Announcements

Handouts - Lecture Outline, CMOS scaling (2 items)
Final - Wednesday, May 24, 1:30-4:30 pm, Johnson Ice Rink

Review - Intrinsic high frequency limits for transistors: $\omega_T$

- Short-circuit current gain: best can do from CE or CS
- Unity gain frequency:
  - BJT: $\omega_T \approx 2D_e/w_B^2 = 2\mu_e V_{th}/w_B^2$ (used Einstein rel.)
  - MOSFET: $\omega_T \approx 3\mu_e (V_{GS} - V_T)/2L^2$

Revisiting the quasi-static assumption (^ same form; ain't it neat!!)

CMOS scaling rules

Review of Lecture 15 results:

- Gate Delay = $12 n L_{min}^2 V_{DD}/\mu_n (V_{DD} - V_T)^2$
- $P_{ave} @ max. f \propto C_L V_{DD}^2/GD = K_n V_{DD} (V_{DD} - V_T)^2/4$

Power density issues and challenges

Approaches to a solution:

- Dimension scaling
- Scaling voltages as well

Velocity saturation

Impact on MOSFET i-v

Consequences for $g_m$, $\omega_t$, and CMOS performance
Intrinsic $\omega_{Hl}$'s for BJTs and MOSFETs - short-circuit current gain

$\beta_{sc}(BJT) \equiv \frac{i_c(j\omega)}{i_b(j\omega)}$

$\beta_{sc}(BJT) = \frac{g_m - j\omega C_{\mu}}{g_\pi + j\omega(C_\pi + C_{\mu})}$

$\beta_{sc}(BJT) = \frac{g_m - j\omega C_{gd}}{j\omega(C_{gs} + C_{gd})}$

$\beta_{sc}(MOSFET) \equiv \frac{i_d(j\omega)}{i_s(j\omega)}$

The common-emitter and common-source short-circuit current gain, $\beta(j\omega)$:
BJT short-circuit current gain, $\beta_{sc}(j\omega)$

Note: $\omega_z > \omega_t >> \omega_\beta$ (= $\omega_t / \beta_F$)

Low frequency value: $\beta_F$

Zero, $\omega_z$: $\omega_z = g_m / C_\mu$

3dB point, $\omega_\beta$: $\omega_\beta = g_\pi / (C_\pi + C_\mu)$

Unity gain point, $\omega_t$: $\omega_t \approx g_m / (C_\pi + C_\mu)$
MOSFET short-circuit current gain, $\beta_{sc}(j\omega)$

- **Unity gain point, $\omega_t$:** $\omega_t \approx \frac{g_m}{(C_{gs}+C_{gd})}$
- **Zero, $\omega_z$:** $\omega_z = \frac{g_m}{C_{gd}}$
- **No 3dB point, $\omega_\beta$.**
- **Low frequency value:** infinity

Note: $\omega_z > \omega_t$
BJT $\omega_t$:

$$\omega_t (\text{BJT}) \approx \frac{g_m}{(C_\pi + C_\mu)}$$

$$= \left[ \left( \frac{qI_C}{kT} \right) \tau_b + C_{eb,dp} + C_{cb,dp} \right]$$

$$= \frac{1}{\left( \frac{qI_C}{kT} \right) \tau_b + kT \left( C_{eb,dp} + C_{cb,dp} \right) / qI_C}$$

$$\lim_{I_C \to \infty} \omega_t (\text{BJT}) \approx \frac{1}{\tau_b} = \frac{2D_{min,B}}{W_B^2} = \frac{2\mu_{min,B}V_{thermal}}{W_B^2}$$

Base transit time

MOSFET $\omega_t$:

$$\omega_t (\text{MOSFET}) = \frac{g_m}{(C_{gs} + C_{gd})} \approx \frac{g_m}{C_{gs}} = \frac{W}{L} \mu_{Ch} C_{ox}^* |V_{GS} - V_T|$$

$$= \frac{3}{2L} \mu_{Ch} \frac{V_{DS,sat}}{L} = \frac{3}{2L} \mu_{Ch} E_{Ch} = \frac{3}{2} \frac{s_{Ch}}{L} = \frac{1}{\tau_{Ch}}$$

Channel transit time

Lessons:
- Bias at large current level
- Make active length short
- Base device on electrons
**CMOS**: switching speed; minimum cycle time

The load capacitance, \( C_L \)
- Assume to be linear
- Is proportional to MOSFET gate area
- In channel: \( \mu_e = 2\mu_h \) so to have \( K_n = K_p \) we must have \( W_p/L_p = 2W_n/L_n \)
  Typically \( L_n = L_p = L_{\text{min}} \) and \( W_n = W_{\text{min}} \), so we also have \( W_p = 2W_{\text{min}} \)

\[
C_L \approx n \left[ W_n L_n + W_p L_p \right] C_{ox}^* = n \left[ W_{\text{min}} L_{\text{min}} + 2W_{\text{min}} L_{\text{min}} \right] C_{ox}^* = 3nW_{\text{min}}L_{\text{min}}C_{ox}^*
\]

Charging cycle, \( v_{\text{IN}} \): HI to LO; \( Q_n \) off, \( Q_p \) on; \( v_{\text{OUT}} \): LO to HI
- Assume charged by constant \( i_{\text{D, sat}} \)

\[
i_{\text{Charge}} = -i_{\text{D}} \approx \frac{K_p}{2} \left[ V_{DD} - |V_{T_p}| \right]^2 = \frac{K_n}{2} \left[ V_{DD} - V_{T_n} \right]^2
\]

\[
q_{\text{Charge}} = C_L V_{DD}
\]

\[
\tau_{\text{Charge}} = \frac{q_{\text{Charge}}}{i_{\text{Charge}}} = \frac{2C_L V_{DD}}{K_n \left[ V_{DD} - V_{T_n} \right]^2} = \frac{6nW_{\text{min}}L_{\text{min}}C_{ox}^*V_{DD}}{W_{\text{min}}\mu_e C_{ox}^* \left[ V_{DD} - V_{T_n} \right]^2} = \frac{6nL_{\text{min}}^2V_{DD}}{\mu_e \left[ V_{DD} - V_{T_n} \right]^2}
\]
**CMOS:** switching speed; minimum cycle time, cont.

**Charging cycle, \( v_{IN} \):** LO to HI; \( Q_n \) on, \( Q_p \) off; \( v_{OUT} \): HI to LO

- Assume discharged by constant \( i_{D,\text{sat}} \)

\[
i_{\text{Disch arg} e} = i_{Dn} \approx K_n \frac{V_{DD} - V_{Tn}}{2}
\]

\[
q_{\text{Disch arg} e} = C_L V_{DD}
\]

\[
\tau_{\text{Disch arg} e} = \frac{q_{\text{Disch arg} e}}{i_{\text{Disch arg} e}} = \frac{2C_L V_{DD}}{K_n [V_{DD} - V_{Tn}]^2}
\]

\[
= \frac{6nW_{\text{min}}L_{\text{min}}C_{ox}^* V_{DD}}{W_{\text{min}} L_{\text{min}} \mu e C_{ox}^* [V_{DD} - V_{Tn}]^2} = \frac{6nL_{\text{min}}^2 V_{DD}}{\mu e [V_{DD} - V_{Tn}]^2}
\]

**Minimum cycle time:**  
\( v_{IN} \): LO to HI to LO;  
\( v_{OUT} \): HI to LO to HI

\[
\tau_{\text{Min. Cycle}} = \tau_{\text{Ch arg } e} + \tau_{\text{Disch arg } e} = \frac{12nL_{\text{min}}^2 V_{DD}}{\mu e [V_{DD} - V_{Tn}]^2}
\]

**Compare to \( \tau_{Ch} \) of the FET's:**

\[
\tau_{Ch} = \frac{2 L^2}{3 \mu Ch (V_{DD} - V_T)}
\]
**CMOS:** power dissipation - total and per unit area

**Average power dissipation**

All dynamic

$$P_{ave} = \frac{E_{Dissipated\ per\ cycle}}{f} = C_L V_{DD}^2 = 3nW_{\text{min}}L_{\text{min}}C_{ox}^* V_{DD}^2 f$$

**Average power at maximum data rate**

Maximum $f$ will be $1/\tau_{\text{Gate\ Delay\ Min.}}$

$$P_{ave@Max.f} = \frac{3nW_{\text{min}}L_{\text{min}}C_{ox}^* V_{DD}^2}{\tau_{\text{Min.Cycle}}} = 3nW_{\text{min}}L_{\text{min}}C_{ox}^* V_{DD}^2 \cdot \frac{\mu_e [V_{DD} - V_Tn]^2}{12nL_{\text{min}}^2 V_{DD}}$$

$$= \frac{1}{4} \frac{W_{\text{min}}}{L_{\text{min}}} \mu_e C_{ox}^* V_{DD} \left[V_{DD} - V_Tn\right]^2$$

**Average power density at maximum data rate**

Assume that the area per inverter is proportional to $W_{\text{min}}L_{\text{min}}$

$$PD_{ave@Max.f} = \frac{P_{ave@Max.f}}{\text{InverterArea}} \propto \frac{P_{ave@Max.f}}{W_{\text{min}}L_{\text{min}}} = \frac{\mu_e C_{ox}^* V_{DD} \left[V_{DD} - V_Tn\right]^2}{L_{\text{min}}^2}$$

Clif Fonstad, 5/16/06  
Lecture 25 - Slide 8
**CMOS:** design for high speed

**Maximum data rate**
Proportional to $1/\tau_{\text{Min Cycle}}$

$$\tau_{\text{Min Cycle}} = \tau_{\text{Ch arg e}} + \tau_{\text{Disch arg e}} = \frac{12nL_{\text{min}}^2V_{DD}}{\mu_e[V_{DD} - V_{Tn}]^2}$$

Implies we should reduce $L_{\text{min}}$ and increase $V_{DD}$.
Note: As we reduce $L_{\text{min}}$ we must also reduce $t_{\text{ox}}$, but $t_{\text{ox}}$ doesn't enter directly in $f_{\text{max}}$ so it doesn't impact us here.

**Average power density at maximum data rate**
Assume that the area per inverter is proportional to $W_{\text{min}}L_{\text{min}}$

$$PD_{\text{ave @ Max. f}} \propto \frac{P_{\text{ave @ Max. f}}}{W_{\text{min}}L_{\text{min}}} = \frac{\mu_e\varepsilon_{\text{ox}}V_{DD}[V_{DD} - V_{Tn}]^2}{t_{\text{ox}}L_{\text{min}}^2}$$

Shows us that PD increases very quickly as we reduce $L_{\text{min}}$ unless we also reduce $V_{DD}$ (which will also reduce $f_{\text{max}}$).
Note: Now $t_{\text{ox}}$ appears in the expression so the impact of reducing $L_{\text{min}}$ is even more dramatic because $t_{\text{ox}}$ must be reduced similarly.

**How do we make $f_{\text{max}}$ larger without melting the silicon?**
By following CMOS scaling rules - our next topic.
Scaling Rules - making CMOS faster without melting Si

- General idea:
  Reduce dimensions and/or voltages by factor 1/s:  $s > 1$
  Evaluate impact on speed, power, power density

- Scaling dimensions alone:
  $L_{\text{min}} \rightarrow L_{\text{min}}/s$
  $w \rightarrow w/s$
  $t_{\text{ox}} \rightarrow t_{\text{ox}}/s$

  this yields
  $K \rightarrow sK$
  $C_{\text{ox}}^* \rightarrow sC_{\text{ox}}^*$

  and ultimately
  $\tau \rightarrow \tau/s^2$
  $P_{\text{ave}} \rightarrow sP_{\text{ave}}$
  $P_{\text{density}} \rightarrow s^3P_{\text{density}}$!!!

  **Scaling dimensions alone** can yield melted silicon!!

- Scaling dimensions and voltages in concert:
  Add:
  $V_{\text{DD}} \rightarrow V_{\text{DD}}/s$
  $V_T \rightarrow V_T/s$

  still have
  $K \rightarrow sK$
  $C_{\text{ox}}^* \rightarrow sC_{\text{ox}}^*$

  but now
  $\tau \rightarrow \tau/s$
  $P_{\text{ave}} \rightarrow P_{\text{ave}}/s^2$
  $P_{\text{density}} \rightarrow P_{\text{density}}$

  **When scale both get:** faster, lower power, same power density!!
  **Note:** scaling the voltages is not as easy as scaling the dimensions.
### An historical scaling example - Inside Intel

<table>
<thead>
<tr>
<th>Parameter</th>
<th>386</th>
<th>486</th>
<th>Pentium</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaling factor, $s$</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>$L_{\text{min}}$ ($\mu$m)</td>
<td>1.5</td>
<td>0.75</td>
<td>0.5</td>
</tr>
<tr>
<td>$W_n$ ($\mu$m)</td>
<td>10</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>$t_{\text{ox}}$ (nm)</td>
<td>30</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>$V_{\text{DD}}$ (V)</td>
<td>5</td>
<td>3.3</td>
<td>2.2</td>
</tr>
<tr>
<td>$V_T$ (V)</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Fan out</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$K$ ($\mu$A/V²)</td>
<td>230</td>
<td>450</td>
<td>600</td>
</tr>
<tr>
<td>GD (ps)</td>
<td>840</td>
<td>400</td>
<td>250</td>
</tr>
<tr>
<td>$f_{\text{max}}$ (MHz)</td>
<td>29</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>$P_{\text{ave}}$/gate (mW)</td>
<td>92</td>
<td>23</td>
<td>10</td>
</tr>
<tr>
<td>Density</td>
<td>220</td>
<td>880</td>
<td>2,000</td>
</tr>
</tbody>
</table>

**Sources:** Prof. Jesus del Alamo and Intel
### An second look inside Intel - a slightly different perspective

<table>
<thead>
<tr>
<th>Parameter</th>
<th>486</th>
<th>Pentium generations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaling factor, $s$</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>$L_{min}$ (µm)</td>
<td>1.0</td>
<td>0.8</td>
</tr>
<tr>
<td>SRAM cell area (µm²)</td>
<td>-</td>
<td>111</td>
</tr>
<tr>
<td>Die size (mm²)</td>
<td>170</td>
<td>295</td>
</tr>
<tr>
<td>$f_{mzx}$ (MHz)</td>
<td>38</td>
<td>66</td>
</tr>
<tr>
<td>$t_{ox}$ (m)</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>Metal layers</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Planarization</td>
<td>SOG</td>
<td>CMP</td>
</tr>
<tr>
<td>Poly type</td>
<td>n</td>
<td>n,p</td>
</tr>
<tr>
<td>Transistors</td>
<td>CMOS</td>
<td>BiCMOS</td>
</tr>
</tbody>
</table>

**Source:** Dr. Leon D. Yau, Intel, 10/8/96

---

For the very latest see the International Technology Roadmap for Semiconductors at
[<http://public.itrs.net>](http://public.itrs.net) *(click on “The 2005 ITRS has been finalized and can be found [here.](http://public.itrs.net)*

Clif Fonstad, 5/16/06  
Lecture 25 - Slide 12
Velocity saturation models* - impact on MOSFET's

Model A

\[ s_y(E_y) = \mu_e E_y \text{ if } E_y \leq E_{\text{crit}} \]
\[ = \mu_e E_{\text{crit}} \equiv s_{\text{sat}} \text{ if } E_y \geq E_{\text{crit}} \]

Model B

\[ s_y(E_y) = \frac{\mu_e E_y}{1 + \frac{E_y}{E_{\text{crit}}}} \]

** See pp 281ff and 307ff in course text.
Impact of velocity saturation - Model A*

Previous results hold until $s_y$ at some point in the channel equals $s_{sat}$; after that the current saturates:

\[
s_y(E_y) = \mu_e E_y \quad \text{if} \quad E_y \leq E_{crit}
\]

\[
s_y(E_y) = \mu_e E_{crit} \equiv s_{sat} \quad \text{if} \quad E_y \geq E_{crit}
\]
Impact of velocity saturation - Model A'

Consider a MOSFET with such a short channel that the carriers reach their saturation velocity at very small $v_{DS}$, so small that $i_D$ is still increasing approximately linearly with $v_{DS}$:

$$i_D(v_{GS}, v_{DS}, v_{BS}) = \frac{W}{L} \mu_e C_{ox}^* \left[ v_{GS} - V_T(v_{BS}) - \frac{v_{DS}}{2} \right] v_{DS}$$

$$\approx \frac{W}{L} \mu_e C_{ox}^* [v_{GS} - V_T(v_{BS})] v_{DS} \quad \text{for} \quad v_{DS} \leq E_{crit} L$$

For $v_{DS} \geq E_{crit} L$, the current will saturate at the value it has when $v_{DS} = E_{crit} L$:

$$i_D(v_{GS}, v_{DS}, v_{BS}) = \frac{W}{L} \mu_e C_{ox}^* \left[ v_{GS} - V_T(v_{BS}) \right] E_{crit} L \quad \text{for} \quad v_{DS} \geq E_{crit} L$$

$$= W s_{sat} C_{ox}^* \left[ v_{GS} - V_T(v_{BS}) \right]$$

We first plot these results, and then use them to calculate $g_m$ in the small signal linear equivalent circuit and to find $\omega_t$. 

Clif Fonstad, 5/16/06

Cont. on next foil...
Impact of velocity saturation - Model A', cont.

The output characteristic of a MOSFET in which the electrons in the channel experience such high electric fields that their velocity saturates at $s_{\text{sat}}$ for relatively low values of $v_{DS}$:

Note that in the forward active region where $v_{DS} \geq E_{\text{crit}} L$, the curves in the output family are evenly spaced, indicating a constant $g_m$:

$$g_m \equiv \frac{\partial i_D}{\partial v_{GS}}|_Q = W s_{sat} C_{ox}^*$$
Impact of velocity saturation - Model A', cont.

Small signal linear equivalent circuit: only $g_m$ is changed

$$g_m \equiv \frac{\partial i_D}{\partial v_{GS}} = W s_{sat} C_{ox}^*$$

Short circuit current gain unity gain frequency:

$$\omega_t \approx \frac{g_m}{C_{gs}} = \frac{W s_{sat} C_{ox}^*}{W L C_{ox}^*} = \frac{s_{sat}}{L} = \frac{1}{\tau_{Ch}}$$

CMOS minimum cycle time and power density at $f_{max}$:

$$\tau_{Min.Cycle} \propto \frac{L_{min} V_{DD}}{s_{sat} [V_{DD} - V_{Tn}]}$$

$$PD_{ave @ Max.f} \propto \frac{s_{sat} \varepsilon_{ox} V_{DD} [V_{DD} - V_{Tn}]}{t_{ox} L_{min}}$$

Lessons: Still gain by reducing $L$, but not as quickly. Scaling of both dimensions and voltage is still required. Channel transit time still rules!
Impact of velocity saturation - Model B*

It is easy to show** that when this model for velocity saturation is used (the model is repeated below) the expression we derived originally for the MOSFET drain current, $i_D$, in the linear region with no velocity saturation becomes the following using this model:

$$i_D = \frac{W}{L} \mu_e C_{ox}^*[\left((v_{GS} - V_T) - v_{DS}/2\right)v_{DS} \left(\frac{1}{1 + v_{DS}/E_{crit}L}\right)]$$

when $0 \leq v_{DS} \leq (v_{GS} - V_T)$

This is exactly our original expression for $i_D$ multiplied by the factor $1/(1+v_{DS}/E_{crit}L)$. This model yields results different in detail, but very similar in essence, to those obtained using Model A.

* Model B for $s_y(E_y)$: $s_y(E_y) = \frac{\mu_e E_y}{1 + E_y/E_{crit}}$

** See pp 281ff in course text.
One final model observation - Insight on $g_m$

We in general want an FET with as large a $g_m$ as possible. We can get insight on how to achieve this by looking at our expression for $\omega_t$, and using what we have learned about it being related to the transit time:

$$\omega_t = \frac{1}{\tau_{tr}} \quad \text{and} \quad \omega_t = \frac{g_m}{C_{gs}}$$

Setting these two expressions for $\omega_t$ equal, and solving for $g_m$:

$$g_m = \frac{C_{gs}}{\tau_{tr}}$$

This result teaches us that to get a large $g_m$ we must have:

1. The shortest possible transit time
2. The largest possible coupling between the gate electrode and the channel charge (that is, the largest possible $C_{gs}$).

Pretty neat isn't it?! Useful, too.
CMOS gate delay and power

Three key performance metrics: (We want to make them all smaller)

- Gate Delay = \(12 n L_{\text{min}}^2 V_{\text{DD}}/\mu_n (V_{\text{DD}} - V_T)^2\)
- \(P_{\text{ave}} @ \text{max. } f \propto C_L V_{\text{DD}}^2/GD = (W_n/L_{\text{min}}) \mu_n C_{\text{ox}}* V_{\text{DD}} (V_{\text{DD}} - V_T)^2/4\)
- \(P_{\text{density, max}} \propto P_{\text{ave, max}}/W_n L_{\text{min}} = \mu_n x \varepsilon_{\text{ox}} V_{\text{DD}} (V_{\text{DD}} - V_T)^2/4 t_{\text{ox}} L_{\text{min}}^2\)

CMOS scaling rules

Summary of rules: best to scale all dimensions and all voltages by 1/s

Scaling as: \(L_{\text{min}} \rightarrow L_{\text{min}}/s\) \hspace{1cm} Results in: \(K \rightarrow sK\)
\(w \rightarrow w/s\) \hspace{1cm} \(C_{\text{ox}}* \rightarrow sC_{\text{ox}}*\)
\(t_{\text{ox}} \rightarrow t_{\text{ox}}/s\) \hspace{1cm} \(\tau \rightarrow \tau/s\)
\(V_{\text{DD}} \rightarrow V_{\text{DD}}/s\) \hspace{1cm} \(P_{\text{ave}} \rightarrow P_{\text{ave}}/s^2\)
\(V_T \rightarrow V_T/s\) \hspace{1cm} \(P_{\text{density}} \rightarrow P_{\text{density}}\)

Impact of velocity saturation

Quicker saturation; less current; more uniform, but smaller \(g_m\)

Improvements go as 1/L, no longer as 1/L^2

Still have to scale voltages and dimensions in unison