MOS Transistor Models
Electrical Effects
Propagation Delay

N-type Transistor
Another Cutaway View

Vgs Forms a Channel

Vt : The threshold voltage to turn on the transistor
MOS Capacitor

- Gate and body form MOS capacitor
- Operating modes
  - Accumulation
  - Depletion
  - Inversion

![Diagram of MOS Capacitor]

Transistor Characteristics

- Three conduction characteristics
  - Cutoff Region
    - No inversion layer in channel
    - $I_{ds} = 0$
  - Nonsaturated, or linear region
    - Weak inversion of the channel
    - $I_{ds}$ depends on $V_{gs}$ and $V_{ds}$
  - Saturated region
    - Strong inversion of channel
    - $I_{ds}$ is independent of $V_{ds}$
  - As an aside, at very high drain voltages:
    - “avalanche breakdown” or “punch through”
    - Gate has no control of $I_{ds}$…
nMOS Cutoff: $V_{gs} < V_t$

- No channel
- $I_{ds} = 0$

nMOS Linear: $V_{gs} > V_t$, small $V_{ds}$

- Channel forms
- Current flows from d to s
  - $e^{-}$ from s to d
- $I_{ds}$ increases with $V_{ds}$
- Similar to linear resistor
nMOS Saturation: \(V_{ds} > V_{gs} - V_t\)

- Channel pinches off
  - Conduction by drift because of positive drain voltage
  - Electrons are injected into depletion region
- \(I_{ds}\) independent of \(V_{ds}\)
- We say that the current saturates
- Similar to current source

```
Basic N-Type MOS Transistor

- Conditions for the regions of operation
  - Cutoff: If \(V_{gs} < V_t\), then \(I_{ds}\) is essentially 0
    - \(V_t\) is the “Threshold Voltage”
  - Linear: If \(V_{gs} > V_t\) and \(V_{ds} < (V_{gs} - V_t)\)
    then \(I_{ds}\) depends on both \(V_{gs}\) and \(V_{ds}\)
    - Channel becomes deeper as \(V_{gs}\) goes up
  - Saturated: If \(V_{gs} > V_t\) and \(V_{ds} > (V_{gs} - V_t)\) then \(I_{ds}\)
    is essentially constant (Saturated)```
Transistor Gain ($\beta$)

- $\beta = (\mu \varepsilon / t_{ox})(W/L)$
  - Process-dependent

  - $\mu = \text{mobility of carriers (cm}^2 / \text{V•s})$
    - Note that N-type is ~3X as good as P-type
  - $\varepsilon = \text{permittivity of gate insulator (oxide)}$
    - $\varepsilon = 3.9 \varepsilon_0$ for SiO$_2$ ($\varepsilon_0 = 8.85 \times 10^{-14}$ F/cm)
  - $t_{ox} = \text{thickness of gate insulator (oxide)}$
  - Also, $\varepsilon / t_{ox} = C_{ox}$ The oxide capacitance
    - $\beta = (\mu C_{ox})(W/L) = k'(W/L) = K_P(W/L)$
  - *Increase W/L to increase gain*

Example

- We will be using an old 0.5/0.6 $\mu$m process for your project
  - From ON Semiconductor
  - $t_{ox} = 100$ Å
  - $\mu = 350$ cm$^2$/V•s
  - $V_t = 0.7$ V

- Plot $I_{ds}$ vs. $V_{ds}$
  - $V_{gs} = 0, 1, 2, 3, 4, 5$
  - Use $W/L = 4/2 \lambda$

$$\beta = \mu C_{ox} \frac{W}{L} = (350) \left( \frac{3.9 \times 8.85 \times 10^{-14}}{100 \times 10^{-4}} \right) \left( \frac{W}{L} \right) = 120 \frac{W}{L} \mu A/V^2$$
Example

- We will be using an old 0.5/0.6 μm process for your project
  - From ON Semiconductor
  - $t_{ox} = 100$ Å
  - $\mu = 350$ cm$^2$/V*s
  - $V_t = 0.7$ V
  - $C_{ox} = \varepsilon/t_{ox}$
  - $\beta = \mu C_{ox} \frac{W}{L} = (350) \left( \frac{3.9 \cdot 8.85 \cdot 10^{-14}}{100 \cdot 10^{-8}} \right) \left( \frac{W}{L} \right) = 120 \frac{W}{L} \mu A/V^2$

“Saturated” Transistor

- In the $V_{ds} > (V_{gs} - V_t) > 0$ case
  - $I_{ds}$ Current is effectively constant
  - Channel is “pinched off” and conduction is accomplished by drift of carriers
  - Voltage across pinched off channel (i.e. $V_{ds}$) is fixed at $V_{gs} - V_t$
    - This is why you don’t use an N-type to pass 1’s!
    - High voltage is degraded by $V_t$
    - Depletion region is lost at $V_{ds} = (V_{gs} - V_t)$
Aside: N-type Pass Transistors

- If it weren’t for the threshold drop, N-type pass transistors (without the P-type transmission gate) would be nice
  - 2-way Mux Example…

N-type Pass Transistors

- Is this a good design using an nMOS pass transistor?
N-type Pass Transistors

- One one hand, the degraded high voltage from the pass transistor will be restored by the inverter
- On the other hand, the P-device may not turn off completely resulting in extra power being used

- One option is a “keeper” transistor fed back from the output
  - This pulls the internal node high when the output is 0
  - But is disconnected when output is high
- Make sure the size is right…(i.e. weak)
N-type Pass Transistors

- In practice, they are used fairly often, but be aware of what you’re doing
  - For example, read/write circuits in a Register File

Back to the Saturated Transistor

- What influences the constant $I_{ds}$ in the saturated case?
  - Channel length
  - Channel width
  - Threshold voltage $V_t$
  - Thickness of gate oxide
  - Dielectric constant of gate oxide
  - Carrier mobility $\mu$
  - Velocity Saturation
Back to the Saturated Transistor

- What influences the constant $I_{ds}$ in the saturated case?
  - *Channel length*
  - *Channel width*
  - Threshold voltage $V_t$
  - Thickness of gate oxide
  - Dielectric constant of gate oxide
  - Carrier mobility $\mu$
  - Velocity Saturation

Threshold Voltage: $V_t$

- The $V_{gs}$ voltage at which $I_{ds}$ is essentially 0
  - $V_t = .77v$ for nmos and -.92v for pmos in our process
  - Tiny $I_{ds}$ is exponentially related to $V_{gs}$, $V_{ds}$
  - Take 6770 & 6720 for “subthreshold” circuit ideas

- $V_t$ is affected by
  - Gate conductor material
  - Gate insulator material
  - Gate insulator thickness
  - Channel doping
  - Impurities at Si/insulator interface
  - Voltage between source and substrate ($V_{sb}$)
Basic DC Equations for \( I_{ds} \)

- **Cutoff Region**
  - \( V_{gs} < V_t \), \( I_{ds} = 0 \)

- **Linear Region**
  - \( 0 < V_{ds} < (V_{gs} - V_t) \)
  - \( I_{ds} = \beta [(V_{gs} - V_t)V_{ds} - V_{ds}^2/2] \)
  - Note that this is only “linear” if \( V_{ds}^2/2 \) is very small, i.e. \( V_{ds} \ll V_{gs} - V_t \)

- **Saturated Region**
  - \( 0 < (V_{gs} - V_t) < V_{ds} \)
  - \( I_{ds} = \beta [(V_{gs} - V_t)^2/2] \)

---

Ids Curves

\[ \beta = \frac{\mu_e}{t_{ox}} \left( \frac{W}{L} \right) = \mu C_{ox} \left( \frac{W}{L} \right) \]

Cutoff Region
\( V_{gs} < V_t \)
\( I_{ds} = 0 \)

Triode (Linear) Region
\( V_{gs} - V_t > V_{ds} > 0 \)
\( I_{ds} = \beta \left( (V_{gs} - V_t)V_{ds} - \frac{V_{ds}^2}{2} \right) \)

Saturation Region
\( V_{ds} > (V_{gs} - V_t) > 0 \)
\( I_{ds} = \beta \left( \frac{(V_{gs} - V_t)^2}{2} \right) \)
P-type Transistor

Poly Gate
Gate Oxide
Diffusion

N-doped substrate

-\( V_{gs} \)
\( G \)
\( S \)
\( D \)
\( + \)
\( - \)
holes

P-type Transistor

Source
Gate
Drain

Gate Oxide
Channel

Substrate
(Usually Vdd)
P-type Transistors

- Source is Vdd instead of GND
  - \( V_{sg} = (Vdd - Vin) \),
  - \( V_{sd} = (Vdd - Vout) \), \( V_t \) is negative
- Cutoff: \( (Vdd-Vin) < -V_t \), \( I_{ds} = 0 \)
- Linear Region
  - \( (Vdd-Vout) < (Vdd - Vin + V_t) \)
  - \( I_{ds} = \beta [(Vdd-Vin+V_t)(Vdd-Vout) - (Vdd-Vout)^2/2] \)
- Saturated Region
  - \( ((Vdd - Vin) + V_t) < (Vdd - Vout) \)
  - \( I_{ds} = \beta [(Vdd - Vin + V_t)^2/2] \)

Pass Transistor Ckts
Pass Transistor Ckts

\[ V_{DD} \quad V_s = V_{DD} - V_{in} \]

\[ V_{SS} \]

Pass Transistor Ckts

\[ V_{DD} \quad V_s = V_{DD} - V_{in} \]

\[ V_{SS} \]
Pass Transistor Ckts

\[ V_s = V_{DD} - V_{tn} \]

\[ V_s = |V_p| \]

\[ V_{SS} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]

\[ V_{DD} \]
2\textsuperscript{nd} Order Effects

- Quick introduction to effects that degrade the “digital” assumptions and models we have presented about our transistors so far.
  - This will be covered in more detail in Advanced VLSI 6770
  - You will learn how to relate them to designs
- Introductory material in this class
  - \textit{In a nutshell – nothing works as well as you think it should! 😊}

2\textsuperscript{nd} Order Effect: Velocity Saturation

- With weak fields, current increases linearly with lateral electric field
- At higher fields, carrier drift velocity rolls off and saturates
  - Due to carrier scattering
  - Result is less current than you think!
  - For a 2\(\mu\) channel length, effects start around 4v Vdd
  - For 180nm, effects start at 0.36v Vdd!
2nd Order Effect: Velocity Saturation

When the carriers reach their speed limit in silicon...

- Channel lengths have been scaled so that vertical and horizontal EM fields are large and interact with each other.

\[ V_{gs} = 0 \]

\[ V_{gd} \]

Vertical field \( \sim 5 \times 10^6 \) V/cm

Horizontal field \( \sim 10^5 \) V/cm

2nd Order Effect: Velocity Saturation

- When the carriers reach their speed limit in silicon...

  - Means that relationship between \( I_{ds} \) and \( V_{gs} \) is closer to linear than quadratic
  - Also the saturation point is smaller than predicted
  - For example, 180nm process
    - 1st order model = 1.3v
    - Really is 0.6v
2nd Order Effect: Velocity Saturation

- This is a basic difference between long- and short-channel devices
  - The strength of the horizontal EM field in a short channel device causes the carriers to reach their velocity limit early
  - Devices saturate faster and deliver less current than the quadratic model predicts

---

2nd Order Effect: Velocity Saturation

- Consider two devices with the same W/L ratio in our process ($V_{gs}=5v$, $V_{dd}=5v$)
  - 100/20 vs 3/0.6
  - They should have the same current...
  - Because of velocity saturation in the short-channel device, it has ~50% less current!
2nd Order Effect: Velocity Saturation

A second order effect that raises $V_t$.

Recall that $V_t$ is affected by $V_{sb}$ (voltage between source and substrate):

- Normally this is constant because of common substrate.
- But, when transistors are in series, $V_{sb}$ ($V_s - V_{substrate}$) may be different.
2nd Order Effect: Body Effect

- Body Effect -
  $V_t$ is a function of voltage between source and substrate

$$V_t = V_{to} + kT \sqrt{(2\varphi_b + |V_{th}|) + 2\varphi_b}$$

$$\varphi_b = \frac{kT}{q} \ln \left( \frac{N_A}{N_i} \right)$$

$$\gamma = \frac{I_{ox}}{I} \sqrt{2q\varepsilon_{si}N_A} = \frac{1}{C_{ox}} \sqrt{2q\varepsilon_{si}N_A}$$

2nd Order Effect: Body Effect

- Consider an nmos transistor in a 180nm process
  - Nominal $V_t$ of 0.4v
  - Body is tied to ground
  - How much does the $V_t$ increase if the source is at 1.1v instead of 0v?

  - Because of the body effect, $V_t$ increases by 0.28v to be 0.68v!
Channel Length Modulation

- **Channel Length Modulation** - Channel length is a function of Vds. When Vds increase, the depletion region of the pinch off at drain shorten the channel length.

\[ L_{\text{eff}} = L = L_{\text{short}} \]
\[ L_{\text{short}} = \sqrt{\frac{2 C_{\text{sd}}}{qN_A} (V_{\text{ds}} - (V_{\text{gs}} - V_I))} \]
\[ I_{\text{ds}} = \frac{kW}{2L} (V_{\text{gs}} - V_I)^2 (1 + \lambda V_{\text{ds}}) \]
Mobility Variation

- **Mobility Variation** -
  The mobility of the carrier decreases when the carrier density increases. Therefore, when $V_{gs}$ is large. The density of the carrier in the channel increases. As a result, the mobility decreases.

  \[
  \mu = \frac{Average \ _ \ carrier \ _ \ drift \ _ \ velocity(V)}{Electrical \ _ \ Field \ (E)}
  \]

  $\mu_n = 600 \ cm^2 / \ V \cdot sec$

  $\mu_p = 250 \ cm^2 / \ V \cdot sec$

Other 2\textsuperscript{nd} Order Effects

- **Fowler-Nordheim Tunneling**
  When the gate oxide is very thin, a current can flow from gate to source by electron tunneling through the gate oxide.

  \[
  I_{FN} = C_W L E_{ox}^2 \ e^{\frac{-E_o}{E_{ox}}}
  \]

  \[
  E_{ox} = \frac{V_{gs}}{ \ e^{\frac{E_o}{E_{ox}}}}
  \]

- **Drain Punchthrough**
  When the drain voltage is high enough, the depletion region around the drain may extend to the source. Thus, causing current to flow irrespective of the gate voltage.
Other 2\textsuperscript{nd} Order Effects

- **Impact Ionization - Hot Electrons**
  When the source-drain electric field is too large, the electron speed will be high enough to break the electron-hole pair. Moreover, the electrons will penetrate the gate oxide, causing a gate current.

- **Subthreshold Region**
  The cutoff region is also referred to as the subthreshold region, where \( I_{ds} \) increase exponentially with \( V_{ds} \) and \( V_{gs} \).

\[ \frac{\beta_n}{\beta_p} = \frac{\mu_n}{\mu_p} = \text{electron mobility} / \text{hole mobility} \]

This ratio is usually between 2 and 3

Means ratio of \( \frac{W_{ptree}}{W_{ntree}} \) needs to be between 2 and 3 for \( \frac{\beta_n}{\beta_p} = 1 \)

For this class, we’ll use \( \frac{W_{ptree}}{W_{ntree}} = 2 \)

Inverter Switching Point

- Inverter switching point is determined by ratio of \( \beta_n/\beta_p \)
  - If \( \beta_n/\beta_p = 1 \), then switching point is \( V_{dd}/2 \)
  - If \( W/L \) of both \( N \) and \( P \) transistors are equal
    - Then \( \beta_n/\beta_p = \mu_n/\mu_p = \text{electron mobility} / \text{hole mobility} \)
    - This ratio is usually between 2 and 3
    - Means ratio of \( W_{ptree}/W_{ntree} \) needs to be between 2 and 3 for \( \beta_n/\beta_p = 1 \)
    - For this class, we’ll use \( W_{ptree}/W_{ntree} = 2 \)
Inverter Switching Point

Inverter Operating Regions

<table>
<thead>
<tr>
<th>Region</th>
<th>NMOS</th>
<th>PMOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Off</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>Sat</td>
<td>Linear</td>
</tr>
<tr>
<td>C</td>
<td>Sat</td>
<td>Sat</td>
</tr>
<tr>
<td>D</td>
<td>Linear</td>
<td>Sat</td>
</tr>
<tr>
<td>E</td>
<td>-</td>
<td>Off</td>
</tr>
</tbody>
</table>
Gate Sizes

- Assume minimum inverter is \( \frac{W_p}{W_n} = 2/1 \) (\( L = L_{\text{min}}, W_n = W_{\text{min}}, W_p = 2W_n \))
  - This becomes a 1x inverter
- To drive larger capacitive loads, you need more gain, more \( I_{ds} \)
  - Double \( W_n \) and \( W_p \) to get 2x inverter
  - \( W_p/W_n \) is still 2/1, but inverter has twice the gain (current drive)
  - Not always a linear relationship…

Inverter \( \beta \) Ratios
Inverter $\beta$ Ratios

$$V_{DD}$$

$V_{out}$

$V_{in}$

$\frac{\beta_p}{\beta_n} = 10$

2

1

0.5

$\frac{\beta_p}{\beta_n} = 0.1$

Inverter Noise Margin

How much noise can a gate see before it doesn’t work right?
Inverter Noise Margin

- To maximize noise margins, select logic levels at:
  - Unity gain point of DC transfer characteristic

\[ \begin{align*}
V_{\text{DD}} & & \text{Unity Gain Points} \\
V_{\text{in}} & & \text{Slope} = -1 \\
V_{\text{out}} & & \\
V_{\text{OH}} & & \beta_p / \beta_n > 1 \\
V_{\text{IL}} & & V_{\text{tn}} \\
V_{\text{IH}} & & \\
V_{\text{tn}} & & \\
\end{align*} \]

Performance Estimation

- First we need to have a model for resistance and capacitance
  - Delays are caused (to first order) by RC delays charging and discharging capacitors
- All these layers on the chip have R and C associated with them
- Low level analog circuit simulations
  - Spectre or HSPICE
- High level PrimeTime simulations
  - In 6770…
Resistance

- \( R = \frac{\rho}{t} \frac{L}{W} = R_s \frac{L}{W} \)
  - \( \rho \) = resistivity of the material
  - \( t \) = thickness
  - \( R_s \) = sheet resistance in \( \Omega \)/square

Typical values of \( R_s \) in our process

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Typ</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>M3</td>
<td>0.04</td>
<td>0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>M1, M2</td>
<td>0.07</td>
<td>0.08</td>
<td>0.1</td>
</tr>
<tr>
<td>Poly</td>
<td>20</td>
<td>25</td>
<td>40</td>
</tr>
<tr>
<td>Poly2</td>
<td>40</td>
<td>50</td>
<td>60</td>
</tr>
<tr>
<td>N(P)-active</td>
<td>60 (70)</td>
<td>90 (120)</td>
<td>120 (160)</td>
</tr>
<tr>
<td>Nwell</td>
<td>1k</td>
<td>2k</td>
<td>5k</td>
</tr>
</tbody>
</table>

Capacitance

- **Three main forms:**
  - Gate capacitance (gate of transistor)
  - Diffusion capacitance (drain regions)
  - Routing capacitance (metal, etc.)

\[ C_g = C_{gb} + C_{gs} + C_{gd} \]

Approximated by

\[ C = C_{ox}A \]

\( C_{ox} \) = thin oxide cap
\( A \) = area of gate
Routing Capacitance

- First order effect is layer->substrate
  - Approximate using parallel plate model
  - \( C = \frac{\varepsilon}{t}A \)
    - \( \varepsilon \) = permittivity of insulator
    - \( t \) = thickness of insulator
    - \( A \) = area
  - Fringing fields increase effective area
- Capacitance between layers becomes very complex to simulate!
  - Crosstalk issues…

Distributed RC on Wires

- Wires look like distributed RC delays
  - Long resistive wires can look like transmission lines
  - Inserting buffers can really help delay
  - \( T_n = \frac{RC_n(n+1)}{2} \)
  - \( T = KRCL^2/2 \) as the number of segments becomes large
    - \( K \) = constant (i.e. 0.7) (accounts for rise/fall times)
    - \( R \) = resistance per unit length
    - \( C \) = capacitance per unit length
    - \( L \) = length of wire
**RC Wire Delay Example**

- \( R = 20 \Omega/\text{sq} \)
- \( C = 4 \times 10^{-4} \text{ pF/um} \)
- \( L = 2\text{mm} \)
- \( K = 0.7 \)
- \( T = KRCL^2/2 \)
- \( T = (0.7)(20)(4 \times 10^{-15})(2000)^2 / 2 \text{ s} \)
  - delay = 11.2 ns

**RC Wire/Buffer Delay Example**

- Now split into 2 segments of 1mm with a buffer
- \( T = 2 \times (0.7)(20)(4\times10^{-15})(1000)^2)/2 + T_{\text{buf}} \)
  - = 5.6ns + \( T_{\text{buf}} \)
- Assuming \( T_{\text{buf}} \) is less than 5.6ns
  (which it will be), the split wire is a win
Another Example: Clock

- 50pF clock load distributed across 10mm chip in 1um metal
  - Clock length = 20mm
  - $R = 0.05\Omega/\text{sq}$, $C = 50\text{pF}/20\text{mm}$
  - $T = (0.7)(RC/2)L^2 = (0.7)(6.25 \times 10^{-17})(20,000)^2$ = 17.5ns

Different Distribution Scheme

- Put clock driver in the middle of the chip
- Widen clock line to 20um wires
  - Clock length = 10mm
  - $R = 0.05\Omega/\text{sq}$, $C = 50\text{pF}/20\text{mm}$
  - $T = (0.7)(RC/2)L^2 = (0.7)(0.31 \times 10^{-17})(10,000)^2$ = 0.22ns
  - Reduces R by a factor of 20, L by 2
  - Increases C a little bit (clock load still 50pF)
Capacitance Design Guide

- Get a table of typical capacitances per unit square for each layer
  - Capacitance to ground
  - Capacitance to another layer
- Add them up…
- See, for example, Figure 6.12 in your text

**FIGURE 6.12** Capacitance of metal2 line as a function of width and spacing
Wire Length Design Guide

- How much wire can you use in a conducting layer before the RC delay approaches that of a unit inverter?
  - Metal3 = 2,500u
  - Metal2 = 2,000u
  - Metal1 = 1,250u
  - Poly = 50u
  - Active = 15u

Propagation Delay

- Recall that it takes time to charge capacitors
- Recall that the gate of a transistor looks like a capacitor
- Wires have resistance and capacitance also!
Inverter Propagation Delay

Signal rise time

Signal fall time

Non-Inverting Delay

\[ t_{ph} \]

\[ t_{plh} \]
Propagation Delay

Where to Measure Delay?

If use 50% point (input) to 50% point (output), can produce negative delays (slow input slope, fast output slope).

A better way is to use the 30% and 70% points on the signals.
Example Non-Inverting Gate

What Affects Gate Delay?

- Environment
  - Increasing Vdd improves delay
  - Decreasing temperature improves delay
  - Fabrication effects, fast/slow devices

- Usually measure delay for at least three cases:
  - Best - high Vdd, low temp, fast N, Fast P
  - Worst - low Vdd, high temp, slow N, Slow P
  - Typical - typ Vdd, room temp (25C), typ N, typ P
Process Corners

- When parts are specified, under what operating conditions?
  - **Temp:** three ranges
    - Commercial: 0 C to 70 C
    - Industrial: -40 C to 85 C
    - Military: -55 C to 125 C
  - **Vdd:** Should vary ± 10%
    - 4.5 to 5.5v for example
  - **Process variation:**
    - Each transistor type can be slow or fast

What Else Affects Gate Delay?

Input slew and output load both effect timing. For a FIXED input slope, FIXED environment, a simple timing model is:

\[
\text{delay} = T_{\text{load}} + K \times \text{Load}
\]

\(T_{\text{load}}\) is the delay of the gate with no external load.

\(K\) is different for TPLH, TPHL since it represents the channel resistance. Same equation is used for Slew values.
Inv_Test Schematic

Closeup of Inv-Test

Note the sizes I used for this example…
Analog Simulation Output

Note different waveforms for different sizes of transistors

Effective Resistance

Shockley models have limited value
- Not accurate enough for modern transistors
- Too complicated for much hand analysis

Simplification: treat transistor as resistor
- Replace $I_{ds}(V_{ds}, V_{gs})$ with effective resistance $R$
  - $I_{ds} = \frac{V_{ds}}{R}$
  - $R$ averaged across switching of digital gate

Too inaccurate to predict current at any given time
- But good enough to predict RC delay
RC Delay Model

- Use equivalent circuits for MOS transistors
  - Ideal switch + capacitance and ON resistance
  - Unit nMOS has resistance R, capacitance C
  - Unit pMOS has resistance 2R, capacitance C
- Capacitance proportional to width
- Resistance inversely proportional to width

![Diagram showing equivalent circuits for MOS transistors]

RC Values

- Capacitance
  - \( C = C_g = C_s = C_d = 2 \text{ fF/\mu m of gate width} \)
  - Values similar across many processes
- Resistance
  - \( R \approx 6 \text{ K\Omega*\mu m in 0.6\mu m process} \)
  - Improves with shorter channel lengths
- Unit transistors
  - May refer to minimum contacted device (1.2\mu/0.6\mu)
  - Or maybe 1 \mu wide device
  - Doesn’t matter as long as you are consistent
Inverter Delay Estimate

- Estimate the delay of a fanout-of-1 inverter
Estimate the delay of a fanout-of-1 inverter

d = 6RC
**What’s a Standard Load?**

- Step or high slew
- Realistic input slope
- Must include realistic load, realistic input slew
- Fanout 4 (FO4) load

**DUT (device under test)**

**What About Gates in Series**

- Basically we want every gate to have the delay of a “standard inverter”
  - Standard inverter starts with 2/1 P/N ratio
- Gates in series? Sum the conductance to get the series conductance
- $\beta_{n-eff} = 1/(1/\beta_1 + 1/\beta_2 + 1/\beta_3)$
  - $\beta_{n-eff} = \beta n/3$
- Effect is like increasing L by 3
  - Compensate by increasing W by 3
Power Dissipation

- Three main contributors:
  1. Static leakage current ($P_s$)
  2. Dynamic short-circuit current during switching ($P_{sc}$)
  3. Dynamic switching current from charging and discharging capacitors ($P_d$)

- Becoming a HUGE problem as chips get bigger, clocks get faster, transistors get leakier!
  - Power typically gets dissipated as heat…

Static Leakage Power

- Small static leakage current due to:
  - Reverse bias diode leakage between diffusion and substrate (PN junctions)
  - Subthreshold conduction in the transistors

- Leakage current can be described by the diode current equation
  - $I_o = i_s(e^{qV/kT} - 1)$
  - Estimate at $0.1\text{nA} - 0.5\text{nA}$ per device at room temperature
Static Leakage Power

- That’s the leakage current
- For static power dissipation:
  - $P_s = \text{SUM of } (I \times V_{dd}) \text{ for all n devices}$
  - For example, inverter at 5v leaks about 1-2 nW in a .5u technology
  - Not much…
  - …but, it gets MUCH worse as feature size shrinks!

Short-Circuit Dissipation

- When a static gate switches, both N and P devices are on for a short amount of time
  - Thus, current flows during that switching time
Short-Circuit Dissipation

- So, with short-circuit current on every transition of the output, integrate under that current curve to get the total current
  - It works out to be:
    - \( P_{sc} = \frac{\beta}{12}(Vdd - 2Vt)^3 \left( \frac{Trf}{Tp} \right) \)
  - Assume that \( Tr = Tf, Vtn = -Vtp, \) and \( \beta_n = \beta_p \)
  - Note that \( P_{sc} \) depends on \( B \), and on input waveform rise and fall times

In practice — unless input edge rate is a LOT slower than the output edge rate — this is a small fraction of the total current, and is usually negligible...
Dynamic Dissipation

- Charging and discharging all those capacitors!
  - By far the largest component of power dissipation
  - \( P_d = C_L V_{dd}^2 f \)

- Watch out for large capacitive nodes that switch at high frequency
  - Like clocks…

Total Power

- These are pretty rough estimates
- It’s hard to be more precise without CAD tool support
  - It all depends on frequency, average switching activity, number of devices, etc.
  - There are programs out there that can help
- But, even a rough estimate can be a valuable design guide
- \( P_{total} = P_s + P_{sc} + P_d \)
Power Dissipation

- Lead microprocessor power continues to increase

- Power delivery and dissipation will be prohibitive

Source: Birkan, De Intel®

Heat Dissipation

- 60 W light bulb has surface area of 120 cm²
- Core2 Duo die dissipates 75 W over 1.4 cm²
  - Chips have enormous power densities
  - Cooling is a serious challenge
- Graphics chips even worse
  - NVIDIA GTX480 – 250 W in ~3 cm²
- Package spreads heat to larger surface area
  - Heat sinks may increase surface area further
  - Fans increase airflow rate over surface area
  - Liquid cooling used in extreme cases ($$$)
**Chip Power Density**

\[ P = VI \]
\[ 75W \times 1.5V = 50\, \text{A} \]

**Thermal Solutions**

- Heat sink
  - Mounted on processor package
- Passive cooling
  - Remote system fan
- Active cooling
  - Fan mounted on sink
- Heat spreaders
  - Increase surface area
  - Example: Metal plate under laptop keyboard

---

Source: [Borkar, D. Intel](http://example.com)

"Thermal Challenges during Microprocessor Testing", Intel Technology Journal, Q3 2000
Alternative View of “Computing Power”

Power Management on Pentium 4

- Over 400 power-saving features!
  - 20% of features = 75% of saved power
- Clock throttling
  - Thermal diode temperature sensor
  - Stop clock for a few microseconds
  - Output pin can be used by system to trigger other responses
- SpeedStep technology for mobile processors
  - Switch to lower frequency and voltage
  - Depends on whether power source is battery or AC
  - Can be manually overridden by Windows control panel

“Managing the Impact of Increasing Microprocessor Power Consumption”, Intel Technology Journal, Q1 2001
Pentium 4 Multi-level Powerdown

- Level 0 = Normal operation (includes thermal throttle)
- Level 1 = Halt instructions (less processor activity)
- Level 2 = Stop Clock (internal clocks turn off)
- Level 3 = Deep sleep (remove chip input clock)
- Level 4 = Deeper sleep (lower Vdd by 66%)

  - For “extended periods of processor inactivity”
  - QuickStart technology – resume normal operation from Deeper Sleep

- Note: We haven’t even talked about system powerdown modes, like removing power from processor, stopping hard disks, dimming or turning off the display...

"Managing the Impact of Increasing Microprocessor Power Consumption", Intel Technology Journal, Q1 2001

---

Graphics Card Power

http://www.xbitlabs.com/articles/video/display/ati-vs-nv-power.html

![Graphics Card Power Consumption Chart]

\[ P=VI: \ 75W @ 1.5V = 50\ A! \]
Graphics Card Power

- NVIDIA GeForce FX5900 (2004): 53 GFLOPS
  - 128 FP units in parallel at 450MHz
  - 192 FP units at 550 MHz, 80 watts
- NVIDIA GeForce GTX 480 (2010): 1.35 TFLOPS
  - 480 cores, 1.4GHz, 250 watts... 105°C
  - 1.5 GB GDDR5, 384 bit interface, 177.4 GB/sec
  - 3B transistors in 40nm CMOS

Graphics Chip Architecture

<table>
<thead>
<tr>
<th></th>
<th>Fermi GF100</th>
<th>Fermi GF104</th>
<th>Kepler GK100</th>
<th>Kepler GK110</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute Capability</td>
<td>2.0</td>
<td>2.1</td>
<td>3.0</td>
<td>3.5</td>
</tr>
<tr>
<td>Threads / Warp</td>
<td>82</td>
<td>82</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Max Warps / Multiprocessor</td>
<td>48</td>
<td>48</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>Max Threads / Multiprocessor</td>
<td>1536</td>
<td>1536</td>
<td>2048</td>
<td>2048</td>
</tr>
<tr>
<td>Max Thread Blocks / Multiprocessor</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>32-bit Registers / Multiprocessor</td>
<td>32768</td>
<td>32768</td>
<td>65536</td>
<td>65536</td>
</tr>
<tr>
<td>Max Registers / Thread</td>
<td>63</td>
<td>63</td>
<td>63</td>
<td>255</td>
</tr>
<tr>
<td>Max Threads / Thread Block</td>
<td>1024</td>
<td>1024</td>
<td>1024</td>
<td>1024</td>
</tr>
<tr>
<td>Shared Memory Size Configurations (Bytes)</td>
<td>10K</td>
<td>10K</td>
<td>10K</td>
<td>10K</td>
</tr>
<tr>
<td></td>
<td>48K</td>
<td>48K</td>
<td>32K</td>
<td>32K</td>
</tr>
<tr>
<td></td>
<td>48K</td>
<td>48K</td>
<td>48K</td>
<td>48K</td>
</tr>
<tr>
<td>Max X Grid Dimension</td>
<td>2^16-1</td>
<td>2^16-1</td>
<td>2^32-1</td>
<td>2^32-1</td>
</tr>
<tr>
<td>Hyper-Q</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Dynamic Parallelism</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Compute Capability of Fermi and Kepler GPUs.
Graphics Chips Transistors

- Nvidia GK110 Kepler
- 10-core Xeon Westmere

Nvidia GK110 Kepler
Graphics Cards