide capacitance per area} \]

\[ C_{ox} = \frac{C_{ox}}{A_{ox}} \quad C_{ox} = 3 \times 10^{-10} \, \text{F/cm}^2 \]

We said that in saturation the channel is pinched off and does not change shape with increasing \( V_{GS} \). But...

\[ V_{GS, sat} = V_{GS} - V_t \]

\[ V_{GS, sat} \uparrow \rightarrow V_{GS} \]

\[ V_{GS} \downarrow \rightarrow V_{GS} \]

Source \quad \text{drain}
The excess voltage \( V_{BS} - V_{DS, sat} \) appears across the depletion region between the end of the channel and the drain.

An electron travelling from source to drain is accelerated first by \( V_{DS, sat} \), then \( V_{BS} - V_{DS, sat} \), and is "swept" across the depletion region to the drain.

The channel length \( L \) is reduced to \( L' \).

\[
i_D = \frac{1}{2} \mu_n C (V_{BS} - V_t)^2 \left( 1 + \lambda V_{BS} \right)
\]

\[
L' \sim \frac{L}{V_{BS}}
\]

\[
i_D \sim \frac{1}{L} \sim V_{BS}
\]

\[\therefore \text{ as } V_{BS} \text{ increases, so will } i_D.\]

This effect is called channel-length modulation.

Modify the \( i_D \) equation:

\[
i_D = \frac{1}{2} \mu_n C \left( V_{BE} - V_t \right)^2 \left( 1 + \lambda V_{BE} \right)
\]

\[0 = \frac{1}{2} \mu n C \left( V_{BE} - V_t \right)^2 \left( 1 + \lambda V_{BE} \right)
\]

\[1 + \lambda V_{BE} = 0\]

\[V_{BE} = -\frac{1}{\lambda} = -V_A\]

\(\lambda\) = channel length modulation parameter

\(V_A\) = Early voltage (from bipolar transistor theory)

Note that this effect will be worse for devices with short channels (small \( L \)).
The Body Effect - another look at the substrate

Suppose we build the following circuit in an IC:

What does this circuit do? Consider the following table...

<table>
<thead>
<tr>
<th>$V_i$</th>
<th>$V_B$</th>
<th>$M_A$</th>
<th>$M_B$</th>
<th>$V_o$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5V</td>
<td>0V</td>
<td>off</td>
<td>off</td>
<td>5V</td>
</tr>
<tr>
<td>0V</td>
<td>5V</td>
<td>off</td>
<td>on</td>
<td>5V</td>
</tr>
<tr>
<td>5V</td>
<td>0V</td>
<td>on</td>
<td>off</td>
<td>5V</td>
</tr>
<tr>
<td>5V</td>
<td>5V</td>
<td>on</td>
<td>on</td>
<td>0V</td>
</tr>
</tbody>
</table>

This is the truth table for a NAND gate.

Where do we connect the body terminals of the FETs?

The FETs share the same substrate, which must be connected to the most negative voltage - ground - to keep the p-n junctions reverse biased.

So we REALLY have this:

Look at top FET

$V_{BS} = -V_{DS,B}$

There will be a reverse bias voltage between the source and the body, causing an increase in the threshold voltage $V_t$.

$$V_t = V_{to} + \gamma \left[ \sqrt{2qI_i V_{GS}} - \sqrt{2qI_i^2} \right]$$

threshold voltage

physical parameter

body-effect parameter $\gamma = \frac{\sqrt{2qN_A E_i}}{C_{ox}}$ (process parameter)
In SPICE the MOSFET is represented by a model, where how elaborate the model is depends on its level:

<table>
<thead>
<tr>
<th>LEVEL</th>
<th>MODEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 *</td>
<td>square law (as presented last week and this week)</td>
</tr>
<tr>
<td>2</td>
<td>physical model - not used much</td>
</tr>
<tr>
<td>3</td>
<td>mix of physics (level 1) and measurements</td>
</tr>
<tr>
<td>&gt; 3</td>
<td>BSIM (Berkeley short-channel IGFET model)</td>
</tr>
</tbody>
</table>

* sometimes called the Shockley-Nichols model

For ICs, all the MOSFETs are made at the same time and behave very similarly.

Same $V_t$, $Cox$, $\lambda$

But each device is sized separately - unique $L$ and $W$

PSpice includes devices called MbrackN3 (source tied to body) and MbrackN4 (four terminals = $G, S, D, B$).

For the family of devices (type MbrackN3) a model defines the general device characteristics. Each device needs a unique $W$ and $L$, however.

See the PSpice tutorial on the web page!

---

FET on resistance / output resistance

1) For a MOSFET in the triode region, we said that

$$i_D = k''_n \left( \frac{W}{L} \right) (V_{GS} - V_t) V_{GS} - \frac{1}{2} V_{DS}^2$$

$$= k''_n \left( \frac{W}{L} \right) (V_{GS} - V_t) V_{GS} \quad (V_{DS} \ll 1)$$

$$i_D \propto V_{DS}$$

:: FET acts like a resistor

$$\frac{V_{DS}}{i_D} = \frac{V_{GS}}{i_D} = \frac{1}{k''_n \left( \frac{W}{L} \right) (V_{GS} - V_t)}$$

2) In saturation (neglect CDM)

$$i_D = \frac{1}{2} k''_n \frac{W}{L} (V_{GS} - V_t)^2$$

The small-signal output resistance is defined

$$r_D = \left. \frac{\partial V_D}{\partial V_{GS}} \right|_{V_{DS} \text{ constant}}$$

$$r_D = \frac{1}{\left( \frac{\partial i_D}{\partial V_{DS}} \right)} = \frac{1}{\partial_i} \rightarrow 00$$
If we include CDM

\[
  i_D = \frac{1}{2} \frac{d}{d\nu_{GS}} \left( \nu_{GS} - \nu_T \right)^2 (1 + \lambda \nu_{GS})
\]

\[
  \frac{1}{\nu_0} = \frac{d}{d\nu_{GS}} \bigg|_{\nu_0} = \frac{1}{2} \frac{d}{d\nu_{GS}} (\nu_{GS} - \nu_T)^2 \lambda
\]

\[
  \approx I_D
\]

\[
  \frac{1}{\nu_0} = \lambda I_D
\]

\[
  \nu_0 = \frac{1}{\lambda I_D}
\]

Note that \( I_D = i_D \bigg|_{\nu_0 \text{ constant}} \)

(we fixed \( I_D \) at a dc value when we held \( \nu_{GS} \) constant)

\[
  \frac{\nu_{GS}}{\nu_0} = \frac{\nu_{GS}}{\nu_0}
\]

Incremental slope in saturation

\[
  \left( \frac{\nu_{GS}}{\nu_0} \right) \bigg|_{\nu_{GS} \text{ constant}} = \frac{1}{\nu_0}
\]

\[
  \text{slope in trade region}
\]

\[
  \nu_{GS} \text{ increasing}
\]

If \( \lambda = 0 \) then \( \nu_0 \to \infty \)

The p-channel enhancement MOSFET

To get a conducting channel between the source and drain, we need to attract positive charge (holes) to the surface. We do this by applying a negative voltage to the gate. \( \nu_s < 0 \)

In the schematic world, what we have is

Adding the \( \nu_{GS} \) and \( \nu_{DS} \) sources

Note: \( \nu_{GS} < 0 \)

\( \nu_{DS} < 0 \)
The expressions for $i_p$ are similar to the NMOS:

\[ i_p = k_p \frac{W}{L} \left( \frac{V_{gs} - V_T}{V_T} \right) \left( V_{gs} - \frac{1}{2} V_T \right) \] (trioide)

If $i_p$ is defined as the current coming out of the drain, the signs work out OK.

Transconductance parameter $k_p$:

\[ k_p = \mu_p C_{ox} \]

\[ \mu_p \text{ = hole mobility} \approx 230 \text{ cm}^2/\text{V}\cdot\text{s} \]

\[ \mu_n \text{ = electron mobility} \approx 580 \text{ cm}^2/\text{V}\cdot\text{s} \]

\[ \mu_p = 0.4 \mu_n \]

Recall that holes move like a bubble in a sea of electrons; therefore, they move a little slower — their mobility is less.

Since $V_T < 0$, turning the device on requires a more negative voltage to be applied to the gate.

\[ V_{gs} = V_T \quad \text{cutoff} \rightarrow \text{triode} \]

To saturate the device we need to make $V_{gs}$ more negative:

\[ V_{gs} = V_{gs} - V_T \]

At saturation, $V_{ds} = V_{gs} - V_T$ — triode — saturation

\[ i_p = \frac{k_p}{L} \left( \frac{V_{gs} - V_T}{V_T} \right) \left( V_{gs} - \frac{1}{2} (V_{gs} - V_T) \right) \]

\[ i_p = \frac{1}{2} k_p \frac{W}{L} (V_{gs} - V_T)^2 \] — saturation

Finally, IC devices are fabricated in one piece of Si.
b) Find \( i_D \) if \( V_D = -5V \)

\[-5V \leq 2V \Rightarrow \text{saturated}\]

\[i_D = \frac{1}{2} k_c \left( \frac{V_{GS} - V_t}{V_s} \right)^2\]

\[\Rightarrow \frac{1}{2} \left( \frac{50 \text{ mA}}{V_s} \right) \left( -5 - (-2) \right)^2 = \left( \frac{50 \text{ mA}}{V_s} \right) (-3)^2\]

\[i_D = 450 \text{ mA}\]

Process Note: Complementary MOS (CMOS) combines NMOS and PMOS in a single substrate. This is done by introducing an n-well region in areas where PMOS devices are.

NMOS

PMOS

(in n-well)
The NMOS Inverter (with resistor load)

What does the transfer characteristic (\(V_O\) vs. \(V_I\)) look like?

1. \(V_I < V_T \rightarrow \) cutoff
   \[ V_O = V_{DD} \]

2. \(V_I > V_T, \; V_O = V_{DD} \geq V_T - V_T \rightarrow \) saturation
   \[ i_p = \frac{1}{2} k_n W L (V_T - V_I)^2 \]
   \[ V_O = V_{DD} - \frac{1}{2} k_n W L (V_T - V_I)^2 R_D \]

3. \(V_I > V_T \rightarrow \) saturation/friode
   \[ V_O = V_{DD} - \frac{1}{2} k_n W L V_O^2 R_D \]
   \[ \frac{1}{2} k_n W L V_O^2 + V_O - V_{DD} = 0 \]
   \[ V_O = -\frac{1 + \sqrt{1 + 2 \frac{k_n W L}{k_n W L} V_{DD}}}{\frac{k_n W L}{k_n W L} R_D} \]

4. \(V_I > V_T \rightarrow \) triode
   \[ i_p = k_n W L \left( \frac{V_T - V_I}{V_D - \frac{1}{2} V_T} \right) \]
   \[ V_O = V_{DD} - i_p R_D \]
   \[ V_O = V_{DD} - k_n W L \left( V_T - V_I \right) \left( \frac{V_D}{V_T} - \frac{1}{2} V_T \right) \]
   \[ \frac{1}{2} k_n W L R_D V_O^2 - \left( 1 + k_n W L R_D \right) \left( V_T - V_I \right) V_O + V_{DD} = 0 \]

   \(\Rightarrow\) solve for \(V_O\) as a function of \(V_I\)

Example

\[ V_I = 1 V \]
\[ V_T = 5 V \]
\[ R_D = 5 k\Omega \]

\[ V_O = 1 mA/\mu A \]

\[ V_I \rightarrow V_O \rightarrow V_T \]

\[ 0 \leq V_I \leq 5 V \]

5. \(V_I < V_T \rightarrow \) cut-off
   \[ V_O = 5 V \]

6. \(V_O = V_T - V_I \rightarrow \) saturation/triode

\[ V_O = -\frac{1 + \sqrt{1 + 2 \frac{k_n W L}{k_n W L} V_{DD}}}{\frac{k_n W L}{k_n W L} R_D} \]

\[ V_O = 1.23 V \]

\[ V_T = V_O + V_I = 2.23 V \]
c) $v_3 = 5V$ => triode

$$2.5 \cdot v_3^2 - 24 \cdot v_3 + 5 = 0$$

$$v_3 = \frac{24 \pm \sqrt{24^2 - 4 \cdot 2.5 \cdot 5}}{2 \cdot 2.5} = \frac{24 \pm \sqrt{516}}{5}$$

$$v_3 = 4.8 \pm 4.59$$

$$v_3 = 0.213V$$

This can be done in an Excel spreadsheet - I'll post one on the web page sometime soon.

Note that this implements an inverter function.

Let $v = 5V$ be 1

$v < 0.3V$ be 0

<table>
<thead>
<tr>
<th>$\bar{v}$</th>
<th>$v$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

We would like some way to characterise the static performance of our logic circuit.

Static means that the input is held constant (or varies very slowly), as opposed to dynamic operation where we switch the input quickly.

Consider a typical voltage transfer characteristic (VTC)

a) when $v_1 = 0$, $v_3$ is at its maximum value - the output is high. Call this $V_{OH}$

b) the input is probably being driven by a similar logic gate. Therefore when $v_1 = V_{OH}$, the output will be driven to its minimum value. Call this $V_{OL}$

c) define the points where the output changes as a result of the input by the slope $\frac{dv_3}{dv_1} = -1$. The lower point represents the maximum input that is still considered a zero. Call it $V_{IL}$. The upper point is the minimum input still considered a one. Call it $V_{IH}$. 

$$\frac{dv_3}{dv_1} = 1$$

Graphically:

- $V_{OH}$
- $V_{OL}$
- $V_{IL}$
- $V_{IH}$

$$\frac{V_{OH} - V_{OL}}{V_{IH} - V_{IL}} = -1$$
d) The point where \( V_2 = V_0 \) is called the inverter's threshold \( V_{th} \). Don't confuse this with the MOSFET's threshold voltage \( V_T \).

We want a logic circuit that is not easily fooled by small changes in the levels due to noise. Define the robustness with the noise margins:

\[
NM_H = V_{OH} - V_{IH}
\]

\[
NM_L = V_{IL} - V_{OL}
\]

Ideally we want \( NM_H = NM_L = \frac{1}{2} V_{DD} \)

\[
V_{OH} = V_{DD} \quad V_{IH} = \frac{1}{2} V_{DD}
\]

\[
V_{IL} = \frac{1}{2} V_{DD} \quad V_{OL} = 0
\]

Ideal VTC:

\[\begin{array}{c|c}
V_0 & \frac{1}{2} V_{DD} \\
V_{th} & V_{DD}
\end{array}\]

Back to our example:

\[
V_{OH} = 5V
\]

\[
V_{OL} = 0.33V
\]

\[
\begin{array}{c}
V_0 = V_{DD} - \frac{1}{2} L \left( V_2 - V_T \right) R_D \\
\frac{\partial V_0}{\partial V_2} = -\frac{1}{2} L \left( V_2 - V_T \right) R_D = -1
\end{array}
\]

\[
V_T = \frac{1}{2} \left( \frac{1}{L} \right) R_D + V_T
\]

\[
= \frac{1}{\left( \frac{1}{L} \right) (5000)} + 1 = 1.2V
\]

Can solve for \( V_0 \)

(we don't need to for noise margins)

\[
\begin{bmatrix}
V_0 = 5 - \frac{1}{2} (w^2) (1.2 - 1)^2 (5000) \\
V_{OL} = 1.2V
\end{bmatrix}
\]

In triode:

\[
\begin{array}{c}
2.5 V_0^2 - 6 (V_2 + 1) V_0 + 5 = 0 \\
-6 V_2 V_0 = -5 - 6 \frac{V_2}{6} - 7.5 V_0^2
\end{array}
\]

\[
V_2 = \frac{2.5 V_0^2 + 6 V_0 + 5}{6} \frac{V_0}{6}
\]

\[
\frac{\partial V_2}{\partial V_0} = \frac{6 V_0 (5 V_0 + 6) - (25 V_0^2 + 6 V_0 + 5) 6}{36 V_0^2} = -1
\]

\[
30 V_0^2 + 36 V_0 - 15 V_0^2 - 36 V_0 - 30 = -36 V_0^2
\]

\[
51 V_0^2 - 30 = 0
\]

\[
V_0^2 = \frac{30}{51}
\]

\[
V_0 = 0.59V
\]
\[ V_s = \frac{25 \cdot 5^2 + 6 \cdot 5 + 5}{6 \cdot 5} = 2.66 \text{ V} \]

\[ V_{IH} = 2.66 \text{ V} \]

So we have

\[ V_{OH} = 5 \text{ V} \quad V_{IH} = 2.66 \text{ V} \]

\[ V_{OL} = 1.2 \text{ V} \quad V_{DL} = 0.213 \text{ V} \]

\[ NM_H = 5 - 2.66 = 2.34 \text{ V} \]

\[ NM_L = 1.2 - 0.213 = 0.99 \text{ V} \]

This type of inverter - NMOS with resistor load - tends to have small noise margins, especially \( NM_L \).

**Static Power Dissipation** - How much power does our inverter use up?

a) \( V_I = 0 \), \( V_O = V_{DD} \) - MOSFET cutoff

\[ i_D = 0 \Rightarrow P_OH = 0 \]

b) \( V_I = V_{OH} = V_{DD} \), \( V_O = V_{OL} \)

\[ i_D = \frac{V_{DD} - V_{OL}}{R_D} \Rightarrow P_{OL} = \frac{(V_{DD} - V_{OL})V_{DD}}{R_D} \]

For our example \( P_{OL} = \frac{(5 - 0.213)5}{500} = 4.9 \text{ mW} \)

If the inverter spends some of the time with its output high, the average static dissipation is

\[ P_{OAVG} = \frac{P_{OH} + P_{OL}}{2} = 2.4 \text{ mW} \]

"The NMOS logic can consume a lot of power, which means you'll get a very hot chip!"
There are two points on the transfer characteristic where the slope \( \frac{\partial I_D}{\partial V_T} = -1 \). We defined these to be the input switching points:

\[
v_i < v_{1T} \quad \frac{\partial I_D}{\partial V_T} \bigg|_{v_i = v_{1T}} = -1
\]

\[
v_i > v_{1T} \quad \frac{\partial I_D}{\partial V_T} \bigg|_{v_i = v_{1T}} = -1
\]

The noise margins are then defined as:

\[
NM_H = V_{OH} - V_{IH}
\]

\[
NM_L = V_{OL} - V_{IL}
\]

**Circuit Area**

![Circuit Diagram]

Let \( V_T = 0.8 \text{ V} \)  \( k_n' = 40 \mu A/\text{V}^2 \)

\[
W = 50 \mu m \quad L = 2 \mu m
\]

Then \( I_{D\max} = 1 \text{ mA/\text{V}^2} \)

The area consumed by our device is:

\[
A = W \times (L + 2k_n')
\]

Assume \( L_c = L \)

\[
A = W (3L) = 300 \mu m^2
\]

Now to complete our circuit, we need a resistor. A resistor can be made from a continuous stripe of diffused region, such as n+:

![Resistor Diagram]

To calculate how big our resistor is, we need to know a little more about diffused resistors.
A resistor formed this way in an IC will depend upon a material parameter called the sheet resistance $R_s$, which has the unusual units $\Omega/\text{sq}$ (ohms/square).

Consider two resistors made from ni material with $R_s = 200 \Omega/\text{sq}$:

$$R_A = \begin{array}{c}
\text{[diagram with dimensions]} \\
L = 30 \mu\text{m}
\end{array}$$

$$R_B = \begin{array}{c}
\text{[diagram with dimensions]} \\
W = 5 \mu\text{m}
\end{array}$$

The number of squares can be determined by the ratio

$$S = \frac{L}{W}$$

A) $S_A = \frac{30 \mu\text{m}}{10 \mu\text{m}} = 3 \Omega$

$$R_A = R_s S_A = (200 \Omega/\text{sq}) \times 3 \Omega = 600 \Omega$$

B) $S_B = \frac{15 \mu\text{m}}{5 \mu\text{m}} = 3 \Omega$

$$R_B = R_s S_B = (200 \Omega/\text{sq}) \times 3 \Omega = 600 \Omega$$

Even though the resistors are different sizes, the value of the two resistors is the same!

You can think of 1 square as a resistor $R$:

$$\square \Rightarrow \begin{array}{c}
R \\
0.5 \Omega 0.5 \Omega
\end{array}$$

Combine 4 squares into one bigger square:

$$\downarrow$$

Let's reverse the process to build a pull-up resistor for our NMOS inverter.

$$R_D = 10 \text{k}\Omega$$

Let $R_s = 33 \Omega/\text{sq}$

$$W = 2 \mu\text{m}$$

What is $L$?

$$S = \frac{R_D}{R_s} = \frac{10000 \Omega}{33 \Omega/\text{sq}} = 300 \Omega$$

$$L = SW = (300 \Omega/2) = 600 \mu\text{m}$$
How much area will our resistor take?

\[
A = (L + 2C)W = (L + 2w)W = WL + 2W^2
\]

\[
\begin{align*}
L &= 600 \mu m \\
W &= 2 \mu m
\end{align*}
\]

\[
A = 1208 \mu m^2
\]

\[\therefore\] Our resistor is 4 times bigger than our FET.

Another blow against the NMOS inverter...

**Switching Speed**

Let's look at what happens if \( V_g \) is a series of pulses: \( V_g(t) \)

\[
V_{PP} = 5V \\
V_f = 0.8V \\
V_{be} = V_{ce} - (V_f - V_g) \exp \left(-\frac{t - t_0}{\tau} \right)
\]

\[
\begin{align*}
R_1 &= 10k \Omega \\
R_2 &= 10k \Omega \\
W &= 50 \mu m \\
L &= 2 \mu m
\end{align*}
\]

Two cases - when the input goes high to low, and low to high

1) Input transition high to low

Assume square-wave input.

Output rises exponentially as \( C_m \) of \( M_2 \) changes through \( R_{p1} \).

\[\uparrow\]

\[ M_1 \text{ turns off at } t_0 \]

\[ V_{o0} \]

Expect an exponential:

\[
V_{o1} = V_{o0} - \left( V_{o0} - V_f \right) \exp \left(-\frac{t - t_0}{\tau} \right)
\]

\[
\begin{align*}
V_f &= V_{be} \\
V_{o0} &= V_{be}
\end{align*}
\]

Using the Excel Spreadsheet (it's on the web!)

\[
V_{o0} = 0.118V
\]

\[ t_0 = S = \left( 5 - 0.118 \right) \exp \left( -\frac{t - t_0}{\tau} \right) \]

What is \( C_m \)?

\[
k_m' = \mu_n C_m \Rightarrow C_m = \frac{k_m'}{\mu_n} = \frac{40 \mu A/V^2}{580 \text{ mV/\mu m}} = 69 \text{ nF/cm}^2
\]

\[
C_g = C_m WL = \left( 69 \text{ nF/cm}^2 \right) \left( 50 \mu m \right) \left( 2 \mu m \right) = 69 \text{ FF}
\]
\[ R_c = 40 \times 10^{-6} \times (0.64\text{ns}) = 0.69 \text{ns} \]

\[ n_{b1} = 5 - 4.882 \exp \left( -\frac{t - t_0}{0.64 \text{ns}} \right) \]

Define the rise time as the time it takes the output to go from 10% (0.5V) to 90% (4.5V)

\[ 0.5 = 5 - 4.882 \exp \left( -\frac{t_1 - t_0}{0.64 \text{ns}} \right) \]
\[ t_1 - t_0 = 56 \text{ ps} \]

\[ 4.5 = 5 - 4.882 \exp \left( -\frac{t_2 - t_0}{0.64 \text{ns}} \right) \]
\[ t_2 - t_0 = 1.57 \text{ ns} \]

\[ t_v = t_2 - t_1 = 1.57 \text{ ns} - 56 \text{ ps} = 1.52 \text{ ns} \]

This isn't so bad... but what if our first inverter were driving eight logic gates?

\[ t_v = 8(1.52 \text{ ns}) = 12.2 \text{ ns} \]

This is not desirable - the rise time is getting a bit slow, and it is directly proportional to the fan-out, or the equivalent load on the inverter.

2) Input transition too high

This is harder to solve for, because the FET operates in two regions - saturation, then triode.

\[ i_p = \frac{1}{2} \left( \frac{k}{k_L} \right) (V_T - V_d) \]
\[ V_T = V_{dd} \]

\[ k, k_L, V_T, V_d \] are all constant device parameters

\[ V_T = V_{dd} \] is constant

So, FET looks like constant current source

\[ i_p = i_r \left( \frac{k}{k_L} \right)(5 - 0.8)^k \]

\[ i_p = 8.8 \text{ mA} \]

\[ i_r = \frac{V}{R} = \frac{5}{10 \text{ k\Omega}} = 0.5 \text{ mA} \]

\[ C_g \frac{dV}{dt} = i_p - i_r \]

\[ C_g \frac{dV}{dt} = \frac{i_p - i_r}{0.5} = \frac{8.8 \text{ mA} - 0.5 \text{ mA}}{69 \text{ pF}} = 128 \text{ V/s} \]

This starts out fast.

b) Triode

As the voltage \( V_{dd} \) decreases, the FET moves into the triode region and the current \( i_p \) decreases.

Now the FET looks more like a resistor than a current source.
This becomes a good Spice problem...

We have seen that the NMOS inverter has the following problems:

1) Poor noise margin, especially NM
2) Static power dissipation
3) Large area due to resistor
4) Slow switching speed

The solution: CMOS!

Let's look a little closer at the oxide capacitance:

\[
\text{a) FET in cutoff region } (V_{GS} < V_t)
\]

parallel-plate capacitor \( C_{gb} = C_{ox}WL \)
between gate and body \( C_{ge} = C_{gd} = 0 \)

\[
\text{b) triode region}
\]

The induced channel shields the body contact from the gate, so the capacitance splits between the source and drain:

\[
C_{gs} = C_{gd} = \frac{1}{2} C_{ox}WL
\]

\[
C_{gd} = 0
\]

\[
C_{gs} = \frac{1}{2} C_{ox}WL
\]

There are also parasitic capacitances associated with the drain-body and source-body pn junctions.

Spice can easily include all these effects.
The CMOS inverter

We would like to construct the transfer characteristic. First consider two extremes:

1) \( V_1 = \text{logic zero} = 0 \text{V} \)

\[ V_{SS} = V_1 = 0 \text{V} < V_T \Rightarrow \text{N莫斯 is in cutoff} \]

\[ V_{SS} = V_T - V_{DS} = V_1 - V_{DD} = -V_{DD} < V_T \]

\( \Rightarrow \) CMOS device is turned on.

Saturation or triode?

Because the CMOS device is on and the NMOS device is off, the output will change to \( V_{DD} \).

\[ V_{DS} = V_T - V_{DS} = V_0 - V_{DD} = V_{DS} = 0 \]

\[ V_{DS} = V_T - V_{DS} = V_0 - V_{DD} = 0 \]

\( \Rightarrow \) CMOS is in triode.

The PMOS current equation in triode is

\[ i_p = \frac{k_p}{V_T} \left( \left( V_{GS} - V_T \right) \frac{V_{DS}}{2} \right) \]

\[ \approx \frac{k_p}{V_T} \left( \left( V_{OS} - V_T \right) \frac{V_{DS}}{2} \right) \]

\[ V_{PD} = V_{DS} = i_p V_{DS} \]

\[ V_{PS} = -\frac{V_{DS}}{i_p} = \frac{-V_{DS}}{k_p \left( \left( V_{OS} - V_T \right) \frac{V_{DS}}{2} \right)} \]

\[ V_{DS} = \frac{1}{k_p \left( \left( V_{OS} - V_T \right) \frac{V_{DS}}{2} \right)} \]

For the case \( V_{DS} = -V_{DD} \) (and \( V_T < 0 \))

\[ r_{DS} = \frac{1}{k_p \left( V_{OS} + V_T \right)} = \frac{1}{k_p \left( V_{OS} - 1 - V_T \right)} \]

So for \( V_1 = 0 \) we have

\[ \frac{V_{DD}}{r_{DS}} \longrightarrow 0 \]

\( \frac{V_{DD}}{} \)
The second extreme case...

2) $v_1 = \text{logic one } = V_{DD}$

$v_{GSN} = V_{DD} > V_{TH} \Rightarrow \text{Nmos device is on}$

$v_{DS} = V_{DD} - V_{TH} = V_{DD} - V_{DD} = 0 > V_{TH}$

$\Rightarrow \text{Pmos device in cutoff}$

Saturation or triode?

Since the Pmos is off and the Nmos on, $v_0 \approx 0$

$v_{DS} = 0 < v_{GS} - V_T = V_{DD} - V_T$

$\Rightarrow \text{Nmos is in triode}$

$V_{DS} = \frac{v_{DS}}{I_D} = \frac{1}{V_{Tn}(V_{DD} - V_T)}$

So far $v_1 = V_{DD}$ we have

$V_{DD}$

$V_T$

$V_{DS} = 0$

Now we would like to develop the complete transfer characteristic:

For a typical CMOS process,

$V_{TH} = |V_{TP}|$

To have a symmetric transfer characteristic, we would like the FETs to have equal current-driving capability

$k_P^\prime = k_P^\prime, k_n^\prime = k_n^\prime$

Recall $k_P = \mu_P C_{OX}, k_n = \mu_n C_{OX}, \mu_P > 0.4 \mu_n$

$\Rightarrow k_P^\prime = 0.4 k_n^\prime$

because $C_{OX}$ is the same everywhere.

From these two cases, we deduce the following CMOS facts:

1) $V_{OL} = 0, \quad V_{OH} = V_{DD}\}$ maximum signal swing possible.

2) No static power dissipation.

3) Active pullup/pulldown - low resistance path at output.

4) Infinite input resistance ($\bar{i}_{in} = 0$). Each gate appears as only a capacitive load which does not affect static operation (but does affect dynamic performance).

5) Small area - no large resistors.
Usually the FETs in a digital circuit are designed with minimum $L$, to give the greatest current-driving capability possible for a given size FET.

If $L_p = L_n$ then $\frac{W_n}{W_p} = k_p' = \frac{1}{0.9} = 2.5$

: The PMOS devices are designed $2-3 \times$ wider than the NMOS devices.

With this in mind, our transfer characteristic becomes

$$V_0(V)$$

$$V_{ds} = V_{th}$$

$$\left| V_{ds} \right| = V_{th} = V_t$$

**Regions of Operation**

1) up to $A$ : $V_1 < V_t$

   $V_b = V_{on} = V_{bo}$

2) $A \rightarrow B$ : $V_1 < V_1 < \frac{V_{bo}}{2}$

   $V_b > \frac{V_{on}}{2} + V_t$

   **NMOS**

   $\begin{align*}
   V_{bs} &= V_1 = \frac{V_{bo}}{2} \\
   V_{bs} &= V_b = V_{bo} + V_t
   \end{align*}$

   At point $B$

   $\begin{align*}
   V_b - V_t &= \frac{V_{bo}}{2} \\
   V_{bs} &= \frac{V_{bo}}{2} = V_{bs} - V_t \\
   V_{bs} &= V_{bs} + V_t
   \end{align*}$

   $\Rightarrow$ $V_{bs} > V_{bs} - V_t$ $\Rightarrow$ saturation

   **PMOS**

   $\begin{align*}
   V_{bs} &= V_1 - V_{bo} = \frac{V_{bo}}{2} - V_{bo} = \frac{V_{bo}}{2} \\
   V_{bs} &= V_b = \frac{V_{bo}}{2} + V_t - V_{bo} = V_t - \frac{V_{bo}}{2} \\
   V_{bs} - V_t &= -V_{bo} \\
   V_{bs} &= V_{bs} + V_t
   \end{align*}$

   at point $B$

   $\begin{align*}
   V_b &= V_{bo} + V_t \\
   V_{bs} &= V_{bo} + V_t
   \end{align*}$

   Recall $V_t = \left| V_{tp} \right| = -V_{tp}$ since $V_{tp} < 0$

   $V_{bs} = V_{bs} - V_{tp}$

   $\Rightarrow$ saturation/triode boundary
3) \( B \rightarrow C: \quad v_3 = \frac{V_{bo}}{2} \)

both transistors in saturation

\[
\frac{V_{bo}}{2} - v_t < v_b < \frac{V_{bo}}{2} + v_t
\]

4) \( C \rightarrow D: \quad \text{NMOS in triode} \)

PMOS in saturation

(by symmetry with \( #3: \ A \rightarrow B \))

5) \( D \rightarrow \): \quad \text{NMOS in triode} \)

PMOS in cutoff

\[ v_b = V_{bo} = 0 \]

To calculate noise margins, we need to let \( \frac{\partial \gamma}{\partial v_y} = -1 \)

for \( v_y > \frac{V_{bo}}{2} \), between \( C \) and \( D \),

PMOS in saturation

\[ i_{pp} = \frac{1}{2} k_p \left( \frac{\gamma}{\gamma} \right) p \left( v_3 - V_{bo} - v_t \right) \]

NMOS in triode

\[ i_{pn} = k_m \left( \frac{\gamma}{\gamma} \right) m \left( v_3 - V_{in} \right) v_0 - \frac{1}{2} v_0^2 \]

Let \( k_m \left( \frac{\gamma}{\gamma} \right) m = k_p \left( \frac{\gamma}{\gamma} \right) p \)

matched devices

\[ v_{in} = [v_{in} = v_t] \quad \text{devices} \]

From our circuit we have \( i_{pp} = i_{in} \)

\[
\frac{\partial i_{pp}}{\partial v_y} = \frac{\partial i_{in}}{\partial v_y}
\]

\[
\frac{\partial i_{pp}}{\partial v_y} = \frac{1}{2} k_p \left( \frac{\gamma}{\gamma} \right) p \left( v_3 - V_{bo} - v_t \right) = \frac{k_p}{p} \left( v_3 - V_{bo} - v_t \right)
\]

\[
\frac{\partial i_{in}}{\partial v_y} = k_m \left( \frac{\gamma}{\gamma} \right) m \left( (v_3 - V_{in}) \frac{\partial v_0}{\partial v_y} + v_0 - \frac{1}{2} v_0 \frac{\partial v_y}{\partial v_y} \right)
\]

\[ v_3 - V_{bo} + v_t = v_0 + (v_{in} - V_t - v_0) \frac{\partial v_0}{\partial v_y} + v_0 \]

But \( \frac{\partial v_0}{\partial v_y} = -1 \quad \text{and} \quad v_y = V_{IH} \)

\[ v_{IH} - V_{bo} + v_t = v_0 + (v_{IH} - V_t - v_0)(-1) \]

\[ v_0 = v_{IH} + v_t + v_0 \]

\[ 2 v_0 = 2 v_{IH} - V_{bo} \]

\[ v_{bo} = v_{IH} - \frac{V_{bo}}{2} \]

Substitute for \( v_3 \) and \( v_0 \) into \( i_{pp} = i_{in} \)

\[
\frac{1}{2} \left( (v_3 - V_{bo} + v_t) \right)^2 = \left( (v_{in} - V_t) v_0 - \frac{1}{2} v_0^2 \right)
\]

\[
\frac{1}{2} \left( v_{ih} - v_{bo} + v_t | v_{ih} - v_{bo} + v_t \right) = \left( (v_{ih} - v_t) (v_{ih} - v_{bo}) - \frac{1}{2} (v_{ih} - \frac{V_{bo}}{2})^2 \right)
\]

\[
\frac{1}{2} \left( v_{ih}^2 - v_{ih} v_{bo} - v_{ih} v_t - v_{ih} v_{bo} - v_{ih} v_t - v_{ih} v_{bo} v_t + v_{ih} v_t + v_{ih} v_{bo} + v_{ih} v_t \right)
\]

\[
= \frac{1}{2} \left[ (v_{ih}^2 - 2 v_{ih} v_{bo} - v_{ih} v_t - v_{ih} v_t - v_{ih} v_{bo} + v_{ih} v_t + v_{ih} v_{bo} + v_{ih} v_t) \right] \quad \text{Left Side}
\]

\[
= \frac{1}{2} \left[ v_{ih}^2 - 2 v_{ih} v_{bo} v_t - v_{ih} v_t - v_{ih} v_{bo} + v_{ih} v_t + v_{ih} v_{bo} - v_{ih} v_t + v_{ih} v_{bo} + v_{ih} v_t \right] \quad \text{Left Side}
\]

\[
= \frac{1}{2} \left[ v_{ih}^2 - 2 v_{ih} v_{bo} - v_{ih} v_t - v_{ih} v_t - v_{ih} v_{bo} + v_{ih} v_t + v_{ih} v_{bo} + v_{ih} v_t \right] \quad \text{Right Side}
\]
\[
\begin{align*}
= \frac{1}{4} V_{IH}^2 - V_{IH} V_L + \frac{1}{4} V_{IH} V_L + \frac{1}{2} V_{IH} V_L - \frac{1}{2} V_{IH}^2 & \quad \text{Right Side} \\
- V_{IH} V_{OH}^2 + 2 V_{IH} V_L - \frac{1}{2} V_{IH}^2 + \frac{1}{2} V_{OH}^2 + \frac{1}{2} V_L^2 & = 0 & \quad \text{Move right side to left side and combine} \\
\frac{5 V_{OH}^2 - 2 V_{IH} V_L + 2 V_L^2}{2 (V_{OH} - 2 V_L)} & = V_{IH} \\
\frac{(5 V_{OH} - 2 V_L) (V_{OH} - 2 V_L)}{2 (V_{OH} - 2 V_L)} & = V_{IH} \\
V_{IH} & = \frac{5}{8} V_{OH} - \frac{1}{4} V_L & \quad \text{(Whew!)}
\end{align*}
\]

Since our transfer characteristic is symmetrical about \( V_I = \frac{1}{2} V_{OH} \),
\[
V_{IH} = \frac{1}{2} V_{OH} = \frac{1}{2} V_{OH} - V_{IL} \\
V_{IL} = V_{OH} - V_{IH} \\
V_{IL} = \frac{3}{8} V_{OH} + \frac{1}{4} V_L
\]

Noise margins
\[
NM_H = V_{OH} - V_{IH} = V_{OH} - (\frac{5}{8} V_{OH} - \frac{1}{4} V_L) \\
NM_H = \frac{3}{8} V_{OH} + \frac{1}{4} V_L \\
NM_L = V_{IL} - V_{OH} = V_{IL} - 0 \\
NM_L = \frac{3}{8} V_{OH} + \frac{1}{4} V_L
\]

---

Power dissipation

We've seen that the CMOS inverter has zero static power dissipation, when the input is at \( V_{OH} \) or \( V_{OL} \).

When the input is changing, there are two components of dynamic power dissipation:

a) shoot through
b) cut-off

A) shoot through occurs when both FETs are turned on during the transition.

![Diagram](image)

The maximum shoot-through current occurs when both FETs are saturated.

Shoot through can be minimized by keeping the transition time short - by maintaining short rise times and fall times.
At $t = 0^+$, $v_b = 0$, and the resistor

$\frac{dv}{dt} = 0$

stores no energy.

At $t = 0^+$, $v_b = 0$, turning the MOS device off and the power device on, causing a current to flow through $C$.

Drive the input with a square wave with period $T$.

The average current from $v_b$ is

$\bar{I} = \frac{Q}{C}$

So the power is

$P_b = \bar{I} v_b = \frac{Q v_b}{C}$

Note: compare this to a resistor, where $P = \frac{V^2}{R}$.

$L = \frac{C^2}{R}$

Switched capacitor:

An effective resistance.

$P = \frac{V^2}{R}$

For $0 < t < \frac{T}{2}$ a charge $Q = CV_b$ flows from

the supply onto $C$.
Dynamic operation of the CMOS inverter

a) rise time ($t_r$) and fall time ($t_f$)

$t_r$ and $t_f$ are specified as the time it takes the output to change between the 10% and 90% points of the output swing $V_{OH} - V_{OL}$.

b) propagation delay

$\tau_{PHL}$: high to low

$\tau_{PLH}$: low to high

Propagation delay is measured between the point where the input is at 50% of the signal swing and the point where the output is at 50%.

When calculating propagation delay we will assume that the input voltage has zero rise and fall times.
CMOS Inverter

Matched MOSFETs:

\[ V_{in} = -V_{tp} = V_t \]

\[ k_p'(\frac{W}{L})_n = k_p'(\frac{W}{L})_p \]

\[ \therefore t_{PHL} = t_{PLH} \]

Q: What happens when \( V_2(0^-) = V_{OL} \) becomes \( V_2(0^+) = V_{OH} \) ?

A: \( V_2(0^+) = V_{OH} \) (capacitor charged to \( V_{OH} = V_{DD} \)).

At \( t = 0 \) the PMOS device turns off.

We have:

\[ V_{OH} = \frac{\int_{t_1}^{t_2} k_{PM} C V_2(t) \, dt}{C} \quad \therefore \quad V_2(0) = V_{DD} \]

Initially

\[ V_{BSW} = V_{DD} \]

\[ V_{BSN} = V_{SS} \]

\[ \Rightarrow V_{SW} > V_{BSN} - V_t \quad \text{NMOS is saturated} \]

FET IV characteristic

A) At \( t = 0^- \) FET is off

B) At \( t = 0 \) FET goes into saturation region

\[ V_{DS} = V_1 = V_{DD} \]

C) Capacitor discharges until

\[ V_2 = V_{DD} - V_t \quad \text{FET enters triode region.} \]

D) Output voltage reaches 50% of signal swing.

Point to measure \( t_{PHL} \).

E) Output reaches \( V_{OL} \)
Integration from point C to point D:

\[ v_0(t_c) = v_{eb} - v_t \]

\[ v_0(t_p) = \frac{1}{2} v_{eb} \]

\[ t_c - t_e = \tau_{ph2} \]

\[-\frac{k_n'(\psi)}{C} (v_{eb} - v_t) \int_{t_c}^{t_p} dt = \left( \frac{1}{2} v_{eb} - \frac{dv_e}{v_{eb} - v_t} \right) \]

\[ \int \frac{dv_e}{ax^2 - x} = \mu_e \left( \frac{1}{a} \right) \]

Check:

\[ \frac{d}{dx} \ln \left( 1 - \frac{1}{ax} \right) = \frac{a}{1 - \frac{1}{ax}} \frac{d}{dx} \left( 1 - \frac{1}{ax} \right) \]

\[ = \frac{1}{ax} \frac{1}{ax^2} = \frac{1}{ax^2 - x} \]

\[ + \frac{k_n'(\psi)}{C} (v_{eb} - v_t) \tau_{ph2} = \mu_e \left( 1 - \frac{2(v_e - v_t)}{v_{eb} - v_t} \right) \]

\[ = \mu_e \left( 1 - \frac{2(v_e - v_t)}{v_{eb} - v_t} \right) - \mu_e \left( 1 - \frac{2(v_{eb} - v_t)}{v_{eb} - v_t} \right) \]

\[ = \mu_e \left( \frac{v_{eb} - 2v_e + 2v_t}{v_{eb} - v_t} \right) \]

\[ = \mu_e \left( \frac{3v_{eb} - 4v_e + 2v_t}{v_{eb} - v_t} - \frac{v_{eb} - v_t}{2v_e} \right) \]
\[ t_{PHL2} = \frac{C}{k'(\xi)v_{TH}v_{DD}} \left[ - \frac{\frac{3v_{DD}-4v_I}{v_{DD}}}{v_{DD} - v_I} \right] \]

Finally, add the two components

\[ t_{PHL} = t_{PHL1} + t_{PHL2} \]

\[ t_{PHL} = \frac{C}{k'(\xi)} \left[ \frac{2v_I}{v_{DD} - v_I} + \ln \left( \frac{3v_{DD} - 4v_I}{v_{DD}} \right) \right] \]

For most 5V CMOS processes, \( v_I = 1V \)

Let \( v_I = 0.2v_{DD} \)

\[ t_{PHL} = \frac{C}{k'(\xi)} \left( \frac{0.4v_{DD}}{v_{DD} - 0.2v_{DD}} + \ln \left( \frac{3v_{DD} - 0.8v_{DD}}{v_{DD}} \right) \right) \]

\[ = \frac{C}{k'(\xi)} \left( \frac{1}{0.8} + \ln \left( \frac{2.2}{0.8} \right) \right) \]

\[ t_{PHL} = \frac{1.6C}{k'(\xi)v_{DD}} \]

By symmetry,

\[ t_{QHL} = \frac{1.6C}{k'(\xi)v_{DD}} \]

The overall propagation delay

\[ t_p = \frac{t_{PHL} + t_{QHL}}{2} \]

1) \( t_p \sim C \) \( \Rightarrow \) reduce capacitance
   - Use minimum \( L \) to keep \( C_{ox}WL \) small
   - Minimize wiring and other parasitic capacitances

2) \( t_p \sim \frac{1}{k'} \) \( \Rightarrow \) increase \( k' \)
   - Be careful! \( k' = \mu C_{ox} \) so \( C \) increases with \( k' \)

3) \( t_p \sim \frac{1}{(\xi)^2} \) \( \Rightarrow \) increase \( W \)
   - Again, \( C \) includes \( C_{ox}WL \) ... but this helps
     if \( C \) is dominated by other terms than \( C_{ox} \)

4) \( t_p \sim \frac{1}{v_{DD}} \) \( \Rightarrow \) increase \( v_{DD} \)
   - The process might not tolerate large \( v_{DD} \),
     and remember \( v_p = Cv_{DD}f \)!
CMOS Logic Gates

1) Inverter \( Y = \overline{A} \)

<table>
<thead>
<tr>
<th>A</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Note that we get \( Y = 1 \) when the PMOS device is on and we get \( Y = 0 \) when the NMOS device is on.

2) NAND gate (NOT AND)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

De Morgan’s theorems:

\( \overline{A \cdot B} = \overline{A} + \overline{B} \)
\( \overline{A + B} = \overline{A} \cdot \overline{B} \)

\( Y = \overline{A \cdot B} = \overline{A} + \overline{B} \)

We need a circuit that will do the following:

a) \( Y = 0 \) if \( A = 1 \) and \( B = 1 \)

b) \( Y = 1 \) if either \( A = 0 \) or \( B = 0 \)

Consider this NMOS circuit:

This fulfills the requirements, but is ‘wussy’ because of the resistor.

Consider this PMOS circuit:

This also works, but uses a resistor.

Consider this CMOS circuit:
### 3) NOR gate (NOT OR)

\[
\begin{array}{cc|c}
A & B & Y \\
0 & 0 & 1 \\
0 & 1 & 0 \\
1 & 0 & 0 \\
1 & 1 & 0 \\
\end{array}
\]

This accomplishes our 2-input NAND gate.

\[
Y = A + B = \overline{A \cdot B}
\]

Note:
- This idea can be extended to more inputs.
- Be careful not to design a circuit that has a path from VDD to ground for some combination of inputs.
- The circuit should always have a path to either VDD or ground - otherwise there is no driving force for the output logic level.
- All combinational CMOS logic gates use parallel and series transistors to implement the pull-up and pull-down networks.
CMOS Logic Gates

We've seen how to build

- inverters
- NAND gates
- NOR gates

This is sufficient to build any combinational logic circuit. However, we can build more efficient circuits...

Let's look at the pull-down network:

1)  \[ Y = \overline{A} \]

Since the PDN is made of an NMOS device, the output is low when the input is high.

2)  \[ Y = A \lor B \]

Parallel devices give the OR function.

3)  \[ Y = A \land B \]

Series devices give the AND function.

- When \( Y(A, B, C, \ldots) = 0 \), PDN conducts and PDU is off.
- When \( Y(A, B, C, \ldots) = 1 \), PDN is off and PDU conducts.

At no time should both the PDN and PDU be on, as this would allow a dc current path from \( V_{DD} \) to ground.
1) \[ Y = \overline{A} \]

2) \[ Y = \overline{A} + \overline{B} \]

3) \[ Y = A \cdot B \]

4) \[ Y = \overline{A} + \overline{B} \]

Pull-up network:

With the PMOS device in the pull, it is more natural to think of \( Y \) (active high) as a function of the complemented (active low) inputs.

Parallel devices give the OR function.

The NMOS input is active high — applying \( V_{dd} \) to the gate turns the device on — whereas the PMOS is active low — the device is on when the gate is pulled to ground.

In digital design, the following symbols are often used:

\[ \text{NMOS} \]
\[ \text{PMOS} \]

Remember — source terminals to \( V_{dd} \) and ground!
From our examples of PDNs and PUNs:

- PDN: \( \bar{Y} = f(A, B, C, \ldots) \)
- PUN: \( Y = f(A, B, C, \ldots) \)

We want to understand how to synthesize the PDN and PUN from a logic function.

1) 2-input NOR gate \( Y = \overline{A + B} = A \cdot \overline{B} \)

This is the natural form for the PUN:

![Diagram of a 2-input NOR gate](image)

Note that the PDN and PUN are dual networks:

<table>
<thead>
<tr>
<th>NMOS PDN</th>
<th>PMOS PUN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Series (AND)</td>
<td>Parallel (OS)</td>
</tr>
<tr>
<td>Parallel (OS)</td>
<td>Series (AND)</td>
</tr>
</tbody>
</table>
3) **exclusive OR (XOR)**

\[
\begin{array}{ccc}
A & B & \overline{Y} \\
0 & 0 & 0 \\
0 & 1 & 1 \\
1 & 0 & 1 \\
1 & 1 & 0 \\
\end{array}
\]

\[Y = \overline{AB} + AB\]

\[\overline{Y} = \overline{AB} + AB\]

Note that we don't have \(Y = f(A, B)\) or \(\overline{Y} = f(A, B)\).

\[\vdash \text{we will need to use inverters to create } \overline{A} \text{ and } \overline{B}\]

**PUN:** \[Y = \overline{AB} + AB\]

\[\overline{A} \rightarrow \overline{B} \rightarrow \overline{A} \rightarrow \overline{B} \rightarrow \overline{Y}\]

**PUN:** \[\overline{Y} = \overline{AB} + AB\]

---

A few notes about the XOR implementation:

1. The PUN and PDN are not dual networks.

![Diagram of XOR implementation](image)

The PDN is

\[\overline{Y} = (A + \overline{B})(\overline{A} + \overline{B})\]

\[= A\overline{A} + AB + \overline{B}\overline{A} + \overline{B}B\]

\[= \overline{A} + AB + \overline{B}A + \overline{B}B\]

\[\overline{Y} = AB + \overline{AB}\]

1. The XOR implementation requires 12 transistors:
   - 8 as shown above
   - 2 inverters \((Y = \overline{A})\)
   - 2 inverters \((Y = \overline{B})\)
   - 12
Transistor Sizing

1. Inverter

\[ Y = \overline{A} = \overline{B} = \overline{C} \]

2 transistors

\[ Y = \overline{A + B} = \overline{C} \]

3 transistors

\[ Y = (A + B) \cdot C \]

4 transistors

Finding the minimum implementation results in a smaller, cheaper circuit.

Summary of Logic Synthesis:

- PDA - use \( Y = f(A, B, C, \ldots) \) without extra inverters.
- PDA - use \( Y = \overline{f(A, B, C, \ldots)} \)
- Use duality to create one from the other.
- Both PDA & PDA should never be on simultaneously.

We want the worst-case (weakest - lowest current) output to be the same strength as the inverter.

How do we size the devices?
Recall that in the trade region for small $v_{DS}$

$$
R_{DS} = \frac{1}{w} \frac{1}{(V_{GS}-V_{TH})} \propto \frac{1}{(\%L)} = \frac{A}{(\%L)}
$$

- **FETs in series**

$$
R_{P} = R_{DS1} + R_{DS2} + \ldots
$$

$$
= \frac{A}{(w/L)_{1}} + \frac{A}{(w/L)_{2}} + \ldots \quad \left( A \text{ is constant} \right)
$$

$$
R_{S} = \frac{A}{(w/L)_{eq}}
$$

$$
\therefore \quad \frac{1}{(w/L)_{eq}} = \frac{1}{(w/L)_{1}} + \frac{1}{(w/L)_{2}} + \ldots
$$

- **FETs in parallel**

$$
\frac{1}{R_{P}} = \frac{1}{R_{DS1}} + \frac{1}{R_{DS2}} + \ldots
$$

$$
\frac{(w/L)_{eq}}{A} = \frac{(w/L)_{1}}{A} + \frac{(w/L)_{2}}{A} + \ldots
$$

$$
(w/L)_{eq} = (w/L)_{1} + (w/L)_{2} + \ldots
$$

The bottom line:

Two identical FETs in series act like a single FET with the same width but twice the length.

Two identical FETs in parallel act like a single FET with twice the width but the same length.

---

So for the NOR gate:

Two PMOS in series - each needs to be twice as wide as the inverter.

Two NMOS in parallel - worst case only one is on, so each needs to be equal to the inverter's NMOS.

Note that $p \approx 2.5n$ so the NOR gate requires really big PMOS devices.

The NAND gate requires bigger NMOS devices so the overall gate size is kept smaller and more balanced.

$\therefore$ NAND gates are preferred.
Example: Given a logic family with the following basic inverter

\[ \begin{align*}
(\bar{Y})_P &= \frac{a}{2} \\
(\bar{Y})_N &= \frac{d}{2}
\end{align*} \]

(a) Synthesize the complex function \( Y = \overline{A + B(C + D)} \) and (b) size the transistors.

(a) \[ Y = f(A, B, C, D) \]

Recall \( OR \) or \( \oplus \) parallel and \( \& \) series

(b) size transistors

\[ Y = \overline{A \cdot B} \quad \text{series} \implies \frac{1}{(\%)_\text{av}} = \frac{1}{(\%)_A} + \ldots \]

\[ \frac{1}{(\%)_\text{av}} = \frac{1}{(\%)_B} = \frac{1}{(\%)_A} + \frac{1}{(\%)_C} + \frac{1}{(\%)_D} \]

Let \( (\%)_A = (\%)_B = (\%)_C = (\%)_D \)

\[ \frac{1}{(\%)} = \frac{3}{(\%)_A} \implies (\%)_A = \frac{\text{AV}}{3} \]

This makes sense - 3 devices in series must be 3x wider to match one equivalent device.

Path 1

\[ Y = \overline{A \cdot B} \quad \text{series} \]

\[ \frac{1}{(\%)_\text{av}} = \frac{1}{(\%)_A} = \frac{1}{(\%)_B} = \frac{1}{(\%)_C} + \frac{1}{(\%)_D} \]

\[ \frac{1}{(\%)_\text{av}} = \frac{1}{(\%)_A} = \frac{1}{(\%)_B} = \frac{1}{(\%)_C} + \frac{1}{(\%)_D} \]

\[ (\%)_A = \frac{15}{2} \]
Pass-transistor logic

The concept: instead of using the transistor switches to pull-up or pull-down, use them to create a path from input to output.

Example:

\[
\begin{array}{cccc}
A & B & C & Y \\
0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 \\
1 & 1 & 1 & 1 \\
\end{array}
\]

\[Y = A \cdot B \cdot C\]

Example:

\[
\begin{array}{ccc}
A & B & Y \\
1 & 0 & 0 \\
1 & 0 & 1 \\
1 & 1 & 1 \\
\end{array}
\]

\[Y = A \cdot B + A \cdot C = A(B+C)\]

The (transistor) switches pass the input signal \(A\) to the output \(Y\) — hence the name pass-transistor logic (PTL).
It is important that all nodes have some low-impedance path to $V_{DD}$ or ground at all times! Consider:

\[
A \rightarrow \square \rightarrow B \rightarrow \square \rightarrow Y = \overline{A \cdot B} = A + B
\]

When $B=1$, we have no problem:

\[
A \rightarrow \square \rightarrow \square \rightarrow Y = A
\]

When $B=0$ we have a floating gate:

\[
A \rightarrow \square \rightarrow \square \rightarrow Y
\]

Because of the parasitic capacitance on the gate, the output might be ok for awhile. Over time $V_a$ can drift in and out of the region where one or both devices is on, however, leading to uncertain values of $Y$ and static power dissipation.

Solution:

<table>
<thead>
<tr>
<th></th>
<th>$A$</th>
<th>$B$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

```
::: Y = A + \overline{B}
```

Some logic function

NMOS transistor switches

Let $v_b = V_{DD}$, $v_y(t=0) = 0$

\[
v_a(t) = V_{DD} u(t) \Rightarrow
\]

We expect $v_Y(t)$ to increase
Initially, \( v_f(t) = 0 \) ⇒ transistor in saturation mode

\[
I_g = \frac{1}{2} V_{dd} \left( \frac{V_{GS} - V_t}{L} \right)^2
\]

\[
V_{GS} = V_t - V_f = V_{dd} - V_f
\]

\[
i_g = \frac{1}{2} V_{dd} \left( \frac{V_{dd} - V_f}{L} \right)^2
\]

There are two problems:

1) when \( v_f = V_{dd} - V_t \)

\[
V_{GS} = V_t - V_f = V_{dd} - (V_{dd} - V_t) = V_t
\]

Once output reaches \( V_{dd} - V_t \), the FET turns off - the output will not reach \( V_{dd} \)

2) what is \( V_t \)?

The source voltage is equal to the output voltage.

The body voltage is 0V.

\[
V_{SB} = V_f = 0 = V_{dd} - V_t
\]

Recall the body effect increases \( V_t \)

\[
V_t = V_{dd} + \frac{L}{2} \left( \frac{V_{dd} + 2V_f}{V_{dd} - V_f} \right)
\]

This can add an extra voltage, \( V_{dd} \) to \( V_f \)

So - the output does not reach \( V_{dd} \), and might be quite a bit lower (1.2V).

- poor noise margin
- following gate might have PMOS and PMOS both conducting
- long propagation delays

What about when \( V_f = 0 \)?

\[
\begin{align*}
V_f &= 0 \\
V_{GS} &= 0 \\
V_{GS} &= V_{dd} \\
Y &= V_{dd}
\end{align*}
\]

Let \( v_f(t=0) = V_{dd} \)

Now when \( V_f = 0 \), the left node acts like the source

- body voltage = source voltage
- device pulls \( v_f \) to ground

Summary of NMOS pass device

The transistor has a hard time pulling the output high: it has a "poor 1." However, it easily pulls the output low: it has a "good 0."

What to do? Think complementary!

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMOS</td>
<td>good</td>
<td>poor</td>
<td></td>
</tr>
<tr>
<td>PMOS</td>
<td>poor</td>
<td>good</td>
<td></td>
</tr>
</tbody>
</table>

Where \( V_f = 0 \)

\[
\begin{align*}
V_f &= 0 \\
V_{GS} &= V_{dd} \\
V_{GS} &= V_{dd} \\
V_f &= V_{dd}
\end{align*}
\]
Note that we've added some complexity
- extra pass device
- normal and complemented input needed - may require an extra inverter.

Example: 2:1 multiplexer (MUX)

```
S   MNA   MPA   MNB   MPB   Y
0    on   on    off   off   A
1    off   off    on    on    B
```

Example: What does this do?

![Diagram of a circuit]

```
Y = \overline{A} \cdot B
```

(transistor count = 10
(but A and B might be available without extra inverters)
Latches and Flip Flops

- Combinational logic: \( Y = f(A, B, C, \ldots) \)
  - Output depends only on present values of inputs

- Sequential logic: \( Y = f(A, B, C, \ldots, A', B', C', \ldots) \)
  - Output depends on current and previous input values; the circuit incorporates memory

Sequential logic requires some sort of timing input – a clock.

Memory can be of two types:

- Static - uses a circuit with two stable states ( bistable ) = 0 or 1
- Dynamic - uses charge stored on a capacitor, but the charge must be refreshed periodically

The Latch: cross-coupled inverters

\[
\begin{align*}
W &= 1 & W &= 0 \\
X &= 0 & X &= 1 \\
Y &= 0 & Y &= 1 \\
Z &= 1 & Z &= 0
\end{align*}
\]

This can be easier to see by "untying" the loop:

Now let's break the loop to find the transfer characteristic:
There are three operating points for the circuit - A, B, C.

Point B is unstable because of the gain of the inverters:

\[ G = \frac{\partial v_x}{\partial v_w} \rightarrow \infty \text{ for CMOS} \]

So if our two-inverter loop is operating at point B, a small amount of noise will send it to A or C.

At the end points (L & N) the gain is

\[ G = \frac{\partial v_x}{\partial v_w} \rightarrow 0 \]

So the two-inverter loop will want to stay at points A or C.

The latch is a bistable circuit - it operates at either point A or point C on the transfer characteristic.

The latch can be used as a single bit of memory by assigning one point to be logic 0 and the other point to be logic 1.

Complementary outputs are available.

The latch is static.
Implementation 1: The SR Flip Flop

Note that for a NOR gate...

<table>
<thead>
<tr>
<th>A</th>
<th>R</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

... if one input is logic 0 the gate appears like an inverter.

So if $S=R=0$ the circuit above is our old friend the latch.

<table>
<thead>
<tr>
<th>R</th>
<th>S</th>
<th>$Q_{n+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>$Q_n$</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>not used</td>
</tr>
</tbody>
</table>

This can also be done with NAND gates:

Implementation 2: Clocked SR Flip Flop
Let $Q = 0$ (\(\bar{Q} = 1\))

Next let $S = 1$, $R = 0$

Next let $\phi = 1$

The two NMOS devices pull $\bar{Q}$ low
Inverter $G_1$ pulls $Q$ high
Inverter $G_2$ pulls $\bar{Q}$ low (regenerates $Q$)

\[ \Rightarrow \] the latch is set: $Q = 1$ (\(\bar{Q} = 0\))

even after $\phi = 0$ or $S = 0$

There is an important design problem that must be solved:

\[ Q \quad \text{M}_2 \quad \text{M}_3 \quad \text{M}_1 \quad \text{M}_6 \quad \phi \quad \text{S} \quad \text{M}_5 \quad \text{M}_7 \quad \text{R} \quad \bar{Q} \quad G_1 \quad G_2 \]

Let $Q = 0$ (\(\bar{Q} = 1\))

then $M_2$ and $M_3$ are turned ON

If $\phi = 1$ and $S = 1$ both $M_5$ and $M_6$ turn on

but to set the latch they must be able to overcome the pullup of $M_2$.

\[ \therefore \text{ Need to adequately size } \left(\frac{1}{C}\right) \text{ of } M_5 \text{ and } M_6. \]

---

A few transistors can be saved by using pass-transistor logic (PTL)

This is the famous 6-transistor cell that makes up most static memory (SRAM).

---

The D Flip Flop

\[ D \quad \text{\longrightarrow} \quad \bar{Q} \quad \phi \quad \bar{\bar{Q}} \]

We want an implementation that is edge-triggered:
the flip flop acquires the value of $D$ immediately prior to $\phi = 1$; changes in $D$ when $\phi = 1$
do not affect the stored state.
Implementation #1

Note: The NMOS switches are probably implemented with CMOS transmission gates.

If both switches close together, the D input is shorted to the closed memory loop. To avoid this, use nonoverlapping clocks:

A problem with this implementation is that the output follows the input when $\phi_1$. This isn’t what we wanted— we wanted an edge-triggered gate.

Implementation #2: Master-Slave Flip-Flop

A: $S_1$ closed, $S_2$ open
D passes through to $Q$

B: $S_1$ open, $S_2$ closed
$Q$ regenerates through the two inverters

$\phi_1$ and $\phi_2$ are nonoverlapping clocks.

Slave is clocked on opposite phase from master
Master-Slave Operation:

$\phi_1$: slave latch is isolated
- master latch open, connected to $D$

$\phi_2$: master latch isolated from $D$
- slave latch open, connected to output of master

An additional note on nonoverlapping clocks:

$\phi_1$ [Diagram]

$\phi_2$ [Diagram]

A: $S_1$ open, $S_2$ closed (latched)

B: both open - rely on input capacitance of
- first inverter to hold state

C: $S_1$ closed, $S_2$ open $\Rightarrow$ $A=D$, input cap charged to $D$

D: both open - same as B

Conclusion:
The nonoverlap time must be kept relatively short

Memory is used for storing information - computer data
and program instructions, music, books, ...
- mass storage - magnetic, optical
- main memory - electronic

Random-Access Memory (RAM)
- time required to write (store) and read (retrieve)
  is independent of location
- relatively fast

Sequential (Sequential) Memory
- access time depends upon location
- slower
- can be physical, like a tape, or electronic

Read/Write vs. Read-Only Memory (ROM)
- read/write permits storage and retrieval
- ROM only allows retrieval

Note that computer ROM is random access!
Each bit is stored in a memory cell. Each bit can be individually addressed or addressed in a group (an architectural choice of the designer).

Example: 64M-bit chip.

\[
\begin{align*}
M &= 2^k \times 1k = (1024)^2 = 1048576 \\
64M &= 64 \times 1048576 = 67,108,864 \text{ bits}
\end{align*}
\]

a) 64M words \times 1 bit

requires 26 address lines \((2^{26} = 67,108,864)\)

b) 16M words \times 4 bits

requires 24 address lines \((2^{24} = 16,777,216)\)

---

**Row decoder**: A combinational circuit that takes \(M\) inputs (a binary number) and selects one of \(2^M\) outputs.

Example: \(M = 2\) \(\Rightarrow A_0, A_1\) inputs

\[
\begin{array}{c|cccc}
A_0 & A_1 & Y_0 & Y_1 & Y_2 & Y_3 \\
\hline
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 \\
\end{array}
\]
You can implement this with NOR gates (and inverters).

\[ \overline{A_0} \rightarrow \overline{Y_0} \]
\[ \overline{A_1} \rightarrow \overline{Y_1} \]
\[ \overline{A_0} \rightarrow \overline{Y_2} \]
\[ \overline{A_1} \rightarrow \overline{Y_3} \]

There is a more compact method we'll look at later.

The row decoder activates the word line, enabling the \( n \) storage cells located along it. Each cell puts its contents on its respective bit line. The signal on the bit line can be very small (0.1-0.2 V).

**Sense Amplifier:** one for each bit line (\( n \) total)

Buffers the small signal on the bit line up to full digital swing (0-\( V_{dd} \)).

**Column Decoder:** selects a single bit line to connect to the output.

Similar to the row decoder with transmission gate:

\[ \overline{A_0} \rightarrow \overline{Y_0} \]
\[ \overline{A_1} \rightarrow \overline{Y_1} \]

Storage Cell Array: \( M \times N \) array of memory cells

Typically arranged in a square or nearly-square matrix.

Examples:

1 M-bit array = 1024 x 1024

requires 10 row address lines
and 10 column address lines

Driver: When writing, the driver buffers the input data and stores it in the selected cell in the array.

As we shall see, the word lines usually drive MOS transistor gates and are fabricated using polysilicon (the same material as the gate).

from row decoder \[ \overline{I} \overline{L} \overline{L} \overline{L} \]

This is equivalent to a distributed RC network.

from row decoder \[ \overline{C} \overline{C} \overline{C} \overline{C} \overline{C} \]

This can really slow down access times because of the slow transient response time.
Finally, two memory-related terms:

- **Access time**: Time between starting to read and
  when data actually appears
- **Cycle time**: Smallest time allowed between two
  memory operations

  Typically a few ns to few hundred ns

These times don't necessarily decrease with smaller
die size.

a) **R** follows $R_s$ ($\Omega/A$)
   as $L$ decreases so does $W$

b) **C** follows $C_{ox}$
   as $V_{ox}$ decreases $C_{ox}(\text{cap/area})$ increases

---

**RAM Memory Cell - SRAM**

---

Consider a read operation on a cell with $Q=1$

1) **Precharge**: $E$ and $B$ are set to some equal
   voltage, usually $\frac{1}{2} VDD$

   - neutralizes old data on bit lines
   - helps sense amplifier read output data
   - relies on the bit line's capacitance; must be
     turned off before selecting the word line

2) **Select**: One word line is pulled to $VDD$, turning on
   all the access transistors connected to it

3) **Read**: The $Q=1$ output pulls $B$ high
   $Q=0$ output pulls $E$ low

   The sense amplifier detects the differential voltage
   between $B$ and $E$.

   Note the cell must be designed so the contents
   don't get accidentally written during the
   read operation.

Example: $Q=1 \Rightarrow M_1 \text{ on}, M_2 \text{ off}$

$Q=0 \Rightarrow M_1 \text{ on}, M_2 \text{ off}$

---

$C_B, C_E$ precharged to $\frac{1}{2}VDD$
Now consider a write operation into the cell (change from 1 to 0)

1) set bit lines: \( B = 0 \quad \bar{B} = V_{pp} \)

2) select word line (\( W = 1 \))

3) write

\[
\begin{align*}
\text{pulls } \bar{Q} \text{ to } V_{pp} \\
W = V_{pp} & \quad 1 \\
\bar{B} = V_{pp} & \quad \ldots \quad \bar{Q} & \quad \ldots \\
\end{align*}
\]

Because of the active driver circuitry which charges (discharges) the bit lines during a write, the delay time required for a write operation is much smaller than that required for a read operation.

**Dynamic Memory (DRAM)**

The\overline{ }\text{over} transistor cells:

1) only a single bit line is used

2) \( C_s \) can be charged to \( V_{pp} - V_t \) before the access transistor turns off

3) \( C_s \) will lose charge over time and thus need to be refreshed by a read/write operation (every 5-10 ms)

4) \( C_s \) can be discharged to ground (0V)
Let's look at a read operation.

a) precharge the bit line to \( \frac{1}{2} V_{DD} \)

![Diagram showing transistor state and capacitor connection]

- transistor is off
- \( V_{CB} = \frac{1}{2} V_{DD} \)
- \( C_B \gg C_S \)

b) select the word line

![Diagram showing transistor state and capacitor connection]

- transistor turns on
- \( C_S \) and \( C_B \) share charge

Change is conserved.

Before the switch closes we have

\[
Q_S = C_S V_{CS}
\]

\[
Q_B = C_B V_{CB} = \frac{1}{2} C_B V_{DD}
\]

After the switch closes

\[
Q_F = \left( \frac{C_B + C_S}{C_B + C_S} \right) V_F = \left( \frac{C_B + C_S}{C_B + C_S} \right) \left( \frac{1}{2} V_{DD} + \Delta V \right)
\]

where \( Q_F = Q_S + Q_B \)

\[
\left( \frac{C_B + C_S}{C_B + C_S} \right) \left( \frac{1}{2} V_{DD} + \Delta V \right) = C_S V_{CS} + \frac{1}{2} C_B V_{DD}
\]

\[
\Delta V = \frac{C_S V_{CS} + \frac{1}{2} C_B V_{DD} - \frac{1}{2} C_B V_{DD} - \frac{1}{2} C_S V_{DD}}{C_S + C_S}
\]

\[
\Delta V = \left( \frac{C_S}{C_B + C_S} \right) \left( V_{CS} - \frac{1}{2} V_{DD} \right)
\]

If the cell holds a logic 1, then \( V_{CS} = V_{DD} - V_t \)

\[
\Delta V_1 = \left( \frac{C_S}{C_B + C_S} \right) \left( \frac{1}{2} V_{DD} - V_t \right)
\]

If the cell holds a logic 0, then \( V_{CS} = 0 \)

\[
\Delta V_0 = \left( \frac{C_S}{C_B + C_S} \right) \left( -\frac{1}{2} V_{DD} \right)
\]
Let $C_S = 4.0 \text{ nF}$

$C_B = 1.5 \text{ pF}$

$V_{DD} = 3.3 \text{ V}$

$V_{I} = 0.6 \text{ V}$

\[ \Delta V_0 = \left( \frac{\text{4.0} \times 10^{-12}}{\text{40x10}^{-9} \times \text{4.5x10}^{-9}} \right) \left( -\frac{1}{2} \times 3.3 \right) = -43 \text{ mV} \]

\[ \Delta V_1 = \left( \frac{\text{4.0} \times 10^{-12}}{\text{40x10}^{-9} \times \text{4.5x10}^{-9}} \right) \left( \frac{1}{2} \times 3.3 - 0.6 \right) = +27 \text{ mV} \]

Recall that the bit line was precharged to

\[ V_{CB} = \frac{1}{2} V_{DD} = 1.65 \text{ V} \]

The voltage the sense amp sees will be

\[ V_{CB1} = 1.65 + 0.027 = 1.68 \text{ V} \]

\[ V_{CB0} = 1.65 - 0.043 = 1.61 \text{ V} \]

This is about 2% signal variation.

So far we've looked at SRAM and DRAM, which
are volatile — they lose their contents when power is removed.

A ROM (read-only memory) is used when we wish to keep its programmed state forever.

Let's say we wanted a memory to store a 4-digit number, using a BCD coding sequence.

<table>
<thead>
<tr>
<th>Decimal</th>
<th>BCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0000</td>
</tr>
<tr>
<td>1</td>
<td>0001</td>
</tr>
<tr>
<td>2</td>
<td>0010</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1001</td>
</tr>
</tbody>
</table>

we will need four bits of data out

To access four digits requires two address lines

\[ 2^2 = 4 \]
Example:

Note also that the ROM is simply a way to implement an arbitrary combinational logic function:

\[ Y = f(A, B, C, \ldots) \]

data output address lines

bits

This can be used to implement the row and column decoders.

Programmable memory: PROM and EPROM
- non-volatile
- can have data stored after the manufacturing process
- can be one-time programmable (OTP) - PROM
- can be erasable - EPROM

In the olden days of yore, a PROM might have been made with fuses:

Programming was accomplished by blowing the fuses.

Using resistors to pull up the bit lines is usually avoided. Instead a dynamic circuit can be used to precharge the bit lines.
Modern PROMs/EPROMs use a special floating-gate n-channel transistor:

In its unprogrammed state, the floating gate has no charge on it, and doesn’t affect the transistor’s operation.

Now put a large positive voltage on the gate and drain:

\[ V_g = +25\text{V} \]
\[ V_d = -15\text{V} \]

Electrons flowing from source to drain become very energetic due to the large applied voltages - hot electrons - and accelerate towards the gate. Electrons are trapped on the floating gate, forming a negative charge.

The negative charge offsets the applied gate voltage \( V_g \), effectively raising the threshold voltage.

Let the unprogrammed threshold be \( V_{t0} \) and the programmed threshold be \( V_{t1} \).

Apply \( V_w \) such that \( V_{t0} < V_w < V_{t1} \).

Unprogrammed - transistor turns on, \( B=0 \)
Programmed - transistor stays off, \( B=1 \)

In special applications multiple threshold voltages are possible.
The EPROM is erased by exposing it to ultraviolet (UV) light. The package has a clear window on top.

Note that OTP devices are probably made the same, just in an opaque package.