Lecture 25 - CMOS Scaling Rules - Outline

- **Announcements**

  Handouts - Lecture Outline and Summary; on web: CMOS scaling
  Course evaluation - at the end of lecture today (approx. 11:45)
  Final - Monday, Dec. 15, 9:00 am - noon, duPont Gymnasium

- **Review - Intrinsic high frequency limits for transistors: ω_T**

  Short-circuit current gain: best can do from CE or CS
  Unity gain frequency: BJT: $\omega_T \quad 2D_e/w_B^2 = 2\mu_e V_{th}/w_B^2$ (used Einstein rel.)
  MOSFET: $\omega_T \quad 3\mu_e(V_{GS}-V_T)/2L^2$ (this needs discussion)
  Revisiting the quasi-static assumption (^ same form; ain't it neat!!)

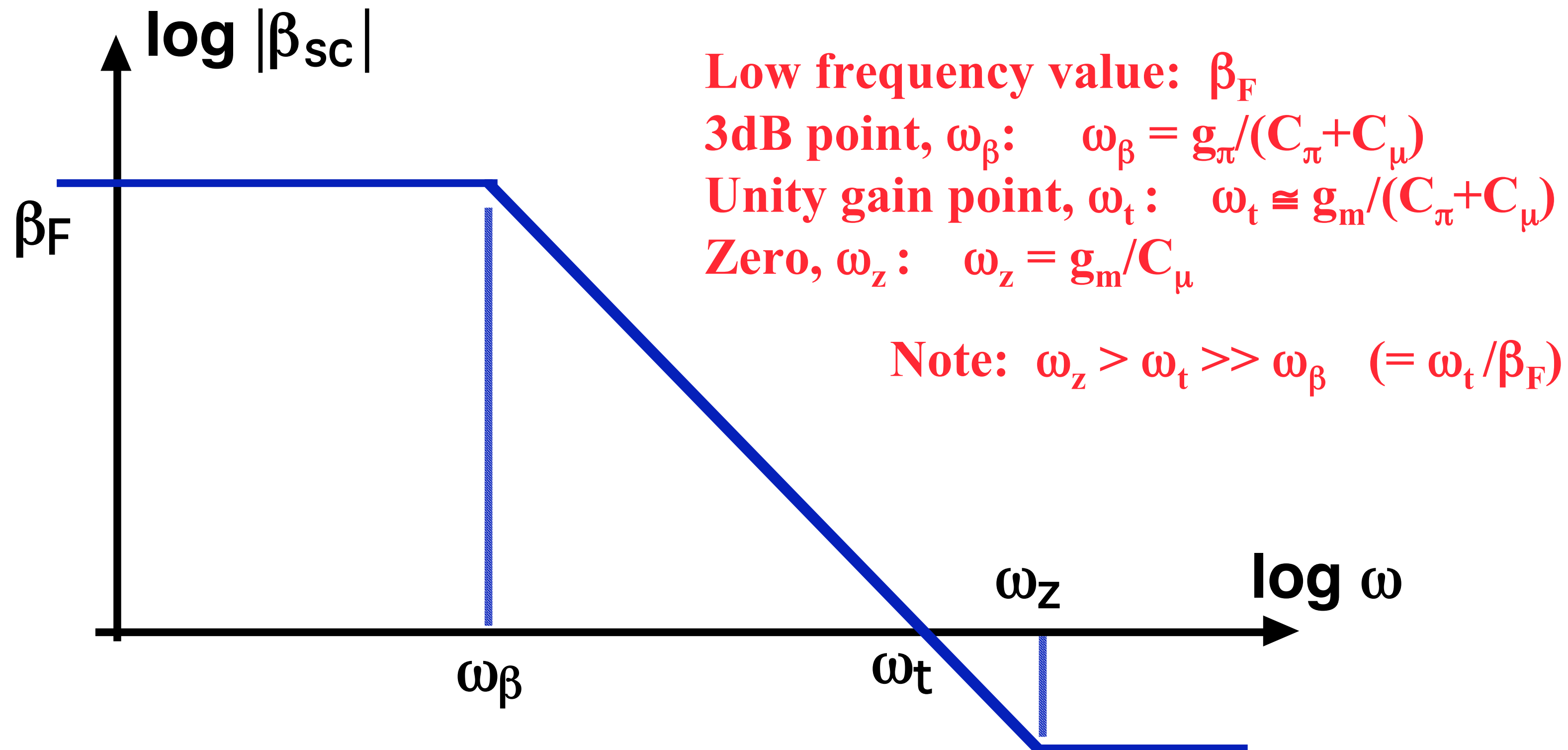
- **CMOS gate delay and power**

  Review of Lecture 16 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 issue: have to add this as well

- **CMOS scaling rules**

  The issues and challenges
  Approaches: Dimension scaling
  Scaling voltages as well
  Summary of rules

BJT short-circuit current gain, $\beta_{sc}(j\omega)$



Low frequency value: β_F

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

Unity gain point, ω_t : $\omega_t \cong g_m/(C_\pi+C_\mu)$

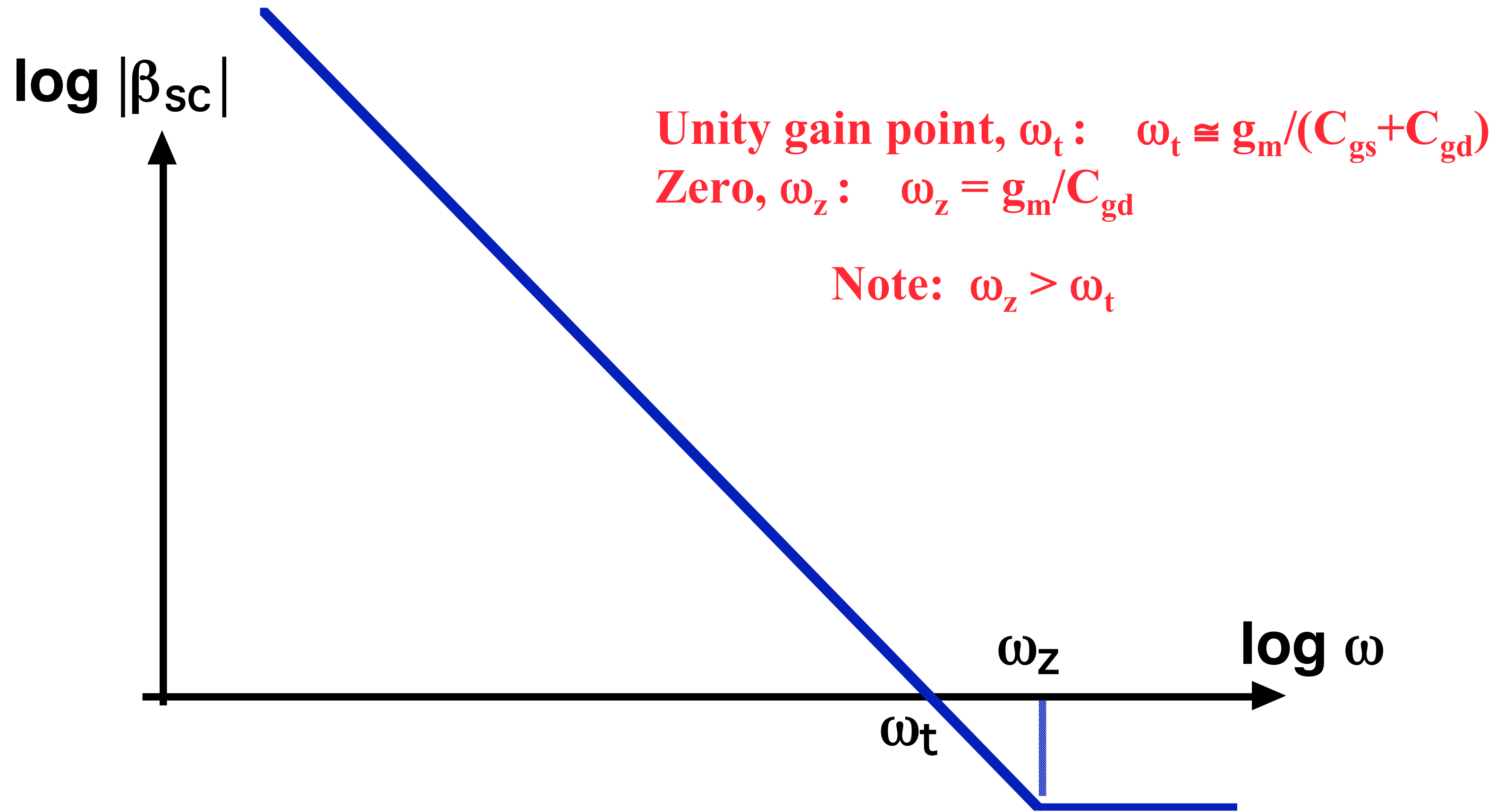
Zero, ω_z : $\omega_z = g_m/C_\mu$

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

$$\omega_t \approx \frac{g_m}{\left[C_\pi + C_\mu\right]} = \frac{g_m}{\left\{g_m \tau_{trB} + C_{eb,dp} + C_{cb,dp}\right\}} = \frac{1}{\left\{\tau_{trB} + kT\left[C_{eb,dp} + C_{cb,dp}\right]/qI_C\right\}}$$

In the limit of large I_C: $\omega_t(\text{BJT}) \approx \dfrac{1}{\tau_{trB}} = \dfrac{2D_{eB}}{w_B^2} = \dfrac{2\mu_e V_{thermal}}{w_B^2}$

MOSFET short-circuit current gain, $\beta_{sc}(j\omega)$



$$\omega_t(\text{MOSFET}) \approx \frac{g_m}{\left[C_{gs} + C_{gd}\right]} \approx \frac{g_m}{C_{gs}} = \frac{\frac{W}{L}\mu_e C_{ox}^*\left[V_{GS} - V_T\right]}{\frac{2}{3}WLC_{ox}^*} = \frac{3}{2}\frac{\mu_e\left[V_{GS} - V_T\right]}{L^2} \approx \frac{1}{\tau_{trCh}}$$

Looking more at ω_T for BJTs and MOSFETs:

For a MOSFET we have

$$\omega_t(\text{MOSFET}) \approx \frac{3}{2}\frac{\mu_e[V_{GS} - V_T]}{L^2}$$

The average E-field in the channel, E_y, is

$$E_y \approx \frac{[V_{GS} - V_T]}{L}$$

So we can also write ω_T as

$$\omega_t(\text{MOSFET}) \approx \frac{3}{2}\mu_e\frac{[V_{GS} - V_T]}{L}\frac{1}{L} \approx \frac{\bar{s}_y}{L} \approx \frac{1}{\tau_{trChannel}}$$

This is identical to the form we have for ω_T in a BJT

$$\omega_t(\text{BJT}) \approx \frac{1}{\tau_{trB}} = \frac{2D_{eB}}{w_B^2} = \frac{2\mu_e V_{thermal}}{w_B^2}$$

****

What happens when we have velocity saturation?

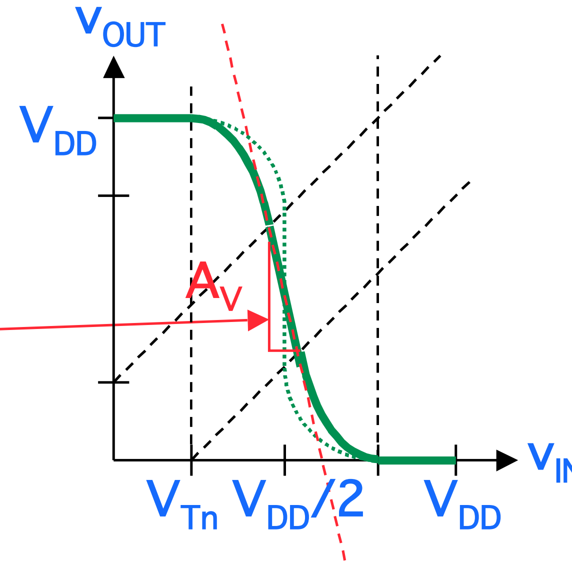
$$\omega_t(\text{MOSFET}) \approx \frac{s_{sat}}{L}, \qquad \omega_t(\text{BJT}) \approx \frac{s_{sat}}{w_B}$$

ω_T still decreases with L and w_B, but not as quickly!

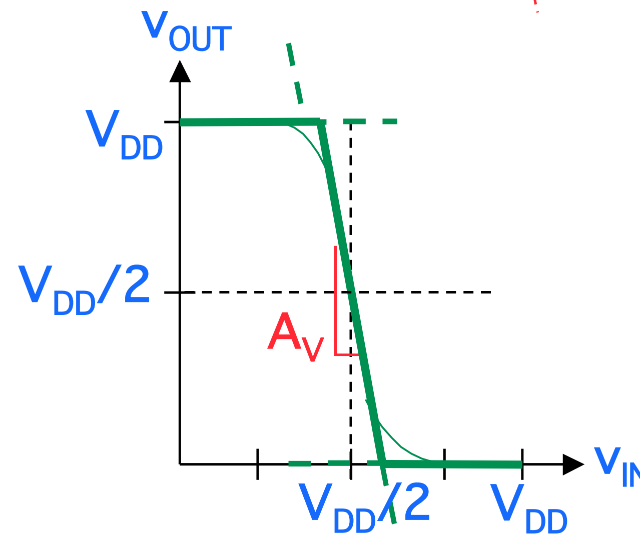
CMOS: transfer characteristic calculation, cont.

Returning to the transfer characteristic, we see that the slope in Region III is not infinite, but is instead:

$$A_v \equiv \left.\frac{\partial v_{OUT}}{\partial v_{IN}}\right|_Q = -\frac{2\sqrt{2K_n}}{\left[\lambda_n + \lambda_p\right]\sqrt{I_{Dn}}}$$



Final comment: A quick and dirty way to approximate the transfer curve of a CMOS gate is to simply draw the three straight line portions in Regions I, III, and V:



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 $L_n = L_p = L_{min}$, and $W_n = W_{min}$, so we also have $W_p = 2\,W_{min}$.

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

Charging cycle

v_{IN}: Hi to Lo; Q_n off, Q_p on; v_{OUT}: Lo to Hi

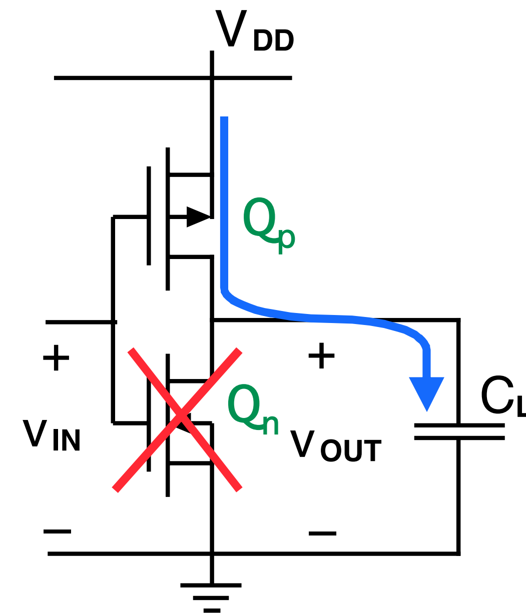
- Assume charged by constant $i_{D,sat}$

$$i_{Charge} = i_{Dp} \approx \frac{K_p}{2}\left[V_{DD} - \left|V_{Tp}\right|\right]^2 = \frac{K_n}{2}\left[V_{DD} - V_{Tn}\right]^2$$

$$q_{Charge} = C_L\,V_{DD}$$

$$\tau_{Charge} = \frac{q_{Charge}}{i_{Charge}} = \frac{2\,C_L\,V_{DD}}{K_n\left[V_{DD} - V_{Tn}\right]^2}$$

$$= \frac{6\,n\,W_{min}L_{min}C_{ox}^*V_{DD}}{\dfrac{W_{min}}{L_{min}}\mu_e C_{ox}^*\left[V_{DD} - V_{Tn}\right]^2} = \frac{6\,n\,L_{min}^2 V_{DD}}{\mu_e\left[V_{DD} - V_{Tn}\right]^2}$$

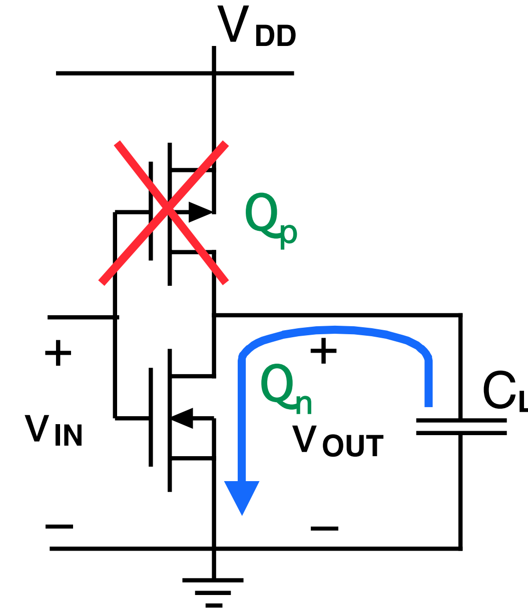


CMOS: switching speed; minimum cycle time, cont.

Discharging cycle

v_{IN}: Lo to Hi; Q_n on, Q_p off; v_{OUT}: Hi to Lo

- Assume discharged by constant $i_{D,sat}$



$$i_{Discharge} = i_{Dn} \approx \frac{K_n}{2}\left[V_{DD} - V_{Tn}\right]^2$$

$$q_{Discharge} = C_L V_{DD}$$

$$\tau_{Discharge} = \frac{q_{Discharge}}{i_{Discharge}} = \frac{2 C_L V_{DD}}{K_n\left[V_{DD} - V_{Tn}\right]^2}$$

$$= \frac{6 n W_{min} L_{min} C^*_{ox} V_{DD}}{\frac{W_{min}}{L_{min}} \mu_e C^*_{ox}\left[V_{DD} - V_{Tn}\right]^2} = \frac{6 n L^2_{min} V_{DD}}{\mu_e\left[V_{DD} - V_{Tn}\right]^2}$$

Minimum cycle time

v_{IN}: Lo to Hi to Lo; v_{OUT}: Hi to Lo to Hi

$$\tau_{Min.Cycle} = \tau_{Charge} + \tau_{Discharge} \approx \frac{12 n L^2_{min} V_{DD}}{\mu_e\left[V_{DD} - V_{Tn}\right]^2}$$

CMOS: power dissipation - total and per unit area

Average power dissipation

All dynamic

$$P_{ave} = C_L V_{DD}^2 f = 3 n W_{\min} L_{\min} C_{ox}^* V_{DD}^2 f$$

Average power at maximum data rate

Maximum f will be $1/\tau_{\text{Min Cycle}}$

$$P_{ave @ Max. f} = 3 n W_{\min} L_{\min} C_{ox}^* V_{DD}^2 \frac{\mu_e [V_{DD} - V_{Tn}]^2}{12 n L_{\min}^2 V_{DD}}$$

$$= \frac{1}{4} \frac{W_{\min}}{L_{\min}} \mu_e C_{ox}^* V_{DD} [V_{DD} - V_{Tn}]^2 = \frac{1}{4} K_n V_{DD} [V_{DD} - V_{Tn}]^2$$

Average power density at maximum data rate

Assume that the area per inverter will be proportional to $W_{\min} L_{\min}$

$$PD_{ave @ Max. f} = \frac{P_{ave @ Max. f}}{\text{Inverter area}} \propto \frac{P_{ave @ Max. f}}{W_{\min} L_{\min}} = \frac{\mu_e C_{ox}^* V_{DD} [V_{DD} - V_{Tn}]^2}{4 L_{\min}^2}$$

CMOS: design for high speed

Maximum data rate

Proportional to $1/\tau_{Min\ Cycle}$

$$f_{max} \propto 1/\tau_{Min.Cycle} = \frac{\mu_e [V_{DD} - V_{Tn}]^2}{12\, n\, L_{min}^2 V_{DD}}$$

Teaches us to make L_{min} small and/or V_{DD} large

Note: As we reduce L_{min} we must also reduce t_{ox}, but t_{ox} doesn't enter directly in f_{max} so it doesn't impact us here.

Average power density at maximum data rate

Assumes area per inverter is proportional to $W_{min}L_{min}$

$$P_{ave} = C_L V_{DD}^2 f \quad \text{and} \quad f_{max} \propto 1/\tau_{Min.Cycle} = \frac{\mu_e [V_{DD} - V_{Tn}]^2}{12\, n\, L_{min}^2 V_{DD}}$$

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

Teaches us PD increases very quickly as we reduce L_{min} (and t_{ox}) unless we also reduce V_{DD} (which reduces f_{max}).

How do we make f_{max} larger without melting the silicon?
Through CMOS scaling rules - the topic of today's lecture.

- **Summary:**

Transfer characteristic: symmetric

$$V_{LO} = 0, \quad V_{HI} = V_{DD}, \quad I_{ON} = 0, \quad I_{OFF} = 0$$

$$N_{ML} = N_{MH} \Rightarrow K_n = K_p \text{ and } |V_{Tp}| = V_{Tn}$$

$$L_n = L_p = L_{\min}, \quad W_p = (\mu_n/\mu_p) W_n, \quad W_n = W_{\min}$$

Gate delay expressions

$$\tau_{Charge} = \tau_{Discharge} \approx \frac{2V_{DD}C_L}{K_n [V_{DD} - V_{Tn}]^2}$$

$$C_L = n(W_n L_n + W_p L_p) C^*_{ox} = 3nW_{\min} L_{\min} C^*_{ox}$$ (Assumes $\mu_n = 2\mu_p$)

$$\tau_{Min.Cycle} = \tau_{Charge} + \tau_{Discharge} \approx \frac{12nL^2_{\min} V_{DD}}{\mu_e [V_{DD} - V_{Tn}]^2}$$

Average power at f_{max}, and Power Density (dissipation per unit area)

$$P_{ave@Max.f} = C_L V^2_{DD} f_{\max} \propto C_L V^2_{DD} / \tau_{Min.Cycle} = \frac{\mu_e W_{\min} C^*_{ox} V_{DD} [V_{DD} - V_{Tn}]^2}{L_{\min}}$$

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

(Note: we've used $C_{ox}^* = \varepsilon_{ox} t_{ox}$)

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:**

  $L_{min} \rightarrow L_{min}/s$ $\quad$ $w \rightarrow w/s$ $\quad$ $t_{ox} \rightarrow t_{ox}/s$

  this yields

  $K \rightarrow sK$ $\quad$ $C_{ox}^* \rightarrow sC_{ox}^*$ $\quad$ (Also stay in grad. channel regime)

  and ultimately

  $\tau \rightarrow \tau/s^2$ $\quad$ $P_{ave} \rightarrow s\,P_{ave}$ $\quad$ $P_{density} \rightarrow s^3\,P_{density}$!!!

  Scaling dimensions alone can yield melted silicon!!

- **Scaling dimensions and voltages in concert:**

  Add: $V_{DD} \rightarrow V_{DD}/s$ $\quad$ $V_T \rightarrow V_T/s$

  still have

  $K \rightarrow sK$ $\quad$ $C_{ox}^* \rightarrow sC_{ox}^*$

  but now

  $\tau \rightarrow \tau/s$ $\quad$ $P_{ave} \rightarrow P_{ave}/s^2$ $\quad$ $P_{density} \rightarrow P_{density}$

  When scale both get: faster, lower power, same power density!!
  Note: scaling the voltages is not as easy as scaling the dimensions.

Lecture 25 - CMOS Scaling Rules - Summary

- **CMOS gate delay and power**

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

$$\tau_{Min.Cycle} = 12\, n\, L_{\min}^2 V_{DD} \big/ \mu_e \left[V_{DD} - V_{Tn}\right]^2$$

$$P_{ave\,@\,Max.f} = C_L V_{DD}^2 f_{\max} \propto \mu_e C_{ox}^* V_{DD}\left[V_{DD} - V_{Tn}\right]^2 \left[W_{\min} / L_{\min}\right]$$

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

- **CMOS scaling rules**

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

Scaling as:

$$L_{\min} \rightarrow L_{\min}/s$$
$$W_{\min} \rightarrow W_{\min}/s$$
$$t_{ox} \rightarrow t_{ox}/s$$
$$V_{DD} \rightarrow V_{DD}/s$$
$$V_T \rightarrow V_T/s$$

Results in:

$$K \rightarrow sK$$
$$C_{ox}^* \rightarrow s\,C_{ox}^*$$
$$\tau_{Min.cycle} \rightarrow \tau_{Min.cycle}/s$$
$$P_{ave} \rightarrow P_{ave}/s^2$$
$$PD_{ave} \rightarrow PD_{ave}$$