Week 2:

MOS RC Model, Delay and Power, CMOS Scaling

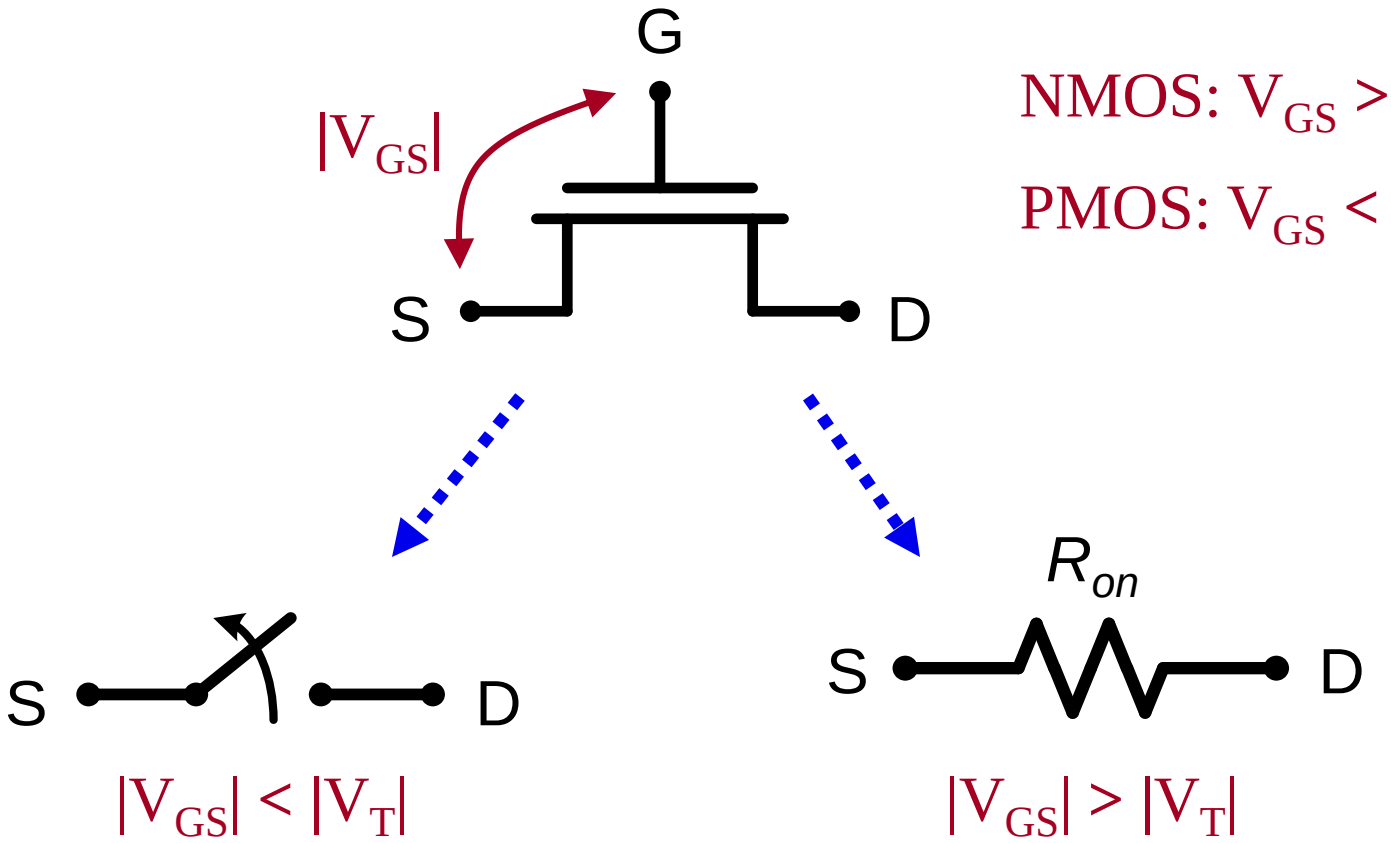
Prof. Dejan Markovic

UCLA

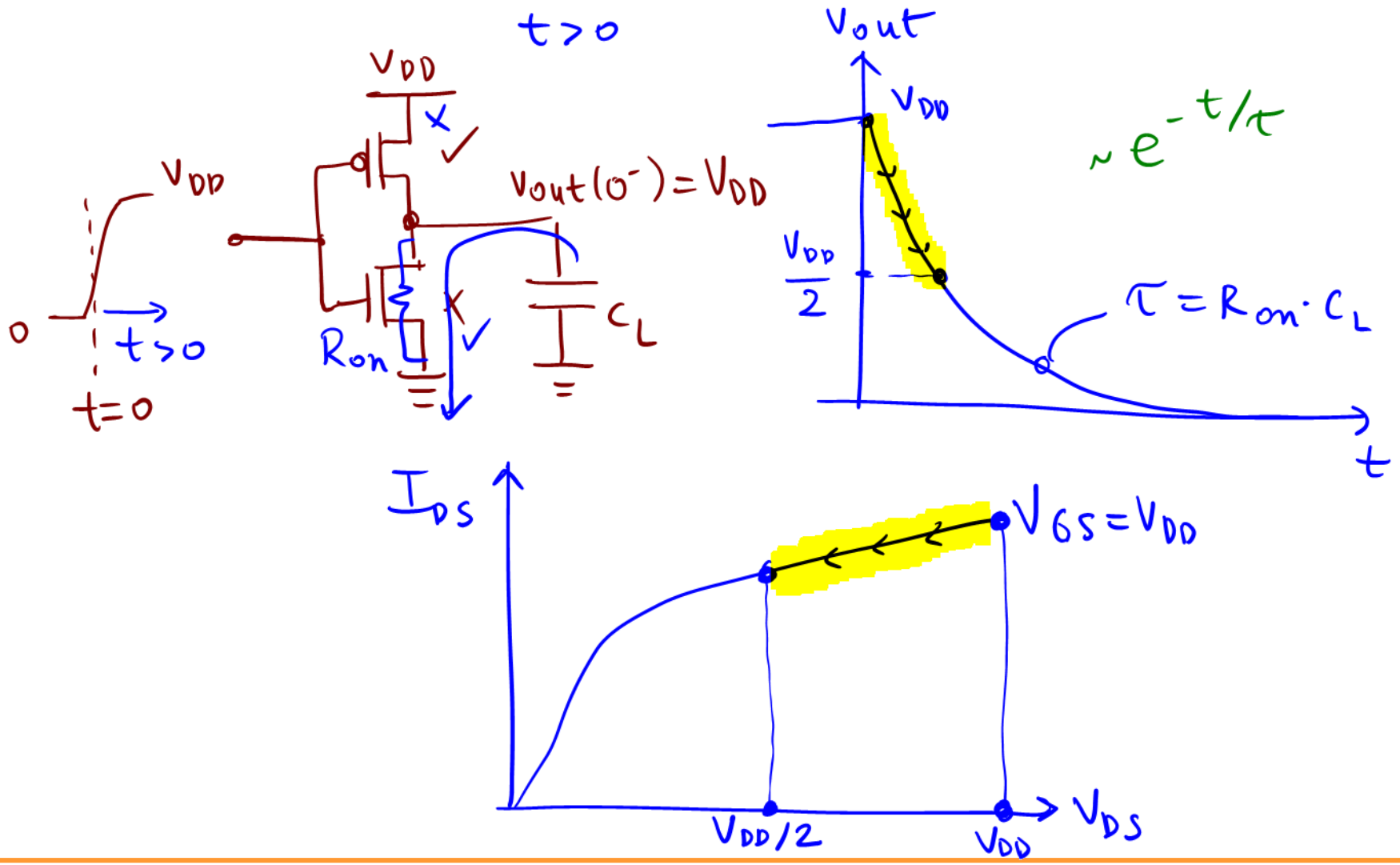
Week 2 Agenda

- MOS RC Model
- Delay Model
- Power Model
- CMOS Scaling

Switch Model of CMOS Transistor



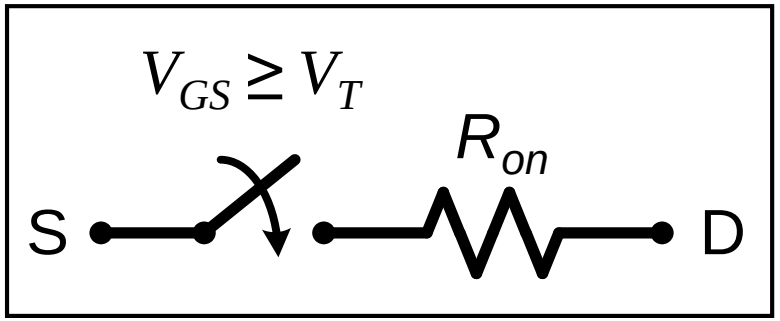
Switching Behavior



The Transistor as a Switch

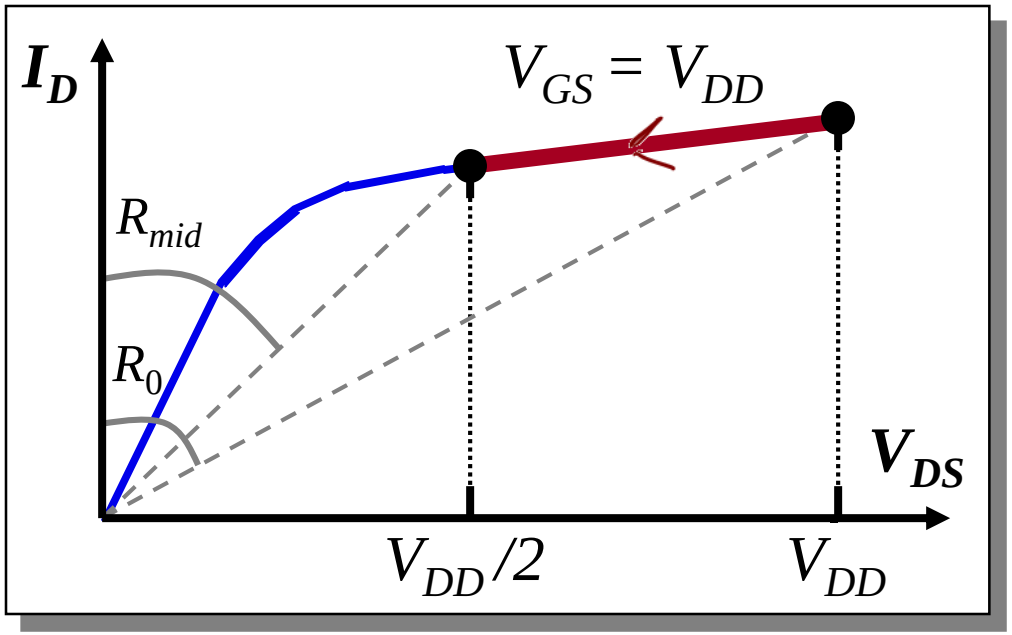
- **MOS can be treated as equivalent resistance**

Calculating MOS resistance: $$R(V_{DS}) = \frac{V_{DS}}{I_{DSAT} \cdot (1 + \lambda V_{DS})}$$



$$R_{on} \approx \frac{1}{2}(R_0 + R_{mid})$$

good approximation (I-V ≈ linear)



- **This model will be used for delay analysis**

Computing Equivalent Resistance (1/2)

- **Method 1 ("exact"): by integration**

$$R_{on} = \frac{1}{-V_{DD}/2}\int_{V_{DD}}^{V_{DD}/2} \frac{V}{I_{DSAT}\cdot(1+\lambda V)}dV \ldots \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1-\frac{7}{9}\lambda V_{DD}\right)$$

$$k'\frac{W}{L}\left[(V_{DD}-V_T)V_{DSAT}-\frac{V_{DSAT}^2}{2}\right]$$

V_{GS}

$$\boxed{R_{on} \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1-\frac{7}{9}\lambda V_{DD}\right)}$$

Computing Equivalent Resistance (2/2)

- **Method 2: simple averaging**
  - The averaging works because of approximately linear dependence of I_{DS} on V_{DS} (recall the CLM model)

$$R_{on} = \frac{1}{2}\left(\underbrace{\frac{V_{DD}}{I_{DSAT}\cdot(1+\lambda V_{DD})}}_{R_0} + \underbrace{\frac{V_{DD}/2}{I_{DSAT}\cdot(1+\lambda V_{DD}/2)}}_{R_{mid}}\right)$$

$$k'\frac{W}{L}\left[(V_{DD}-V_T)V_{DSAT} - \frac{V_{DSAT}^2}{2}\right]$$

(V_{GS})

fixed L

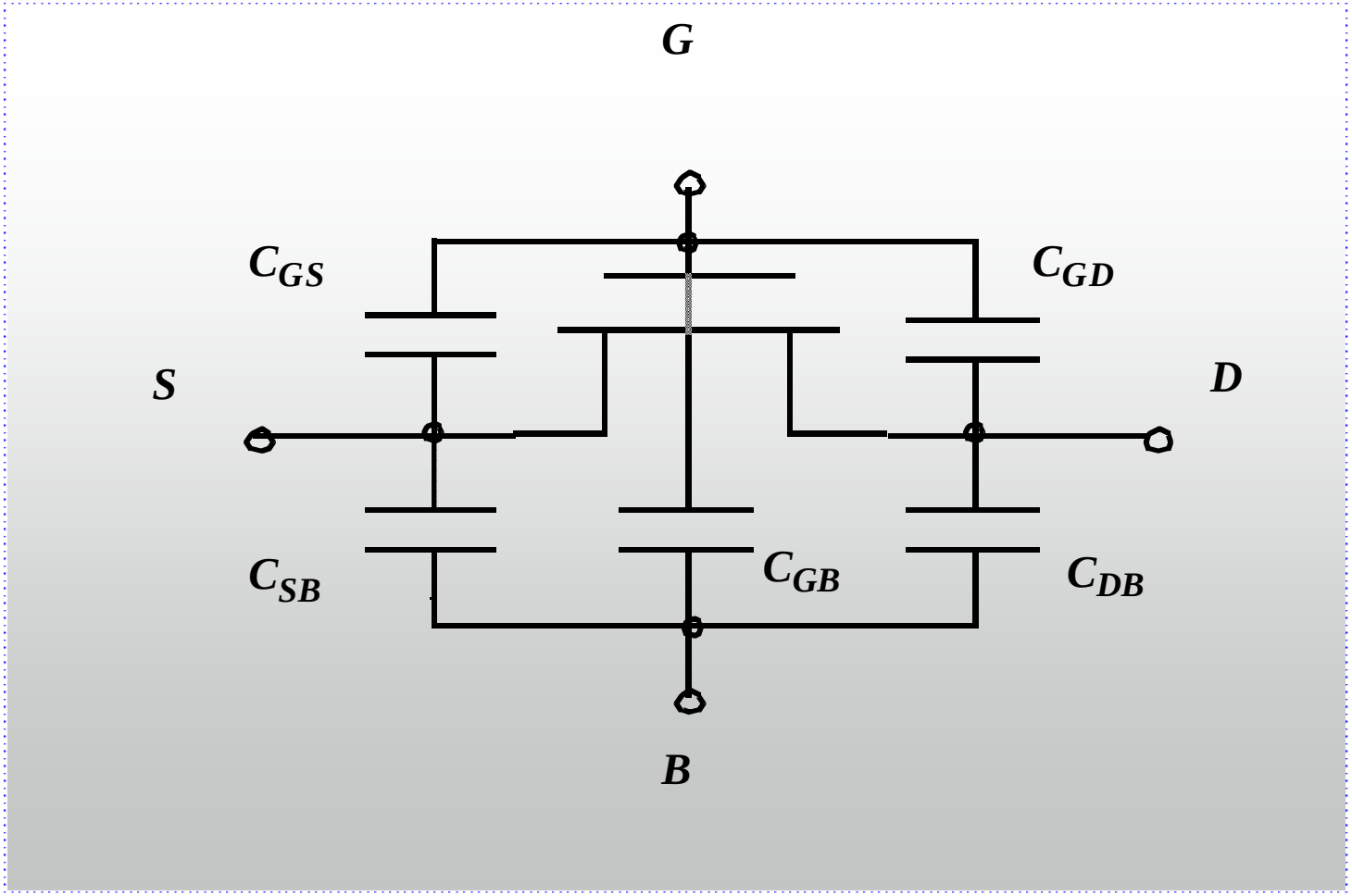
$\rightarrow R_{on} \sim \frac{1}{W}$

$R_{on} \sim \frac{L}{W}$

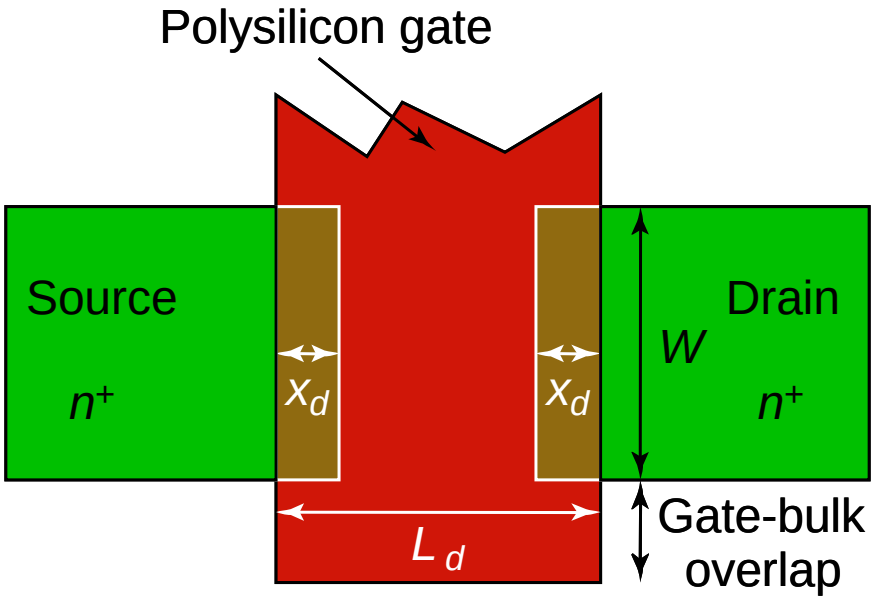
$$R_{on} \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1-\frac{5}{6}\lambda V_{DD}\right)$$

Use this formula for hand analysis

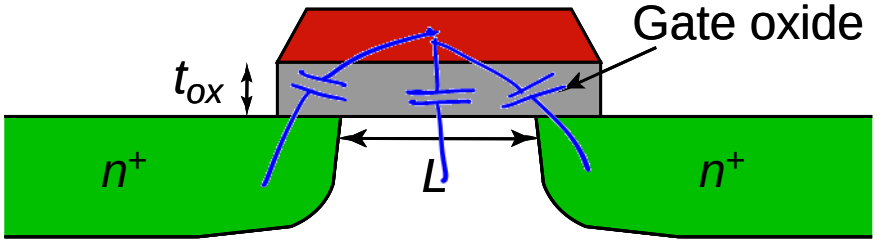
MOS Capacitances



The Gate Capacitance

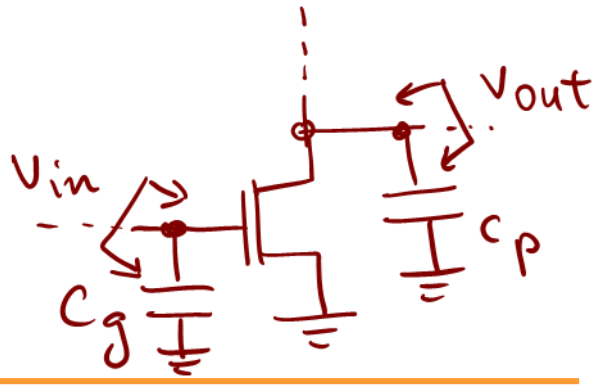


Top view



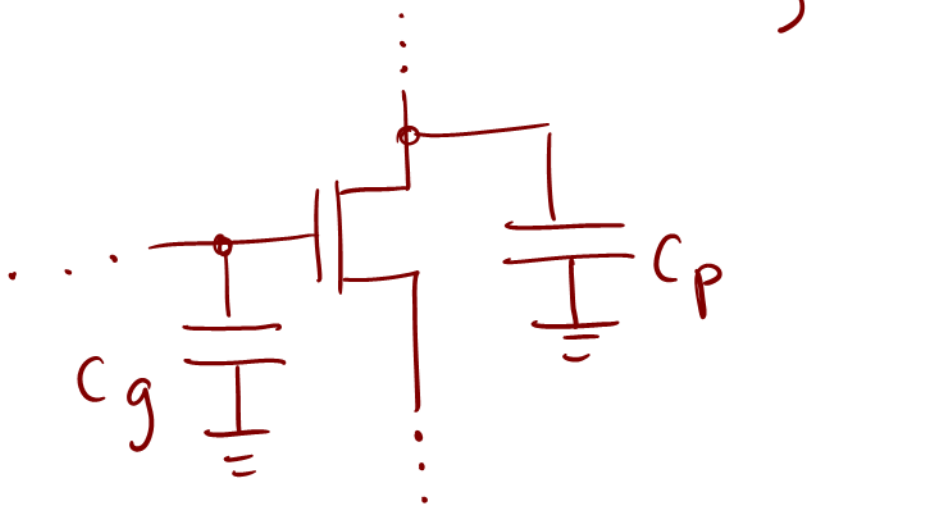
Cross section

$$C_{gate} = \frac{\varepsilon_{ox}}{t_{ox}} WL$$



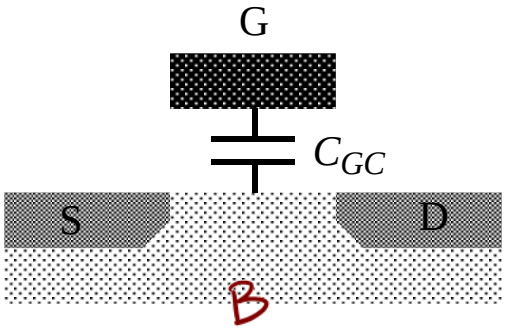
Capacitance Components

- #1: Gate-Channel Capacitance
- #2: Gate Overlap Capacitance
- #3: Junction/Diffusion Capacitance

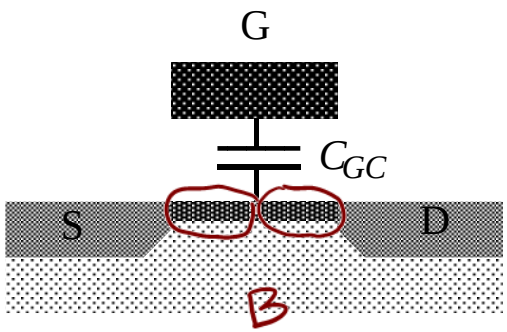


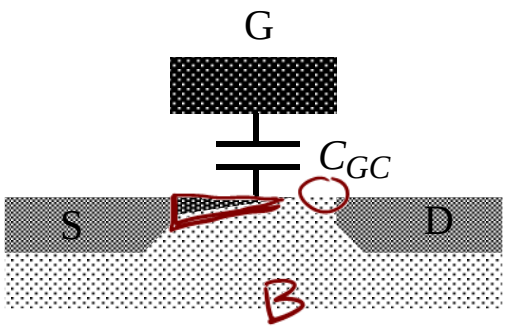
#1: Gate-Channel Capacitance

$C_{GC} = C_{GCB} + C_{GCS} + C_{GCD}$



Cut-off





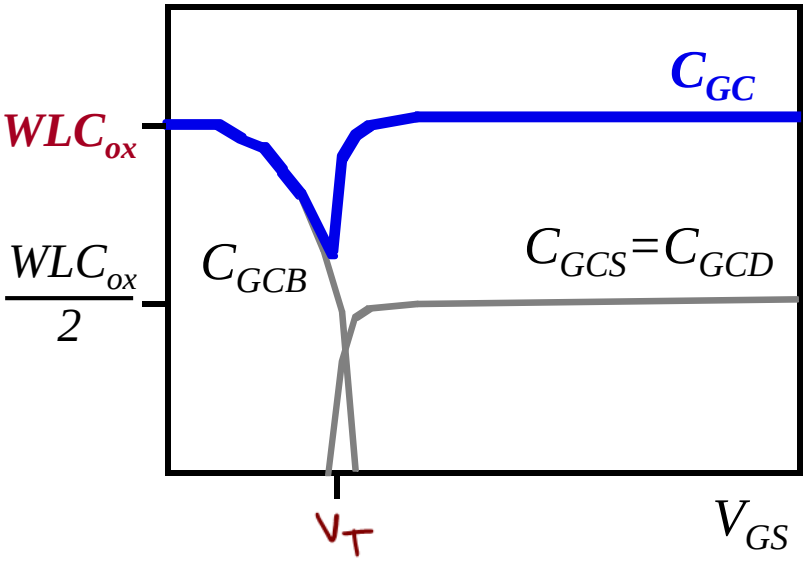
Operation Region	C_{gb}	C_{gs}	C_{gd}
✓ Cutoff	$C_{ox}WL_{eff}$	0	0
linear (Triode)	0	$C_{ox}WL_{eff}/2$	$C_{ox}WL_{eff}/2$
✓ Saturation	0	$(2/3)C_{ox}WL_{eff}$	0

→ $WL\,C_{ox}$

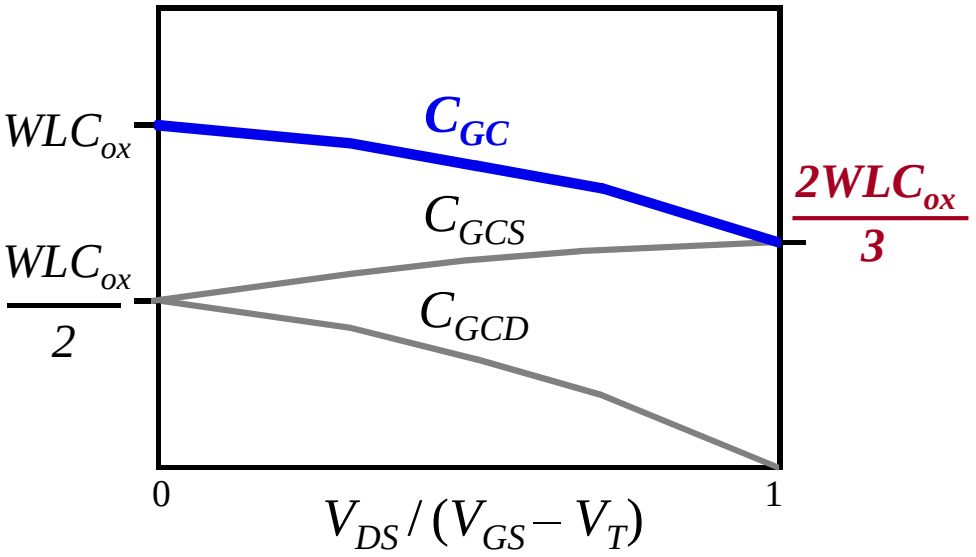
→ $\frac{2}{3}WL\,C_{ox}$

Most important regions in digital design: saturation and cut-off

A Close Look at Gate-Channel Capacitance

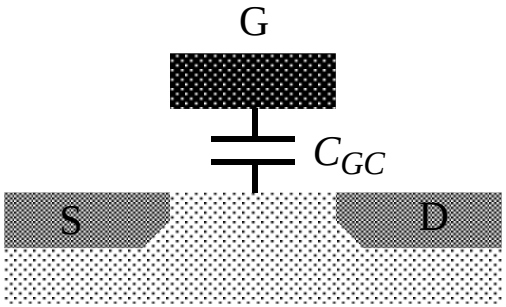


C_{gate} as a function of V_{GS} (with V_{DS} = 0)

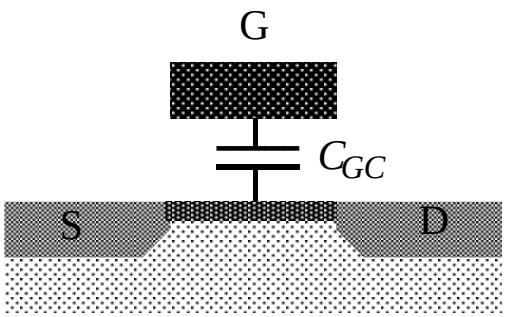


C_{gate} as a function of the degree of saturation

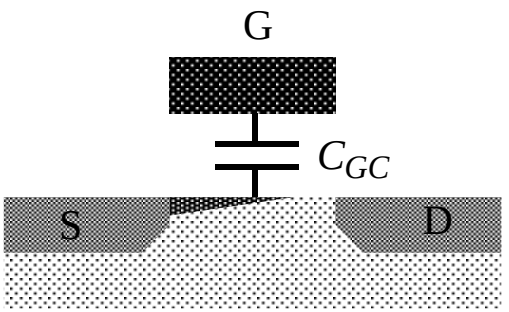
Summary:
#1: Gate-Channel Capacitance



Cut-off



Resistive



Saturation

Operation Region	C_{GCB}	C_{GCS}	C_{GCD}
Cutoff	$C_{ox}WL_{eff}$	0	0
Triode	0	$C_{ox}WL_{eff}/2$	$C_{ox}WL_{eff}/2$
Saturation	0	$(2/3)C_{ox}WL_{eff}$	0

Off/Lin → $C_{gate} = C_{ox} \cdot W \cdot L_{eff}$

Sat → $C_{gate} = (2/3) \cdot C_{ox} \cdot W \cdot L_{eff}$

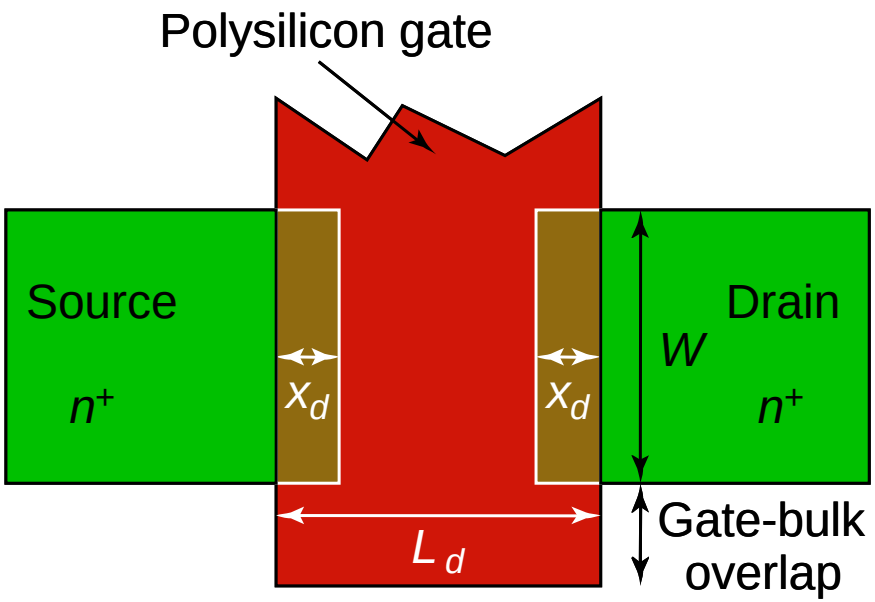
$L_d = 120\,nm$, $x_d = 15\,nm$ → $L_{eff} = 90\,nm$

$W = 240\,nm$

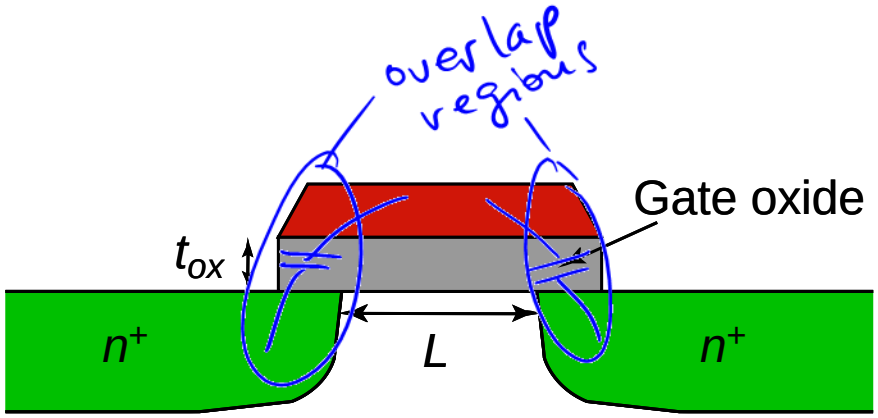
$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} = 15\ \frac{fF}{\mu m^2}$

$C_g = 15 \times 0.24 \times 0.09 = 0.2\,fF$

#2: Gate Overlap Capacitance



Top view



Cross section

$$C_O = C_{ox} \cdot x_d$$

15 0.015

0.225 fF/μm

xd

w

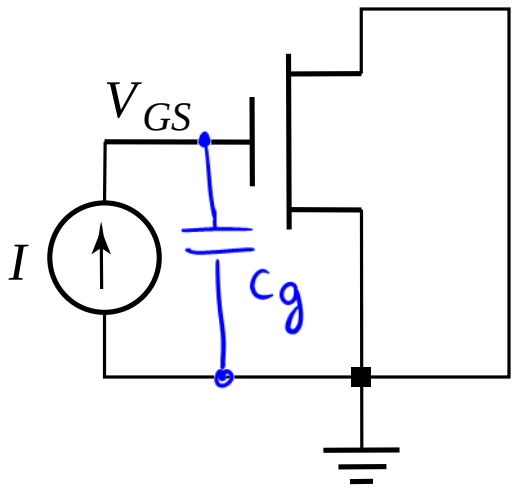
e.g. 0.24μ

Off/Lin/Sat → $C_{GSO} = C_{GDO} = C_O \cdot W$ => 0.225 × 0.24 = 0.05fF

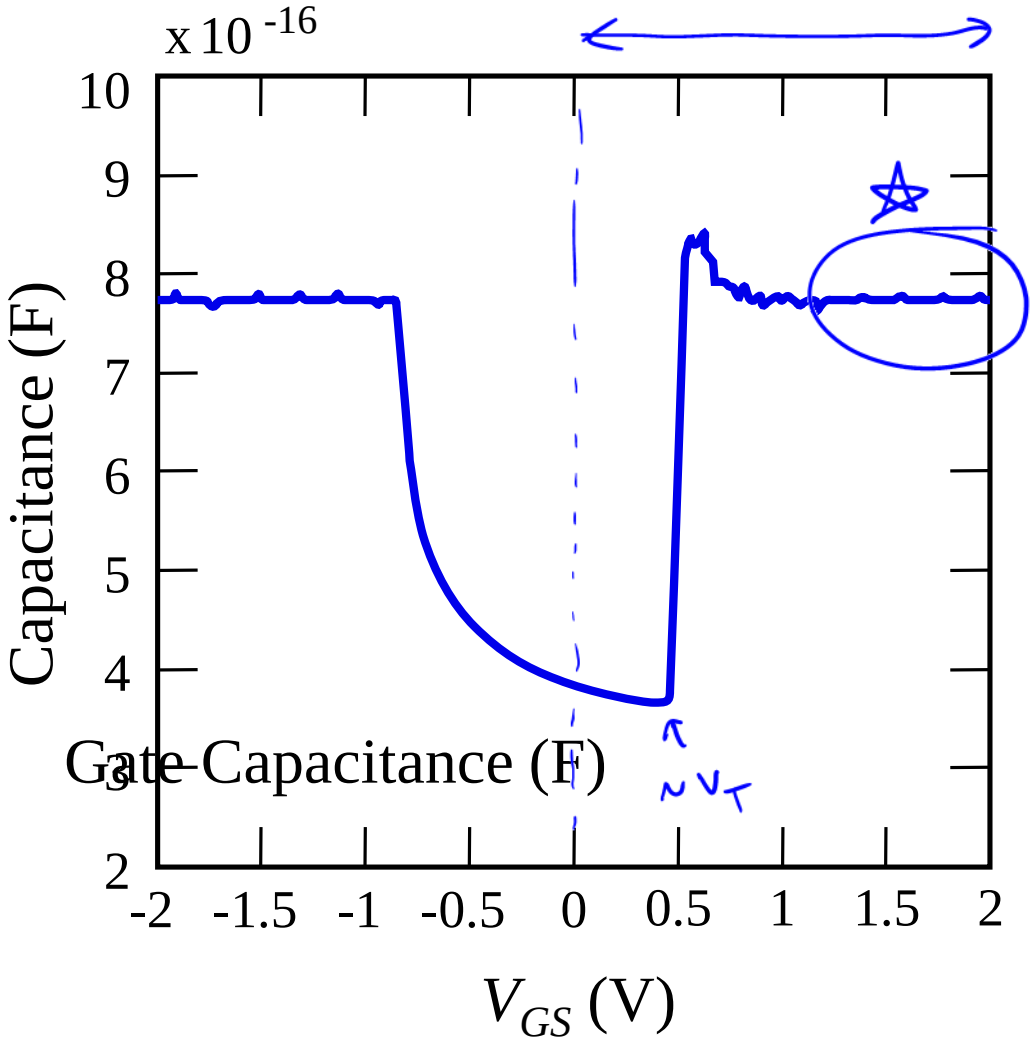
$$C_g = C_{GC} + \underbrace{2}_{S+D} C_{GSO} = 0.2fF + 2 \times 0.05fF = 0.3fF$$

Measuring the Gate Cap

Transient analysis

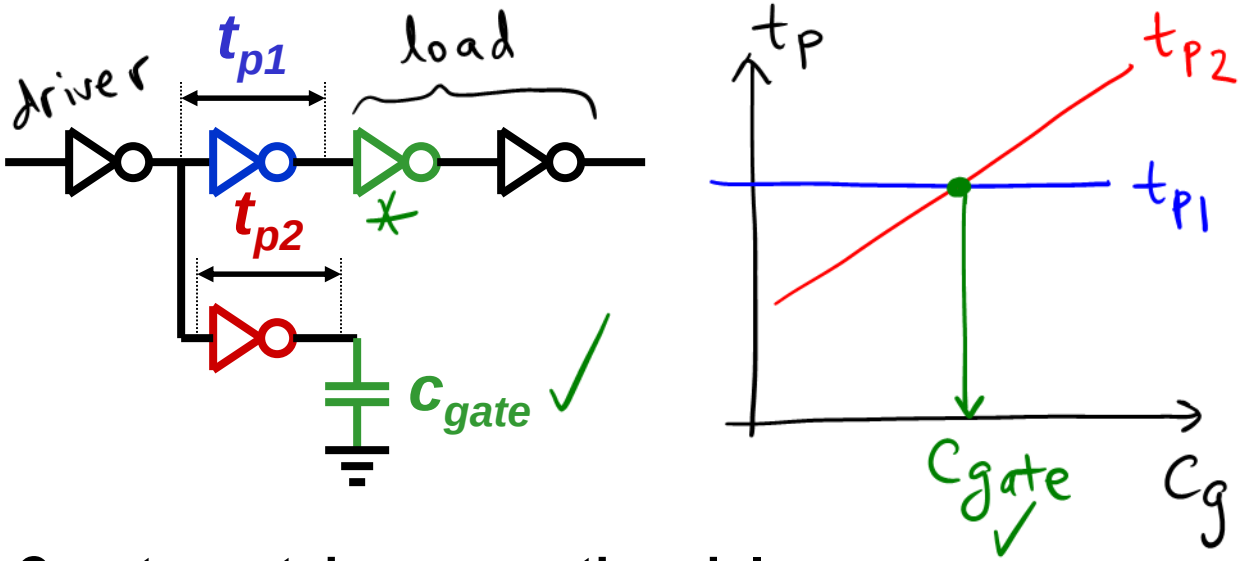


$$C_{gate} = \frac{I \cdot \Delta t}{\Delta V}$$



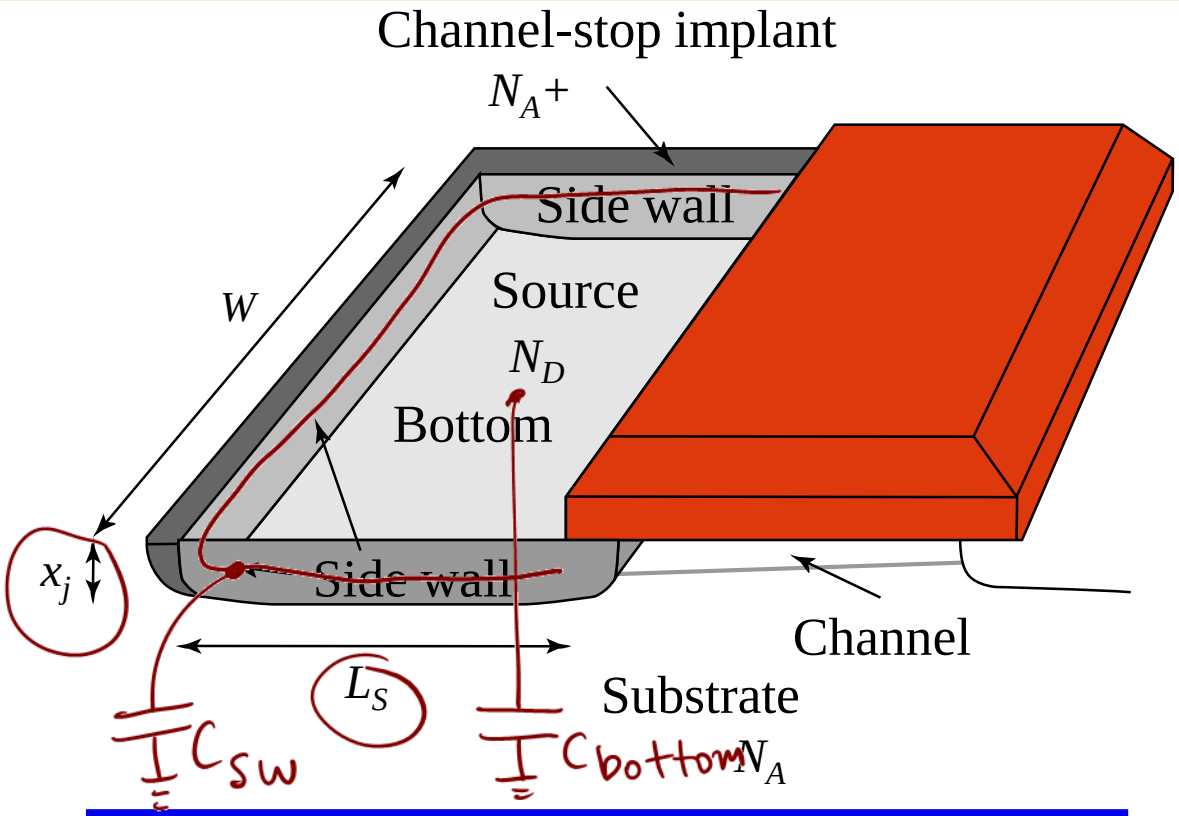
Finding Equivalent Capacitance – Delay

- Curve fitting approach to find a number that works for hand analysis of the gate delay
- Understand the limitations: the model will depend on signal rise times, voltage, temperature, process parameter variation



- Experiment: find C_{gate} to match propagation delays
  - $t_{p1} = t_{p2}$ → C_{gate} is equivalent cap of the green gate

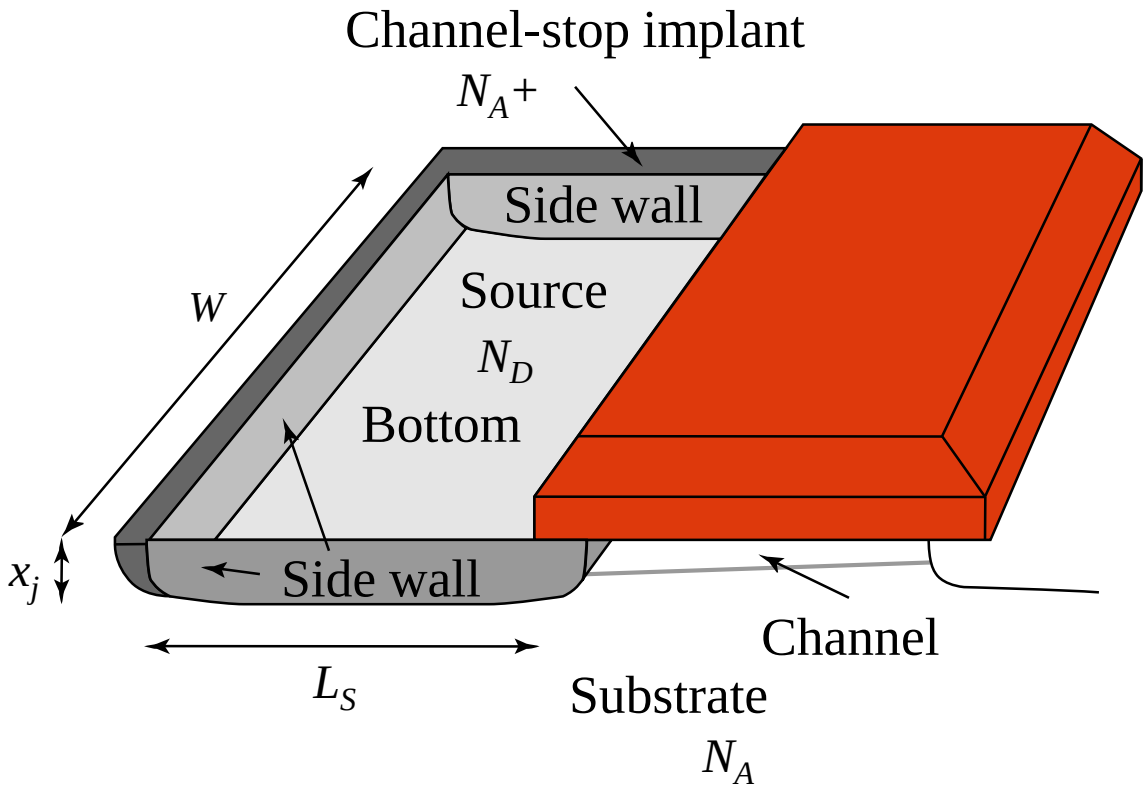
#3: Diffusion Capacitance



$$C_{diff} = C_{bottom} + C_{sw}$$
$$= C_j \cdot AREA + C_{jsw} \cdot PERIMETER$$
$$= C_j \cdot L_S W + C_{jsw}(2L_S + W)$$

$\sim W$

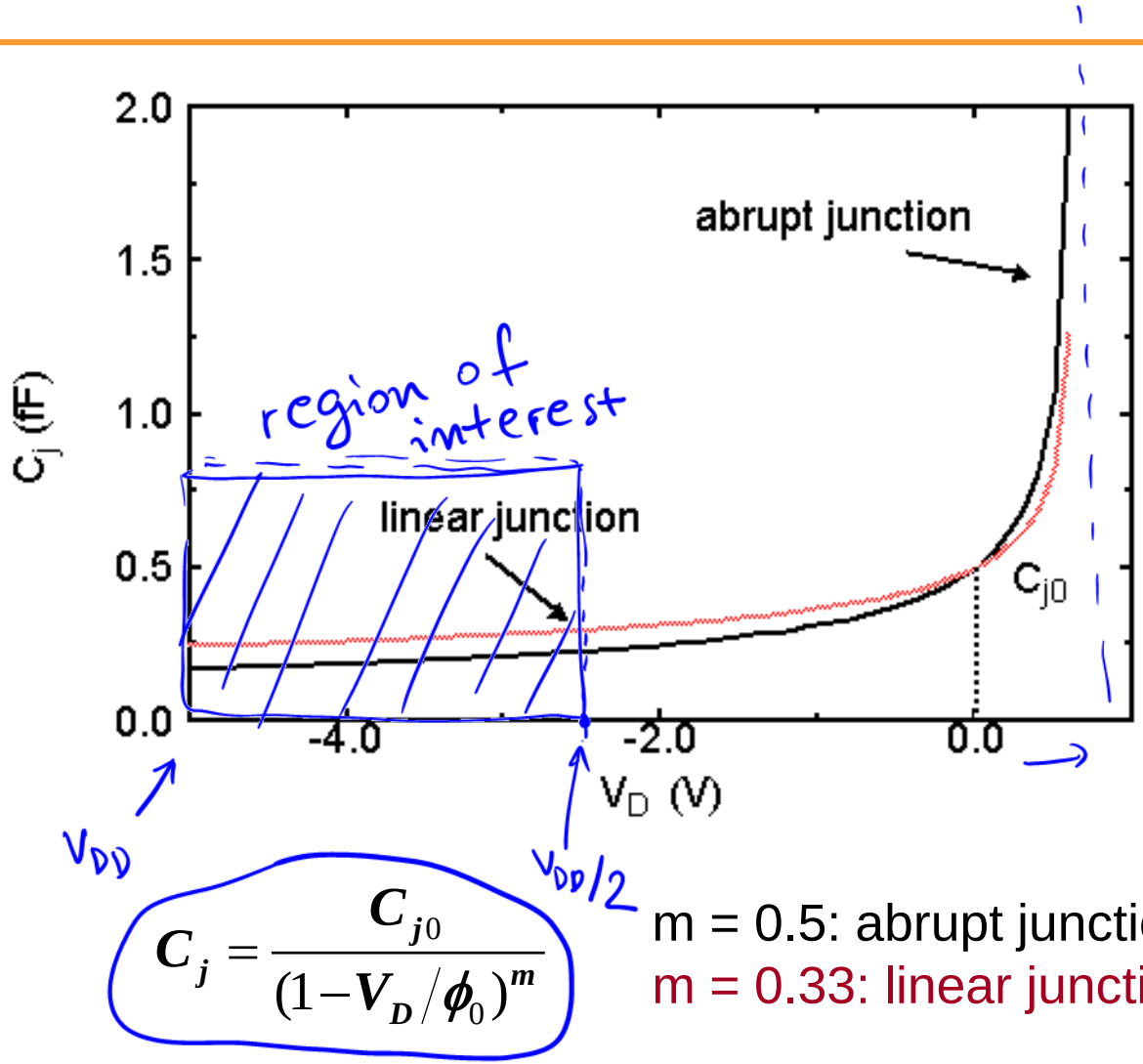
#3: Diffusion Capacitance

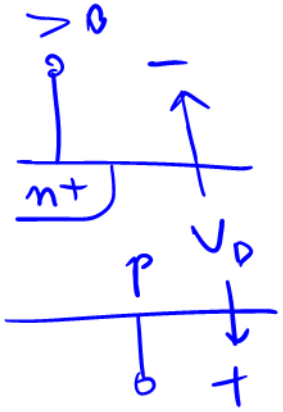


$$C_{diff} = C_{bottom} + C_{sw}$$
$$= C_j \cdot AREA + C_{jsw} \cdot PERIMETER$$

Off/Lin/Sat → $C_{diff} = C_j \cdot L_S \cdot W + C_{jsw} \cdot (2L_S + W)$

Junction Capacitance is Bias-dependent





$$C_j = \frac{C_{j0}}{(1 - V_D/\phi_0)^m}$$

m = 0.5: abrupt junction

m = 0.33: linear junction

#3 Diffusion Capacitance: Summary of Equations

- **Bottom-plate C**

$$C_j = \frac{C_{j0}}{(1 - V_D/\Phi_0)^{m_j}}$$

C_{j0}: 0.8 fF/μm²; m_j: 0.22; Φ_0: 1; V_D: 0.8 V

$= 0.7$ fF/μm²

m = 0.5: abrupt junction

m = 0.33: linear junction

- **Side-wall C**

$$C_{jsw} = \frac{C_{jsw0}}{(1 - V_D/\Phi_{0sw})^{m_{jsw}}}$$

C_{jsw0}: 0.05 fF/μm; m_{jsw}: 0.01; Φ_{0sw}: 0.8; V_D: 1

$= 0.05$ fF/μm

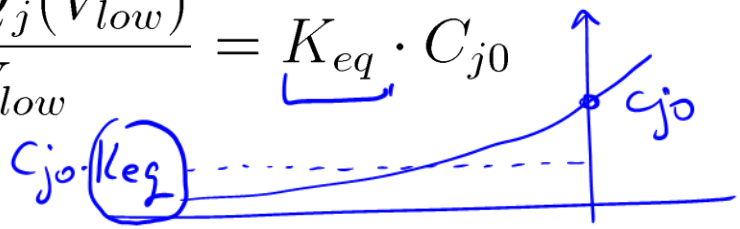
$$C_{diff} = C_j WL + C_{jsw}(2L_s + W)$$

linearize

Linearizing the Junction Cap

- **Replace non-linear capacitance by large-signal equivalent linear capacitance, which displaces equal charge over voltage swing of interest**

$$C_{eq} = \frac{\Delta Q_j}{\Delta V_D} = \frac{Q_j(V_{high}) - Q_j(V_{low})}{V_{high} - V_{low}} = K_{eq} \cdot C_{j0}$$



$$K_{eq} = \frac{-\Phi_0^m}{(V_{high} - V_{low}) \cdot (1-m)} \left[(\Phi_0 - V_{high})^{(1-m)} - (\Phi_0 - V_{low})^{(1-m)} \right]$$

Typical value for K_{eq} around 0.5

Summary: Capacitive Device Model

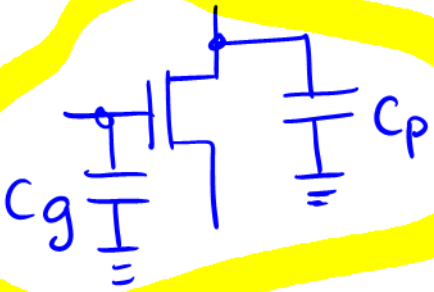
- Gate-Channel Capacitance
  - $C_{GC} = C_{ox} \cdot W \cdot L_{eff}$ (Off, Linear)
  - $C_{GC} = (2/3) \cdot C_{ox} \cdot W \cdot L_{eff}$ (Saturation)
- Gate Overlap Capacitance
  - $C_{GSO} = C_{GDO} = C_O \cdot W$ (Always)
- Junction/Diffusion Capacitance
  - $C_{diff} = C_j \cdot L_S \cdot W + C_{jsw} \cdot (2L_S + W)$ (Always)

Zero-bias → $C_{diff} > C_{gate}$

MOS On → $C_{diff} \leq C_{gate}$

Circuit design

$C_{gate} \approx 2 fF/\mu m$



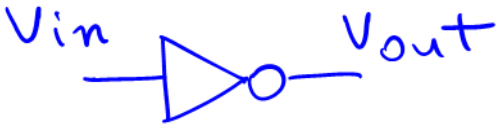
$C_{parasitic} \approx 1.2 fF/\mu m$

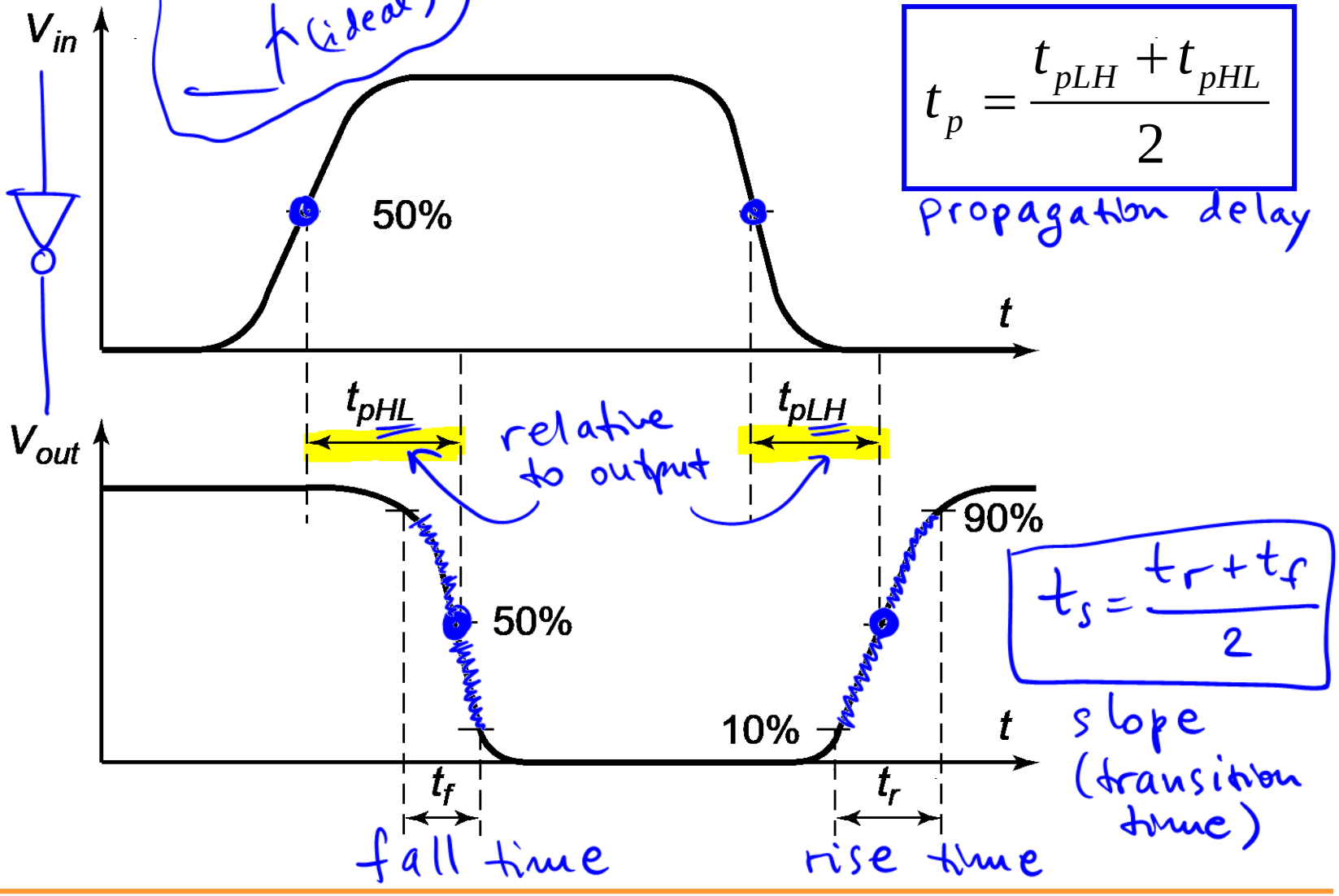
$\gamma = \frac{C_p}{C_g} \approx 0.6$ (technology dependent)

Week 2 Agenda

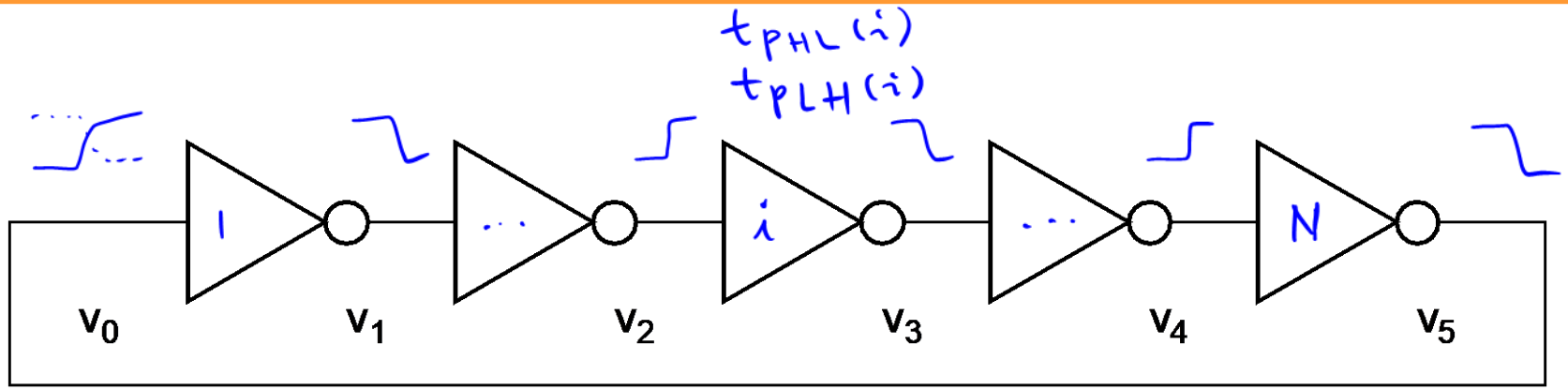
- MOS RC Model
- Delay Model
- Power Model
- CMOS Scaling

Performance: Delay Definitions

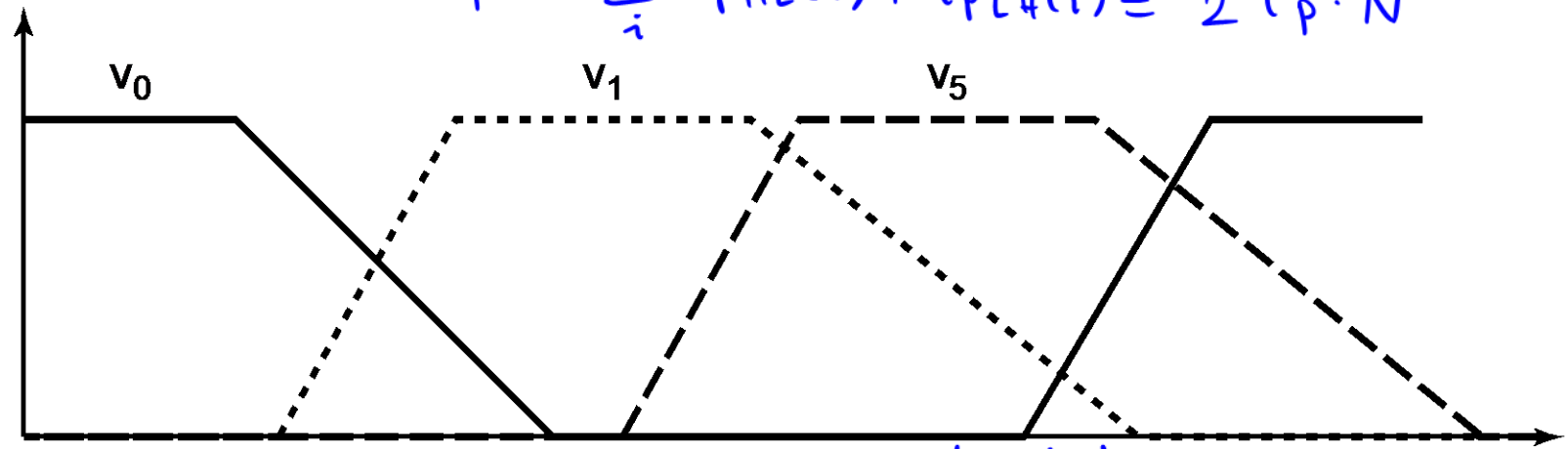




Technology Characterization: Ring Oscillator for t_p



$$T = \sum_i t_{pHL(i)} + t_{pLH(i)} = 2\,t_p \cdot N$$

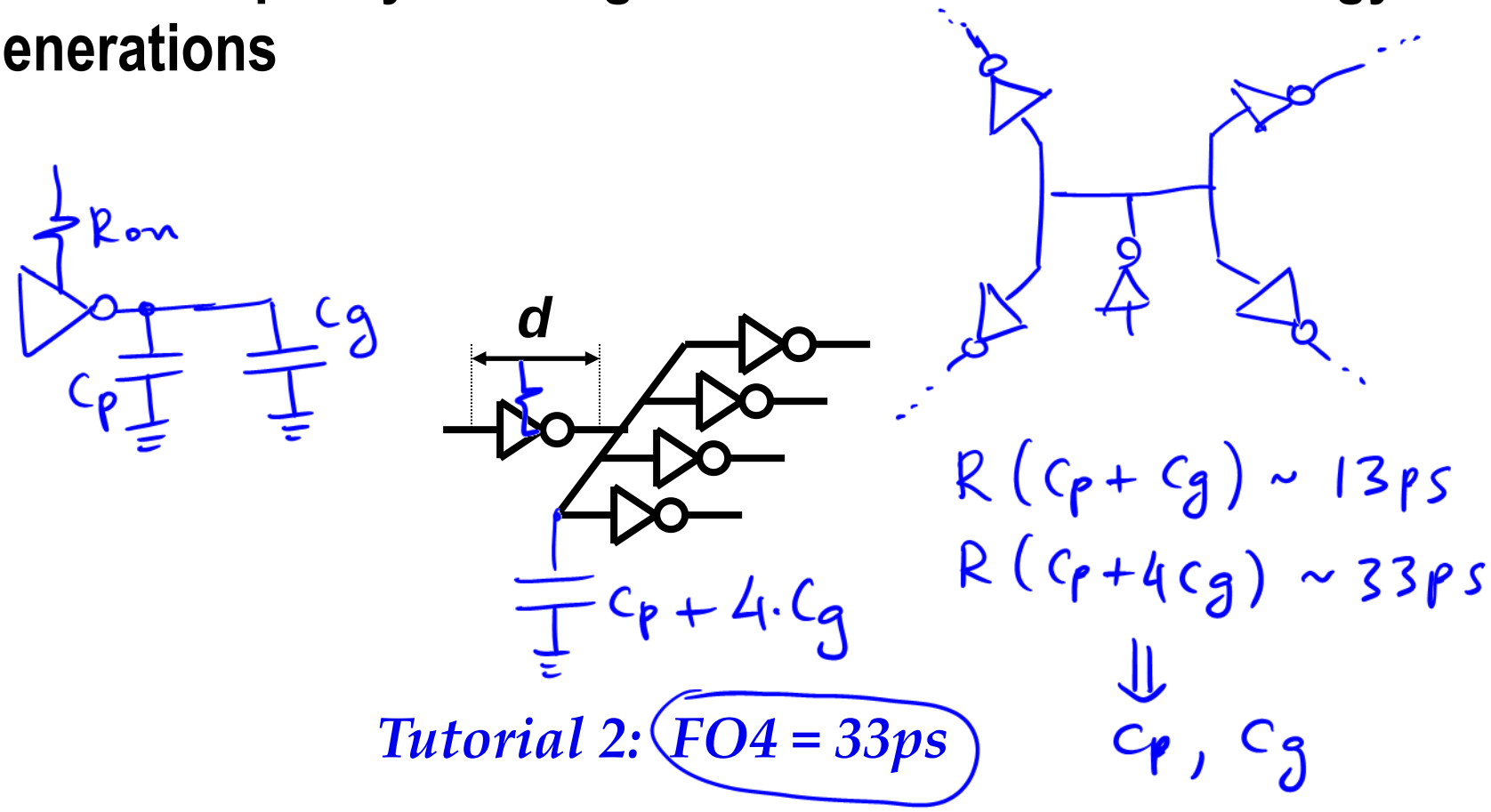


measure T by simulation $\Rightarrow$ $T = 2 \times t_p \times N$ (calculate)

Tutorial 2: $t_p = 13ps$

Performance: FO4 Inverter

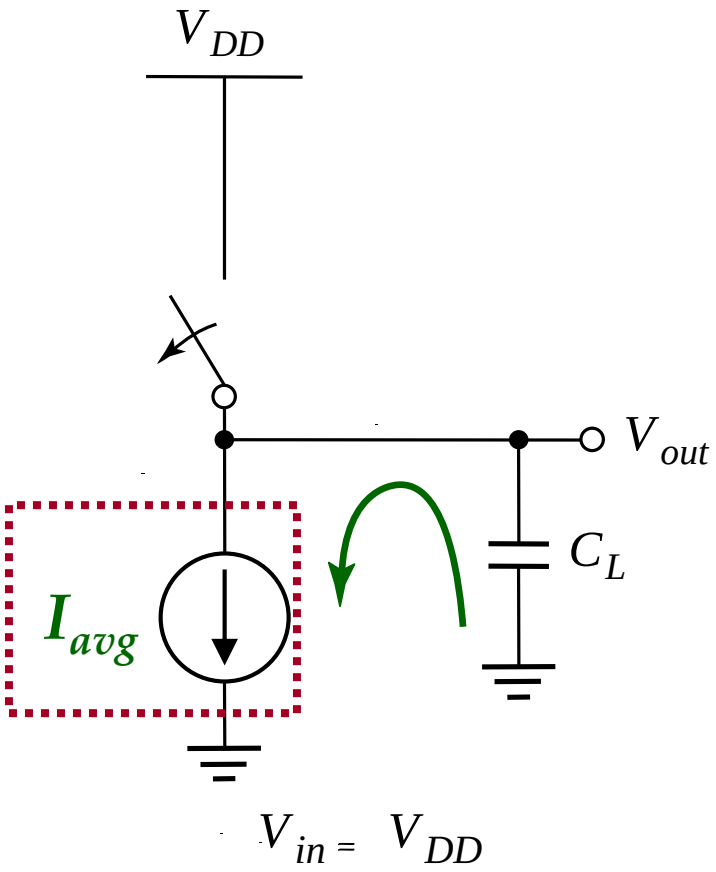
- Measures quality of design across different technology generations



Tutorial 2: FO4 = 33ps

CMOS Inverter Propagation Delay

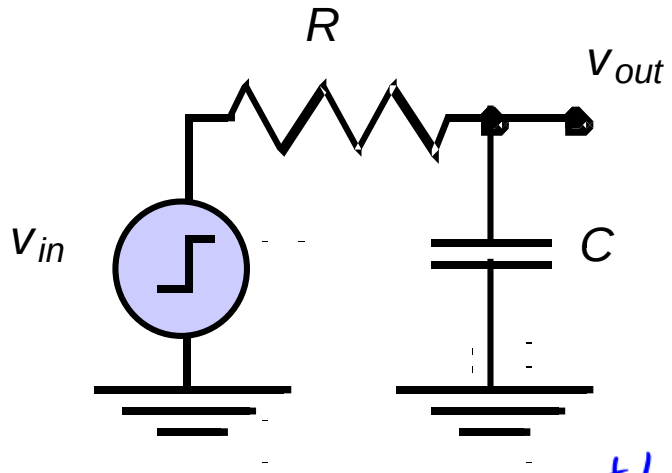
MOS Current Model



$$t_{pHL} = \frac{C_L \cdot V_{swing}/2}{I_{avg}}$$

$$t_{pHL} \sim \frac{C_L}{k_n \cdot V_{DD}}$$

A First-Order RC Network: Step Response



Step response:

$$v_{out}(t) = \left(1 - e^{-t/\tau}\right) \cdot v_{in}$$

RC

1) $v_{out}(0) = V_0$

2) $v_{out}(\infty) = V_\infty$

$V_0, V_\infty \in \{V_{OL}, V_{OH}\}$

$t=\infty: e^{-t/\tau} = 0$

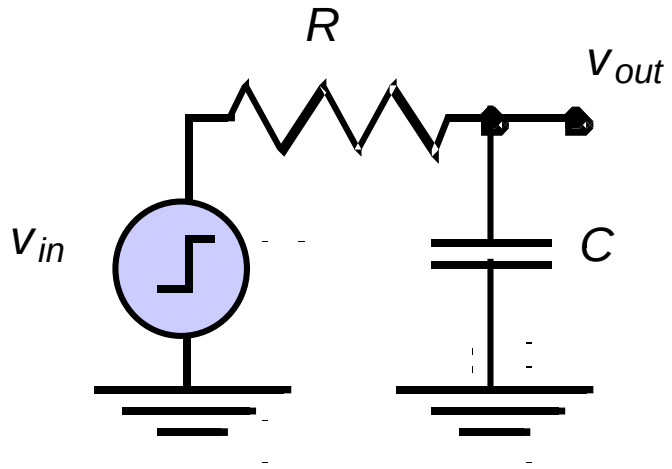
$t=0: e^{-t/\tau} = 1$

$$v_{out}(t) = V_\infty + (V_0 - V_\infty) \cdot e^{-t/\tau}$$

GENERAL FORMULA

Special case $V_\infty = V_{DD}, V_0 = 0 \rightarrow v_{out} = V_{DD}(1 - e^{-t/\tau})$

A First-Order RC Network: Propagation Delay



Propagation delay:

$$t_p = \tau \cdot ln2 = 0.69RC$$

$t_s \sim 2.2\,RC$ 10-90% slope

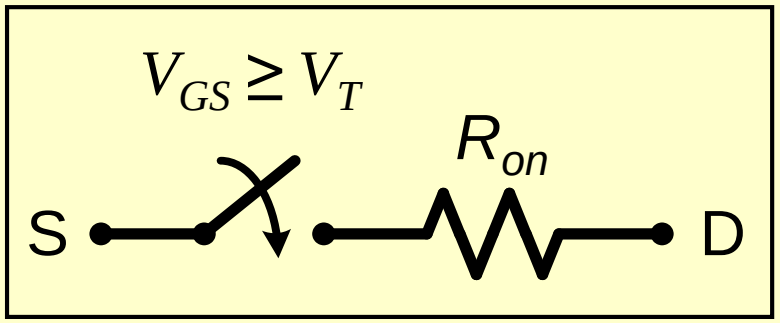
$t = t_p$ $v_{out} = V_M$

step input

$$v_{out}(t_p) = V_M = V_\infty + (V_0 - V_\infty)\, e^{-t_p/\tau}$$

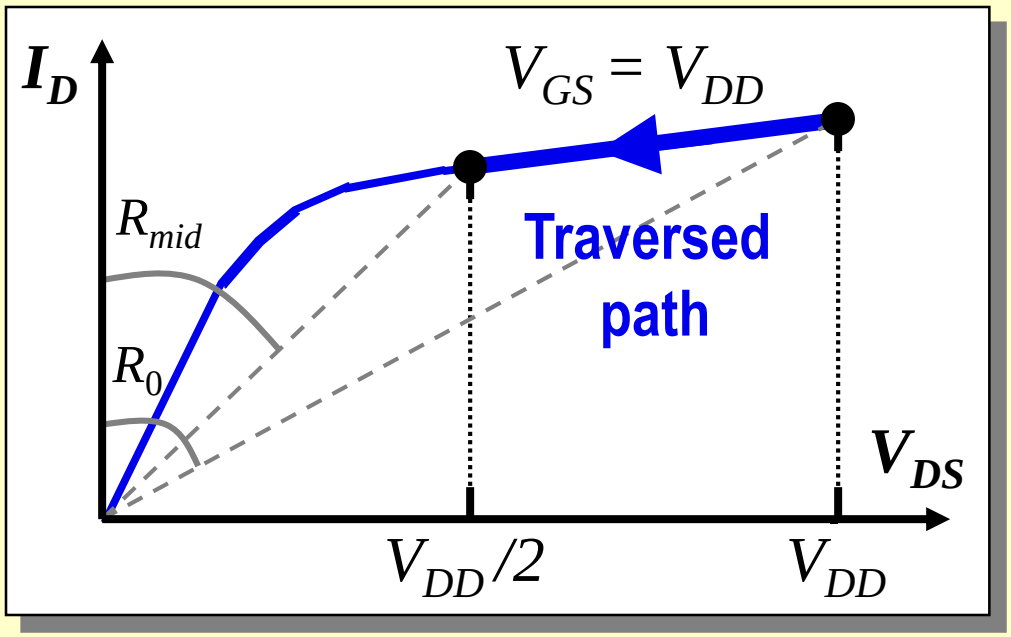
$$\rightarrow t_p = \tau \ln \underbrace{\frac{V_0 - V_\infty}{V_M - V_\infty}}_{2 \text{ for } V_\infty = V_{DD},\ V_0 = 0}$$

Review: Transistor as a Switch



$$R_{on} \approx \frac{1}{2}(R_0 + R_{mid})$$

good approximation (I-V ≈ linear)



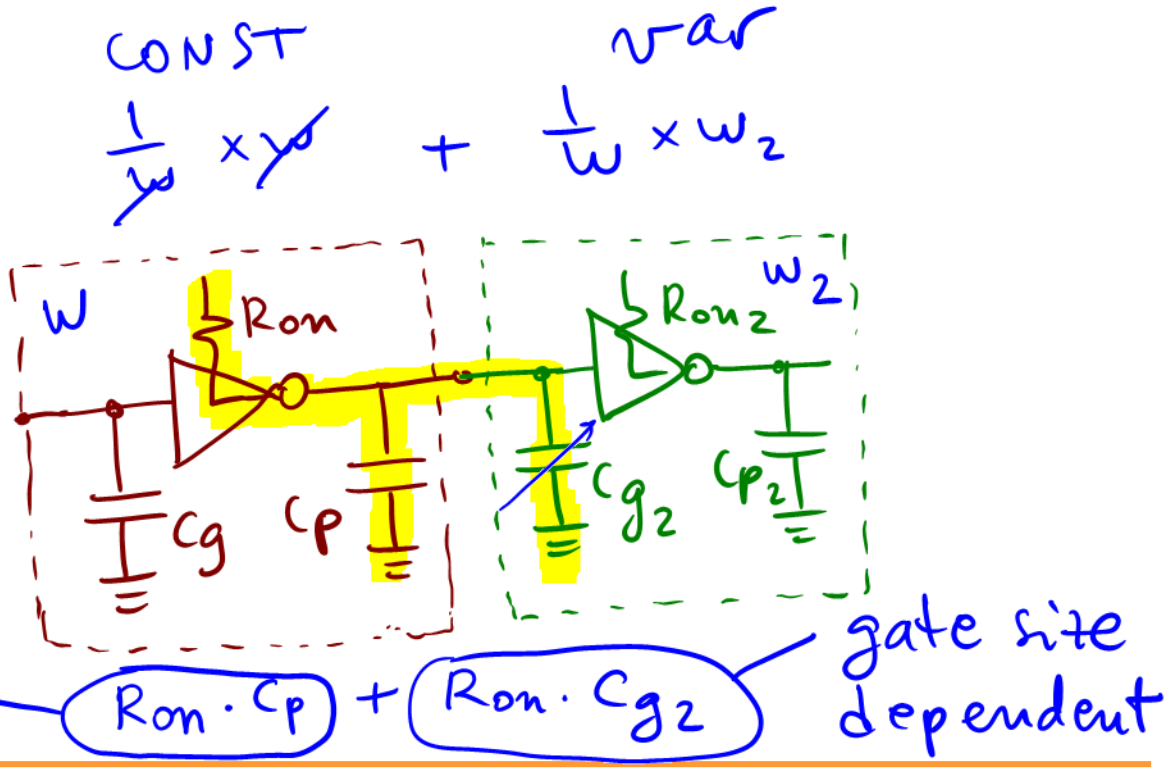
$$R_{on} = \frac{1}{2}\left(\frac{V_{DD}}{I_{DSAT} \cdot (1 + \lambda V_{DD})} + \frac{V_{DD}/2}{I_{DSAT} \cdot (1 + \lambda V_{DD}/2)}\right)$$

$$R_{on} \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1 - \frac{5}{6}\lambda V_{DD}\right)$$

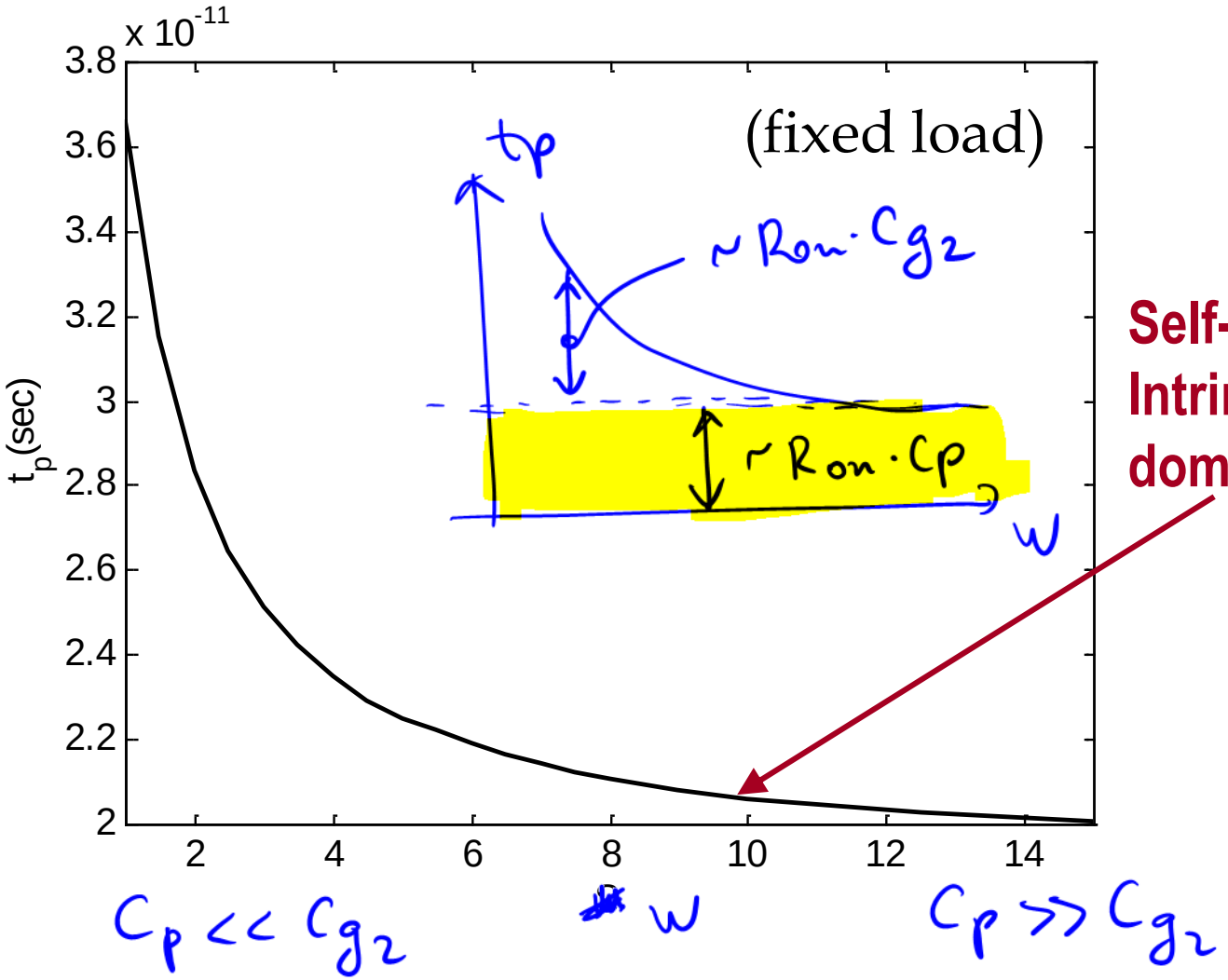
Design for Performance

- **Keep capacitances small**
- **Increase transistor sizes**
  - watch out for self-loading!
- **Increase V_{DD} (?)**

$$t_{pHL} \sim \frac{C_L}{k_n \cdot V_{DD}}$$

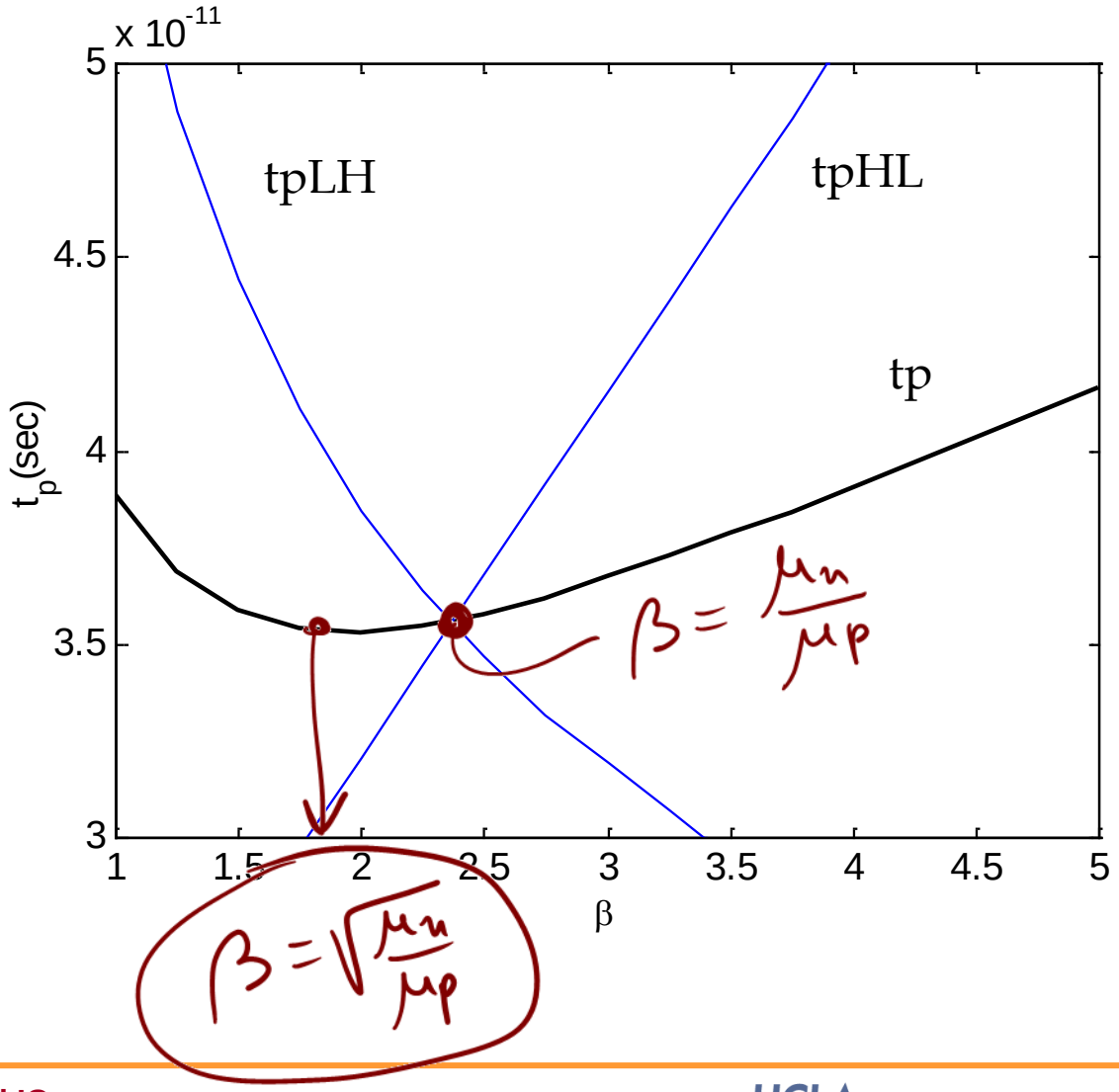


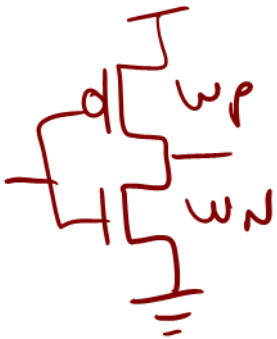
Device Sizing



Self-loading effect: Intrinsic capacitances dominate

NMOS/PMOS Ratio



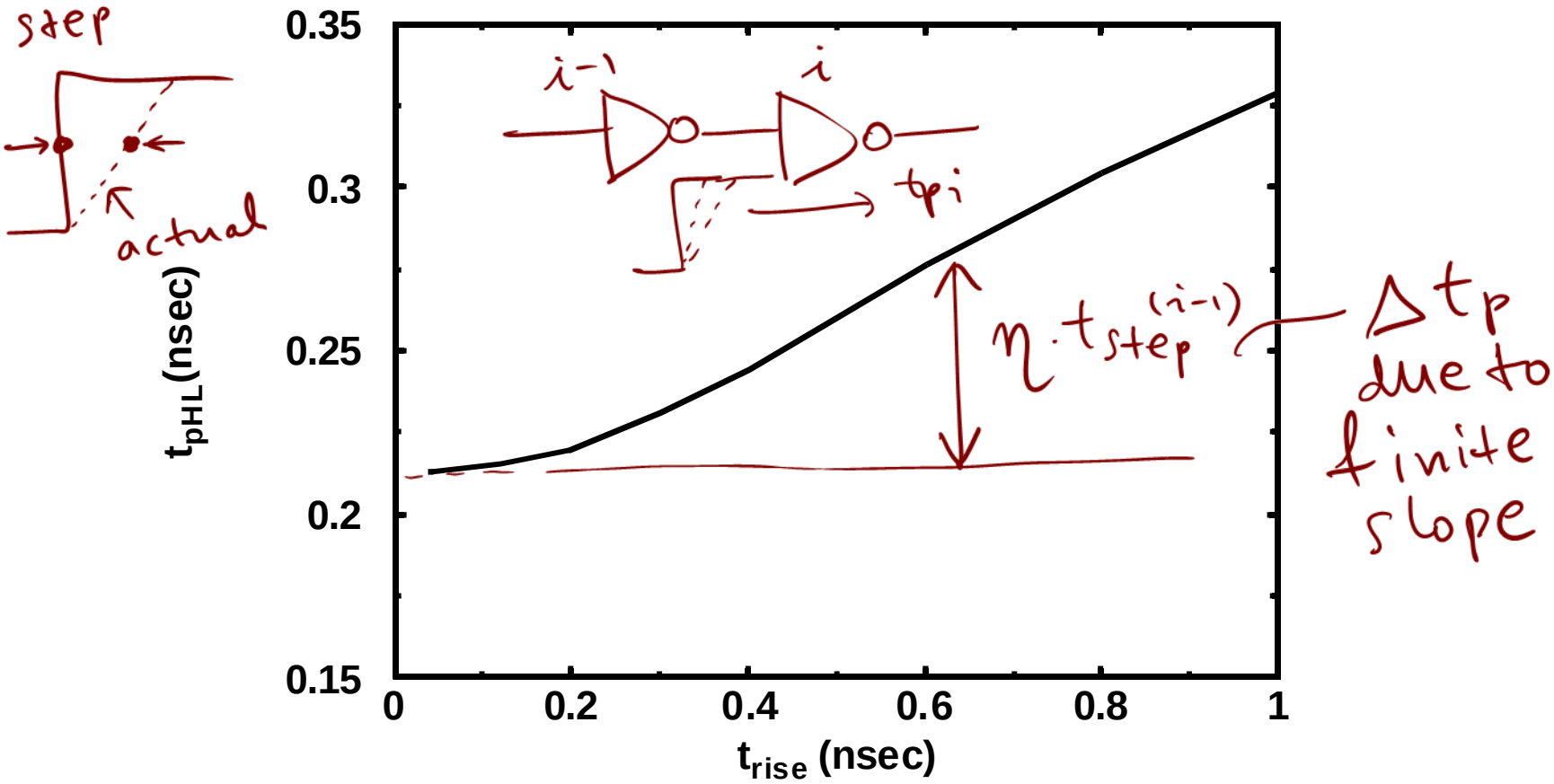


$\beta = W_p / W_n$

std cells

$\frac{W_p}{W_N} \approx 1.3$

Impact of Rise Time on Delay



$$t_p = t_{step(i)} + \eta \cdot t_{step(i-1)}$$

Simplified Macro Model

- **Consider two macro capacitances**
  - Input gate capacitance, C_{in} (or C_{gate})
  - Output parasitic (self-loading capacitance), C_{par}

- **Assume that both capacitances are linearized**
  - C_{in} and C_{par} are proportional to W
    (remember, we keep L at L_{min}, so it is lumped into constant)
  - In our 90nm technology, C_{par} / C_{in} is about 0.6

- **For gate delay analysis, we will use:**

$$C_{in} = 2fF/\mu m \qquad C_{par}/C_{in} = 0.61$$

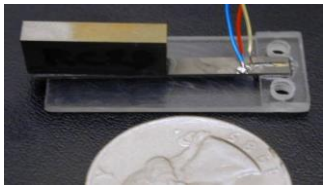
Week 2 Agenda

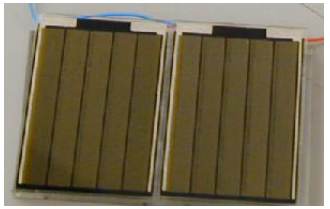
- MOS RC Model
- Delay Model
- Power Model
- CMOS Scaling

Power and Energy Challenges

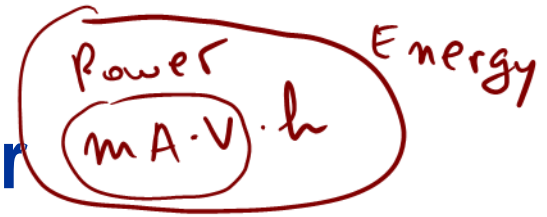
- **1 billion computers in the world**
  - 0.4 PW (PetaWatt = 10^{15} W) of power dissipation
  - Equivalent to 65 nuclear plants!
- **Data centers represent the absolute challenge**
  - 1 single server rack is between 5 and 20 kW
  - 100's of those racks in a single room!

Power and energy management and minimization have emerged as some of the most dominant roadblocks. The best opportunity lies in a **very aggressive scaling and adaptation of supply and threshold values** in concert with a careful orchestration of the system activity.



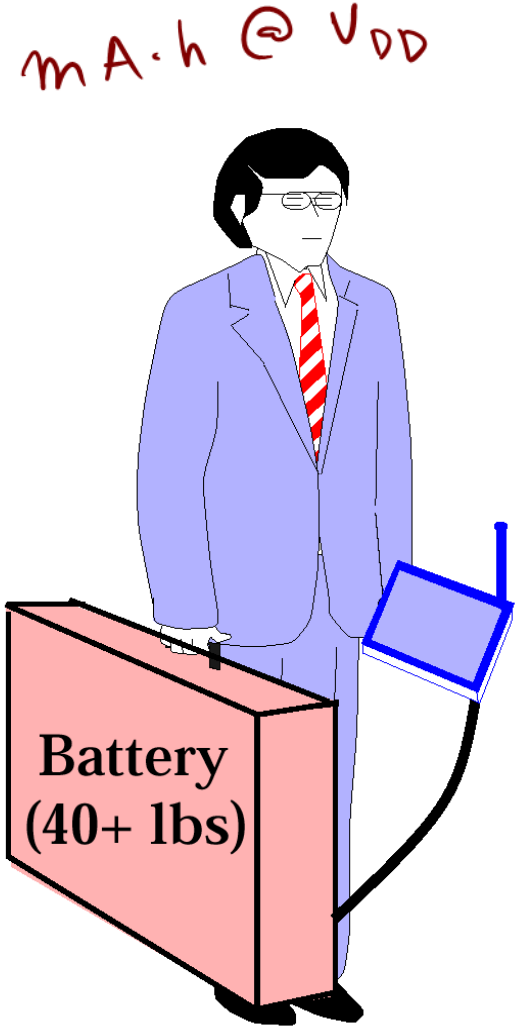


Portability: Battery Storage is the Limiting Factor

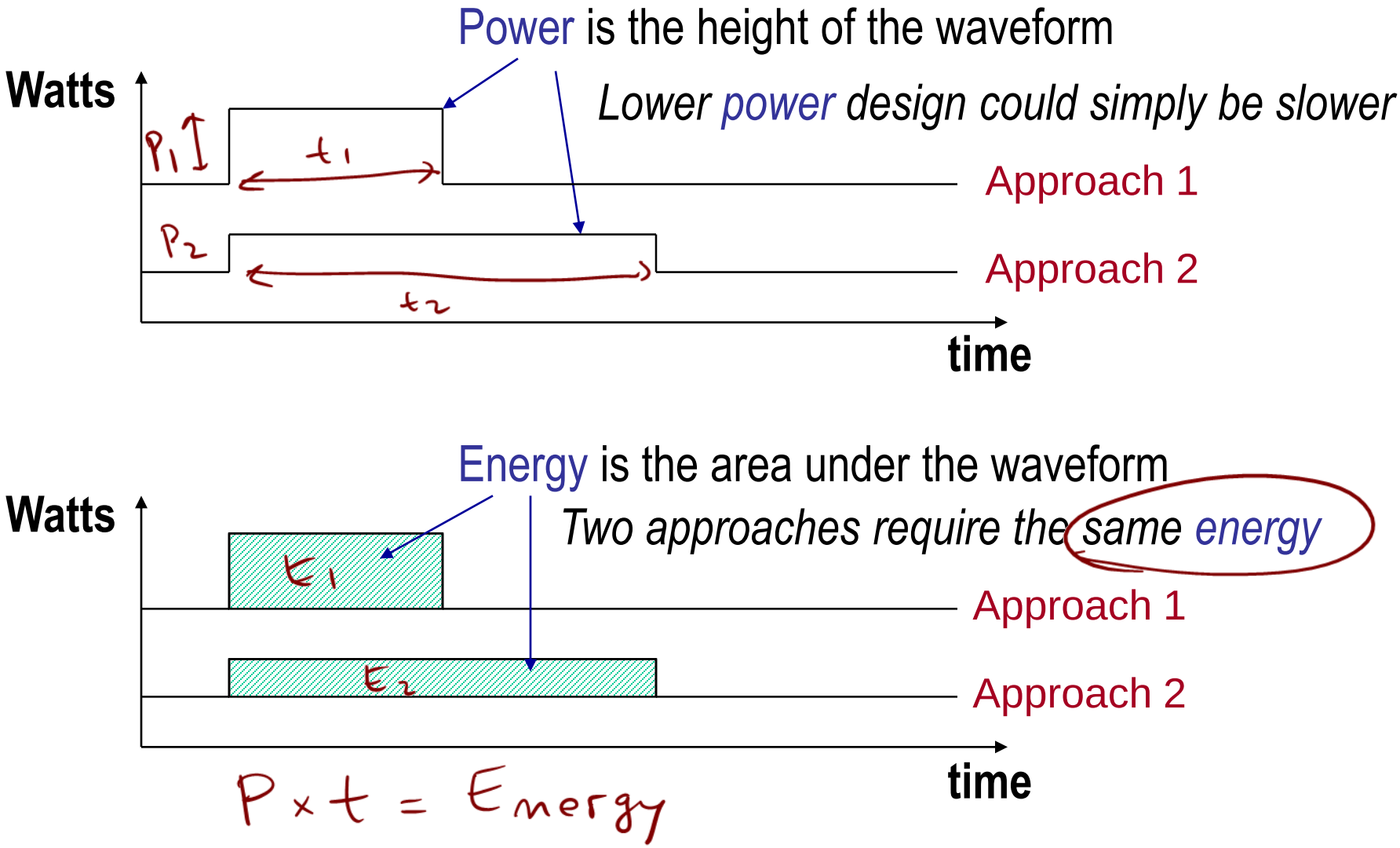


- **Little change in basic technology**
  - store energy using a chemical reaction
- **Battery capacity doubles every 10 years**
  - 4x in the last 10 years
- **Energy density, size, and safe handling are limiting factors**

Energy density of material	***KWH/kg***
Gasoline	14
Lead-Acid	0.04
Li polymer	0.15



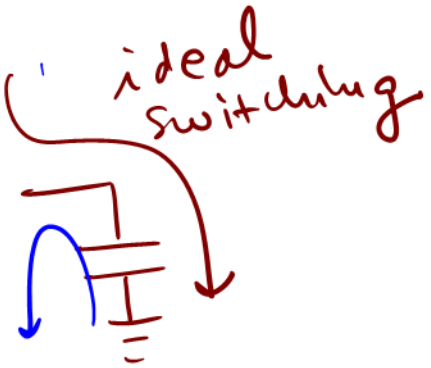
Power versus Energy



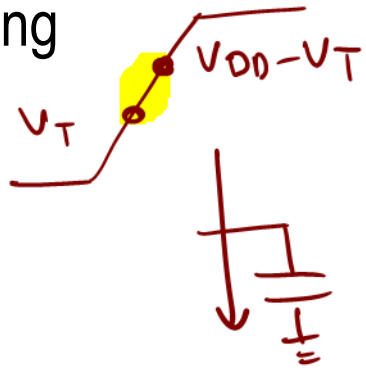
Where Does Power Go in CMOS?

Switching

- **#1: Dynamic Power Consumption** ~75%
  - Charging and discharging capacitors

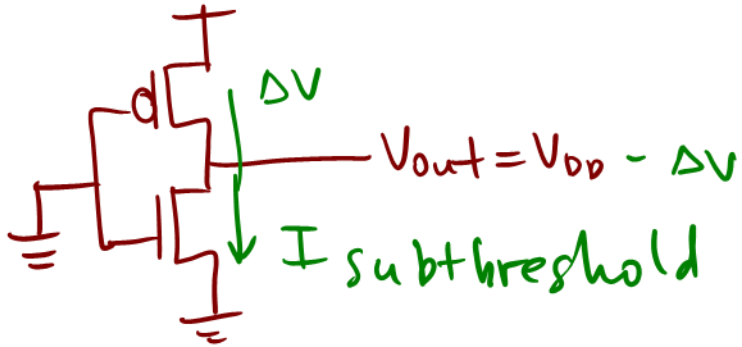


- **#2: Short Circuit Currents** ~5-10%
  - Short-circuit path between supply rails during switching

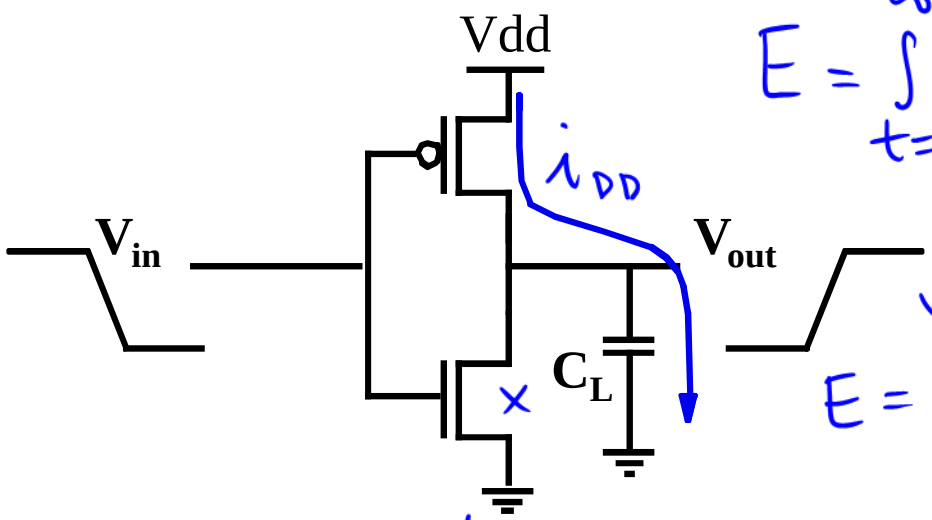


Leakage

- **#3: Leakage Currents** ~15-20%
  - Leaking diodes and transistors



#1: Dynamic Power Dissipation



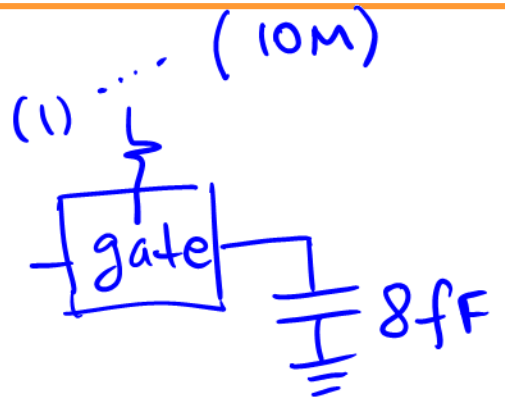
$$\text{Energy/transition} = C_L \cdot V_{dd}^2$$

$$\text{Power} = \text{Energy/transition} \cdot f = f \cdot C_L \cdot V_{dd}^2$$

- **Dynamic power: observations**
  - Not a function of transistor sizes!
  - Need to reduce C_L, V_{dd}, and f to reduce power

Example

- **Parameters** $V_{DD} = 1V$
  - Switched capacitance: 2fF / gate
  - Fanout 4 gates $\rightarrow C_L = 8fF$
  - Clock frequency: 2.5 GHz



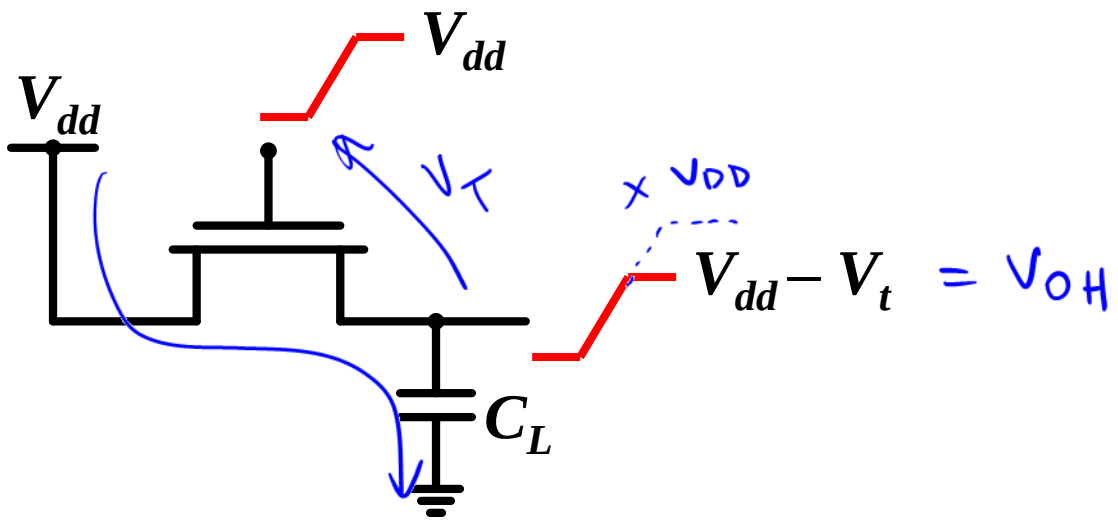
- **Power per gate**

$$P_1 = f_{clk} \cdot C_L \cdot V_{DD}^2 = 2.5 \times 10^9 \cdot 8 \times 10^{-15} = 20\mu W$$

- **Now, with many gates**
  - Activity: 0.1 $\rightarrow \alpha_{0\rightarrow1}$
  - 10 M gates $= N$

$$P_{tot} = \alpha_{0\rightarrow1} \cdot N \cdot P_1 = 20W$$

Modification for Circuits With Reduced Swing



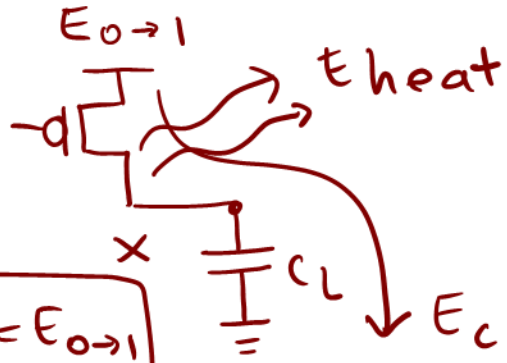
$$E_{0\to1} = C_L \cdot V_{DD} \cdot (V_{DD} - V_T) < C_L V_{DD}^2$$

- Can exploit reduced swing for lower power (e.g., reduced bit-line swing in memory)

General Formulas & Energy Cycle

- **Basic formula**
  - Energy(V_{DD}) = Energy(heat) + Energy(C_L)
- **Components:**

Vout RISING:



$$E_{0\to1} = V_{DD}\cdot C_L \int_{V_{OL}}^{V_{OH}} dV = C_L\cdot V_{DD}\cdot(V_{OH}-V_{OL}) = E_{0\to1}$$

$$E_C = C_L \int_{V_{OL}}^{V_{OH}} V\cdot dV = \frac{1}{2}C_L\cdot(V_{OH}^2 - V_{OL}^2) = E_C$$

$$E_{0\to1} - E_C = E_{heat}$$

special case (CMOS)

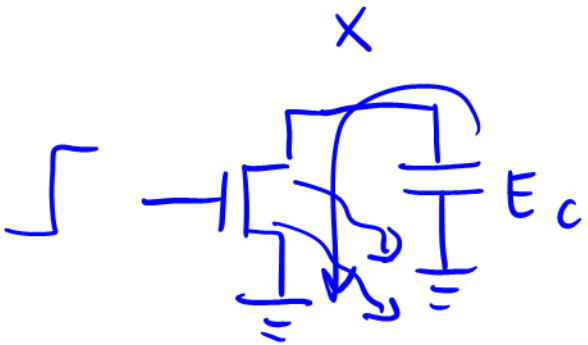
$V_{OH} = V_{DD}$, $V_{OL} = 0$

$$E_{0\to1} = C_L V_{DD}^2$$

$$E_{heat}^{(P)} = E_C = \frac{1}{2}C_L V_{DD}^2$$

$$E_{heat}^{(N)} = \frac{1}{2}C_L V_{DD}^2$$

Vout FALLING:



$$E_{1\to0} = 0$$

$$E_{heat} = E_C$$

Node Transition Activity and Power

- **Consider switching a CMOS gate for N clock cycles**

$$E_N = C_L \cdot V_{DD}^2 \cdot n(N)$$

E_N: the energy consumed for N clock cycles
$n(N)$: the number of 0→1 transitions in N clock cycles

$$P_{avg} = \lim_{N\to\infty} \frac{E_N}{N} \cdot f_{clk} = \left(\lim_{N\to\infty} \frac{n(N)}{N} \right) \cdot C_L \cdot V_{DD}^2 \cdot f_{clk}$$

$$\alpha_{0\to1} = \lim_{N\to\infty} \frac{n(N)}{N}$$

$\alpha_{0\to1} C_L = C_{sw}$

$\alpha_{0\to1} f_{clk} = f_{0\to1}$

$$P_{avg} = \alpha_{0\to1} \cdot C_L \cdot V_{DD}^2 \cdot f_{clk}$$

Lowering Dynamic Power

Capacitance:
Function of fan-out, wire length, transistor sizes

Supply Voltage:
Has been dropping with successive generations

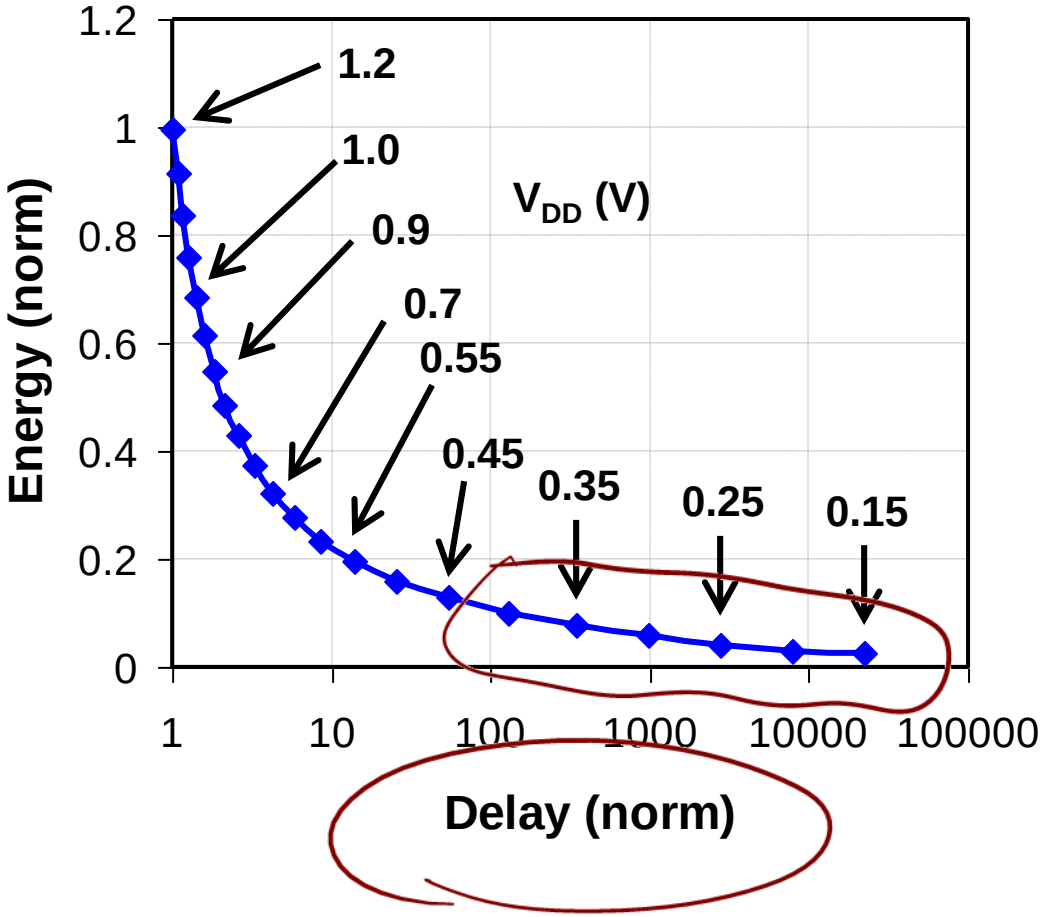
$$P_{dyn} = C_L \cdot V_{DD}^2 \cdot \alpha_{0\rightarrow1} \cdot f_{clock}$$

Activity factor:
How often, on average, do wires switch?

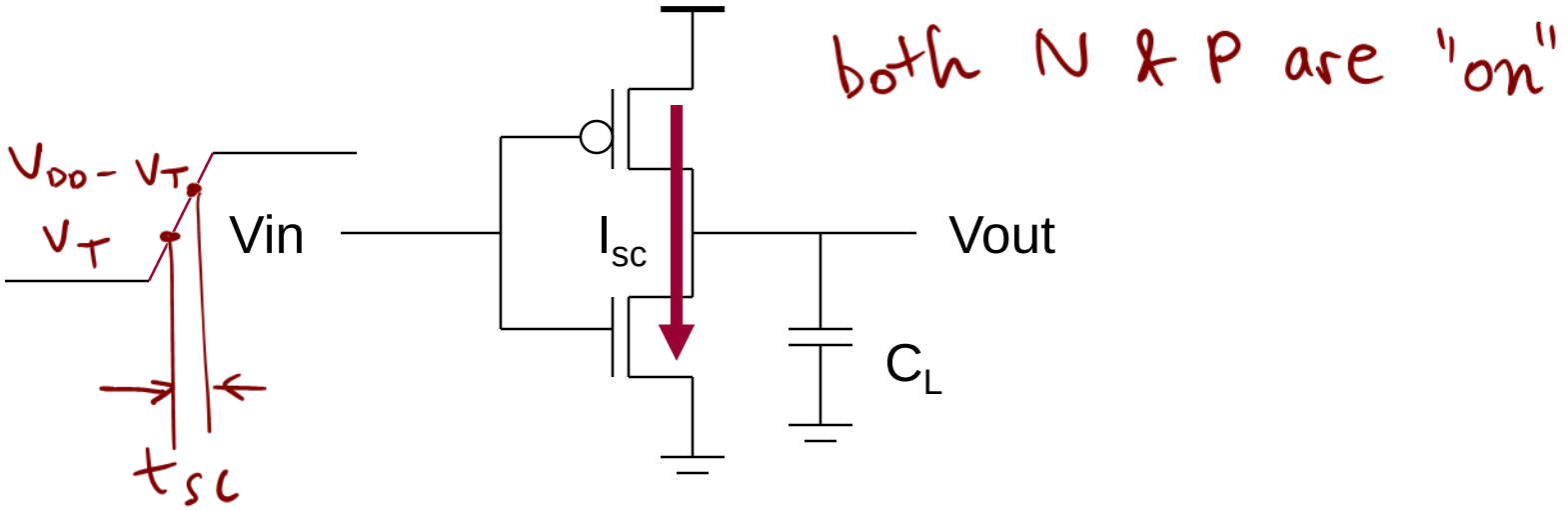
Clock frequency:
Increasing…

Dynamic Power as a Function of V_{DD}

- Decreasing V_{DD} decreases dynamic energy consumption (quadratically)
- But, increases gate delay (decreases performance)
- Scaling into the sub-threshold regime results in very large delays (100-1000x)



#2: Short-Circuit Power Consumption

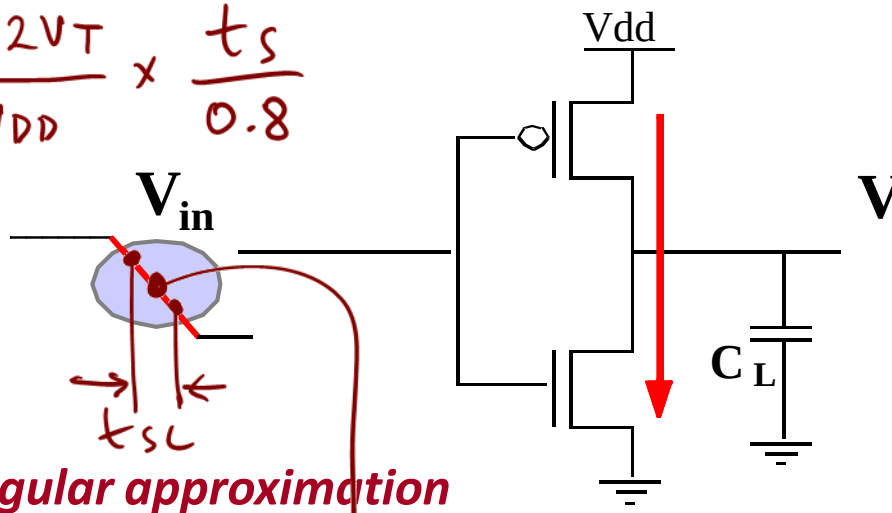


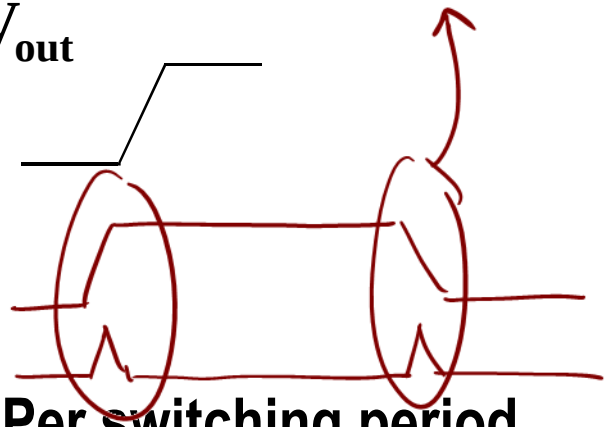
- **Short-circuit current**
  - Finite slope of the input signal causes a direct current path between V_{DD} and GND for a short period of time during switching when both the NMOS and PMOS transistors are conducting
  - **Both LH and HL transitions have short-circuit current**

Calculating Short-Circuit Power

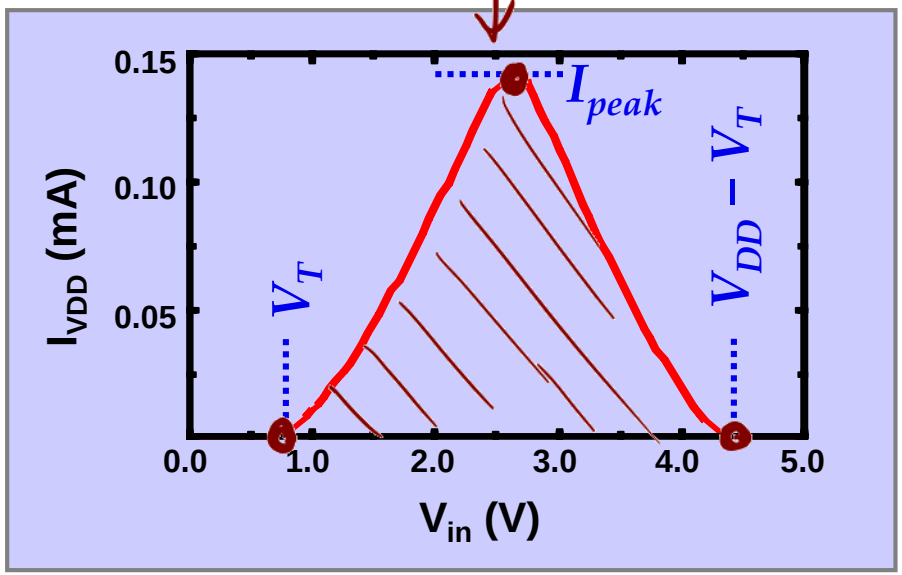
$t_{sc} \approx \frac{V_{DD}-2V_T}{V_{DD}} \times \frac{t_s}{0.8}$

$E_{sc} = \frac{1}{2} I_{peak} \cdot V_{DD} \cdot t_{sc}$





Triangular approximation



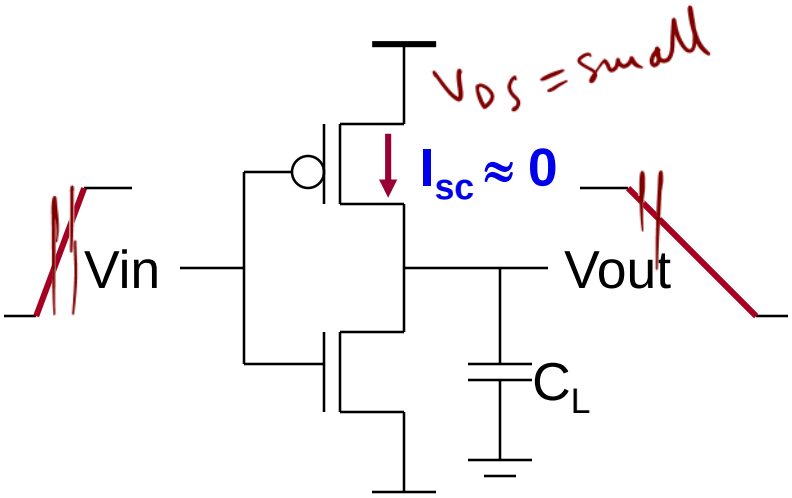
- Per switching period

$$E_{sc} = t_{sc} \cdot V_{DD} \cdot I_{peak}$$

$$P_{sc} = t_{sc} \cdot V_{DD} \cdot I_{peak} \cdot f_{0\rightarrow1}$$

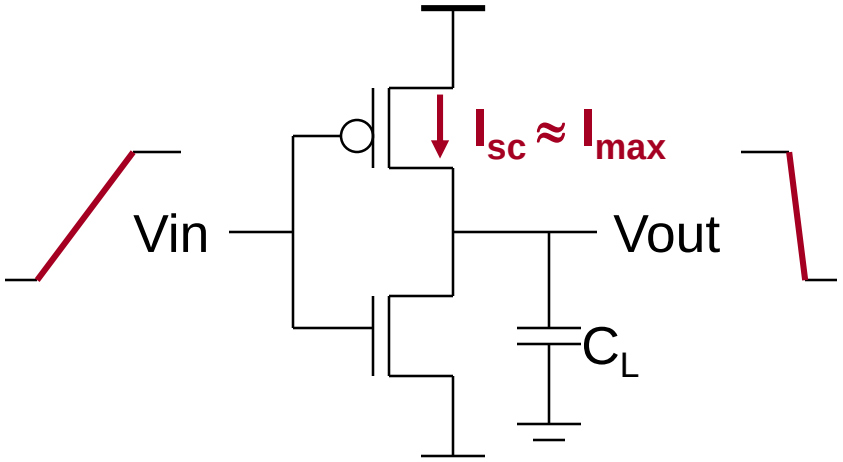
- I_{peak} depends on C_L

Impact of C_L on P_{sc}



Large capacitive load

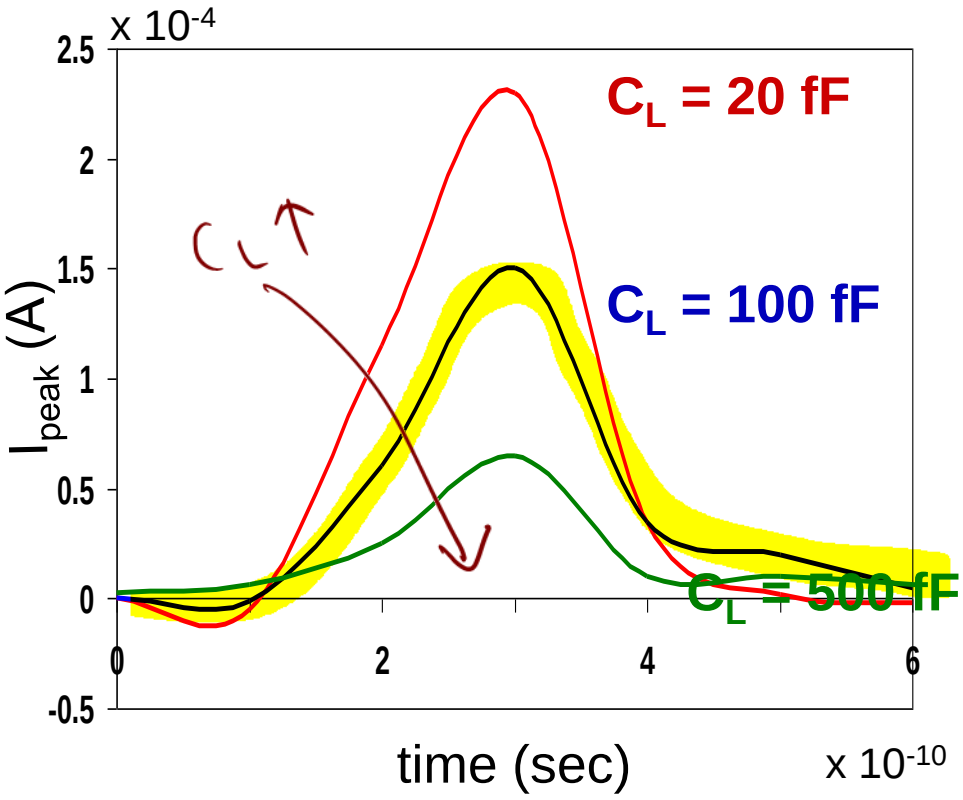
Output fall time significantly larger than input rise time.



Small capacitive load

Output fall time substantially smaller than the input rise time.

I_{peak} as a Function of C_L



500 ps input slope

Small C_L → I_{peak} is large.

⇒ Use small C_{in} and large C_L

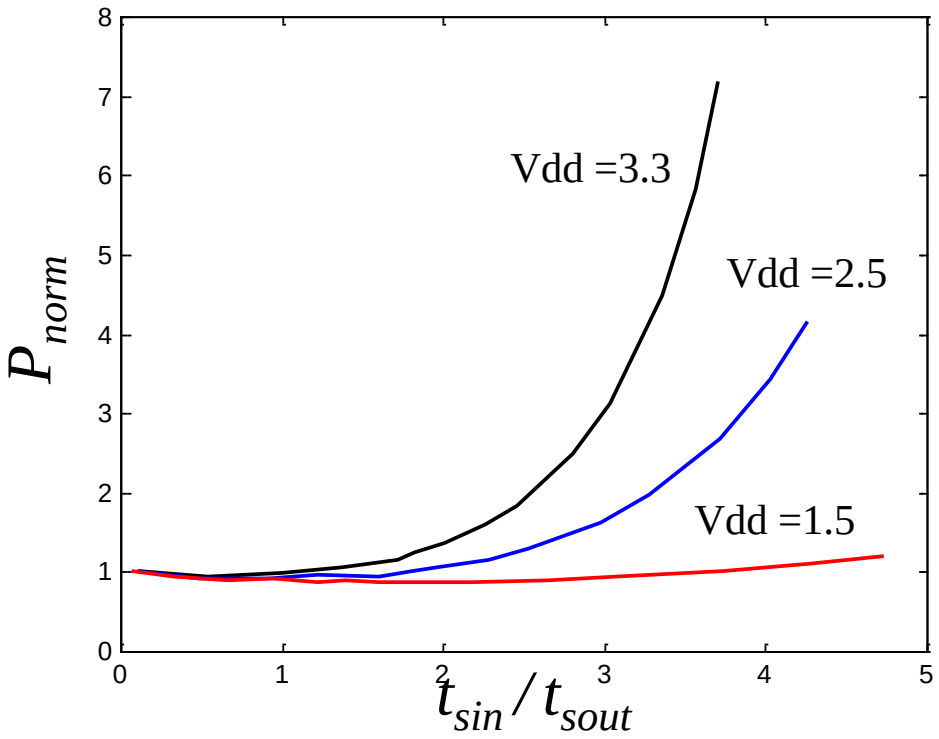
⇒ Gate delay is higher and the input rise time of the next gate (fanout gate) is higher

⇒ Higher I_{sc} in fanout gate!!

⇒ Local optim. is not good

Short circuit dissipation is minimized by matching the rise/fall times of the input and output signals - slope engineering.

Minimizing Short-Circuit Power

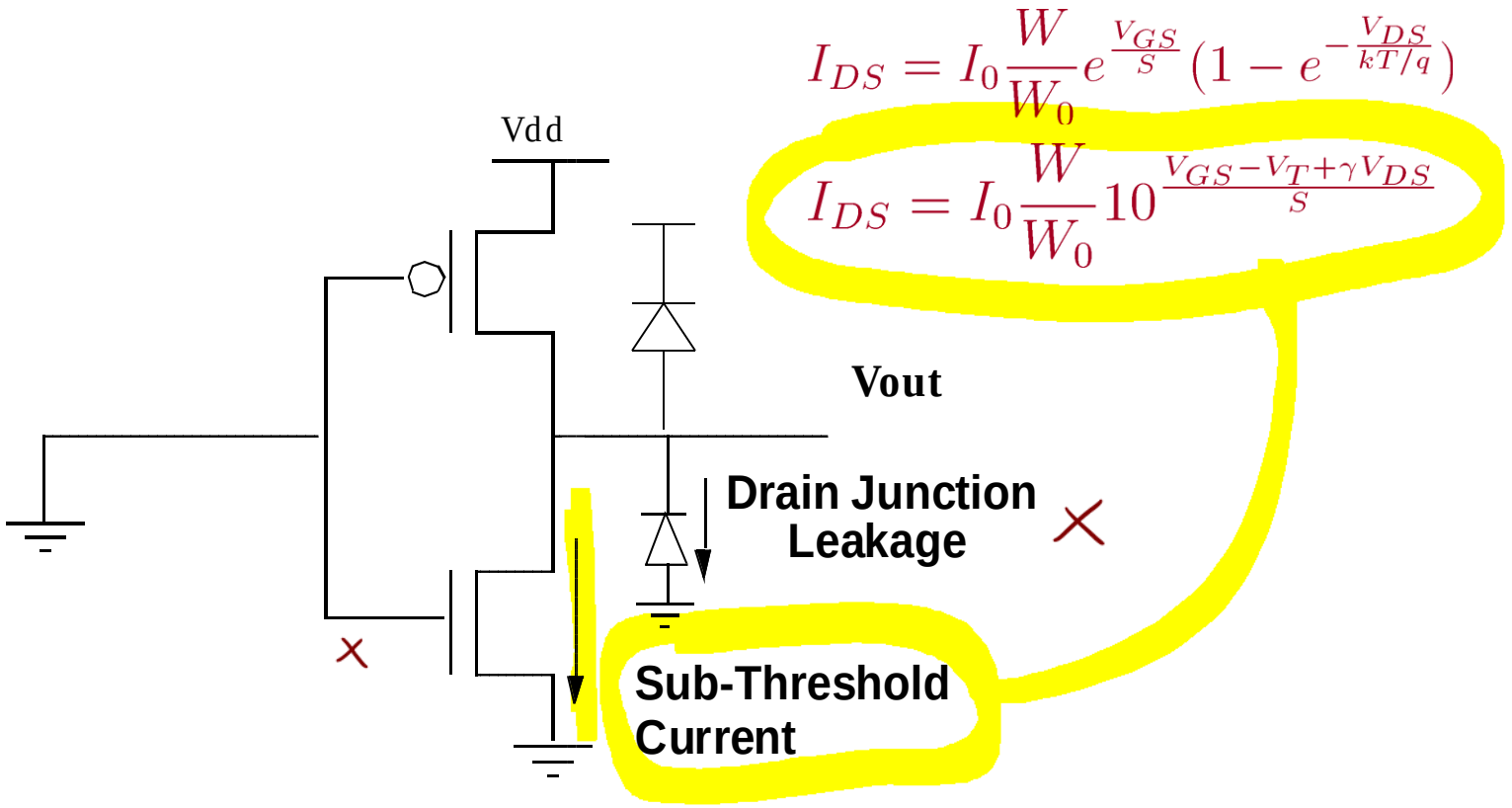


- **Keep the input and output rise/fall times the same (<10% of total consumption)**

  From: Veendrick, IEEE Journal of Solid-State Circuits, Aug'84

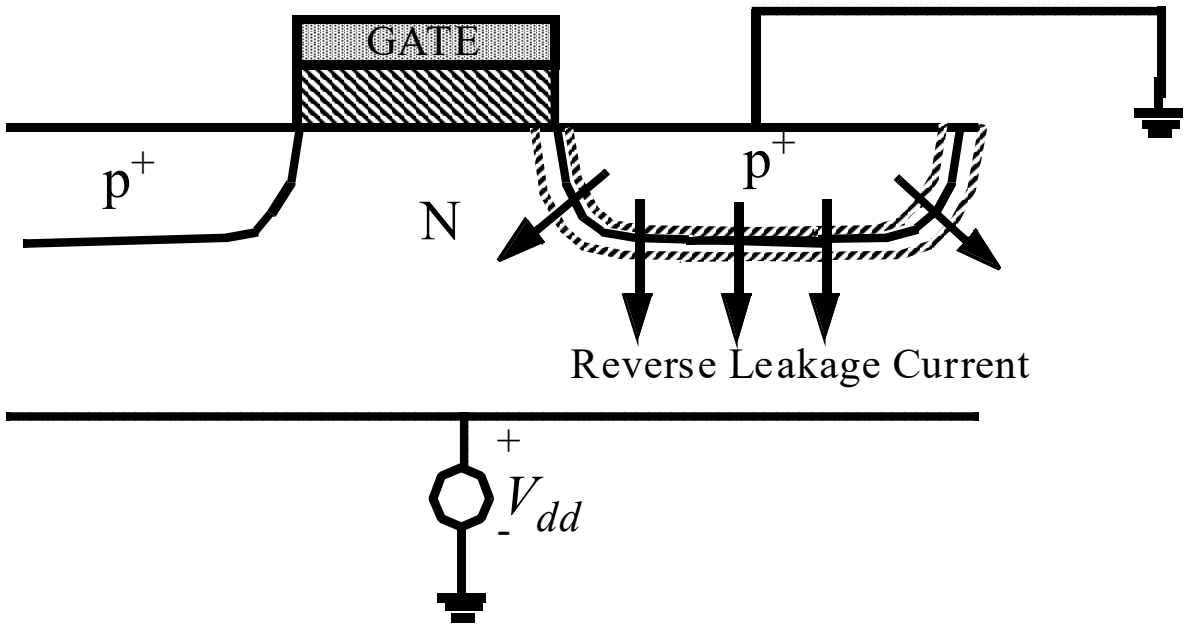
- **If $V_{dd} < V_{Tn} + |V_{Tp}|$ then short-circuit power can be eliminated!**

#3: Leakage Current



Sub-threshold current is one of the most compelling issues in low-energy circuit design!

Reverse-Biased Diode Leakage



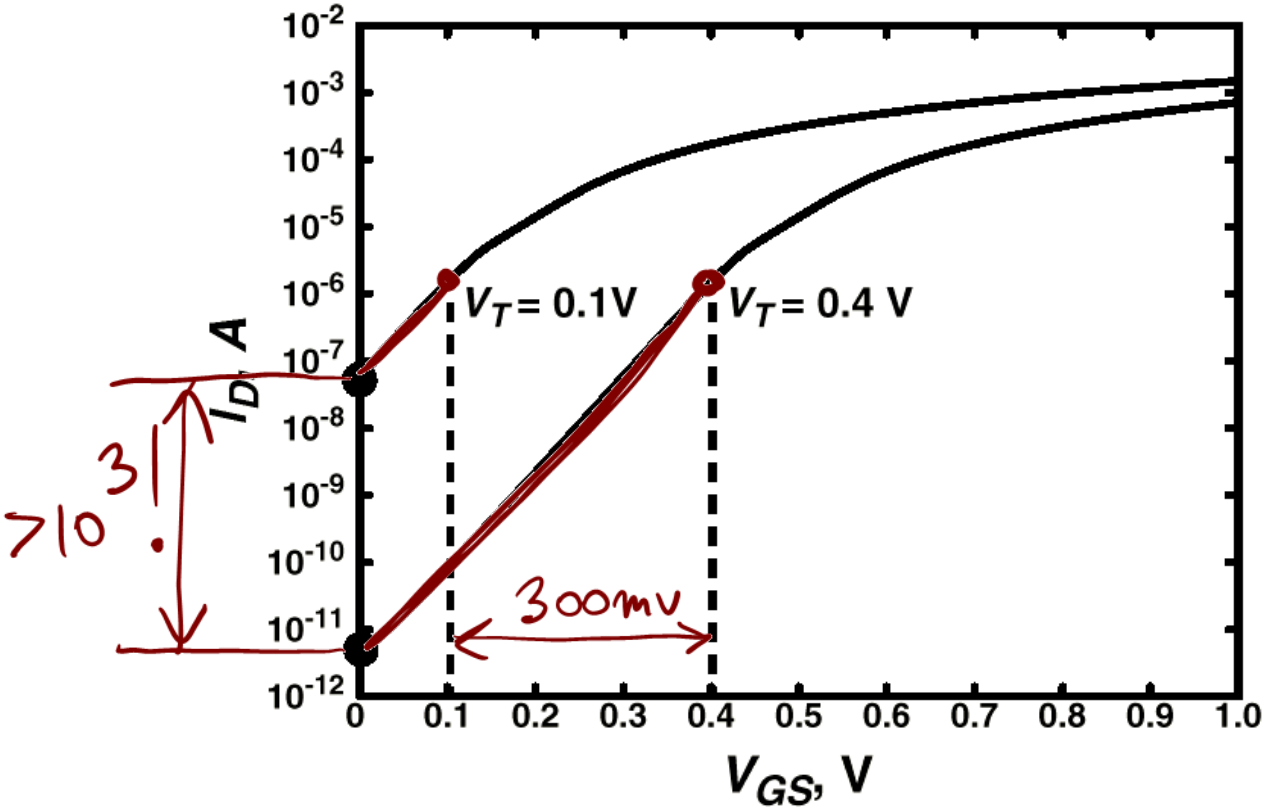
$$I_{DL} = J_S \times A$$

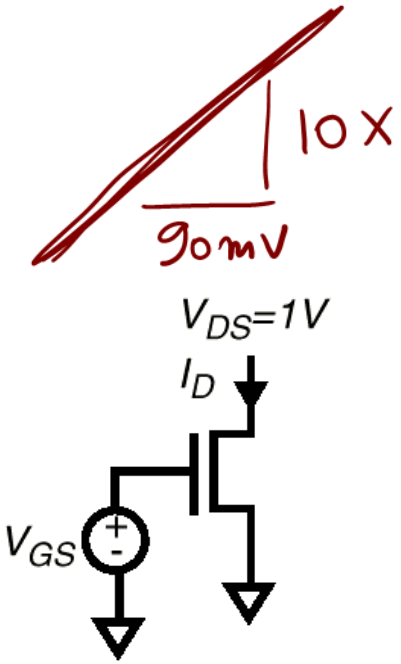
J_S = 10-100 pA/μm^2 at 25 deg C for 0.25μm CMOS

J_S doubles for every 9 deg C!

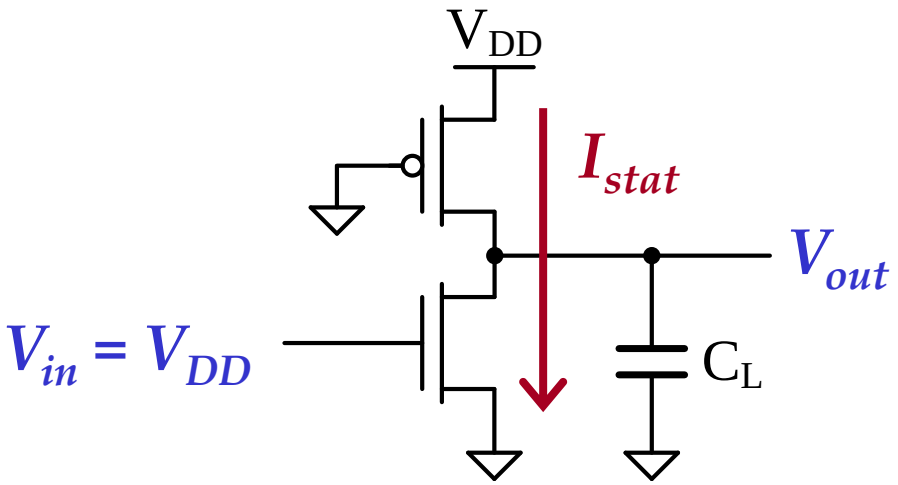
Sub-Threshold Leakage Component

- Leakage control is critical for low-voltage operation





#4: Static Power Consumption



$$P_{stat} = P(in = 1) \cdot V_{DD} \cdot I_{stat}$$

Wasted energy …
Should be avoided in most cases,
but could help reducing energy in others (e.g. sense amps)

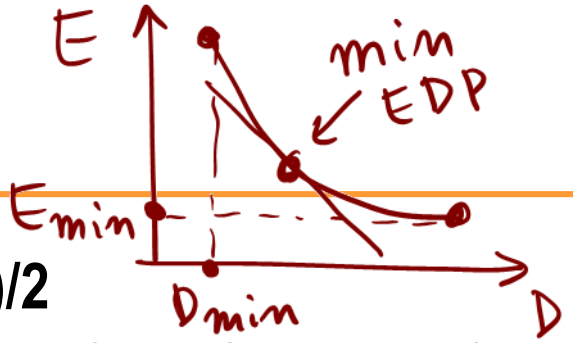
General Principles for Power Reduction

- Prime choice: Reduce voltage!
- Reduce switching activity
- Reduce physical capacitance

Power and Energy Figures of Merit

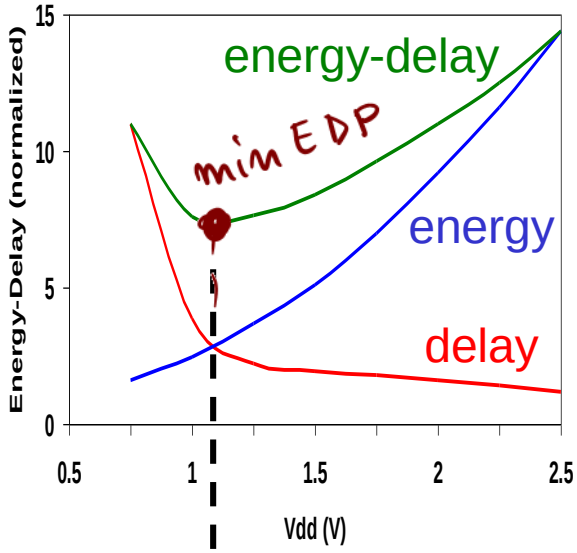
- **Power consumption in Watts**
  - determines battery life in hours
- **Peak power**
  - determines power ground wiring designs
  - sets packaging limits
  - impacts signal noise margin and reliability analysis
- **Energy efficiency in Joules**
  - rate at which power is consumed over time
- **Energy = power * delay**
  - Joules = Watts * seconds
  - lower energy number means less power to perform a computation at the same frequency

PDP and EDP



- **Power-delay product (PDP) = $P_{av} * t_p = (C_L V_{DD}^2)/2$**
  - PDP is the average energy consumed per switching event (Watts * sec = Joule)
  - **Lower power design could simply be a slower design**

- **Energy-delay product (EDP)**
  - EDP = PDP * t_p = P_{av} * t_p^2
  - EDP is the average energy consumed multiplied by the computation time required
  - **Takes into account that one can trade increased delay for lower energy/op (e.g. via V_{DD} scaling)**



CMOS Energy & Power Equations (Summary)

	Dynamic	Short-circuit	Leakage
Energy	$C_L \cdot V_{DD}^2$	$t_{sc} \cdot V_{DD} \cdot I_{peak}$	$V_{DD} \cdot I_{leakage} / f_{clock}$
Power	$C_L \cdot V_{DD}^2 \cdot f_{0\rightarrow1}$	$t_{sc} \cdot V_{DD} \cdot I_{peak} \cdot f_{0\rightarrow1}$	$V_{DD} \cdot I_{leakage}$
	~75% today and decreasing relatively	~5% today and decreasing absolutely	~20% today and slowly increasing

- Switching frequency / activity
  - $f_{0\rightarrow1} = \alpha_{0\rightarrow1} \cdot f_{clock}$

Simplified Model for Circuit Analysis

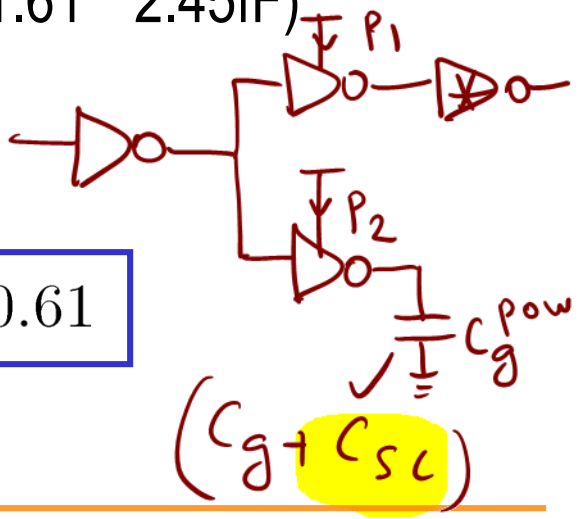
- **Often we assume that switching energy is dominant**
- **Similarly to delay analysis, we can find "equivalent" capacitance for power analysis**
  - It is to expect that this capacitance will be higher, because it includes short-circuit power and leakage
- **In our process, C_{in} (power) = 2.45fF**
  - Including output parasitic $C_{in} + C_{par}$ = 3.95fF (1.61 * 2.45fF)
- **Simplified model for hand analysis:**

power:

$$C_{in} = 2.45fF/\mu m$$

$$C_{par}/C_{in} = 0.61$$

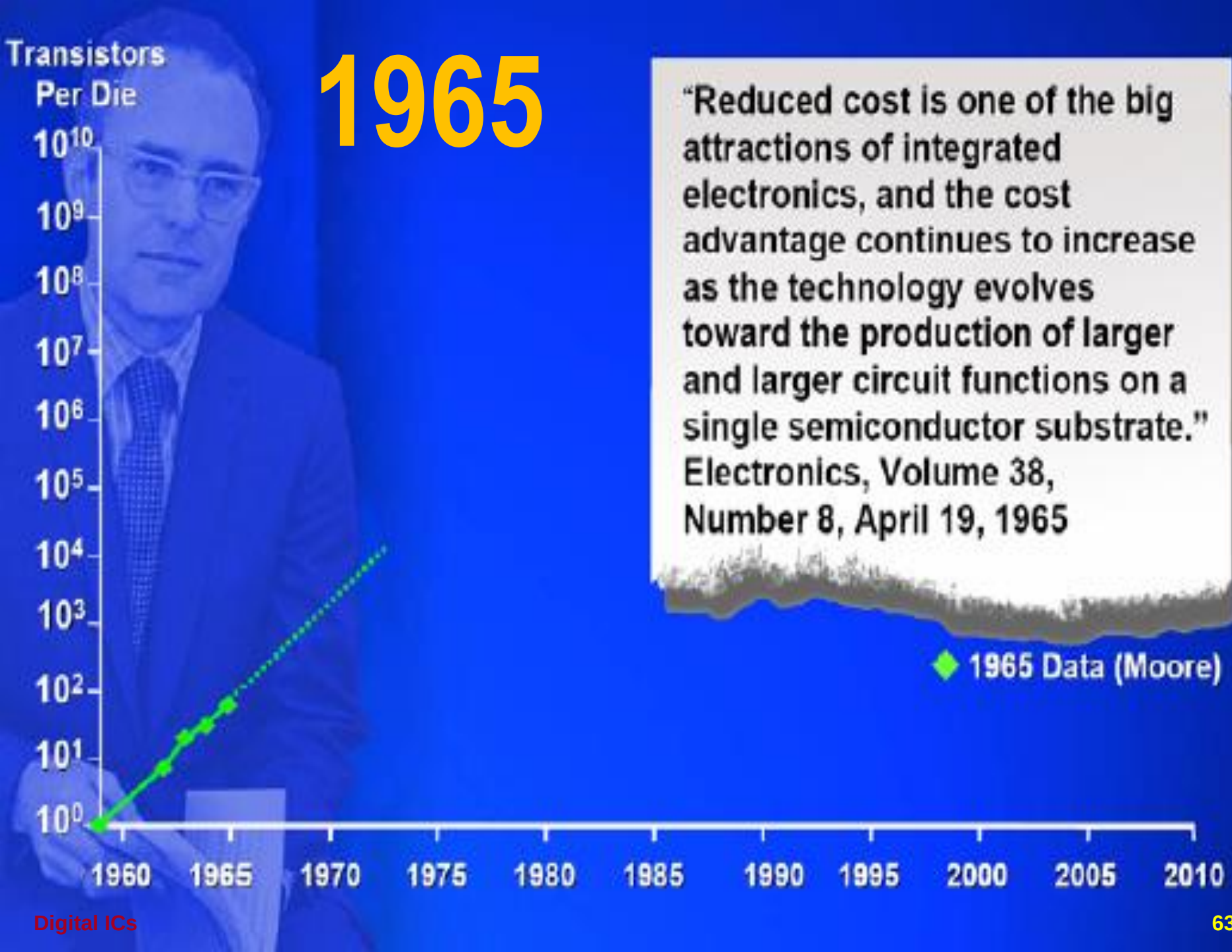
"power" cap ≠ "delay" cap!



Week 2 Agenda

- MOS RC Model
- Delay Model
- Power Model
- CMOS Scaling

1965



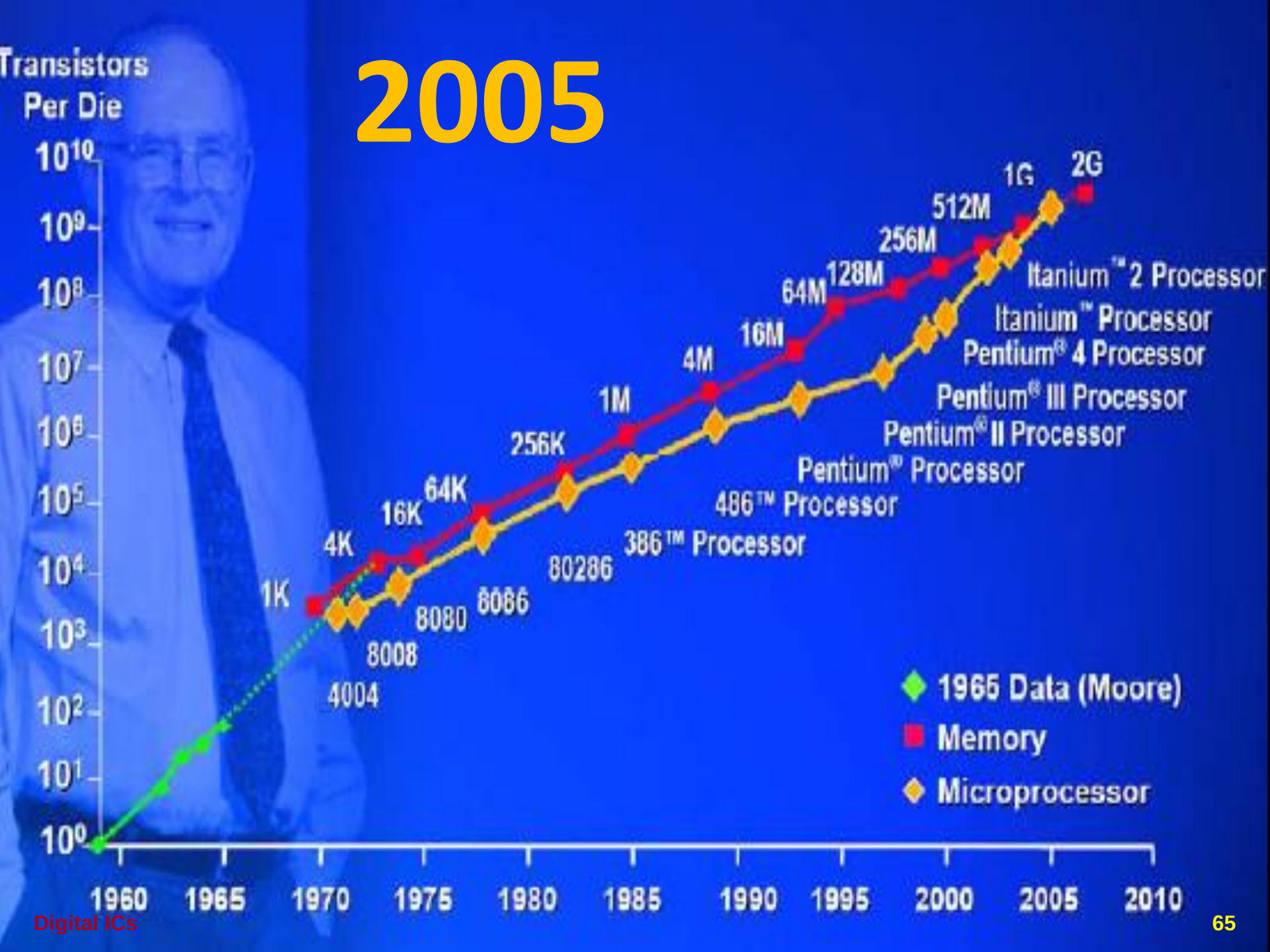
"Reduced cost is one of the big attractions of integrated electronics, and the cost advantage continues to increase as the technology evolves toward the production of larger and larger circuit functions on a single semiconductor substrate."
Electronics, Volume 38, Number 8, April 19, 1965

Moore's Law

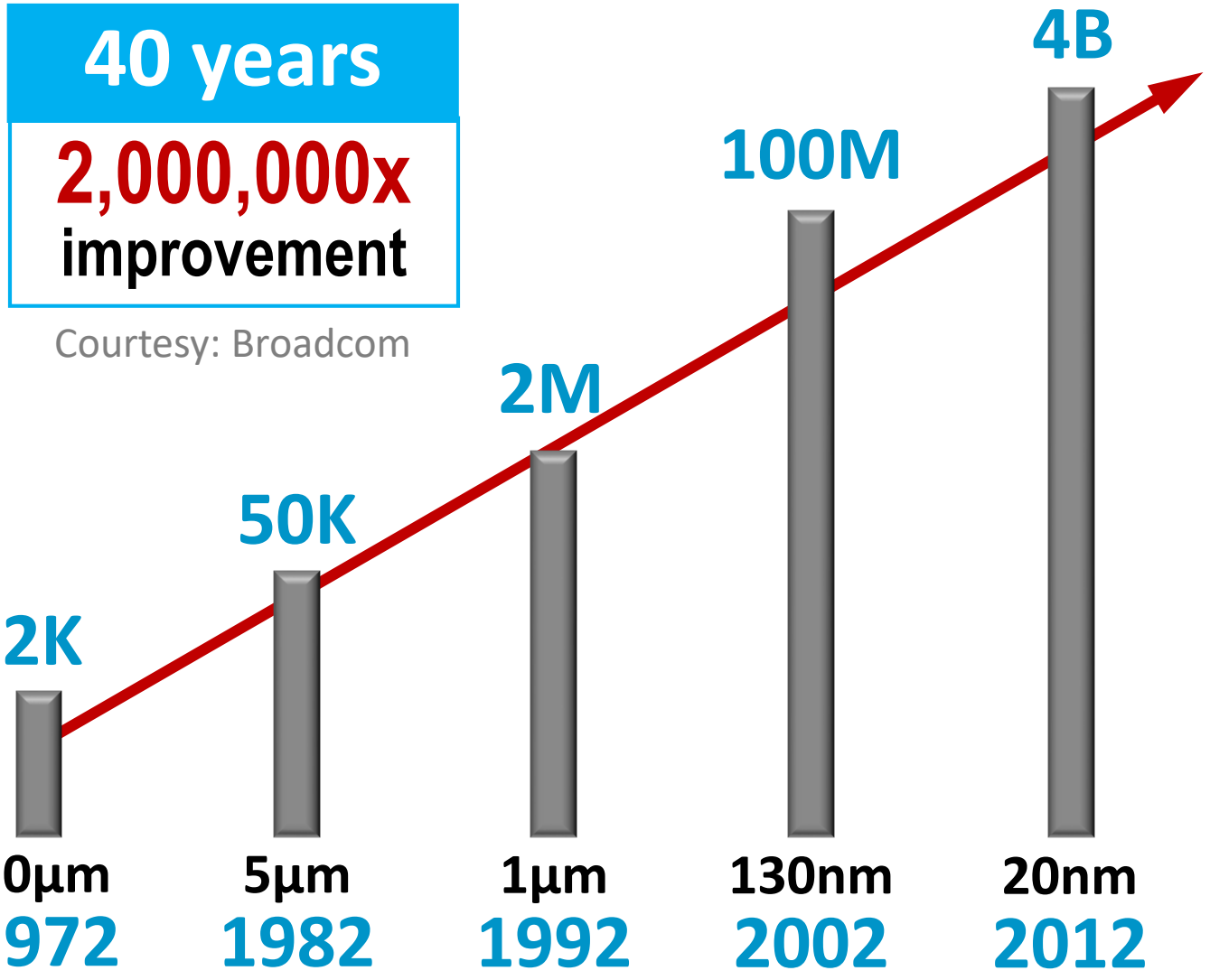
- **In 1965, Gordon Moore noted that the number of transistors on a chip doubled every 18 to 24 months**

"The complexity for minimum component costs has increased at a rate of roughly a factor of two per year. Certainly over the short term, this rate can be expected to continue, if not to increase. Over the longer term, the rate of increase is a bit more uncertain, although there is no reason to believe it will not remain nearly constant for at least 10 years. That means by 1975, the number of components per integrated circuit for minimum cost will be 65,000."

[G. Moore, Electronics, 1965]


2005
Transistors
Per Die
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
10^{5}
10^{4}
10^{3}
10^{2}
10^{1}
10^{0}
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
1K
4K
16K
64K
256K
1M
4M
16M
64M
128M
256M
512M
1G
2G
4004
8008
8080
8086
80286
386™ Processor
486™ Processor
Pentium® Processor
Pentium® II Processor
Pentium® III Processor
Pentium® 4 Processor
Itanium™ Processor
Itanium™ 2 Processor
1965 Data (Moore)
Memory
Microprocessor

Transistors / cm^2



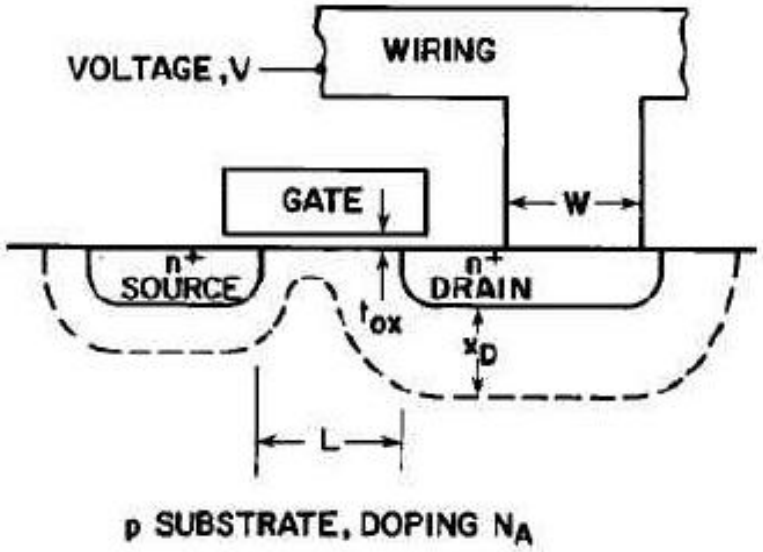
Scaling

Voltage: V_{DD}, V_T

Size: W, L, t_{ox}

Dennard's Classical MOSFET Scaling (1974)

Scaling Factor	Device or Circuit Parameter
1/κ :	Device dimension t_{ox}, L, W
κ :	Doping concentration Na
1/κ :	Voltage V
1/κ :	Current I
1/κ :	Capacitance $\varepsilon A/t_{ox}$
1/κ :	Delay time/circuit VC/I
$1/\kappa^2$:	Power dissipation/circuit VI
1 :	Power density VI/A



R. Dennard, JSSC, Oct 1974.

Constant E-field Scaling

Voltage and size scale by the same factor, S (S > 1)

- E = V/L = constant

Outcomes:

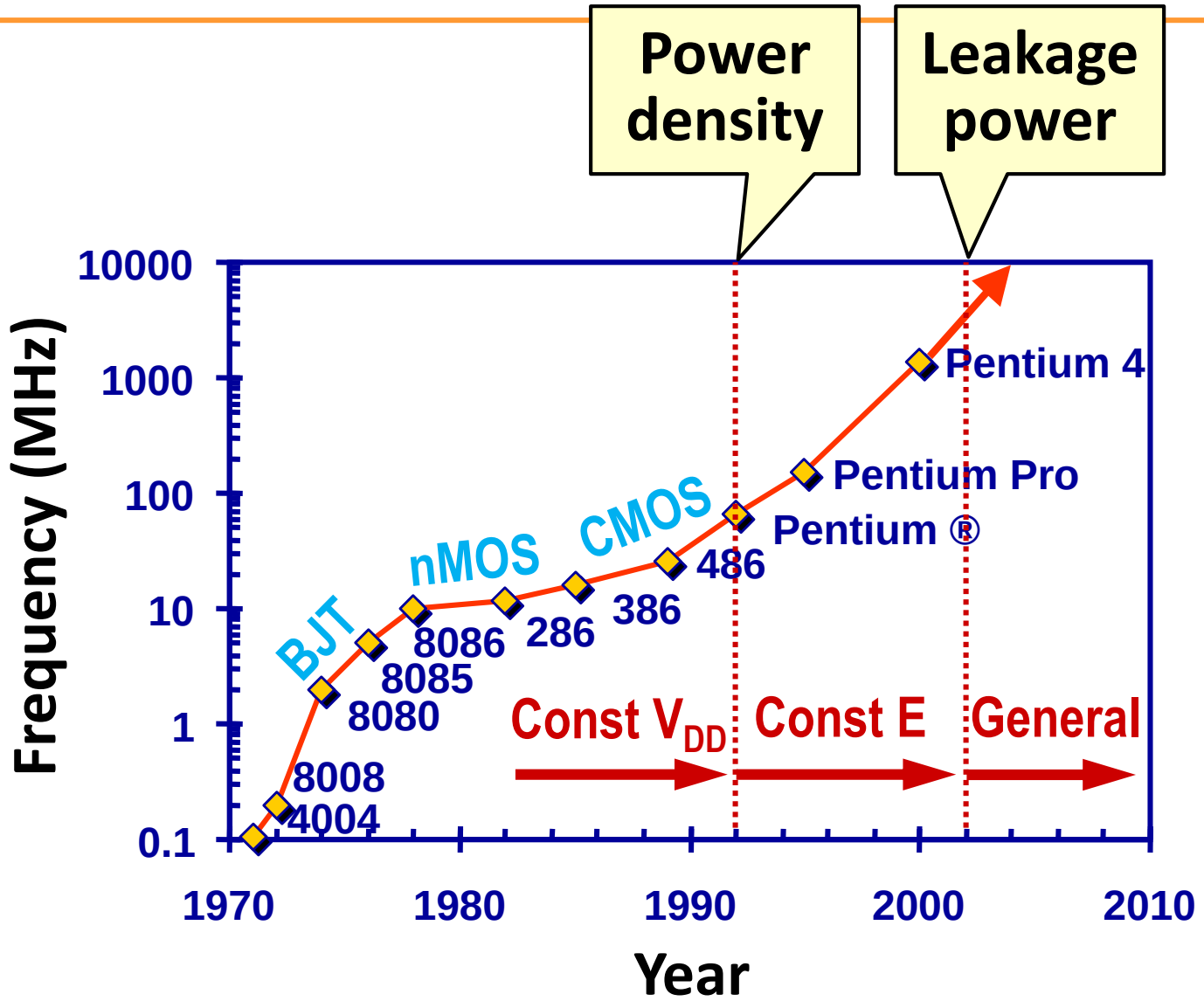
- More transistors/area $1/S^2$
- Faster delay $1/S$
- Lower energy/op $1/S^3$

Problem: V_T scaling (exponential leakage)

Constant E-field Scaling

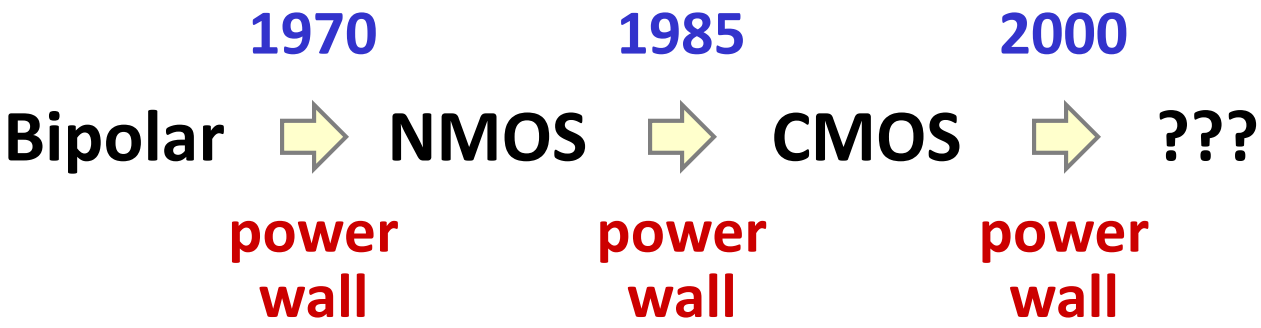
Ended at the 130nm node

Historical Scaling Trends



Courtesy: S. Borkar (Intel)

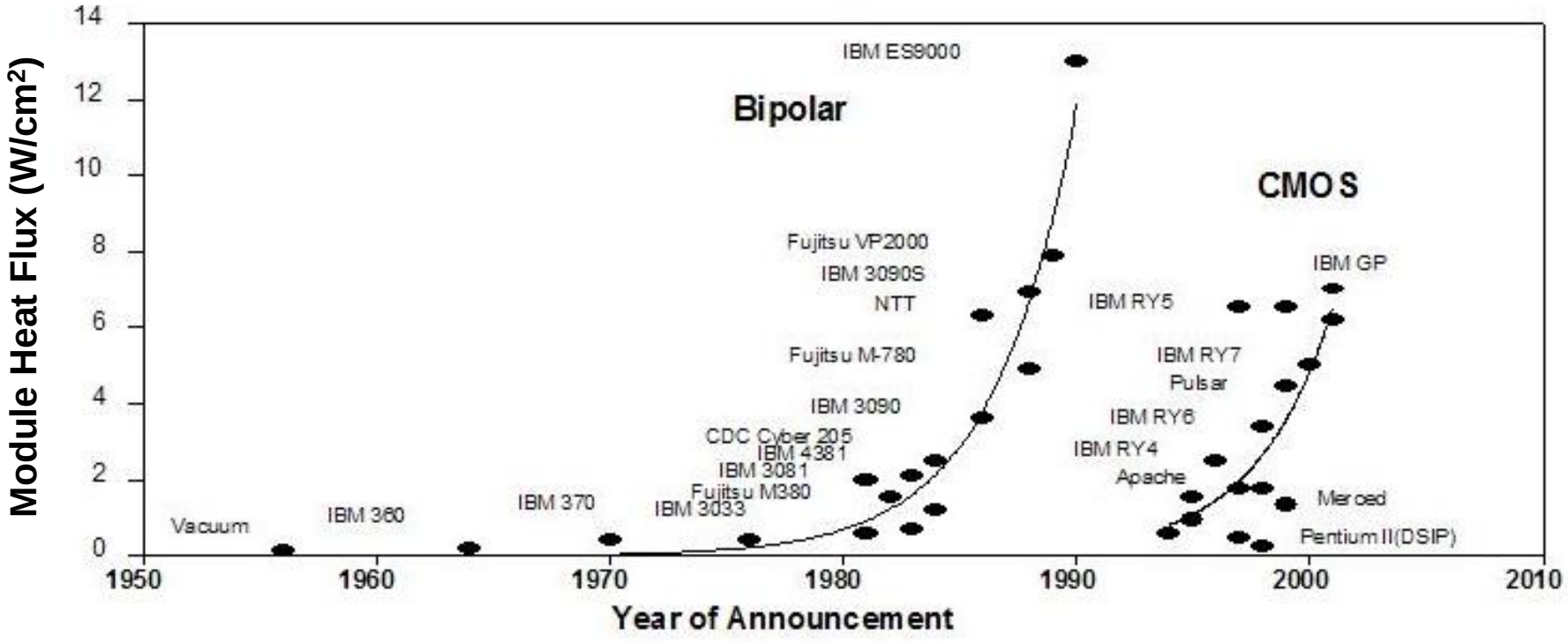
Technology Scaling is Power Driven



- **CMOS delivered better cost performance**
  - It was more energy efficient
  - It improved the integration level

Bipolar → Power Wall → CMOS

- **Technologies:** bipolar, nMOS, CMOS
- **Constant voltage scaling:** increasing power



Courtesy: Roger Schmidt (IBM)

Scaling Scenarios: Fixed V, Fixed E, General

Parameter	Relation	Fixed V	Fixed E	General
W, L, t_{ox}		***1/S***	***1/S***	***1/S***
V_{DD}, V_T		***1***	***1/S***	***1/U***
Area/Device	WL	$1/S^2$	$1/S^2$	$1/S^2$
C_{ox}	$1/t_{ox}$	S	S	S
C_{gate}	C_{ox} WL	1/S	1/S	1/S
k_n, k_p	C_{ox} W/L	S	S	S
I_{sat}	C_{ox} WV	1	1/S	1/U
Current Density	I_{sat} / Area	S^2	S	S^2/U
R_{on}	V / I_{sat}	1	1	1
Intr. Delay	R_{on} C_{gate}	1/S	1/S	1/S
Power	I_{sat} V	1	$1/S^2$	$1/U^2$
P Density	Power/Area	S^2	1	S^2/U^2

General Scaling

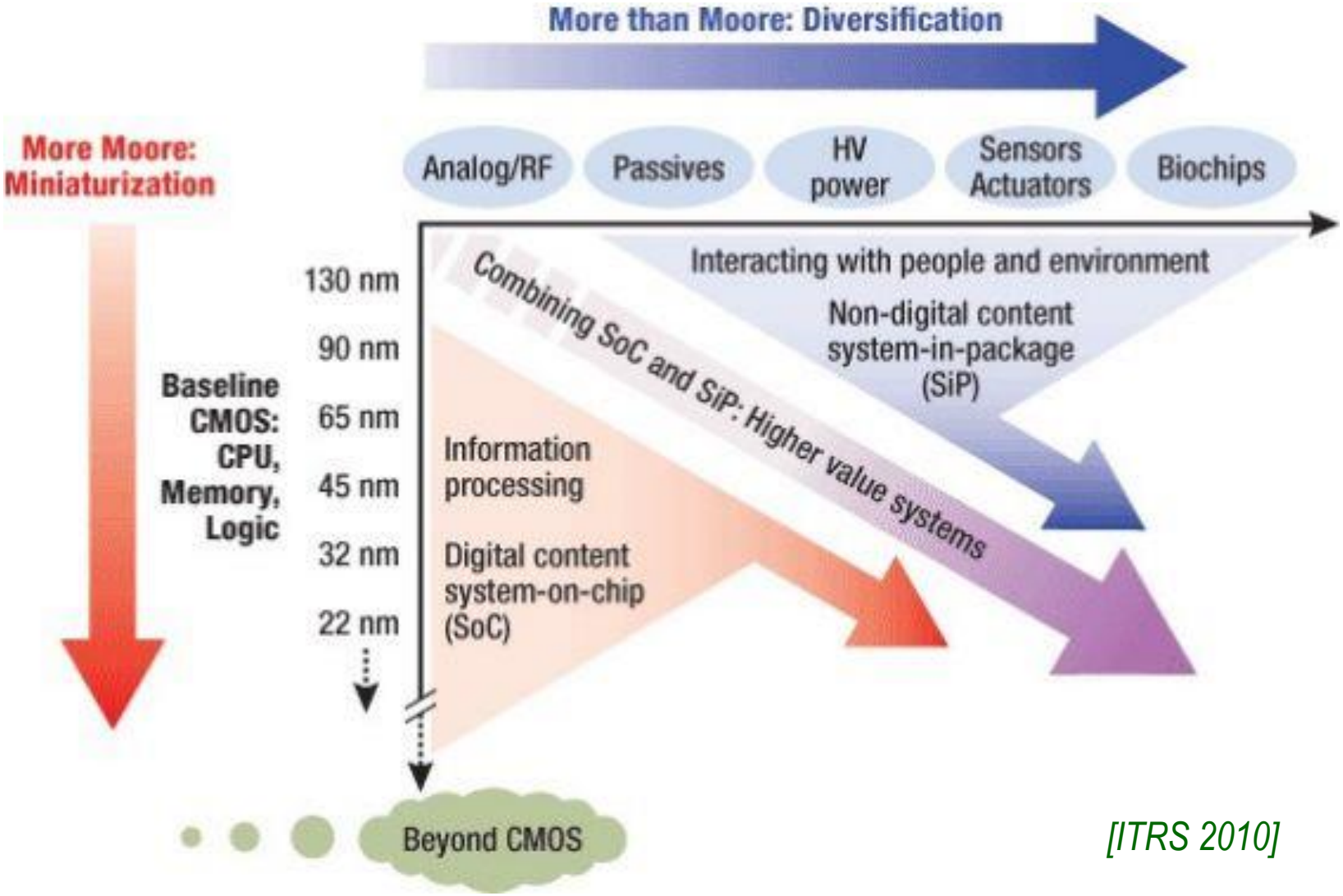
Size scaling S > Voltage scaling U

Voltage scaling slowing down

- V_T determined by leakage
- t_{ox} also set by leakage

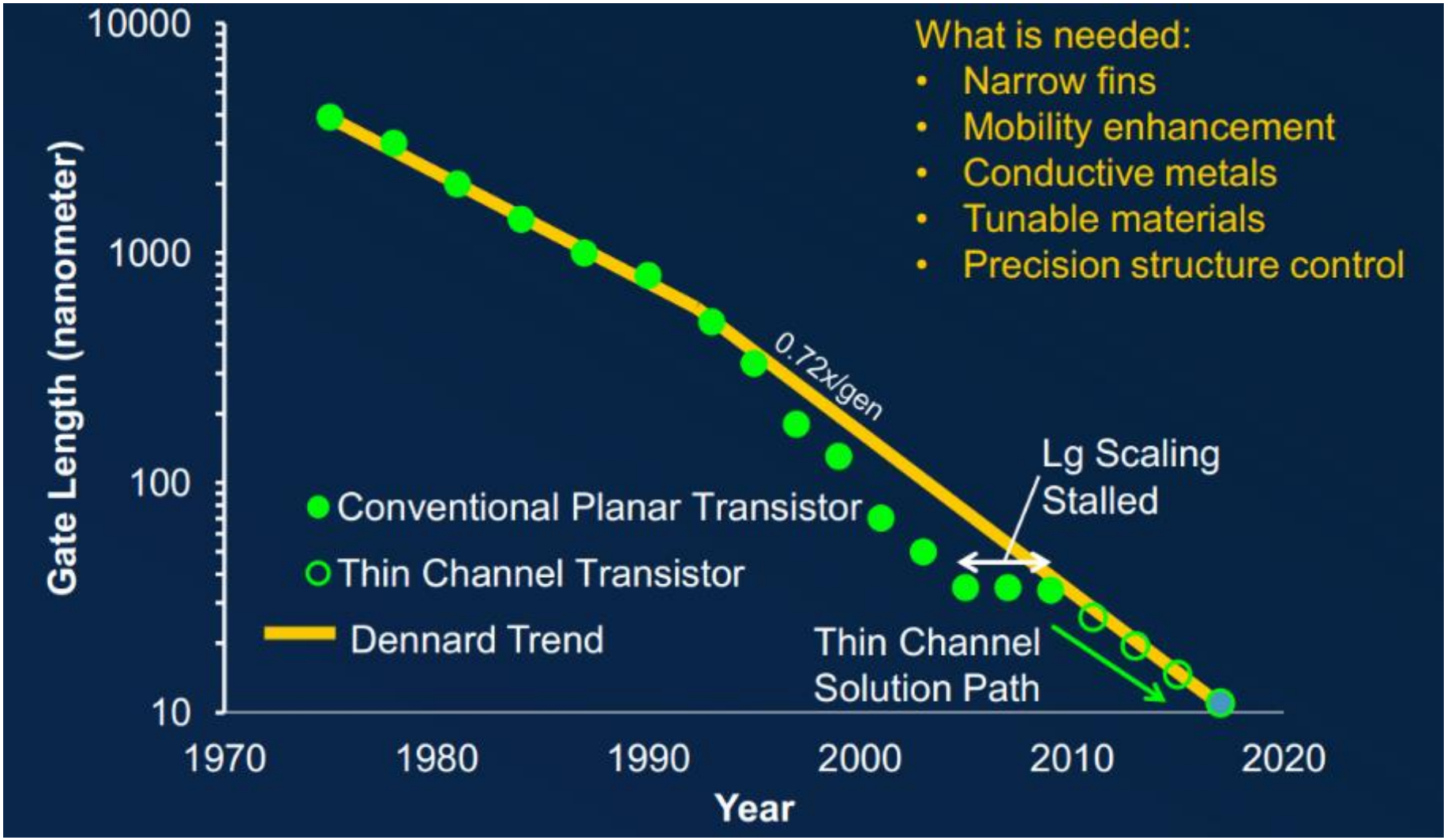
Current increasing by stressing silicon

More than Moore: 2010+



[ITRS 2010]

Channel Length Scaling Trend (Now at 5nm)

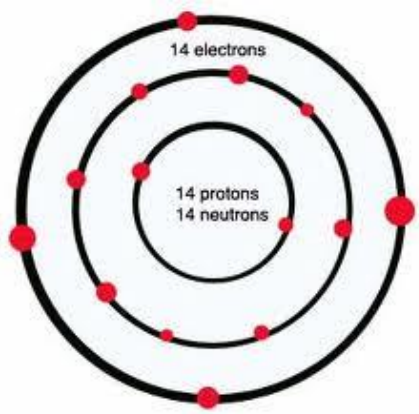


Source: Applied Materials

Approaching Atomic Limits

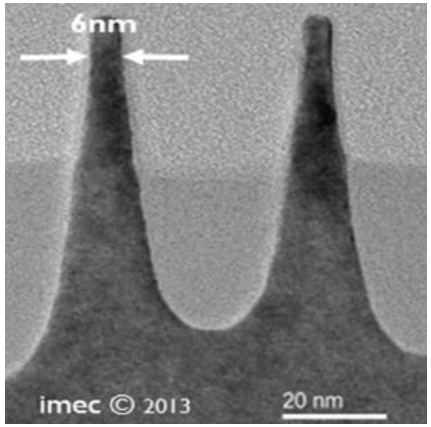
Si atom

0.25nm



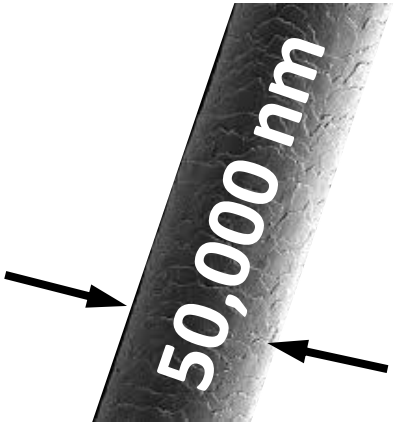
1x

7nm FET



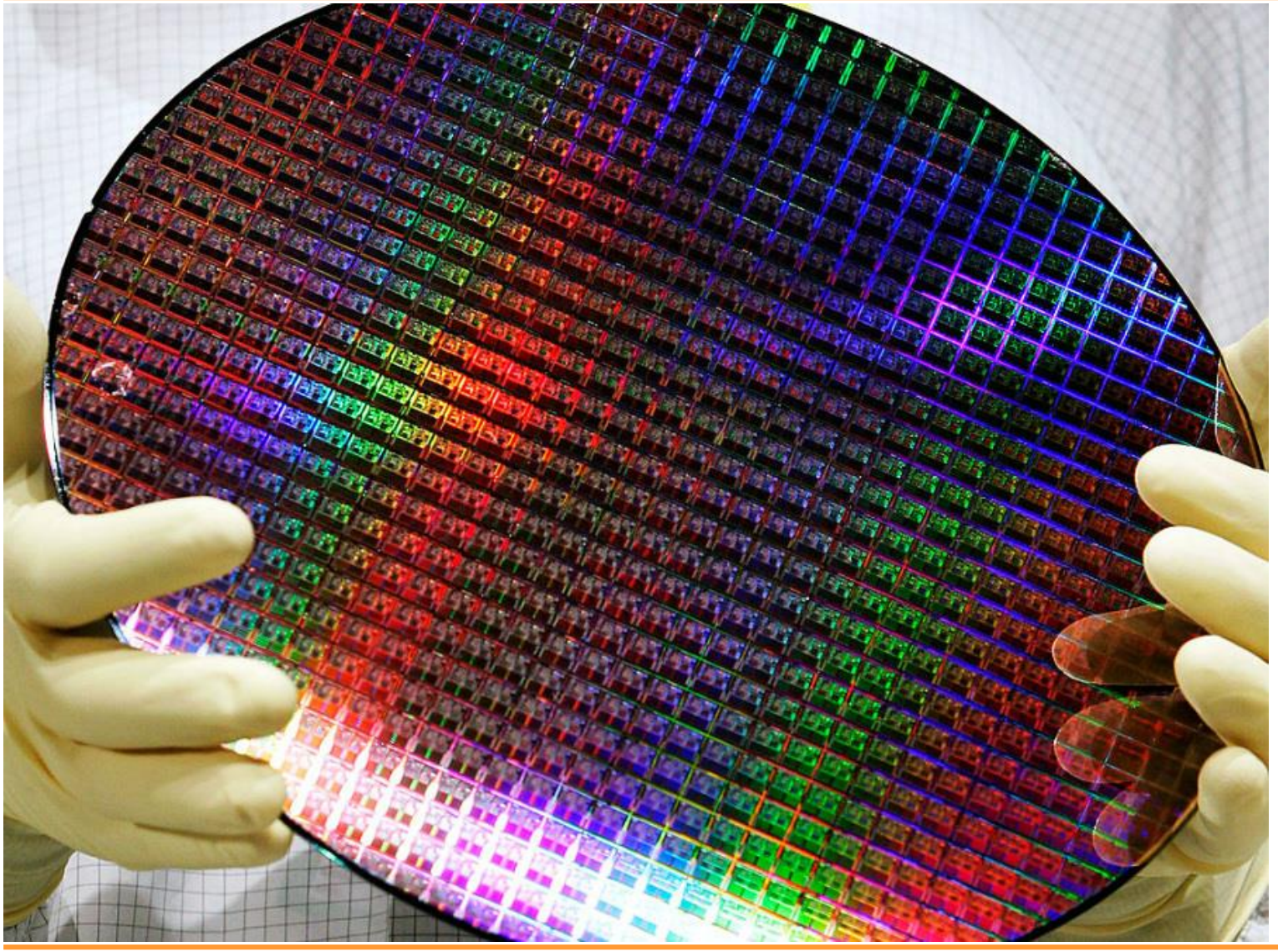
25x

Hair



200,000x

Transistor Scaling is Projected to End at 3nm



300mm wafer diameter