

Lecture 16 - CMOS scaling; The Roadmap - Outline

- **Announcements**

 - PS #9 - Will be due next week Friday; no recitation tomorrow.

 - Postings - CMOS scaling (multiple items)

 - Exam Two - Tonight, Nov. 5, 7:30-9:30 pm

- **Review - CMOS gate delay and power**

 - Lecture 15 results: Gate Delay = $12 n L_{\min}^2 V_{DD} / \mu_n (V_{DD} - V_T)^2$

- $P_{\text{dyn}} @ f_{\text{max}} \propto C_L V_{DD}^2 / \text{GD} = K_n V_{DD} (V_{DD} - V_T)^2 / 4$

 - Velocity Saturation

- **CMOS scaling rules**

 - Power density issues and challenges

 - Approaches to a solution: Dimension scaling alone
Scaling voltages as well

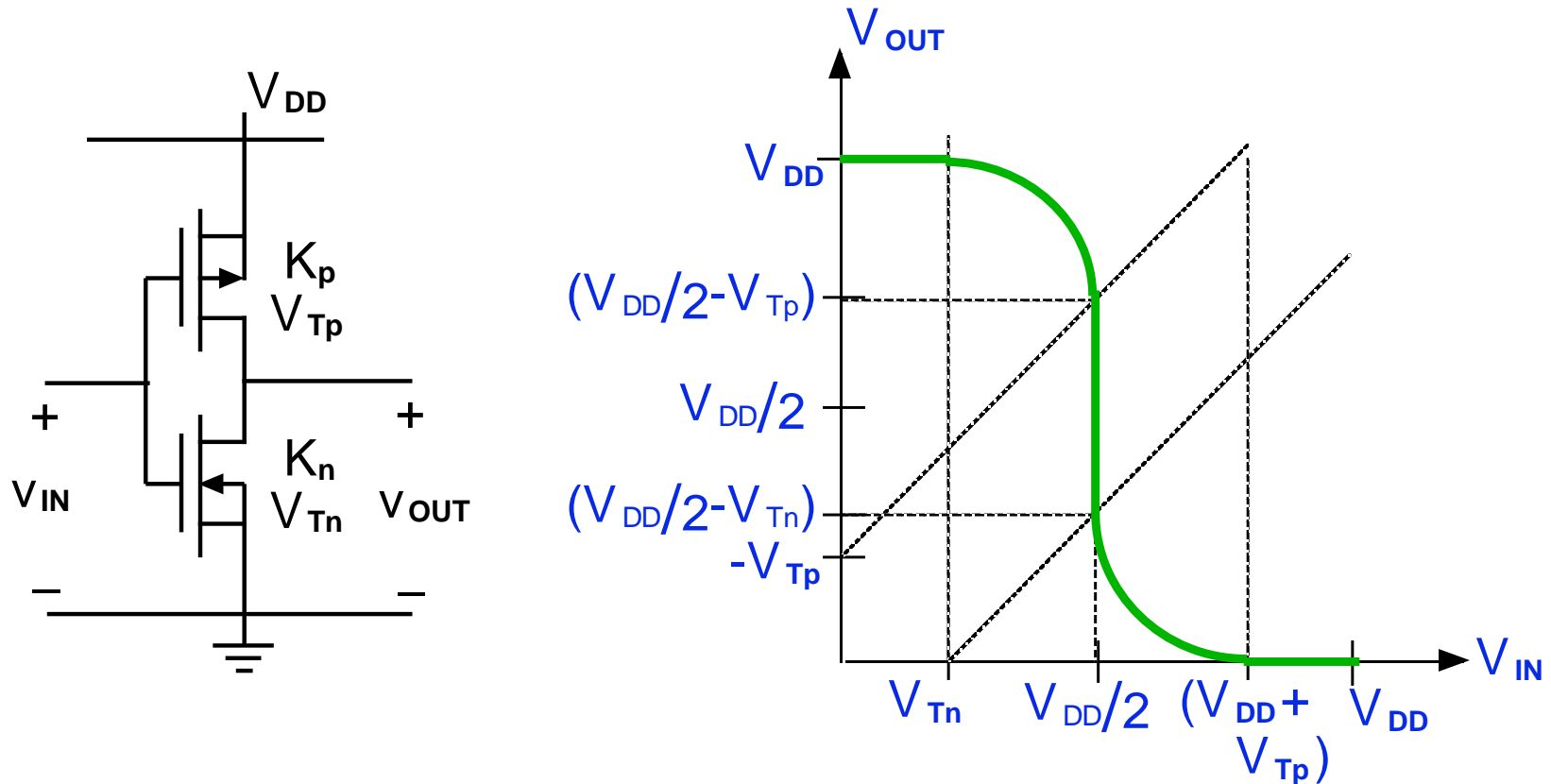
- **The Road Map; the Future**

 - Size and performance evolution with time

 - How long can it go on?

CMOS: transfer characteristic

Complete characteristic w.o. Early effect:



NOTE: We design CMOS inverters to have $K_n = K_p$ and $V_{Tn} = -V_{Tp}$ to obtain the optimum symmetrical characteristic.

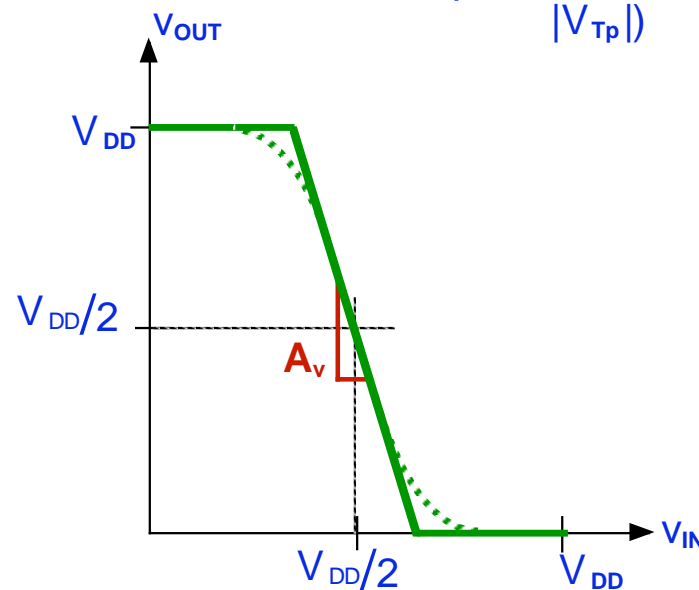
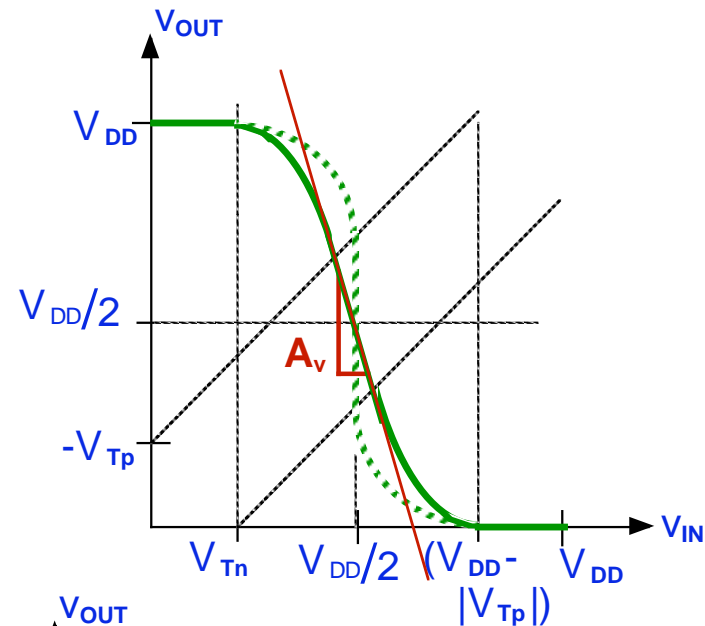
CMOS: transfer characteristic calculation, cont.

We found from an LEC analysis that the slope in Region III is not infinite, but is instead:

$$A_v \equiv \frac{v_{out}}{v_{in}} = \left. \frac{\partial v_{OUT}}{\partial v_{IN}} \right|_{Q(=V_{DD}/2, V_{DD}/2)}$$

$$= - \frac{[g_{mn} + g_{mp}]}{[g_{on} + g_{op}]} = - \frac{2\sqrt{2K_n}}{[\lambda_n + \lambda_p]\sqrt{I_{Dn}}}$$

Quick approximation: An easy way to sketch the transfer characteristic 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 $W_p = 2W_{\min}$

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

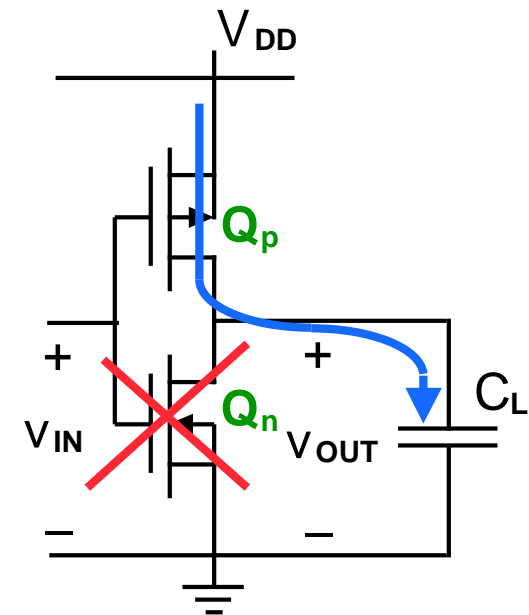
Charging cycle: v_{IN} : HI to LO; Q_n off, Q_p on; v_{OUT} : LO to HI

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

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

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

$$\begin{aligned} \tau_{Charge} &= \frac{q_{Charge}}{i_{Charge}} = \frac{2C_L V_{DD}}{K_n [V_{DD} - V_{Tn}]^2} \\ &= \frac{6nW_{\min} L_{\min} C_{ox}^* V_{DD}}{\frac{W_{\min}}{L_{\min}} \mu_e C_{ox}^* [V_{DD} - V_{Tn}]^2} = \frac{6nL_{\min}^2 V_{DD}}{\mu_e [V_{DD} - V_{Tn}]^2} \end{aligned}$$



CMOS: switching speed; minimum cycle time, cont.

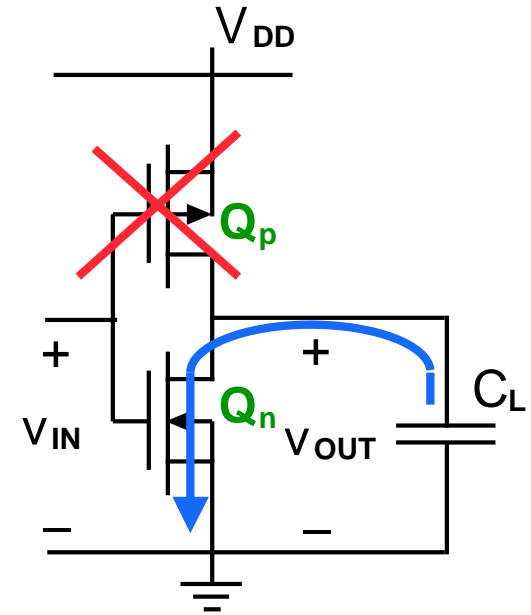
Discharging cycle: v_{IN} : LO to HI; Q_n on, Q_p off; v_{OUT} : HI to LO

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

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

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

$$\begin{aligned} \tau_{Discharge} &= \frac{q_{Discharge}}{i_{Discharge}} = \frac{2C_L V_{DD}}{K_n [V_{DD} - V_{Tn}]^2} \\ &= \frac{6nW_{min} L_{min} C_{ox}^* V_{DD}}{\frac{W_{min}}{L_{min}} \mu_e C_{ox}^* [V_{DD} - V_{Tn}]^2} = \frac{6nL_{min}^2 V_{DD}}{\mu_e [V_{DD} - V_{Tn}]^2} \end{aligned}$$



Minimum cycle time: v_{IN} : LO to HI to LO; v_{OUT} : HI to LO to HI

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

CMOS: switching speed; minimum cycle time, cont.

Discharging and Charging times:

What do the expressions tell us? We have

$$\tau_{MinCycle} = \frac{12nL_{min}^2 V_{DD}}{\mu_e [V_{DD} - V_{Tn}]^2}$$

This can be written as:

$$\tau_{MinCycle} = \frac{12nV_{DD}}{(V_{DD} - V_{Tn})} \cdot \frac{L_{min}}{\mu_e (V_{DD} - V_{Tn})/L_{min}}$$

The last term is the channel transit time:

$$\frac{L_{min}}{\mu_e (V_{DD} - V_{Tn})/L_{min}} = \frac{L_{min}}{\mu_e E_{Ch}} = \frac{L_{min}}{\bar{s}_{e,Ch}} = \tau_{ChTransit}$$

Thus the gate delay is a multiple of the channel transit time:

$$\tau_{MinCycle} = \frac{12nV_{DD}}{(V_{DD} - V_{Tn})} \tau_{ChannelTransit} = n' \tau_{ChannelTransit}$$

CMOS: power dissipation - total and per unit area

Average power dissipation

Only dynamic for now

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

Power at maximum data rate

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

$$\begin{aligned} P_{dyn@f_{max}} &= \frac{3nW_{min} L_{min} C_{ox}^* V_{DD}^2}{\tau_{Min.Cycle}} = 3nW_{min} L_{min} C_{ox}^* V_{DD}^2 \cdot \frac{\mu_e [V_{DD} - V_{Tn}]^2}{12nL_{min}^2 V_{DD}} \\ &= \frac{1}{4} \frac{W_{min}}{L_{min}} \mu_e C_{ox}^* V_{DD} [V_{DD} - V_{Tn}]^2 \end{aligned}$$

Power density at maximum data rate

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

$$PD_{dyn@f_{max}} = \frac{P_{dyn@f_{max}}}{InverterArea} \propto \frac{P_{dyn@f_{max}}}{W_{min} L_{min}} = \frac{\mu_e C_{ox}^* V_{DD} [V_{DD} - V_{Tn}]^2}{L_{min}^2}$$

CMOS: design for high speed

Maximum data rate

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

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

Implies we should reduce L_{min} and increase V_{DD} .

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

Power density at maximum data rate

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

$$PD_{\text{dyn}@f_{\text{max}}} \propto \frac{P_{\text{dyn}@f_{\text{max}}}}{W_{\text{min}}L_{\text{min}}} = \frac{\mu_e \epsilon_{\text{ox}} V_{DD} [V_{DD} - V_{Tn}]^2}{t_{\text{ox}} L_{\text{min}}^2}$$

Shows us that PD increases very quickly as we reduce L_{min} unless we also reduce V_{DD} (which will also reduce f_{max}).

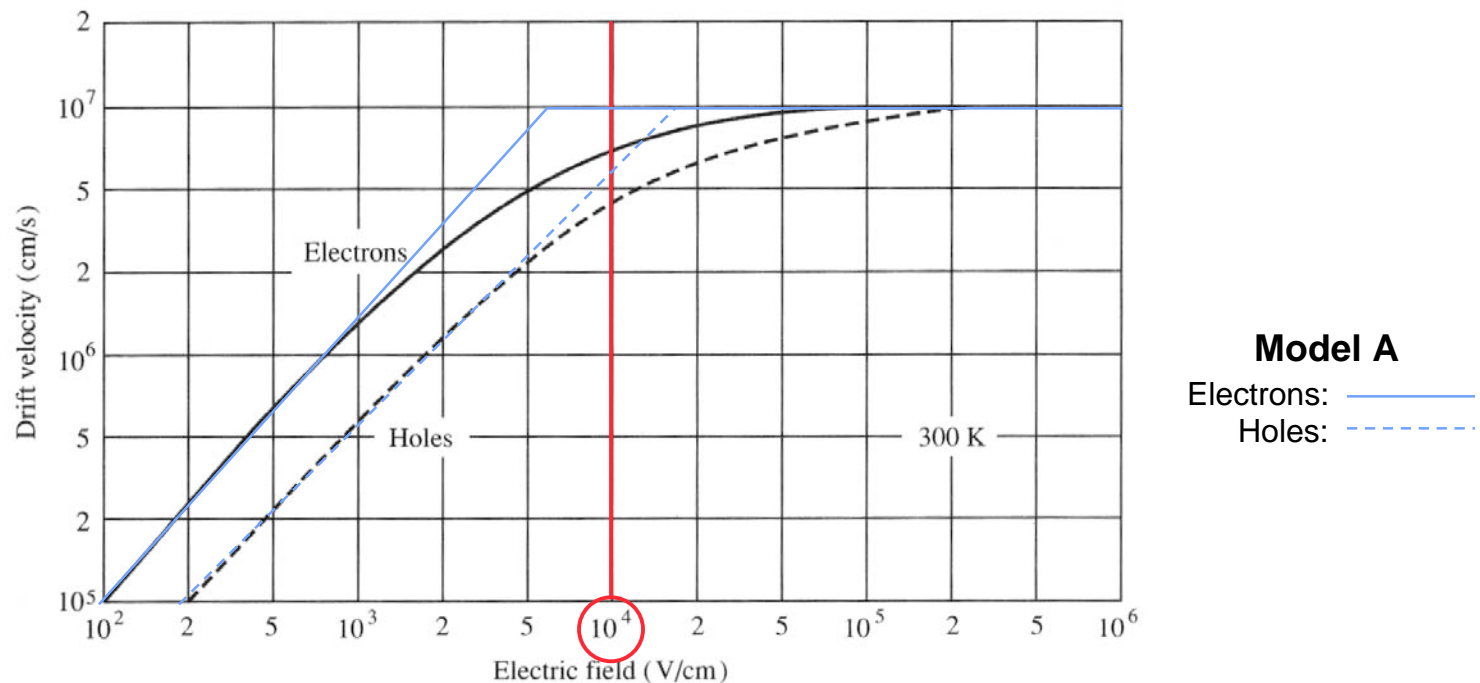
Note: Now t_{ox} does appear in the expression, so the rate of increase with decreasing L_{min} is even greater because t_{ox} must be reduced along with L to stay in the gradual channel regime.

How do we make f_{max} larger without melting the silicon?

CMOS: velocity saturation

Sanity check before looking at device scaling

CMOS gate lengths are now under $0.1 \mu\text{m}$ (100 nm). The electric field in the channel can be very high: $E_y \geq 10^4 \text{ V/cm}$ when $v_{DS} \geq 0.1 \text{ V}$.



Clearly the velocity of the electrons and holes in the channel will be saturated at even low values of v_{DS} !

What does this mean for the device and inverter characteristics?

MOS: Output family with velocity saturation



$$i_D(v_{GS}, v_{DS}, v_{BS}) \approx \begin{cases} 0 & \text{for } v_{GS} < V_T, 0 < v_{DS} & \text{Cutoff} \\ W s_{sat} C_{ox}^* [v_{GS} - V_T(v_{BS})] & \text{for } V_T < v_{GS}, E_{crit} L < v_{DS} & \text{Saturation} \\ \frac{W}{L} \mu_e C_{ox}^* [v_{GS} - V_T(v_{BS})] v_{DS} & \text{for } V_T < v_{GS}, 0 < v_{DS} < E_{crit} L & \text{Linear} \end{cases}$$

This simple model for the output characteristics of a very short channel MOSFET (plotted above) provides us an easy way to understand the impact of velocity saturation on MOSFET and CMOS inverter performance.

CMOS: Gate delay and f_{\max} with velocity saturation

Charge/discharge cycle and gate delay:

The charge and discharge currents, charges, and times are now:

$$i_{Discharge} = i_{Charge} = W_{\min} s_{sat} C_{ox}^* (V_{DD} - V_{Tn})$$

$$q_{Discharge} = q_{Charge} = C_L V_{DD} = 3W_{\min} L_{\min} C_{ox}^* V_{DD}$$

$$\tau_{Discharge} = \tau_{Charge} = \frac{q_{Discharge}}{i_{Discharge}} = \frac{3W_{\min} L_{\min} C_{ox}^* V_{DD}}{W_{\min} s_{sat} C_{ox}^* (V_{DD} - V_{Tn})} = \frac{3nL_{\min} V_{DD}}{s_{sat} (V_{DD} - V_{Tn})}$$

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

$$\tau_{Min.Cycle} = \tau_{Charge} + \tau_{Discharge} = \frac{6nL_{\min} V_{DD}}{s_{sat} [V_{DD} - V_{Tn}]} \quad \text{Note: } \tau_{ChanTransit} = \frac{L}{s_{sat}}$$

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

Lessons: We still benefit from reducing L, but not as quickly.
Channel transit time, L_{\min}/s_{sat} , is still critical.

CMOS: Power and power density with velocity saturation

Average power dissipation

All dynamic

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

Power at maximum data rate

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

$$\begin{aligned} P_{dyn@f_{max}} &= \frac{3nW_{min} L_{min} C_{ox}^* V_{DD}^2}{\tau_{Min.Cycle}} = 3nW_{min} L_{min} C_{ox}^* V_{DD}^2 \cdot \frac{s_{sat} [V_{DD} - V_{Tn}]}{6n L_{min} V_{DD}} \\ &= \frac{1}{2} W_{min} s_{sat} C_{ox}^* V_{DD} [V_{DD} - V_{Tn}] \end{aligned}$$

Power density at maximum data rate

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

$$PD_{dyn@f_{max}} = \frac{P_{dyn@f_{max}}}{InverterArea} \propto \frac{P_{dyn@f_{max}}}{W_{min} L_{min}} = \frac{s_{sat} C_{ox}^* V_{DD} [V_{DD} - V_{Tn}]}{L_{min}}$$

Lesson: Again benefit from reducing L, but again not as quickly.

CMOS: Collected results

Maximum data rate:

No velocity saturation:

$$\tau_{Min.Cycle} \propto \frac{L_{min}^2 V_{DD}}{\mu_e [V_{DD} - V_{Tn}]^2}$$

With velocity saturation:

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

Smaller
is faster

Power density at maximum data rate:

No velocity saturation:

$$PD_{dyn @ f_{max}} = \frac{\mu_e \epsilon_{ox} V_{DD} [V_{DD} - V_{Tn}]^2}{t_{ox} L_{min}^2}$$

With velocity saturation:

$$PD_{dyn @ f_{max}} = \frac{S_{sat} \epsilon_{ox} V_{DD} [V_{DD} - V_{Tn}]}{t_{ox} L_{min}}$$

Smaller also
dissipates
more power
per unit area

Scaling Rules - making CMOS faster without melting Si

General idea:

Reduce dimensions by factor 1/s: $s > 1$

Evaluate impact on speed, power, power density

Assume no velocity saturation for now

Scaling dimensions alone:

$$L_{\min} \rightarrow L_{\min}/s \quad W \rightarrow W/s \quad t_{ox} \rightarrow t_{ox}/s \quad N_A \rightarrow sN_A$$

This yields

$$C_{ox}^* = \frac{\epsilon_{ox}}{t_{ox}} : C_{ox}^* \rightarrow sC_{ox}^* \quad K = \frac{W}{L} \mu_e C_{ox}^* : K \rightarrow sK$$

and thus

$$\tau \propto \frac{L_{\min}^2 V_{DD}}{\mu_e [V_{DD} - V_{Tn}]^2} : \tau \rightarrow \tau/s^2$$

$$P_{dyn} = 3nW_{\min} L_{\min} C_{ox}^* V_{DD}^2 f : P_{dyn} \rightarrow sP_{dyn}$$

$$PD_{dyn @ f_{\max}} = \frac{\mu_e \epsilon_{ox} V_{DD} [V_{DD} - V_{Tn}]^2}{t_{ox} L_{\min}^2} : PD_{dyn @ f_{\max}} \rightarrow s^3 PD_{dyn @ f_{\max}}$$

Scaling Rules, cont. - constant E-field scaling

Observation:

Reducing dimensions alone won't work.

Reduce voltage in concert (constant E-field scaling)

Scaling dimensions and voltages by 1/s:

$$L_{\min} \rightarrow L_{\min}/s \quad W \rightarrow W/s \quad t_{ox} \rightarrow t_{ox}/s \quad N_A \rightarrow sN_A$$

$$V_{DD} \rightarrow V_{DD}/s \quad V_{BS} \rightarrow V_{BS}/s \quad V_T \rightarrow V_T/s$$

We still have

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

but now we find

$$\tau \propto \frac{L_{\min}^2 V_{DD}}{\mu_e [V_{DD} - V_{Tn}]^2} : \quad \tau \rightarrow \tau/s$$

$$P_{dyn} = 3nW_{\min} L_{\min} C_{ox}^* V_{DD}^2 f : \quad P_{dyn} \rightarrow P_{dyn}/s^2$$

$$PD_{dyn @ f_{\max}} = \frac{\mu_e \epsilon_{ox} V_{DD} [V_{DD} - V_{Tn}]^2}{t_{ox} L_{\min}^2} : \quad PD_{dyn @ f_{\max}} \rightarrow PD_{dyn @ f_{\max}}$$

When we scale dimension and voltage we get higher speed and lower power, while holding the power density unchanged.

Scaling Rules, cont. - constant E-field scaling

Threshold voltage:

We've said V_T scales, but this merits some discussion*:

$$V_T(v_{BS}) \equiv V_{FB} + \underbrace{|2\phi_{p-Si}|}_{\text{Small because with n+ poly Si gate, } \phi_m \approx -\phi_p \text{ and } V_{FB} \approx -|2\phi_p|} + \frac{t_{ox}}{\epsilon_{ox}} \sqrt{2\epsilon_{Si}qN_A \left[\underbrace{|2\phi_{p-Si}| + |v_{BS}|}_{\text{Dominated by } |v_{BS}| \text{ if } |v_{BS}| \gg |2\phi_p|} \right]}$$

Small because with n+ poly Si gate, $\phi_m \approx -\phi_p$ and $V_{FB} \approx -|2\phi_p|$

Dominated by $|v_{BS}|$ if $|v_{BS}| \gg |2\phi_p|$

Thus:

$$V_T(v_{BS}) \approx \frac{t_{ox}}{\epsilon_{ox}} \sqrt{2\epsilon_{Si}qN_A |v_{BS}|} \rightarrow \frac{t_{ox}/s}{\epsilon_{ox}} \sqrt{2\epsilon_{Si}q s N_A |v_{BS}|/s} \rightarrow V_T/s$$

It works.

Subthreshold leakage and static power:

Including V_{BS} , I_{Doff} is:

$$I_{D,off} \approx \frac{W}{L} \mu_e V_t^2 \sqrt{\frac{\epsilon_{Si}qN_A}{2[-|2\phi_p| + |V_{BS}|]}} e^{\{-V_T\}/nV_t} \approx \frac{W}{L} \mu_e V_t^2 \sqrt{\frac{\epsilon_{Si}qN_A}{2|V_{BS}|}} e^{\{-V_T\}/nV_t}$$

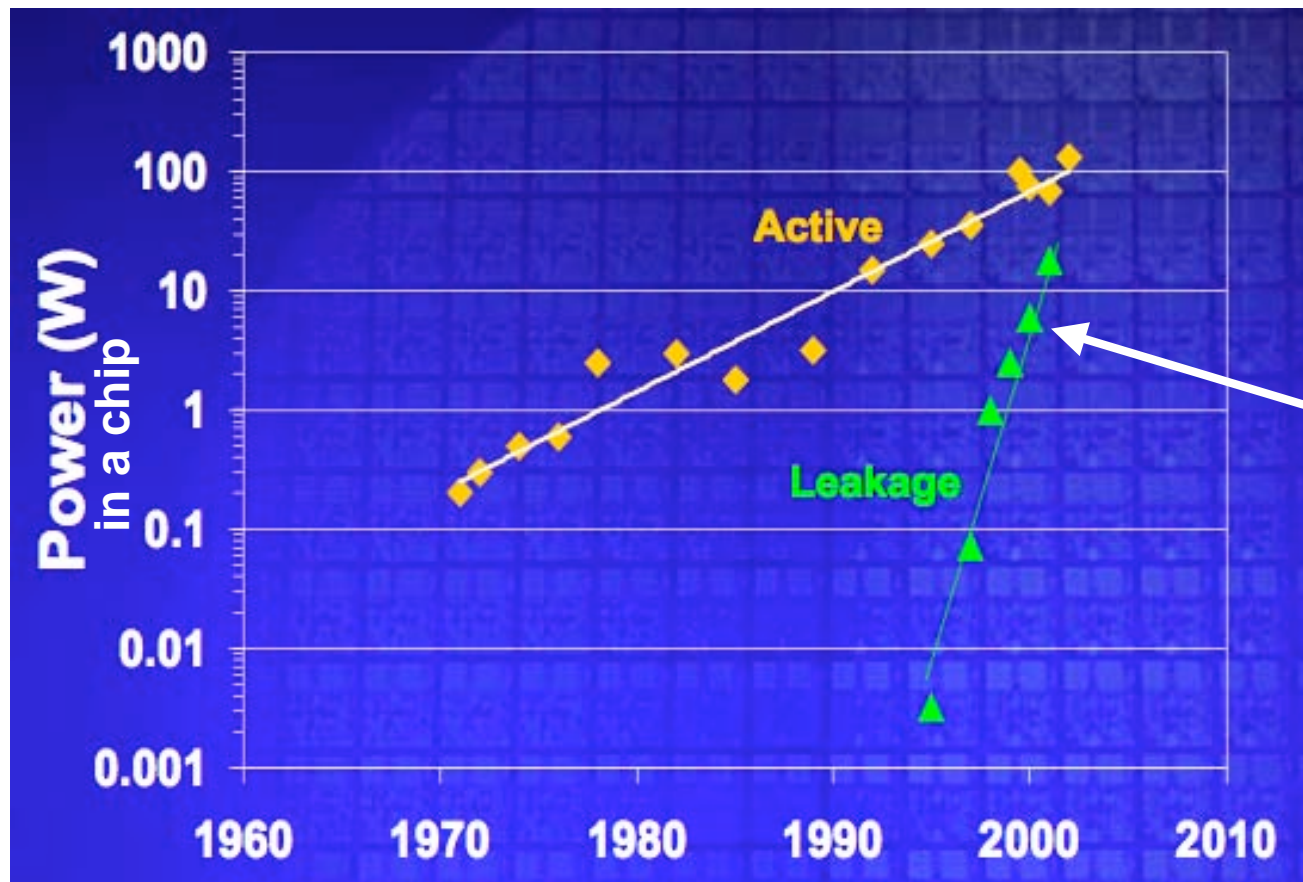
Scaling all the factors, we find that I_{Doff} and P_{static} scale poorly!

$$I_{D,off} \rightarrow s I_{D,off} e^{\left\{ \left(1 - \frac{1}{s}\right) V_T \right\} / nV_t} \quad P_{Static} = V_{DD} I_{D,off} \rightarrow P_{Static} e^{\left\{ \left(1 - \frac{1}{s}\right) V_T \right\} / nV_t}$$

Scaling Rules, cont. - static power scales badly, but...

Static power density's scaling is even worse:

$$PD_{static} = \frac{I_{D,off} V_{DD}}{W_{min} L_{min}} \rightarrow \frac{s I_{D,off} e^{(s-1)V_T/snV_t} V_{DD}/s}{W_{min} L_{min}/s^2} \rightarrow s^2 e^{(s-1)V_T/snV_t} PD_{static}$$



A typical V_T/nV_t is ~10. If $s = \sqrt{2}$, the exponential factor is $\sim e^3$, or about 20!

Bottom Line: Static power can no longer be neglected.

Figure source: Intel Web Site

Scaling Rules, cont. - What about velocity saturation?

Do the same constant E-field scaling by 1/s:

$$L_{\min} \rightarrow L_{\min}/s \quad W \rightarrow W/s \quad t_{ox} \rightarrow t_{ox}/s \quad N_A \rightarrow sN_A$$

$$V_{DD} \rightarrow V_{DD}/s \quad V_{BS} \rightarrow V_{BS}/s \quad V_T \rightarrow V_T/s$$

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

Examining our expressions when velocity saturation is important we find:

$$\tau \propto \frac{L_{\min} V_{DD}}{s_{sat} [V_{DD} - V_{Tn}]} : \quad \tau \rightarrow \tau/s$$

$$P_{dyn} = 3nW_{\min} L_{\min} C_{ox}^* V_{DD}^2 f : \quad P_{dyn} \rightarrow P_{dyn}/s^2$$

$$PD_{dyn @ f_{\max}} = \frac{s_{sat} \epsilon_{ox} V_{DD} [V_{DD} - V_{Tn}]}{t_{ox} L_{\min}} : \quad PD_{dyn @ f_{\max}} \rightarrow PD_{dyn @ f_{\max}}$$

Amazingly, there is no difference in the scaling behavior of the gate delay, average power, or power density in this case!

An historical scaling example - Inside Intel

<u>Parameter</u>	<u>386</u>	<u>486</u>	<u>Pentium</u>
Scaling factor, s	1	2	3
L_{\min} (μm)	1.5	0.75	0.5
W_n (μm)	10	5	3
t_{ox} (nm)	30	15	9
V_{DD} (V)	5	3.3	2.2
V_T (V)	1	-	-
Fan out	3	3	3
K ($\mu\text{A}/\text{V}^2$)	230	450	600
GD (ps)	840	400	250
f_{max} (MHz)	29	50	100
$P_{\text{ave}}/\text{gate}$ (mW)	92	23	10
Density (kgates/cm ² @ 20W/cm ² max.)	220	880	2,000

An second look inside Intel - a slightly different perspective

<u>Parameter</u>	<u>486</u>	<u>Pentium generations</u>		
Scaling factor, s	-	1	1.6	2.3
L_{\min} (μm)	1.0	0.8	0.5	0.35
SRAM cell area (μm^2)	-	111	44	21
Die size (mm^2)	170	295	163	91
f_{mzx} (MHz)	38	66	100	200
t_{ox} (nm)	20	10	8	6
Metal layers	2	3	4	4
Planarization	SOG	CMP	CMP	CMP
Poly type	n	n,p	n,p	n,p
Transistors	CMOS	BiCMOS	BiCMOS	BiCMOS

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

Moore's Law - Everything* doubles every 2 years.

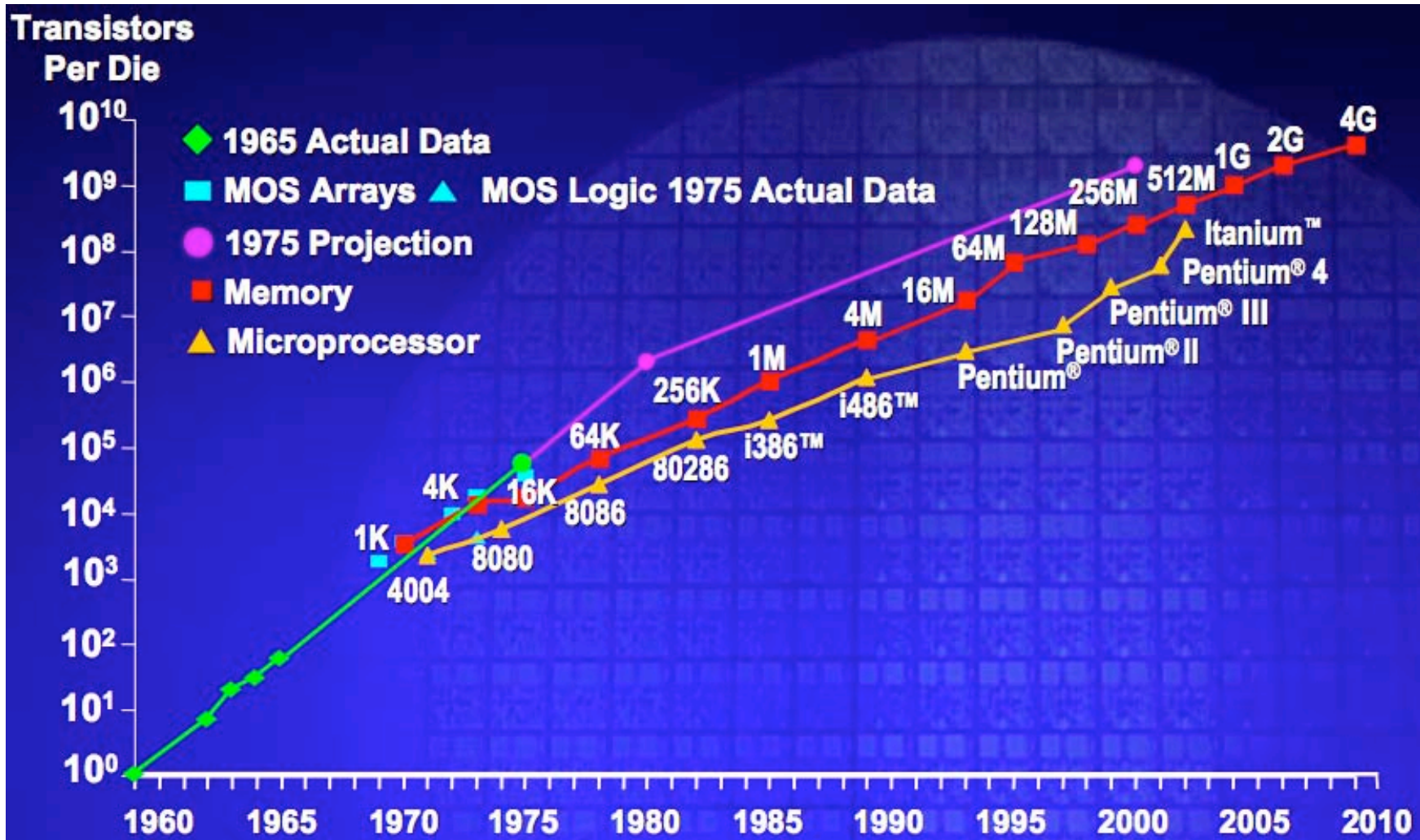


Figure source: Intel Web Site

* Density, speed, performance, transistors per chip, transistors shipped, transistors per cent, revenues, etc. First stated in 1965 as every year; revised to every 2 years in 1975.

Lecture 16 - CMOS scaling; The Roadmap - Summary

- **CMOS gate delay and power**

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

$$\text{Gate Delay} = 12 n L_{\min}^2 V_{DD} / \mu_e (V_{DD} - V_T)^2$$

$$P_{\text{dyn}@f_{\max}} \propto C_L V_{DD}^2 / \text{GD} = (W_n / L_{\min}) \mu_e C_{\text{ox}}^* V_{DD} (V_{DD} - V_T)^2 / 4$$

$$PD_{\text{dyn,max}} \propto P_{\text{dyn}@f_{\max}} / W_n L_{\min} = \mu_e \epsilon_{\text{ox}} V_{DD} (V_{DD} - V_T)^2 / 4 t_{\text{ox}} L_{\min}^2$$

- **CMOS scaling rules**

Summary of rules: Constant E-field - scale all dimensions and all voltages by 1/s

Scaling as: $L_{\min} \rightarrow L_{\min}/s$

$$w \rightarrow w/s$$

$$t_{\text{ox}} \rightarrow t_{\text{ox}}/s$$

$$N_A \rightarrow s N_A$$

$$V_T, V_{BS}, V_{DD} \rightarrow V_T/s, V_{BS}/s, V_{DD}/s$$

Results in: $K \rightarrow sK$

$$C_{\text{ox}}^* \rightarrow s C_{\text{ox}}^*$$

$$\tau \rightarrow \tau/s$$

$$P_{\text{dyn}} \rightarrow P_{\text{dyn}}/s^2$$

$$PD_{\text{dyn}} \rightarrow PD_{\text{dyn}}$$

- **The Roadmap; what's next?**

Stay tuned: 3-D; new semiconductors; performance over size

MIT OpenCourseWare
<http://ocw.mit.edu>

6.012 Microelectronic Devices and Circuits
Fall 2009

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.