



# ***Digital Integrated Circuits***

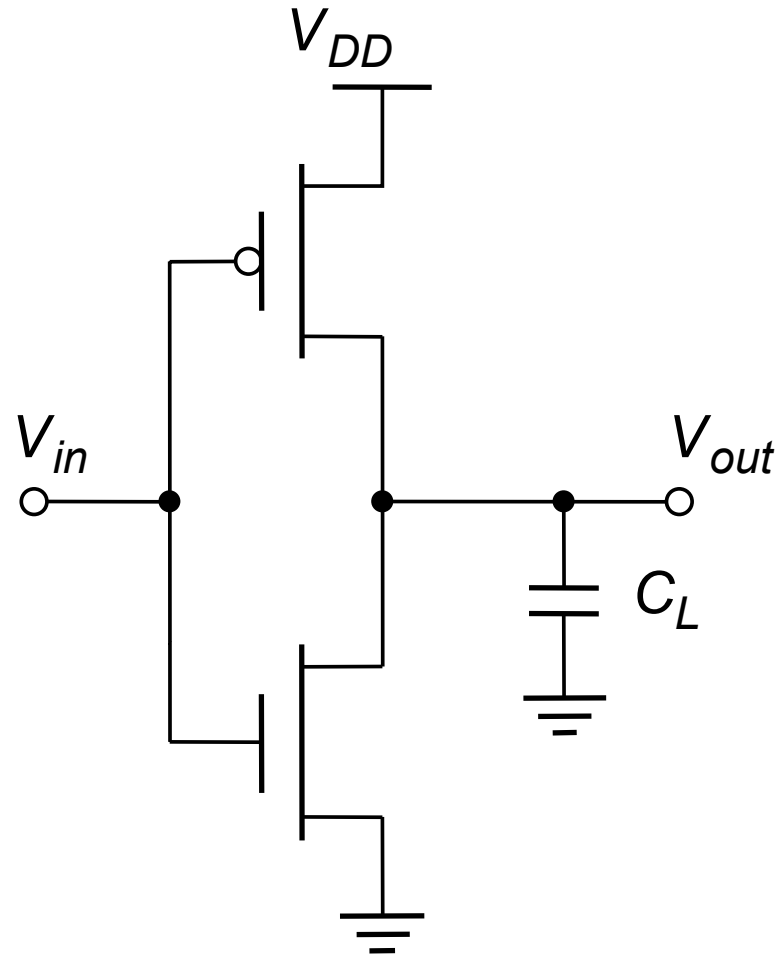
## ***A Design Perspective***

Jan M. Rabaey  
Anantha Chandrakasan  
Borivoje Nikolic

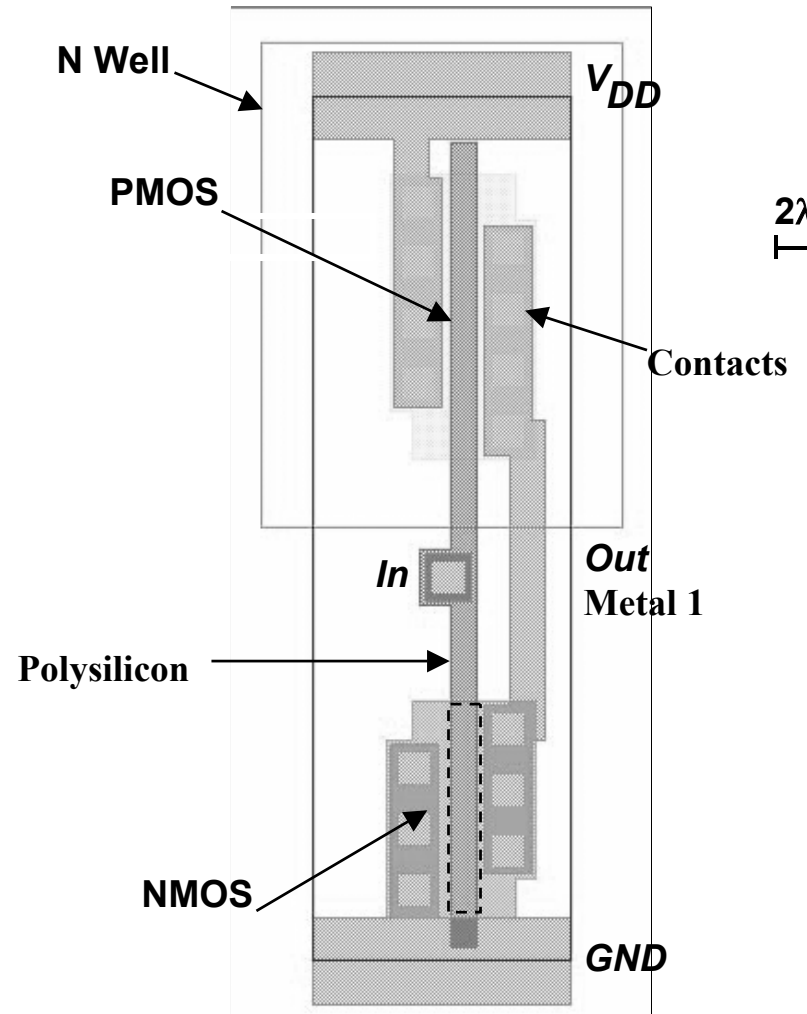
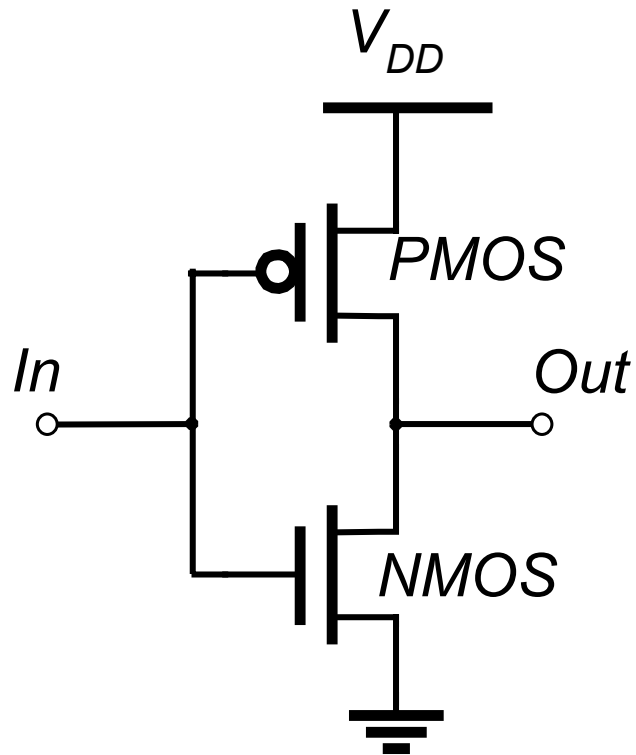
# **The Inverter**

*July 30, 2002*

# ***The CMOS Inverter: A First Glance***



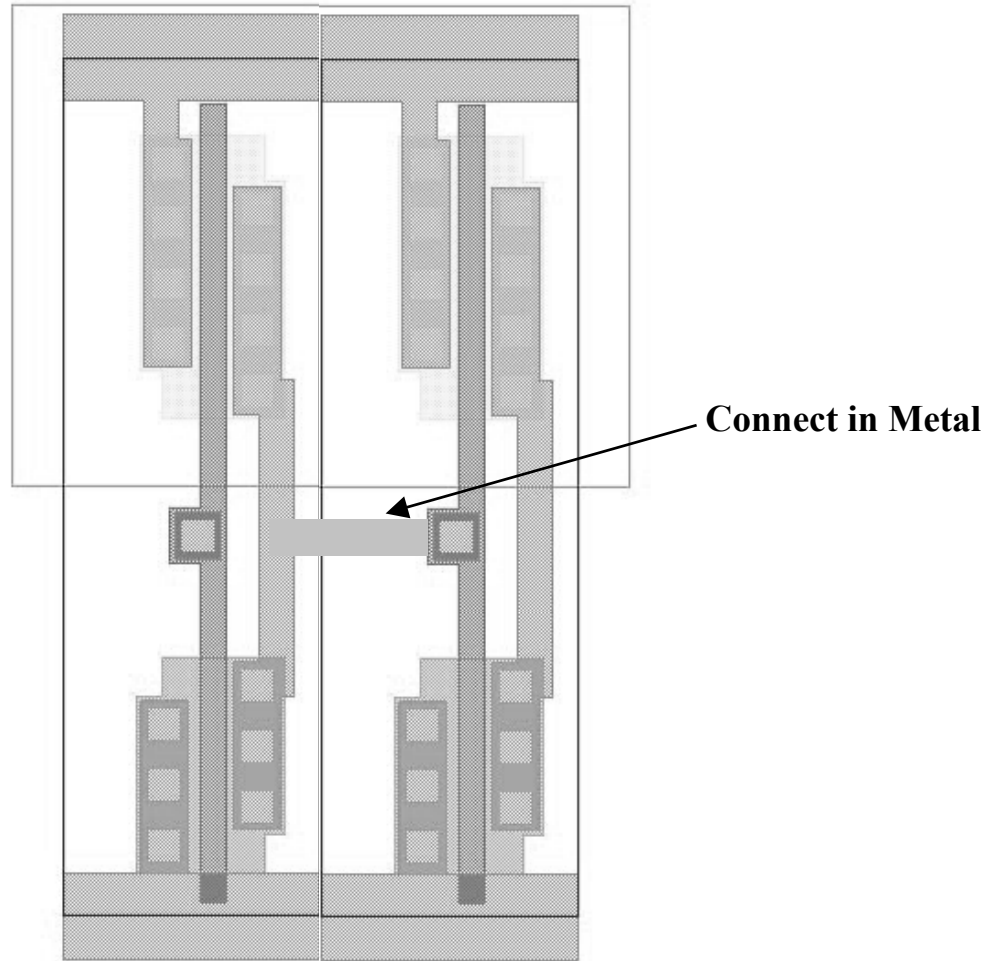
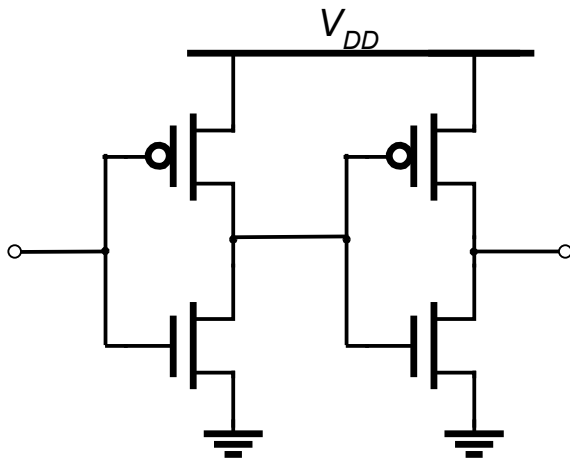
# CMOS Inverter



# Two Inverters

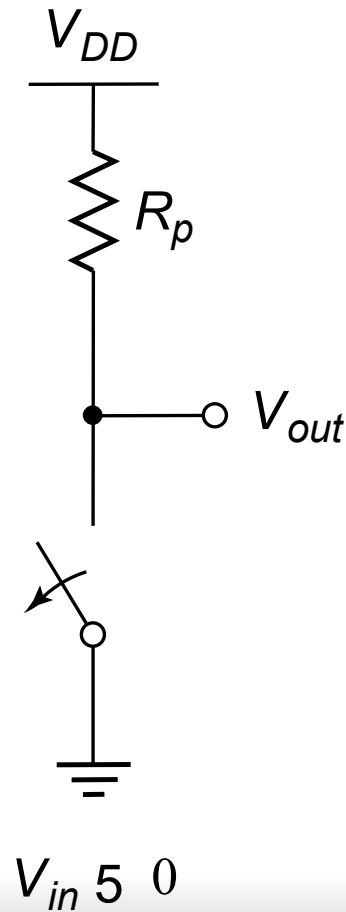
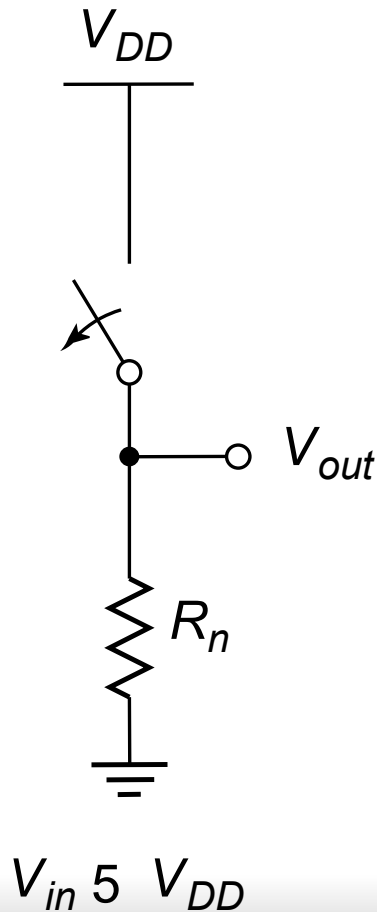
Share power and ground

Abut cells



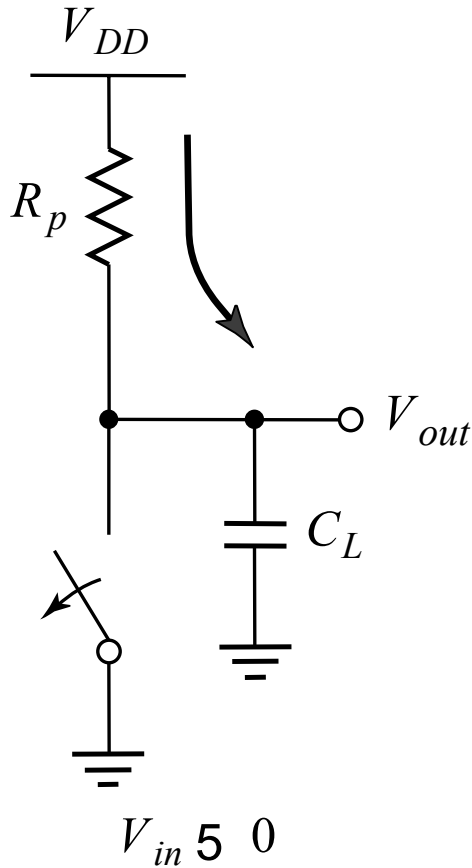
# CMOS Inverter

## First-Order DC Analysis

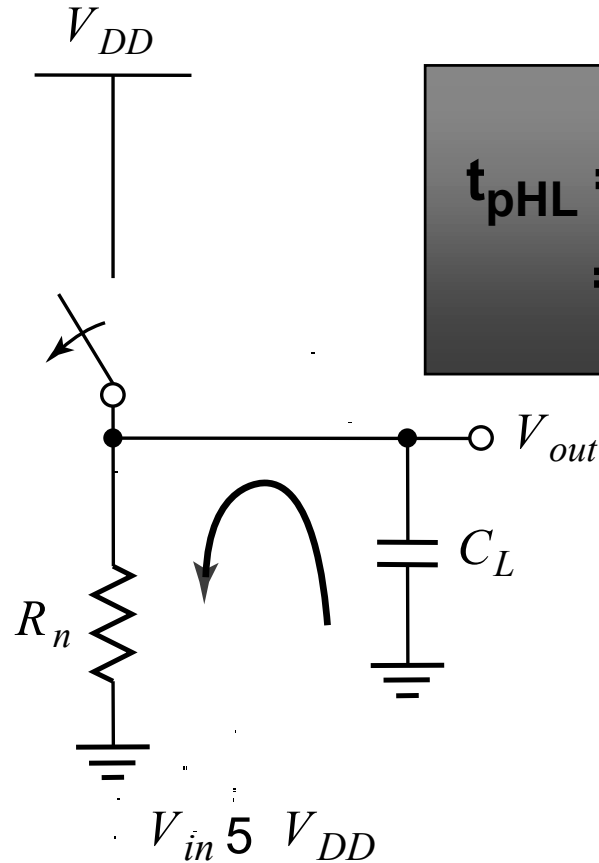


$$\begin{aligned} V_{OL} &= 0 \\ V_{OH} &= V_{DD} \\ V_M &= f(R_n, R_p) \end{aligned}$$

# CMOS Inverter: Transient Response



(a) Low-to-high



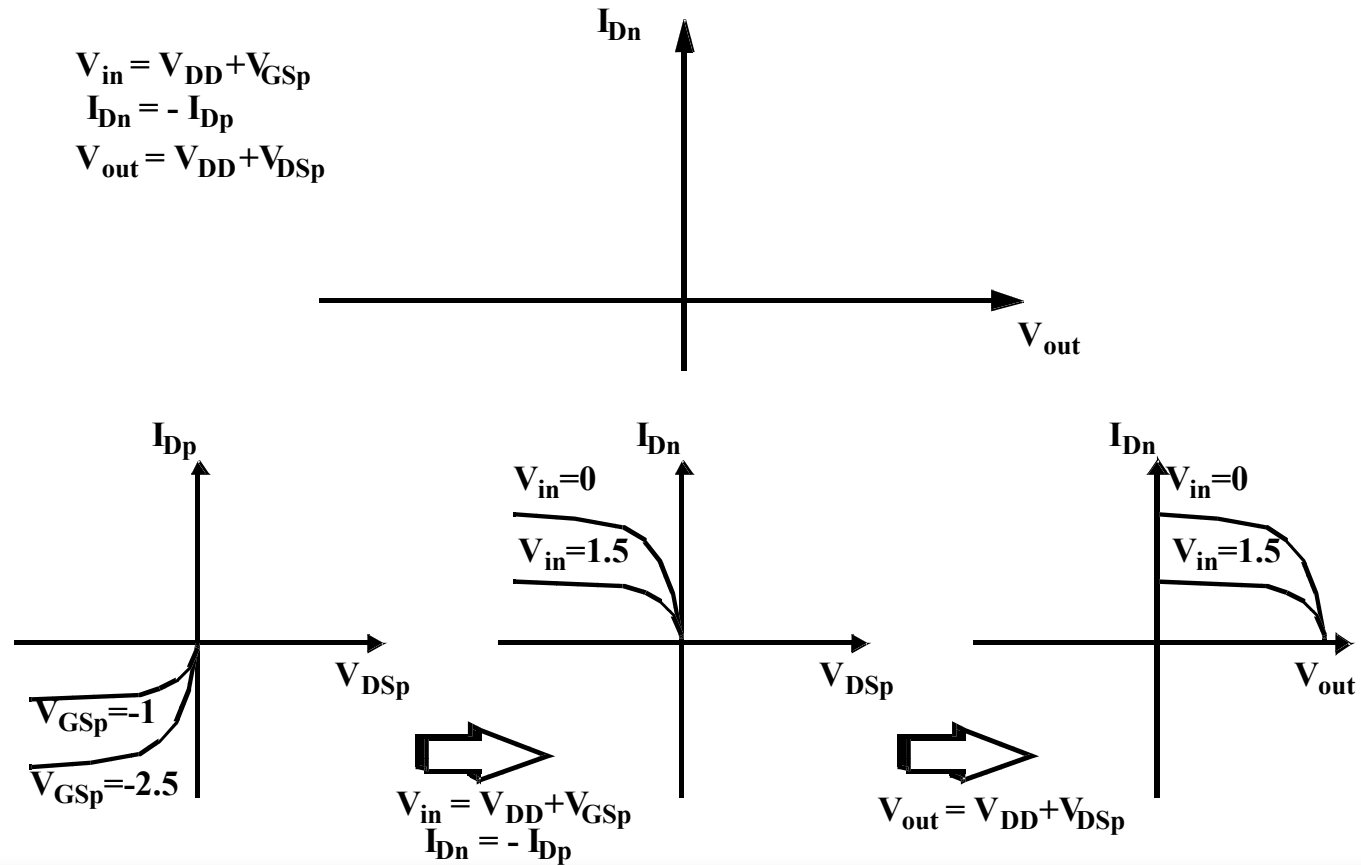
(b) High-to-low

$$t_{pHL} = f(R_{on} \cdot C_L) \\ = 0.69 R_{on} C_L$$



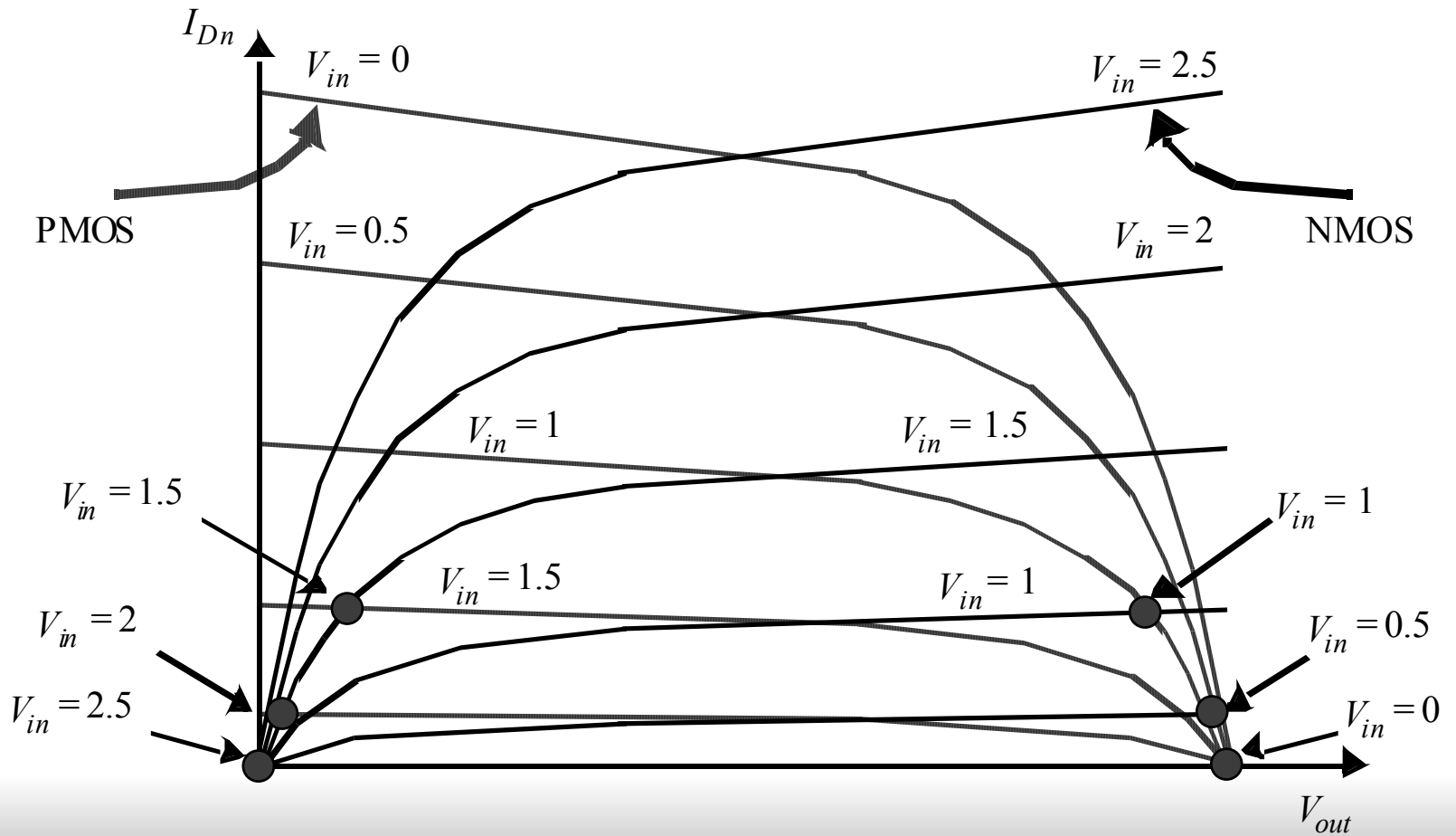
# ***Voltage Transfer Characteristic***

# PMOS Load Lines

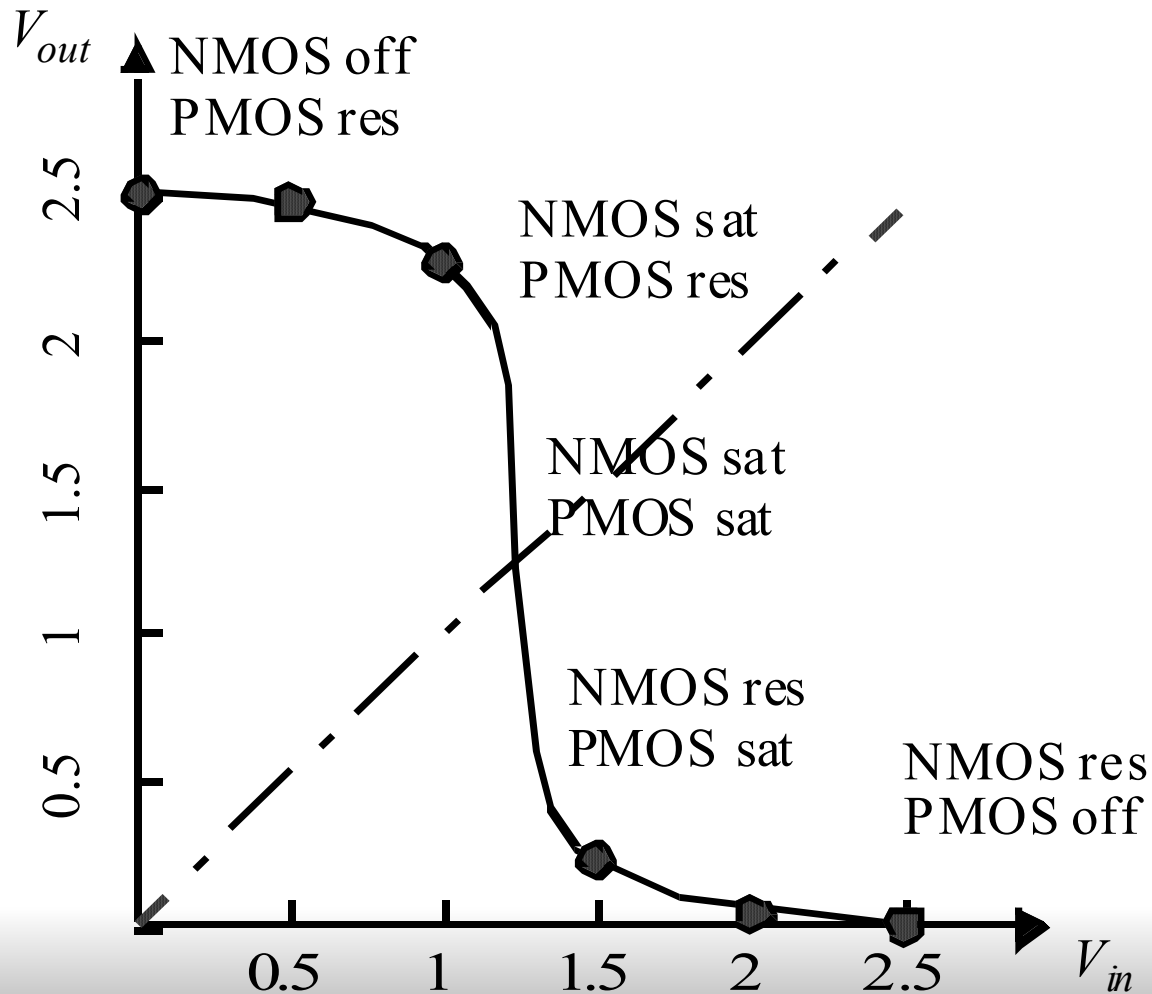




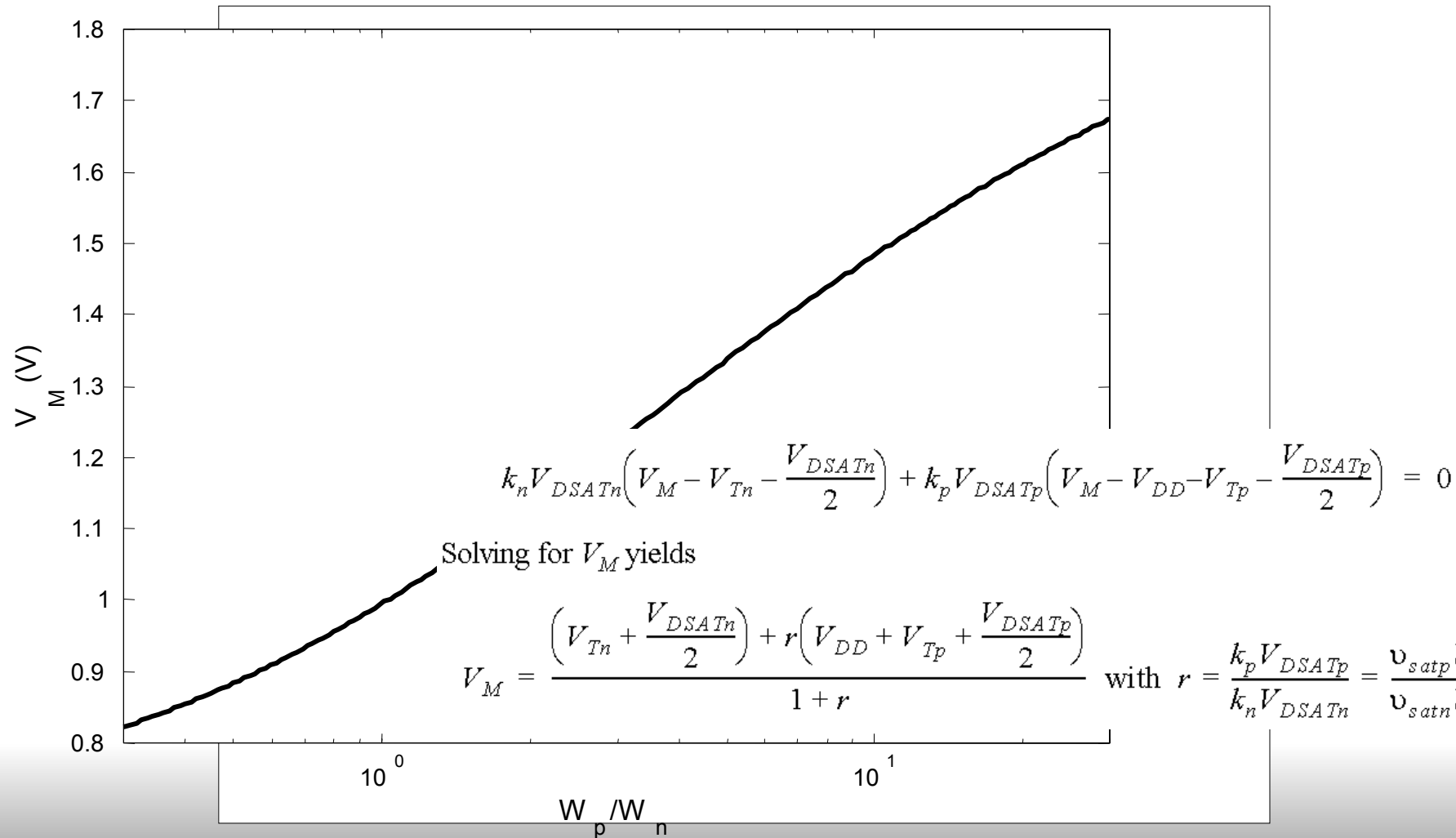
# CMOS Inverter Load Characteristics



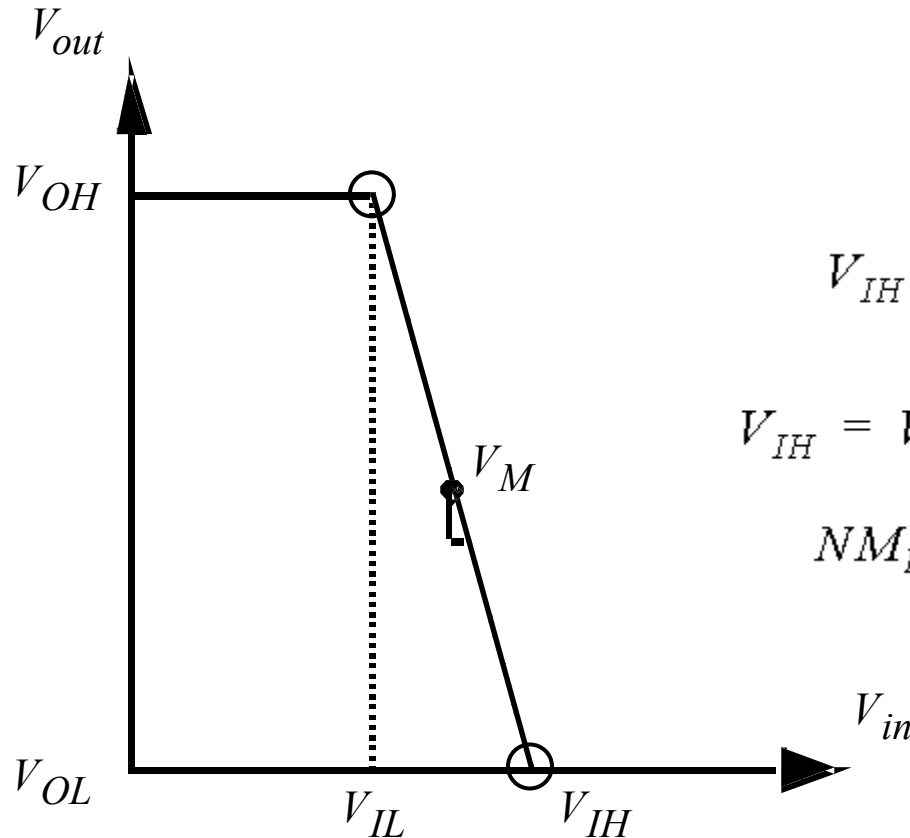
# CMOS Inverter VTC



# Switching Threshold as a function of Transistor Ratio



# Determining $V_{IH}$ and $V_{IL}$



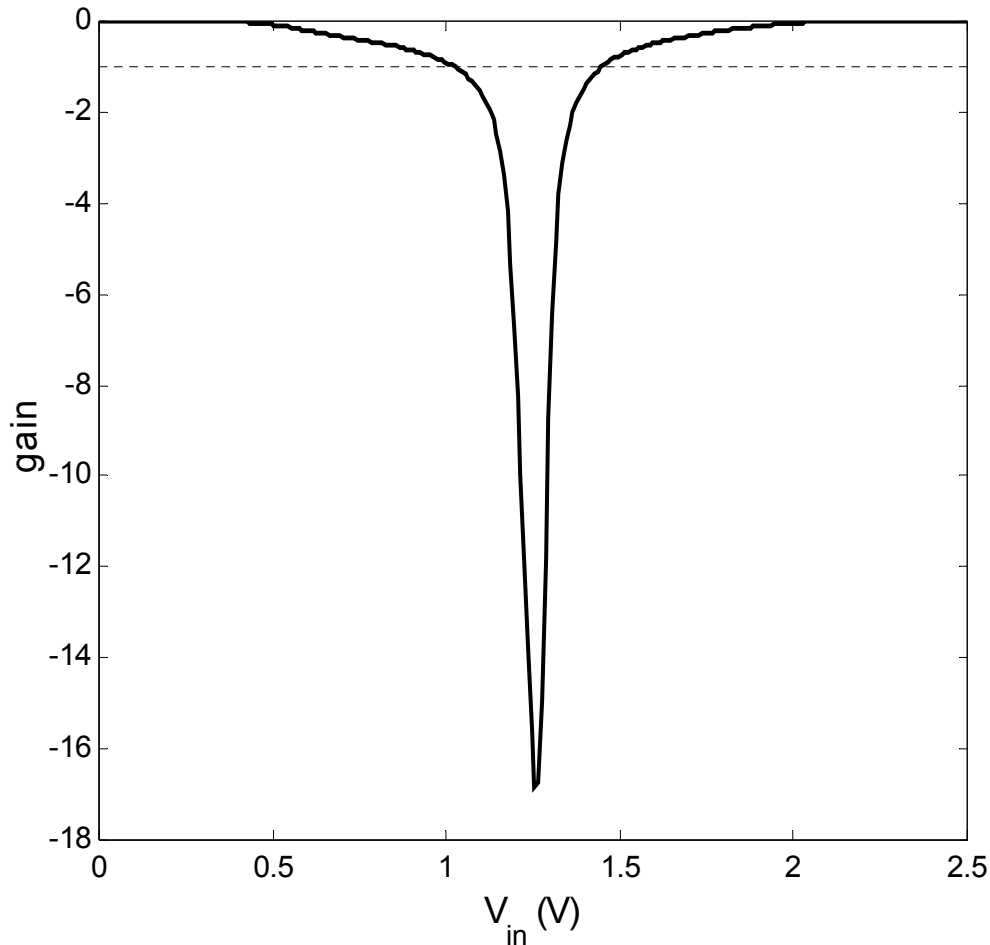
$$V_{IH} - V_{IL} = -\frac{(V_{OH} - V_{OL})}{g} = \frac{-V_{DD}}{g}$$

$$V_{IH} = V_M - \frac{V_M}{g} \quad V_{IL} = V_M + \frac{V_{DD} - V_M}{g}$$

$$NM_H = V_{DD} - V_{IH} \quad NM_L = V_{IL}$$

A simplified approach

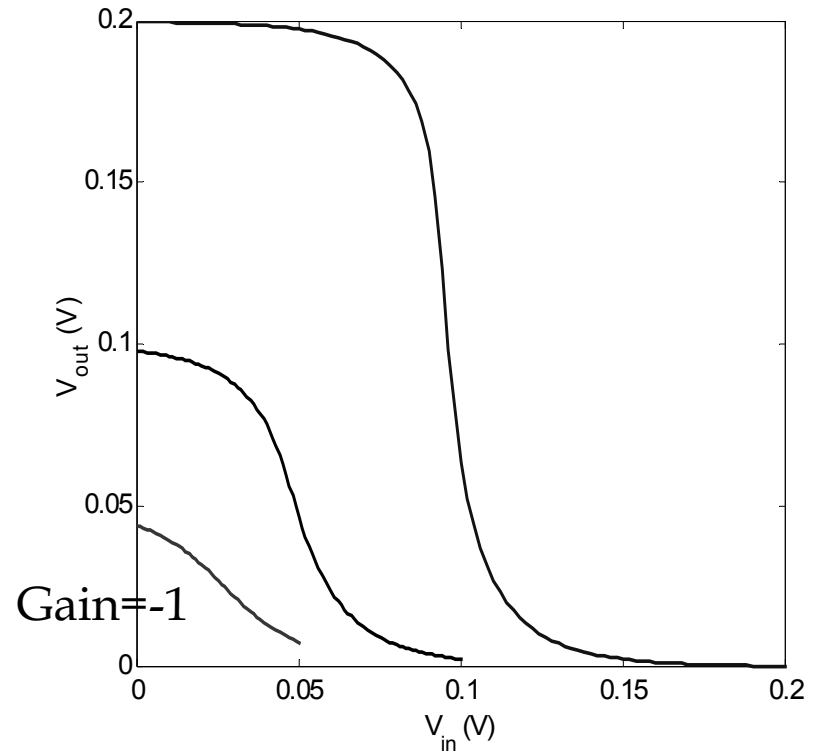
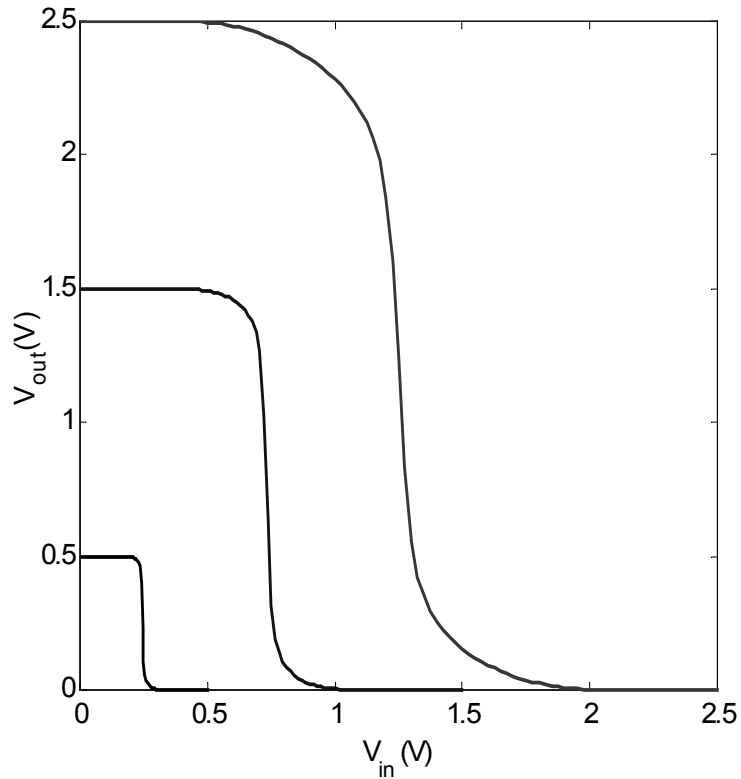
# Inverter Gain



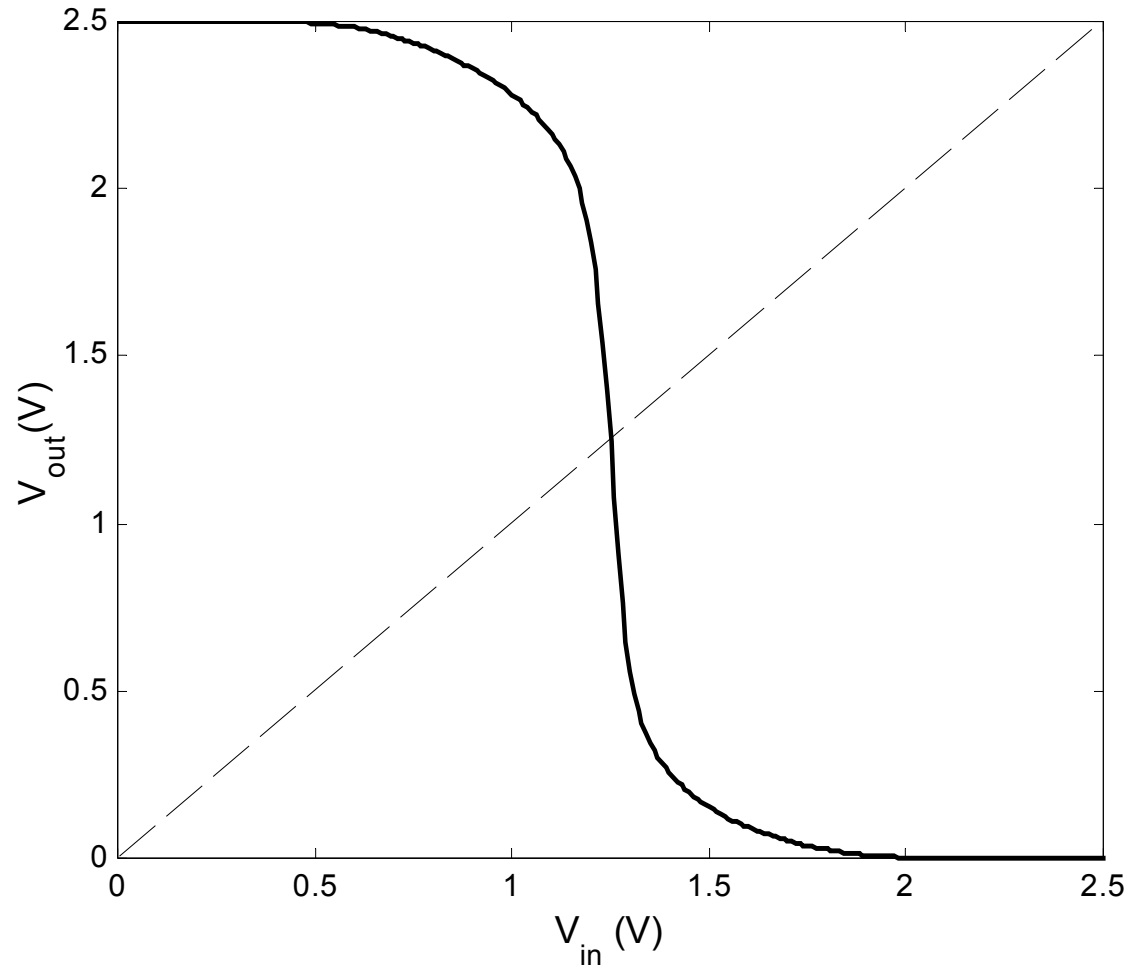
$$g = -\frac{1}{I_D(V_M)} \frac{k_n V_{DSATn} + k_p V_{DSATp}}{\lambda_n - \lambda_p}$$

$$\square \approx \frac{1 + r}{(V_M - V_{Tn} - V_{DSATn}/2)(\lambda_n - \lambda_p)}$$

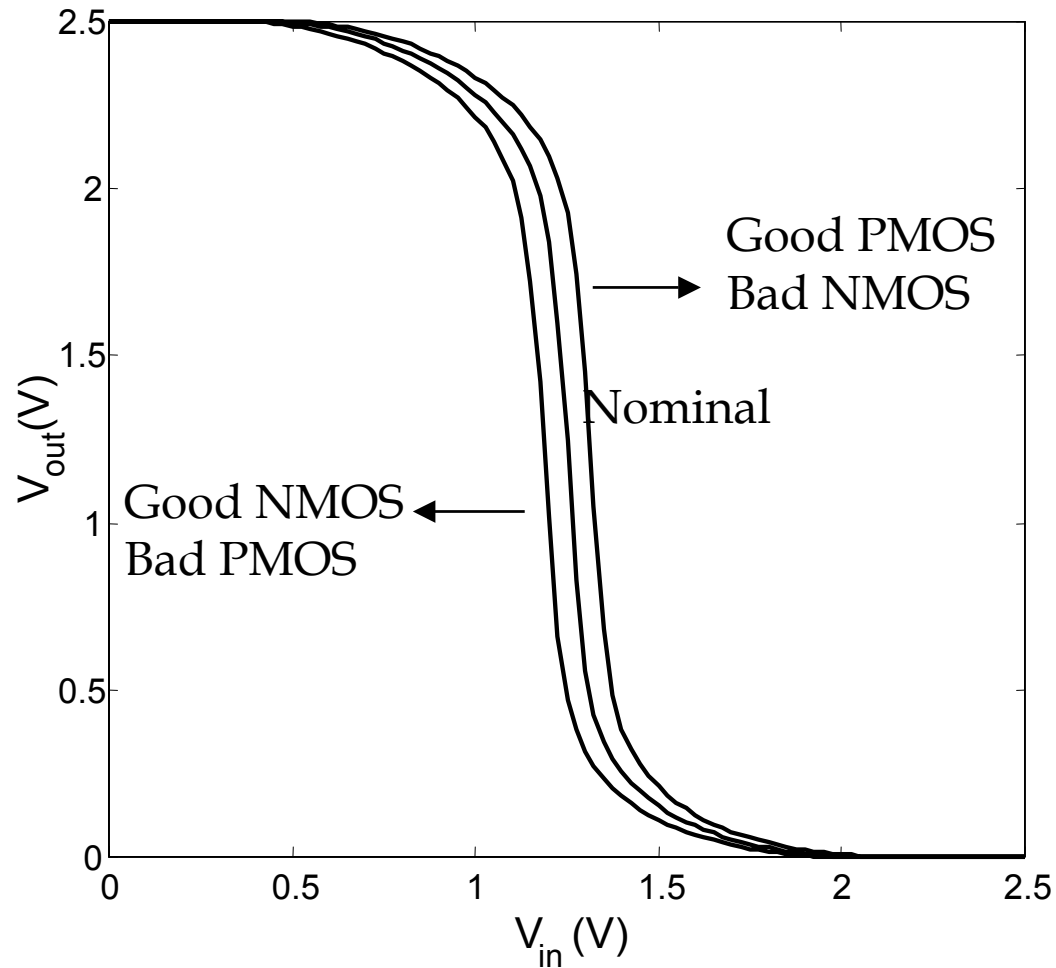
# ***Gain as a function of $V_{DD}$***



# ***Simulated VTC***



# ***Impact of Process Variations***



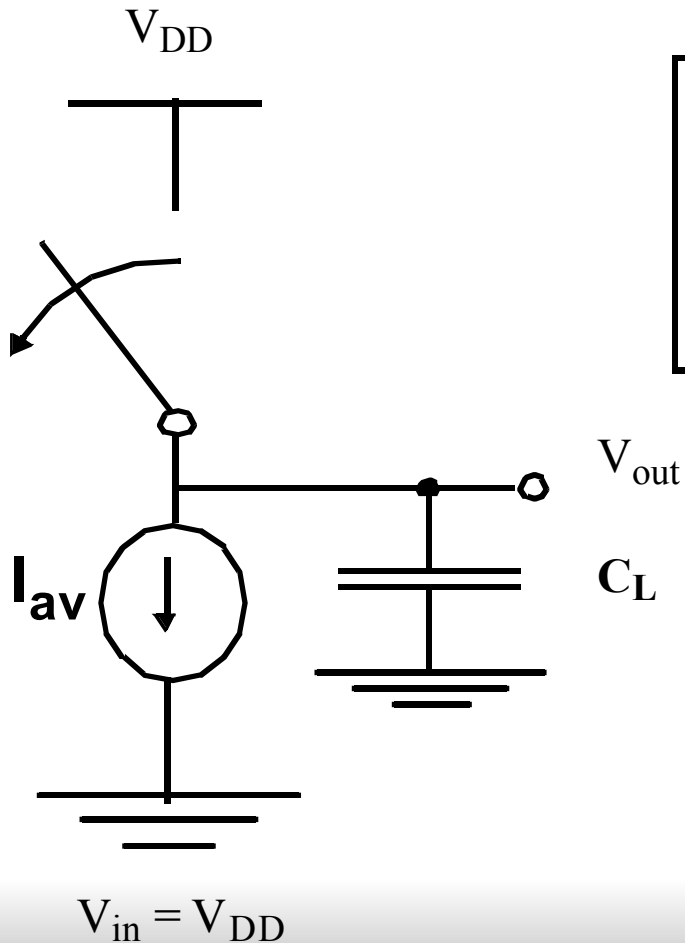




# ***Propagation Delay***

# CMOS Inverter Propagation Delay

## Approach 1

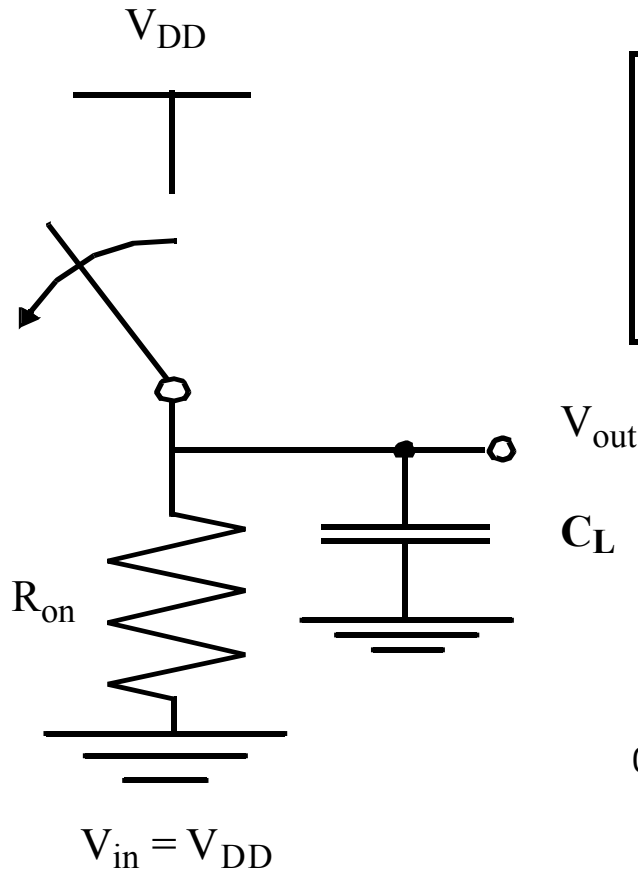


$$t_{pHL} = \frac{C_L V_{swing}/2}{I_{av}}$$

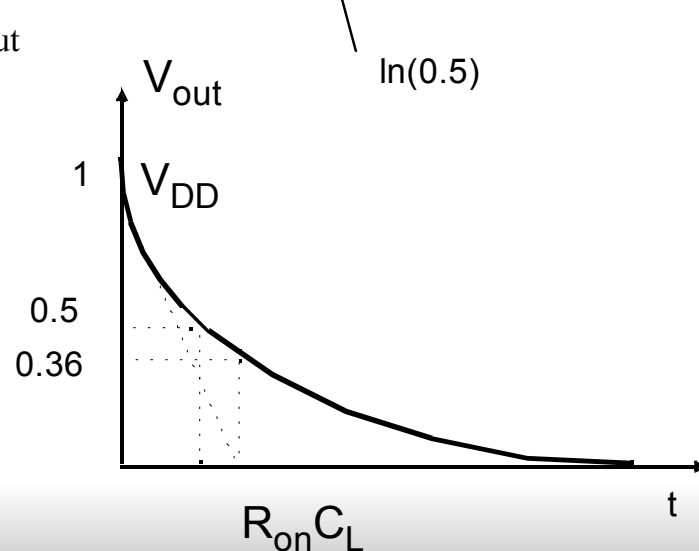
$$\sim \frac{C_L}{k_n V_{DD}}$$

# CMOS Inverter Propagation Delay

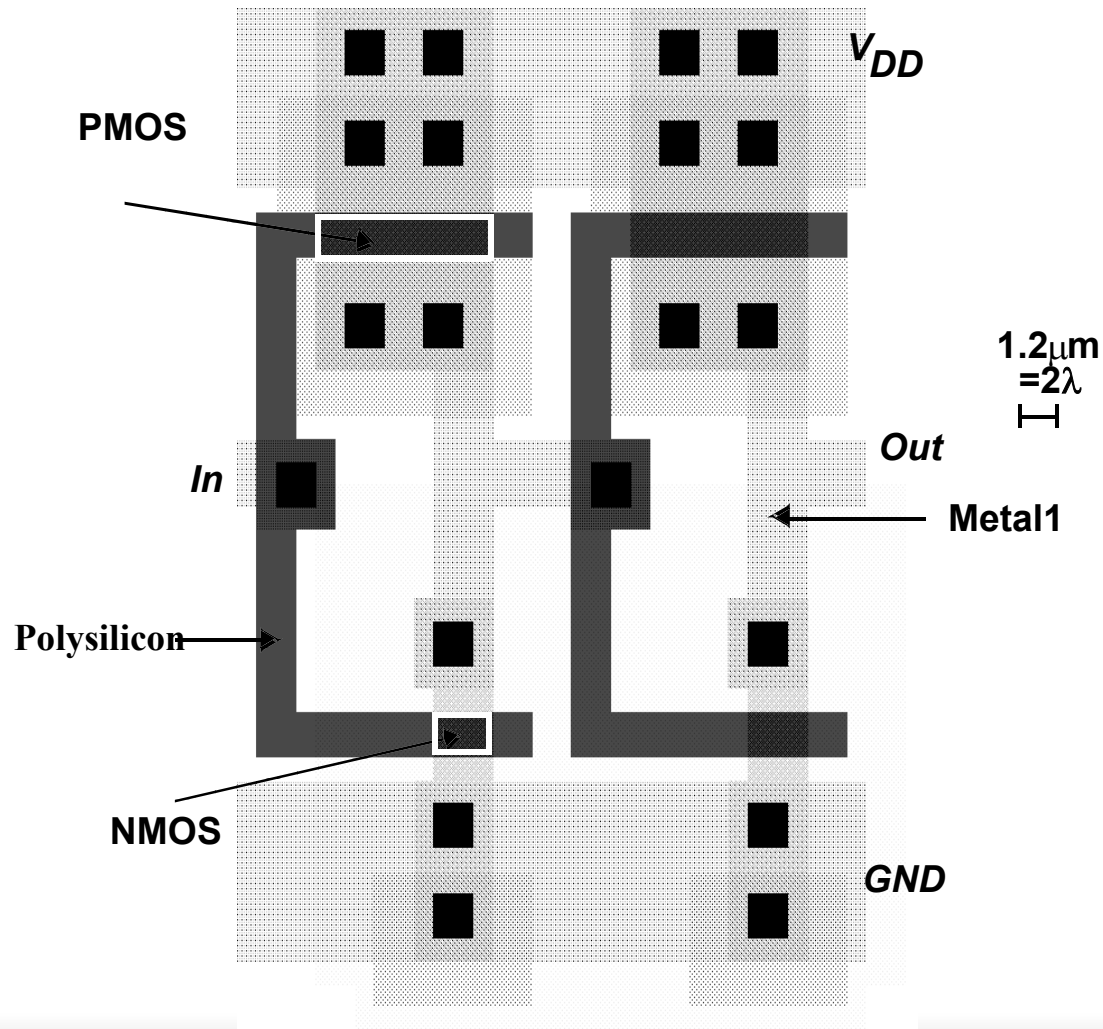
## Approach 2



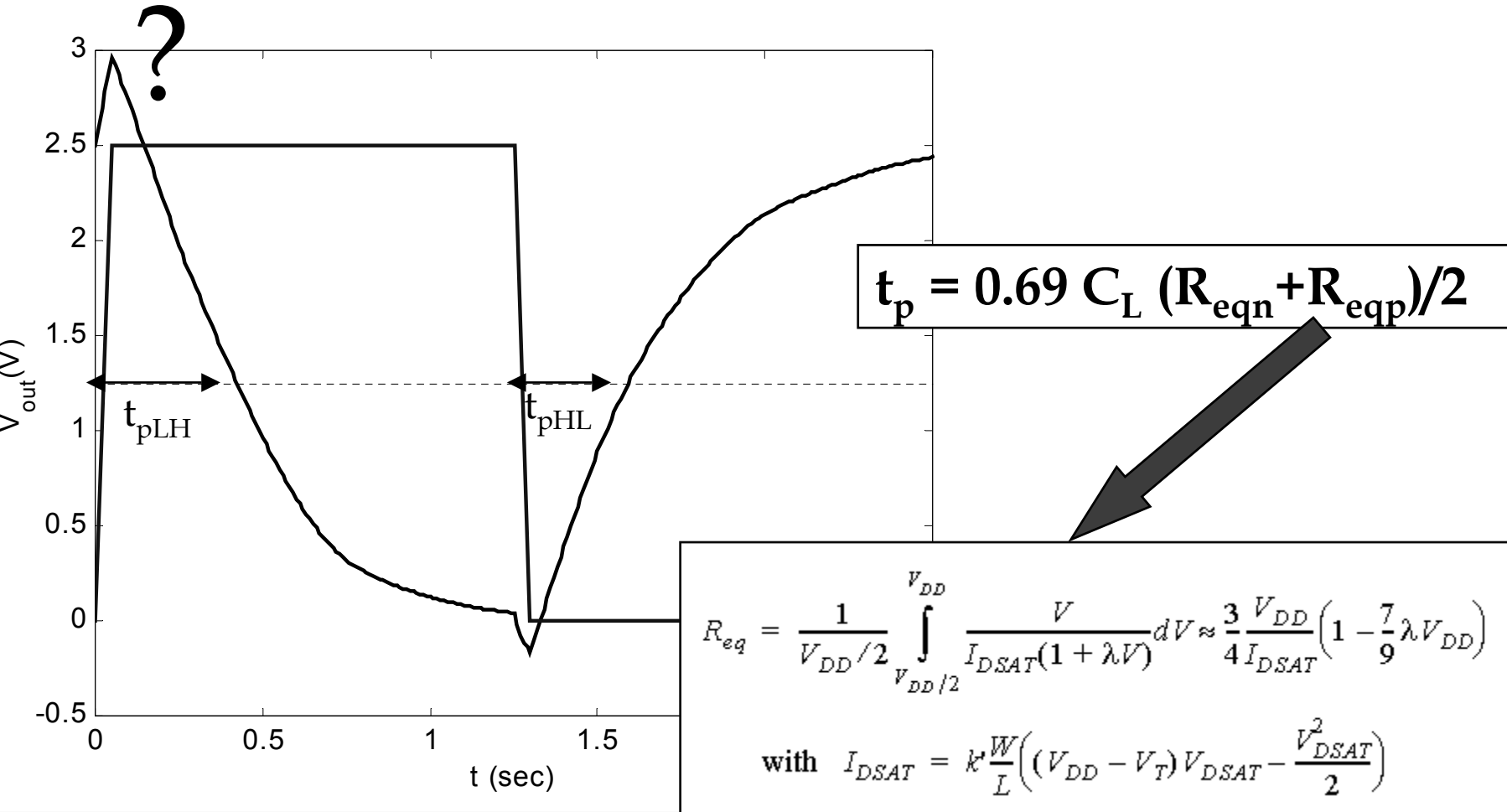
$$t_{pHL} = f(R_{on} \cdot C_L)$$
$$= 0.69 R_{on} C_L$$



# CMOS Inverters



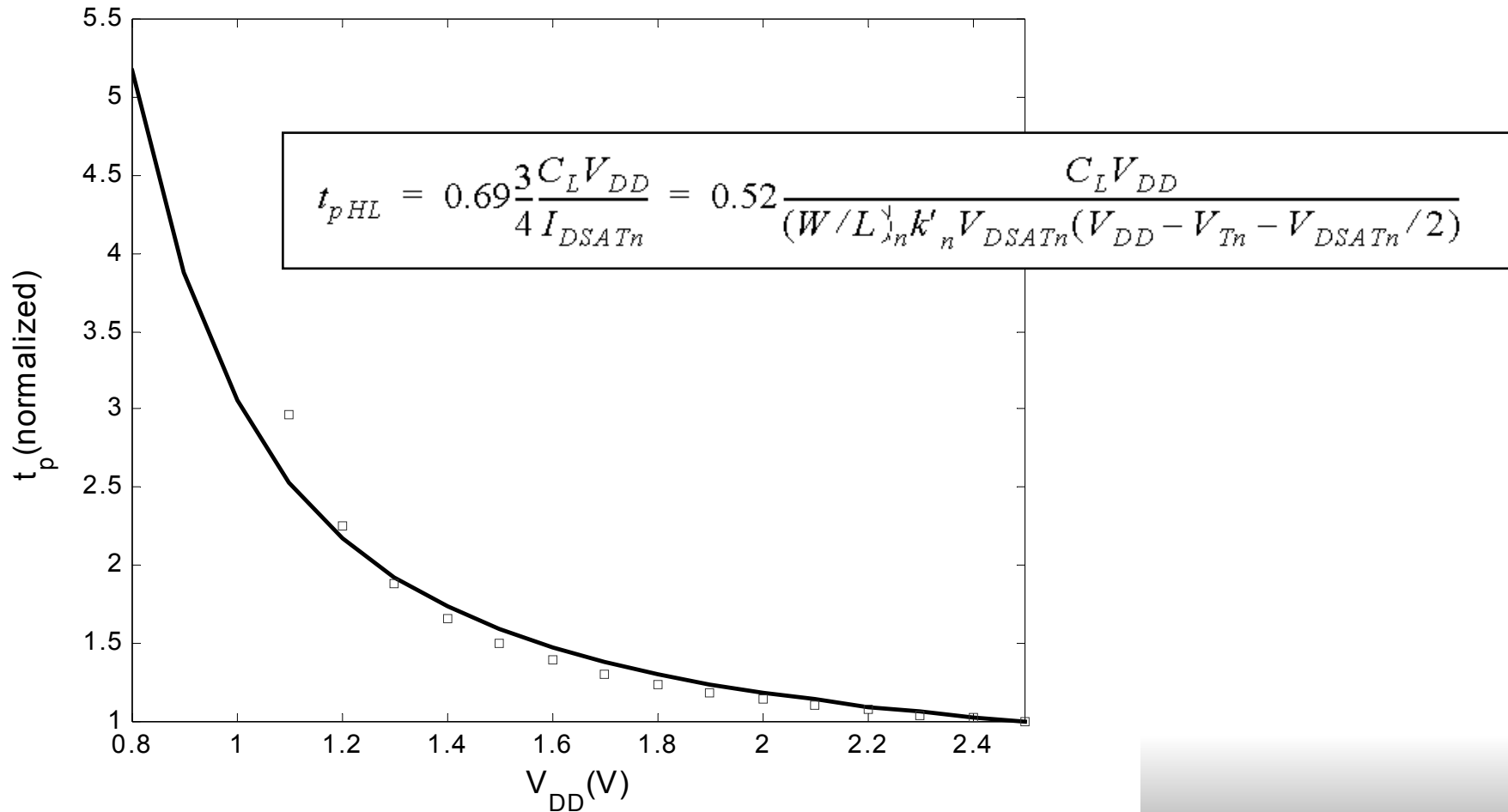
# Transient Response



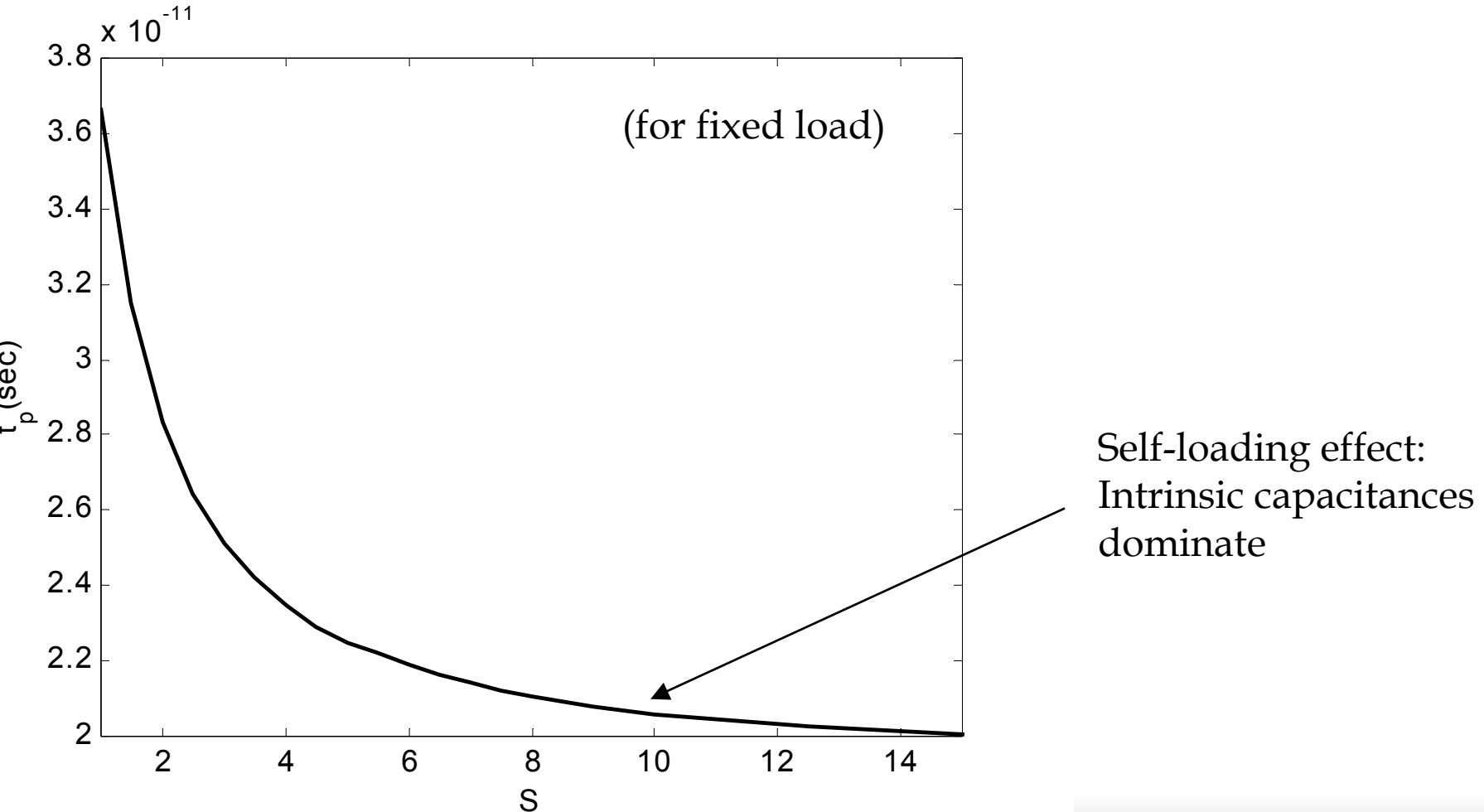
# ***Design for Performance***

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

# Delay as a function of $V_{DD}$

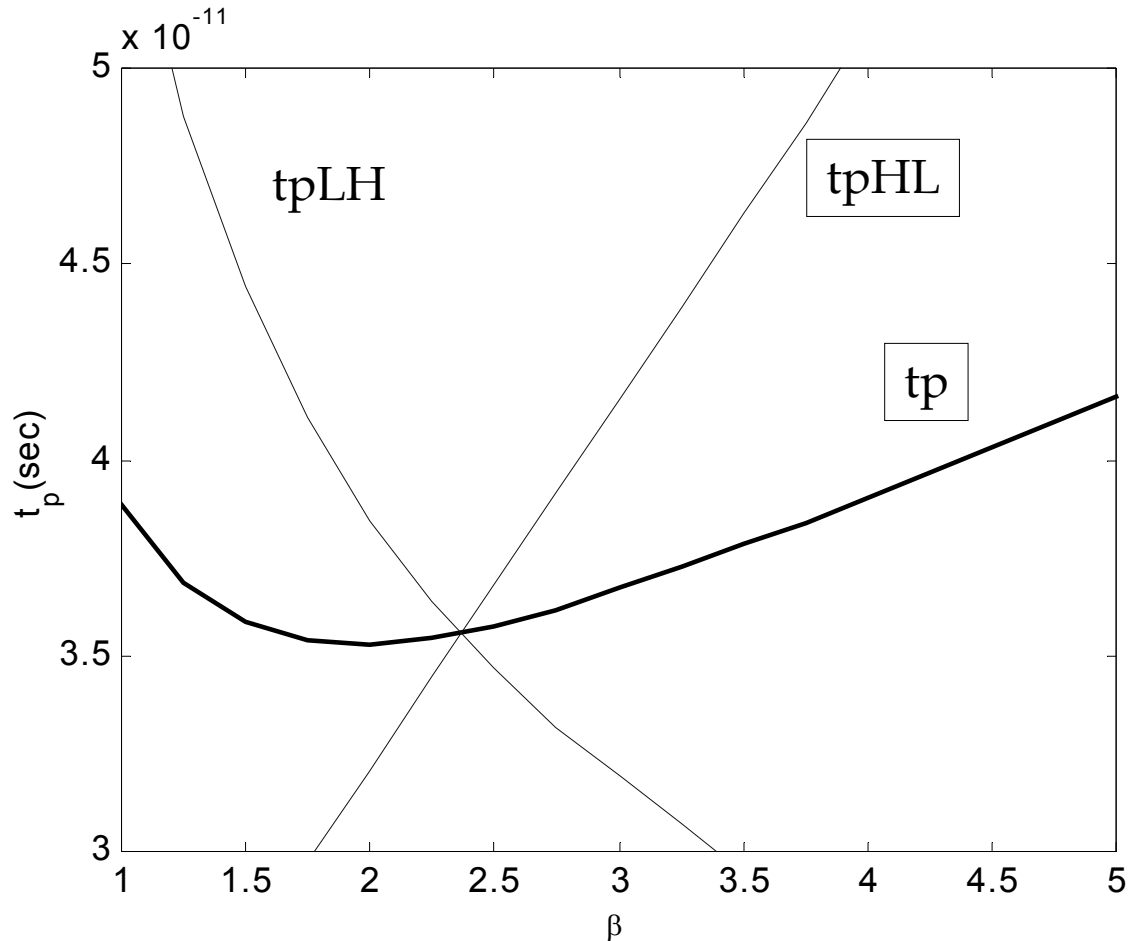


# Device Sizing



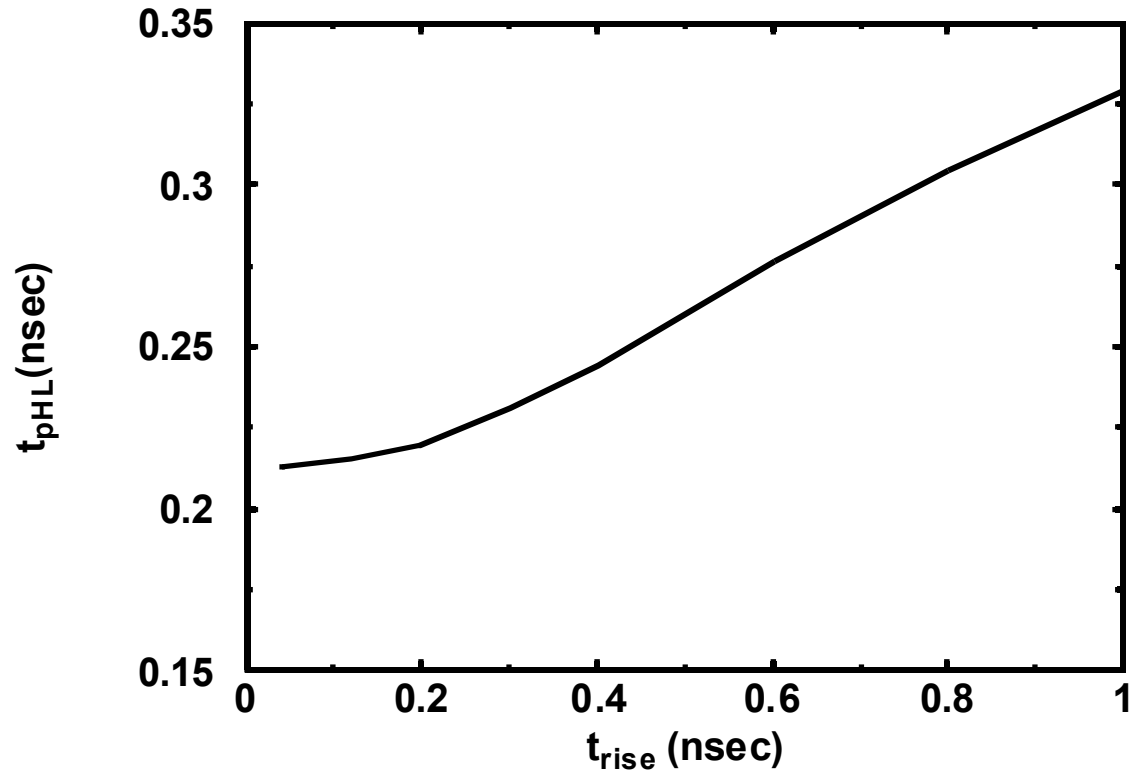


# ***NMOS/PMOS ratio***



$$\beta = W_p/W_n$$

# ***Impact of Rise Time on Delay***

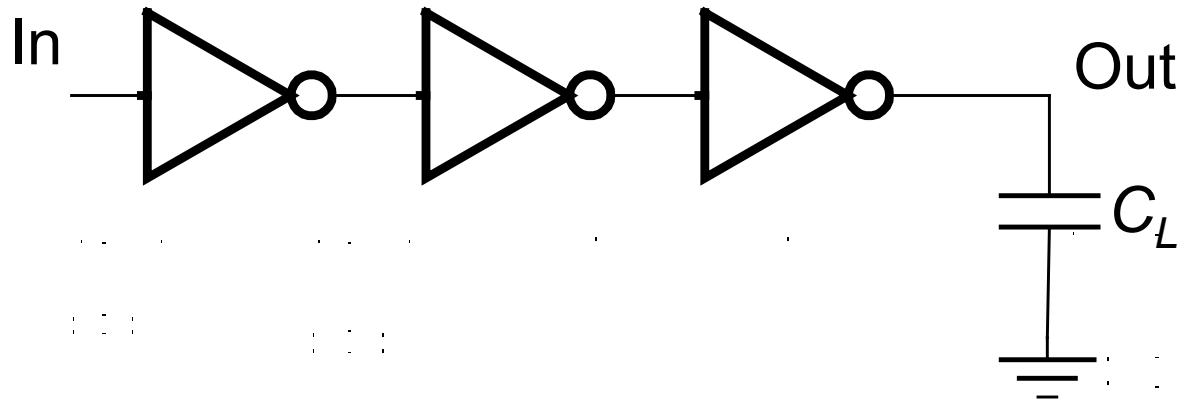


$$t_{pHL} = \sqrt{t_{pHL(step)}^2 + (t_r/2)^2}$$



# ***Inverter Sizing***

# *Inverter Chain*



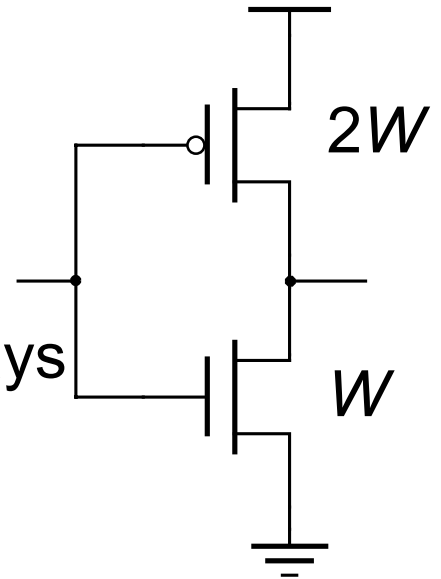
If  $C_L$  is given:

- How many stages are needed to minimize the delay?
- How to size the inverters?

May need some additional constraints.

# Inverter Delay

- Minimum length devices,  $L=0.25\mu\text{m}$
- Assume that for  $W_P = 2W_N = 2W$ 
  - same pull-up and pull-down currents
  - approx. equal resistances  $R_N = R_P$
  - approx. equal rise  $t_{pLH}$  and fall  $t_{pHL}$  delays
- Analyze as an RC network

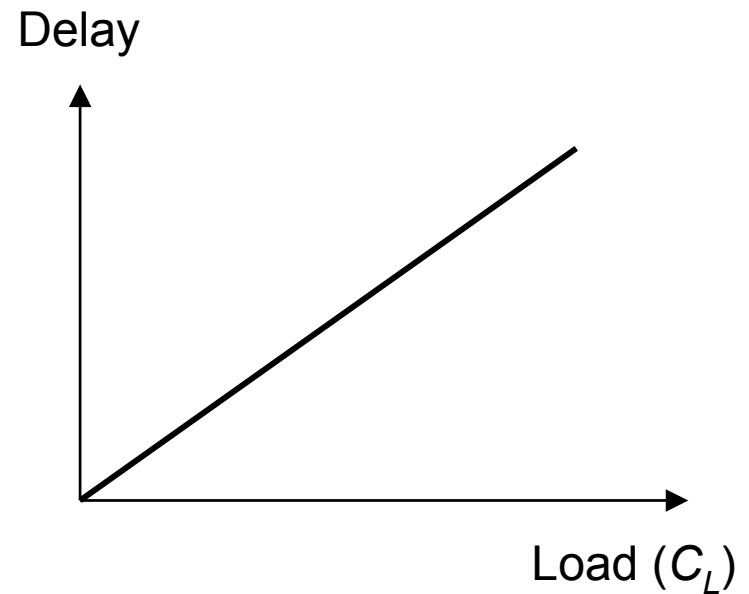
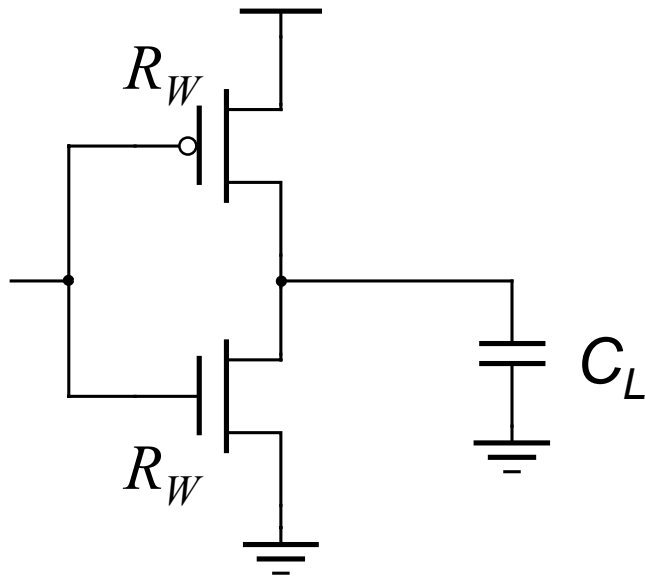


$$R_P = R_{unit} \left( \frac{W_P}{W_{unit}} \right)^{-1} \approx R_{unit} \left( \frac{W_N}{W_{unit}} \right)^{-1} = R_N = R_W$$

$$\text{Delay (D): } t_{pHL} = (\ln 2) R_N C_L \qquad t_{pLH} = (\ln 2) R_P C_L$$

$$\text{Load for the next stage: } C_{gin} = 3 \frac{W}{W_{unit}} C_{unit}$$

# *Inverter with Load*



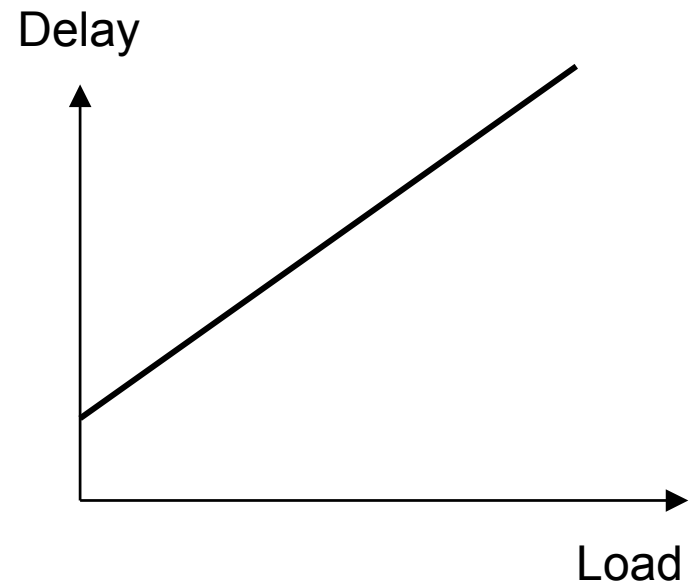
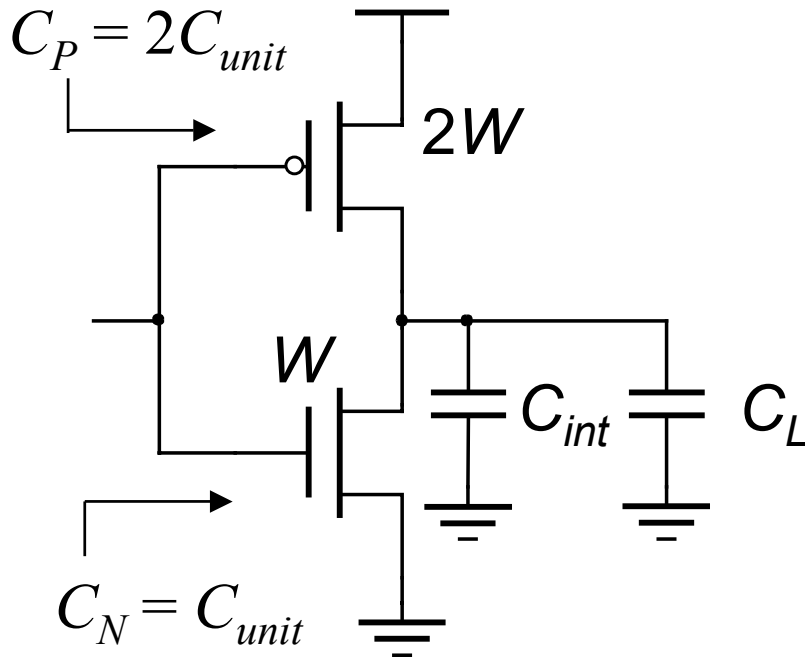
$$t_p = k R_W C_L$$

$k$  is a constant, equal to 0.69

Assumptions: no load  $\rightarrow$  zero delay

$$W_{unit} = 1$$

# ***Inverter with Load***



$$\begin{aligned}\text{Delay} &= kR_W(C_{int} + C_L) = kR_W C_{int} + kR_W C_L = kR_W C_{int}(1 + C_L / C_{int}) \\ &= \text{Delay (Internal)} + \text{Delay (Load)}\end{aligned}$$

# Delay Formula

$$\text{Delay} \sim R_W (C_{int} + C_L)$$

$$t_p = kR_W C_{int} (1 + C_L / C_{int}) = t_{p0} (1 + f / \gamma)$$

$$C_{int} = \gamma C_{gin} \text{ with } \gamma \approx 1$$

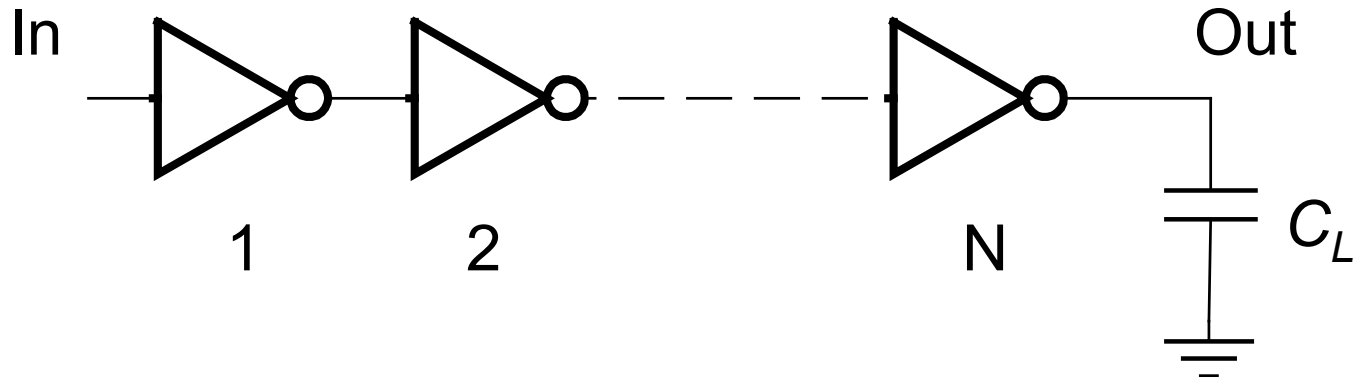
$$f = C_L / C_{gin} - \text{effective fanout}$$

$$R = R_{unit} / W ; C_{int} = WC_{unit}$$

$$t_{p0} = 0.69 R_{unit} C_{unit}$$



# Apply to Inverter Chain



$$t_p = t_{p1} + t_{p2} + \dots + t_{pN}$$

$$t_{pj} \sim R_{unit} C_{unit} \left( 1 + \frac{C_{gin,j+1}}{\gamma C_{gin,j}} \right)$$

$$t_p = \sum_{j=1}^N t_{p,j} = t_{p0} \sum_{i=1}^N \left( 1 + \frac{C_{gin,j+1}}{\gamma C_{gin,j}} \right), \quad C_{gin,N+1} = C_L$$

# *Optimal Tapering for Given N*

Delay equation has  $N - 1$  unknowns,  $C_{gin,2} - C_{gin,N}$

Minimize the delay, find  $N - 1$  partial derivatives

Result:  $C_{gin,j+1}/C_{gin,j} = C_{gin,j}/C_{gin,j-1}$

Size of each stage is the geometric mean of two neighbors

$$C_{gin,j} = \sqrt{C_{gin,j-1} C_{gin,j+1}}$$

- each stage has the same effective fanout ( $C_{out}/C_{in}$ )
- each stage has the same delay

# Optimum Delay and Number of Stages

When each stage is sized by  $f$  and has same eff. fanout  $f$ :

$$f^N = F = C_L / C_{gin,1}$$

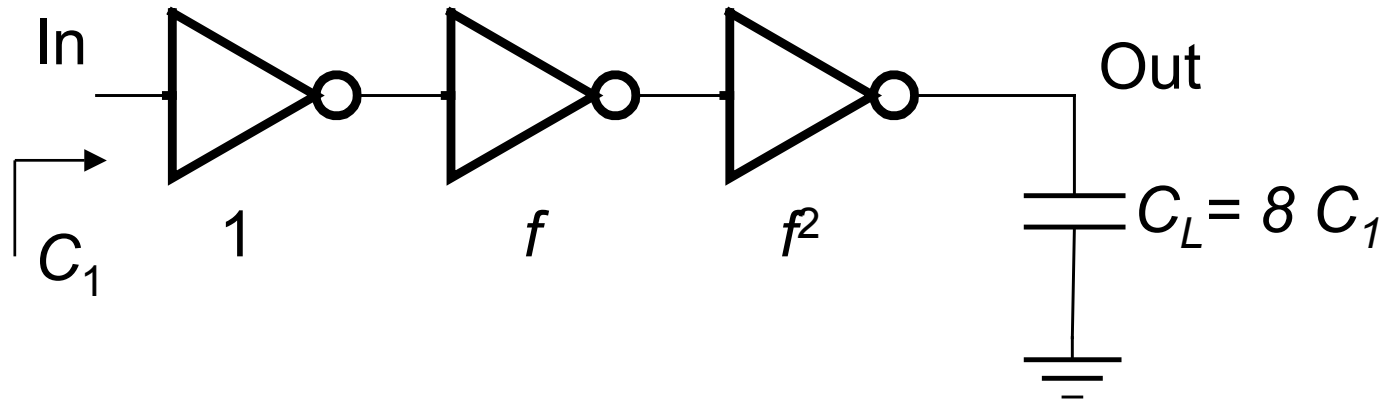
Effective fanout of each stage:

$$f = \sqrt[N]{F}$$

Minimum path delay

$$t_p = Nt_{p0} \left( 1 + \sqrt[N]{F} / \gamma \right)$$

# Example



$C_L/C_1$  has to be evenly distributed across  $N = 3$  stages:

$$f = \sqrt[3]{8} = 2$$

# Optimum Number of Stages

For a given load,  $C_L$  and given input capacitance  $C_{in}$   
Find optimal sizing  $f$

$$C_L = F \cdot C_{in} = f^N C_{in} \quad \text{with} \quad N = \frac{\ln F}{\ln f}$$

$$t_p = N t_{p0} \left( F^{1/N} / \gamma + 1 \right) = \frac{t_{p0} \ln F}{\gamma} \left( \frac{f}{\ln f} + \frac{\gamma}{\ln f} \right)$$

$$\frac{\partial t_p}{\partial f} = \frac{t_{p0} \ln F}{\gamma} \cdot \frac{\ln f - 1 - \gamma/f}{\ln^2 f} = 0$$

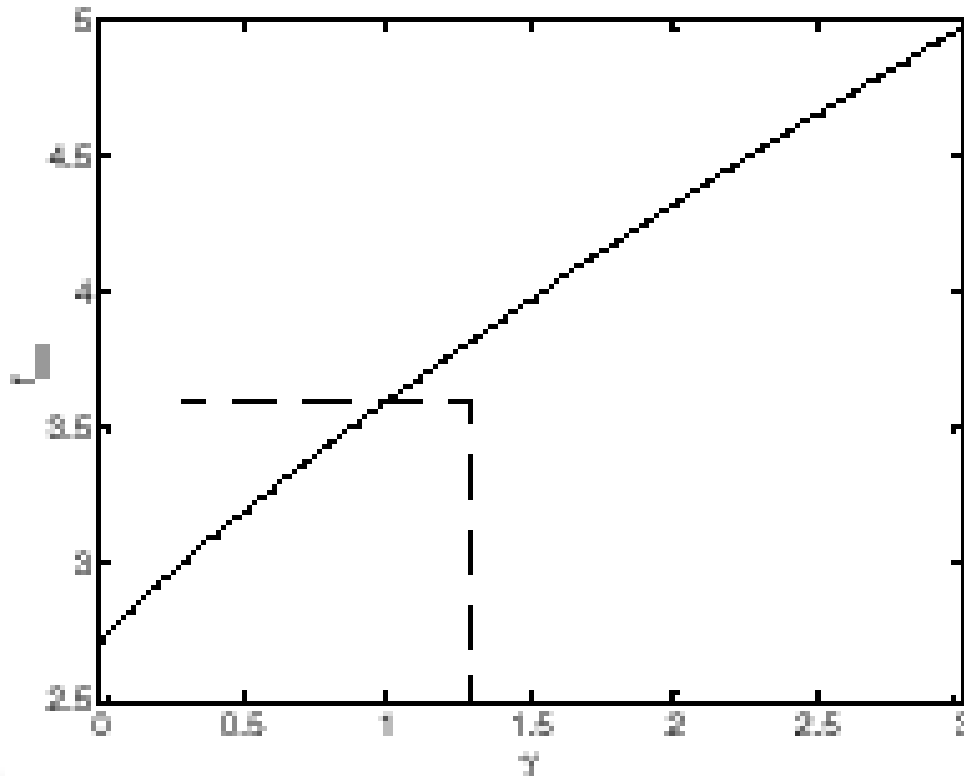
For  $\gamma = 0$ ,  $f = e$ ,  $N = \ln F$

$$f = \exp(1 + \gamma/f)$$

# Optimum Effective Fanout $f$

Optimum  $f$  for given process defined by  $\gamma$

$$f = \exp(1 + \gamma / f)$$

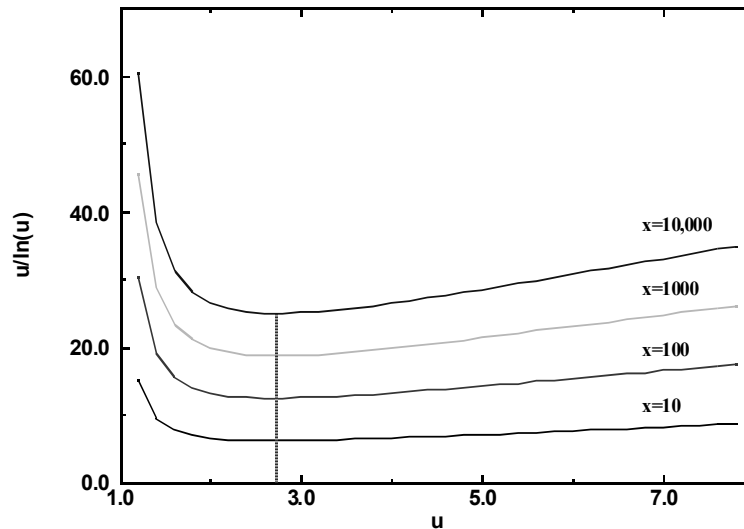


$$f_{opt} = 3.6$$

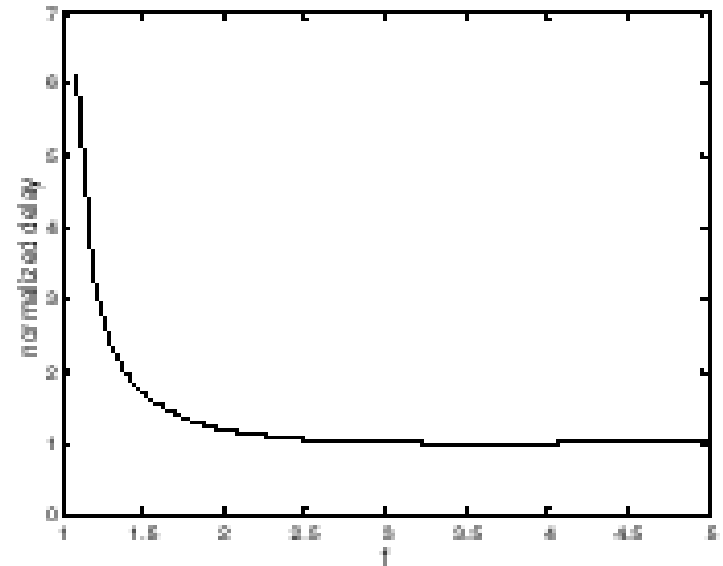
for  $\gamma = 1$

# Impact of Self-Loading on $t_p$

No Self-Loading,  $\gamma=0$



With Self-Loading  $\gamma=1$



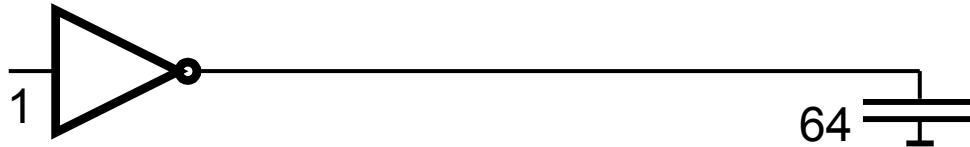
# ***Normalized delay function of F***

$$t_p = Nt_{p0} \left( 1 + \sqrt[N]{F} / \gamma \right)$$

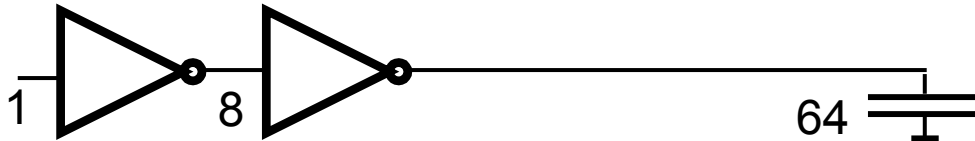
<i>F</i>	Unbuffered	Two Stage	Inverter Chain
10	11	8.3	8.3
100	101	22	16.5
1000	1001	65	24.8
10,000	10,001	202	33.1



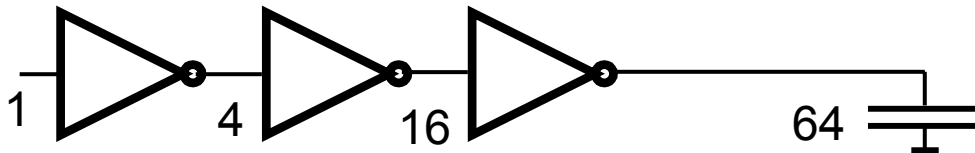
# Buffer Design



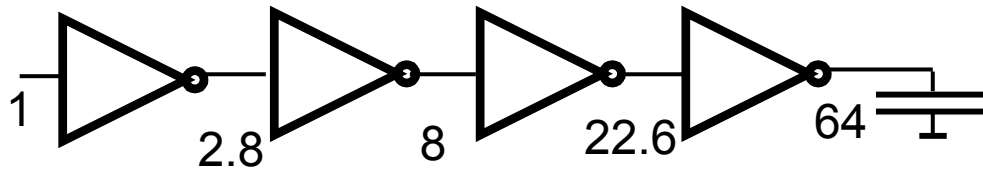
N	f	$t_p$
1	64	65



2	8	18
---	---	----



3	4	15
---	---	----



4	2.8	15.3
---	-----	------



# ***Power Dissipation***

# ***Where Does Power Go in CMOS?***

- **Dynamic Power Consumption**

**Charging and Discharging Capacitors**

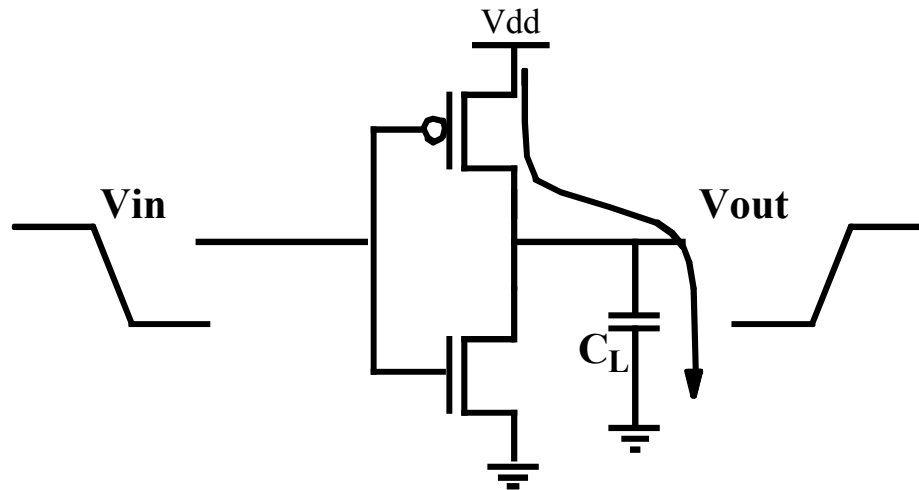
- **Short Circuit Currents**

**Short Circuit Path between Supply Rails during Switching**

- **Leakage**

**Leaking diodes and transistors**

# Dynamic Power Dissipation

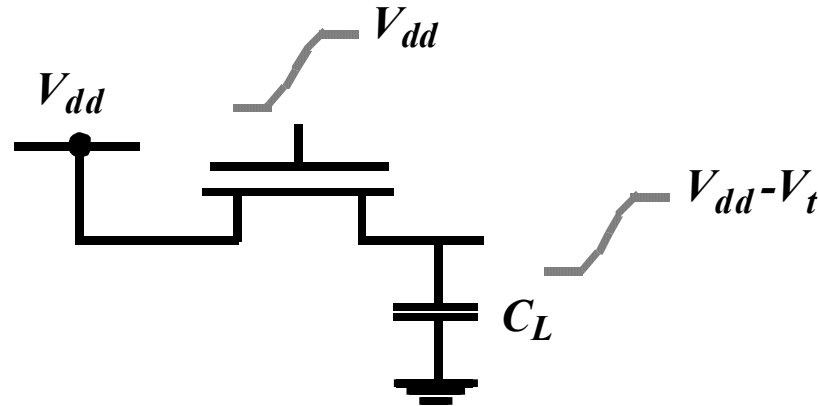


$$\text{Energy/transition} = C_L * V_{dd}^2$$

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

- Not a function of transistor sizes!
- Need to reduce  $C_L$ ,  $V_{dd}$ , and  $f$  to reduce power.

# Modification for Circuits with Reduced Swing

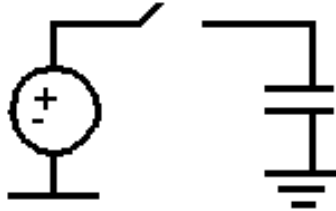


$$E_{0 \rightarrow 1} = C_L \cdot V_{dd} \cdot (V_{dd} - V_t)$$

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

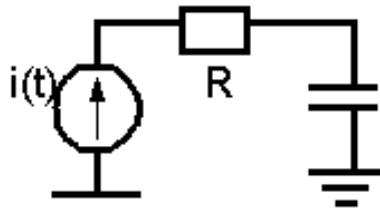
# Adiabatic Charging

Charging a capacitor



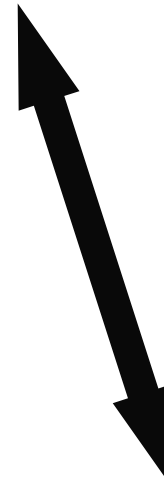
$$CV_{dd}^2/2$$

Consider



$$v_c = \frac{1}{C} \cdot \int_0^T i dt = \frac{1}{C} \cdot I_{av} \cdot T \quad I_{av} = \frac{C \cdot v_c}{T}$$

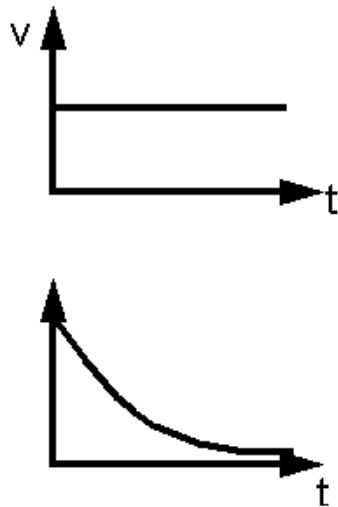
$$E_{dis} = R \cdot \int_0^T i^2(t) dt \geq R \cdot \int_0^T I_{av}^2 dt = R \cdot I_{av}^2 \cdot T = \frac{RC}{T} \cdot C \cdot V_c^2$$



# Adiabatic Charging

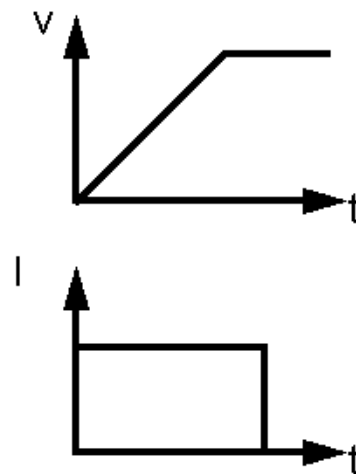
$$V_I = RI + V_c = RC \frac{dv_c}{dt} + V_c$$

$V_I = \text{cst} \rightarrow$  Exponential current



$$E_R = CV_c^2 / 2$$

$I = I_{av} \rightarrow$  Linear ramp on  $V_I$



*wins if  $T > 2RC$*

minimal energy  

$$E_R = RC/T CV_c^2$$

# 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 transition in  $N$  clock cycles

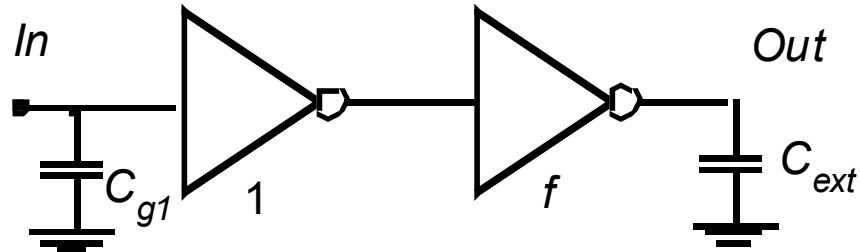
$$P_{avg} = \lim_{N \rightarrow \infty} \frac{E_N}{N} \cdot f_{clk} = \left( \lim_{N \rightarrow \infty} \frac{n(N)}{N} \right) \cdot C_L \cdot V_{dd}^2 \cdot f_{clk}$$

$$\alpha_{0 \rightarrow 1} = \lim_{N \rightarrow \infty} \frac{n(N)}{N}$$

$$P_{avg} = \alpha_{0 \rightarrow 1} \cdot C_L \cdot V_{dd}^2 \cdot f_{clk}$$



# Transistor Sizing for Minimum Energy



- Goal: Minimize Energy of whole circuit
  - Design parameters:  $f$  and  $V_{DD}$
  - $t_p \leq t_{pref}$  of circuit with  $f=1$  and  $V_{DD} = V_{ref}$

$$t_p = t_{p0} \left( \left( 1 + \frac{f}{\gamma} \right) + \left( 1 + \frac{F}{f\gamma} \right) \right)$$

$$t_{p0} \propto \frac{V_{DD}}{V_{DD} - V_{TE}}$$

# Transistor Sizing (2)

## □ Performance Constraint ( $\gamma=1$ )

$$\frac{t_p}{t_{pref}} = \frac{t_{p0}}{t_{p0ref}} \frac{\left(2 + f + \frac{F}{f}\right)}{(3 + F)} = \frac{V_{DD}}{V_{ref}} \frac{V_{ref} - V_{TE}}{V_{DD} - V_{TE}} \frac{\left(2 + f + \frac{F}{f}\right)}{(3 + F)} = 1$$

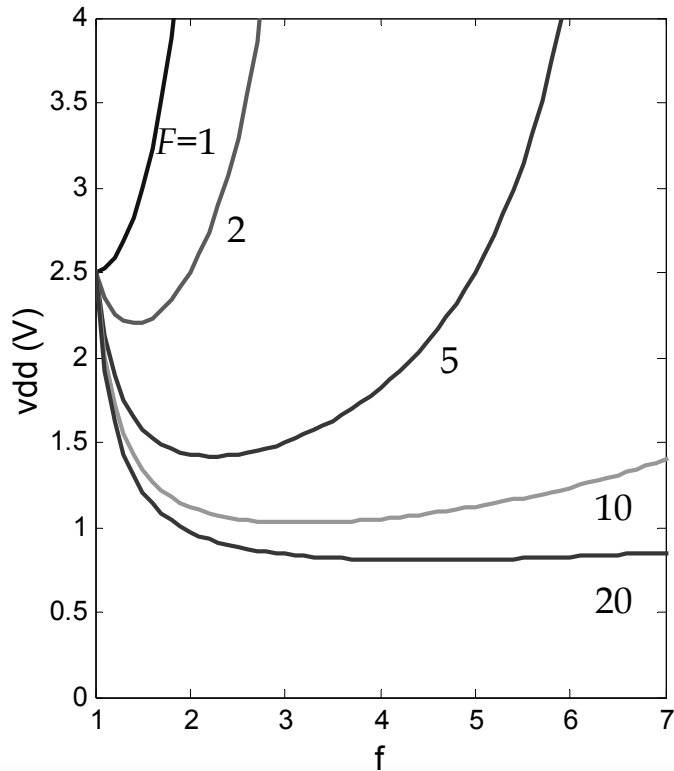
## □ Energy for single Transition

$$E = V_{DD}^2 C_{g1} [(1 + \gamma)(1 + f) + F]$$

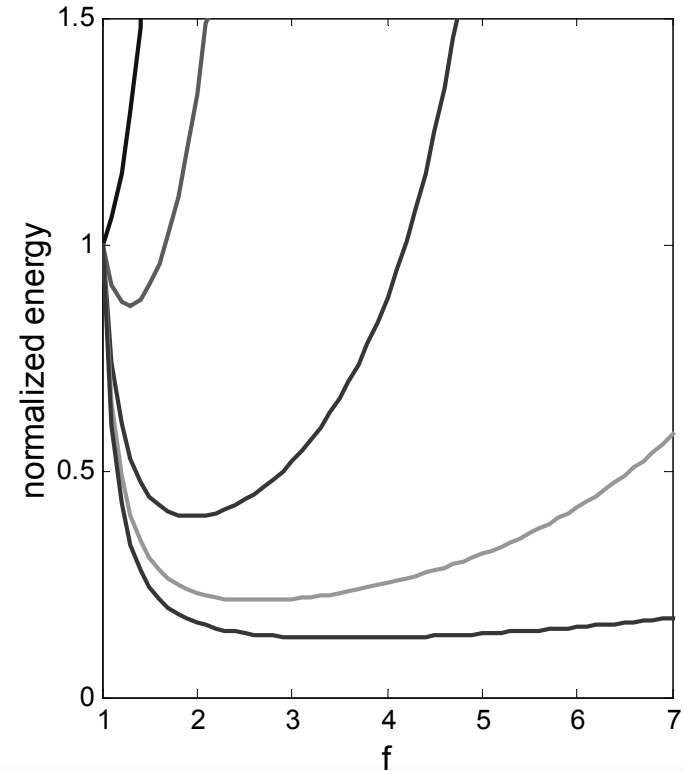
$$\frac{E}{E_{ref}} = \left(\frac{V_{DD}}{V_{ref}}\right)^2 \left(\frac{2 + 2f + F}{4 + F}\right)$$

# Transistor Sizing (3)

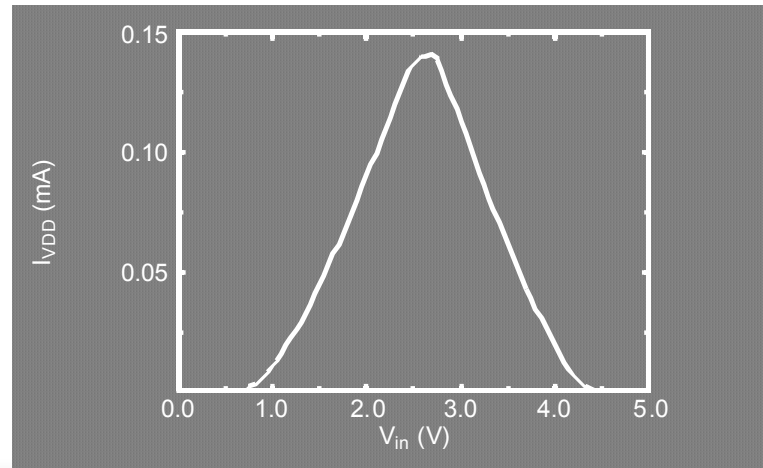
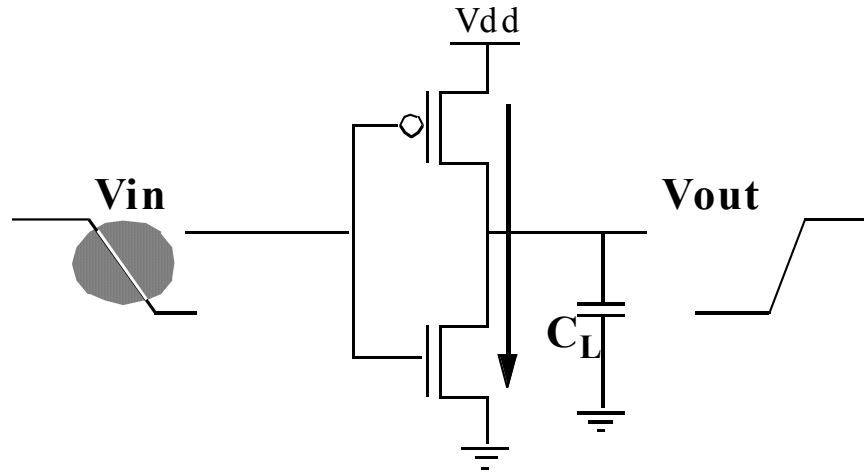
$$V_{DD} = f(f)$$



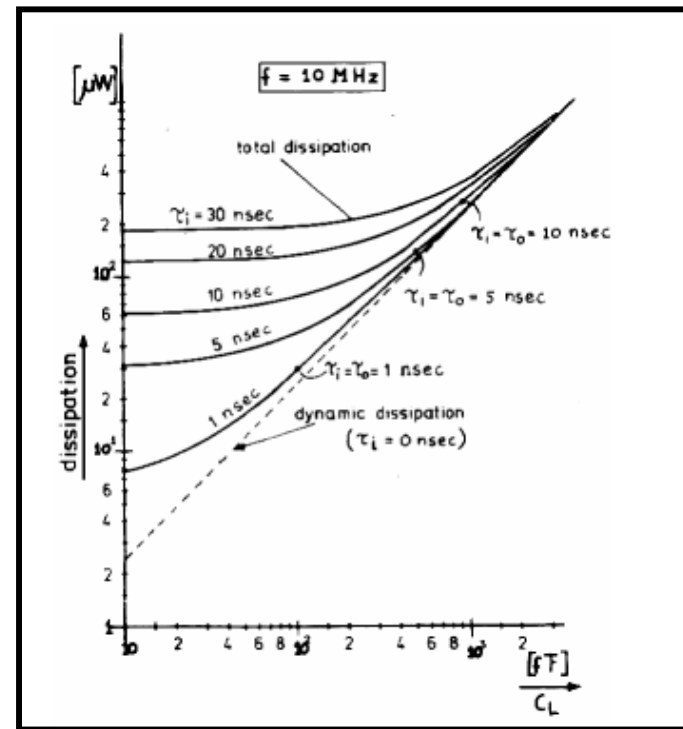
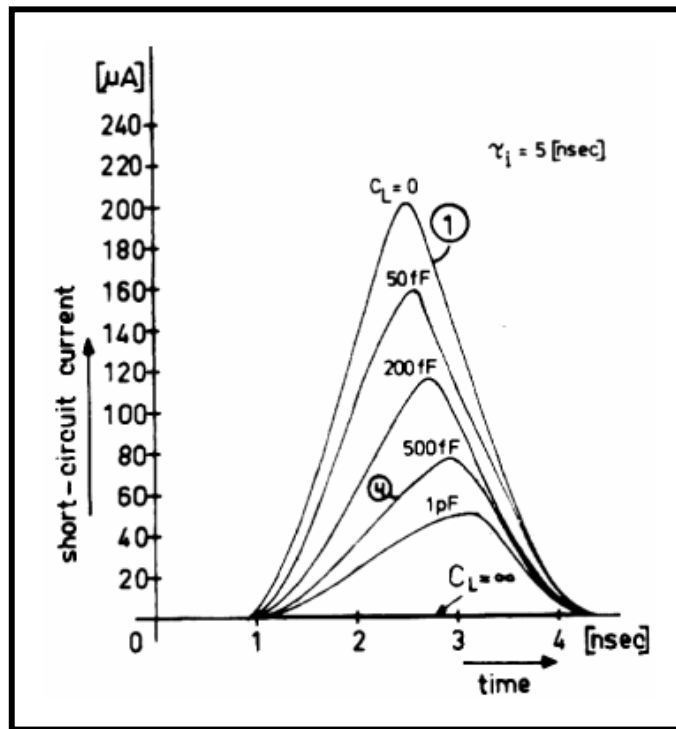
$$E/E_{ref} = f(f)$$



# Short Circuit Currents

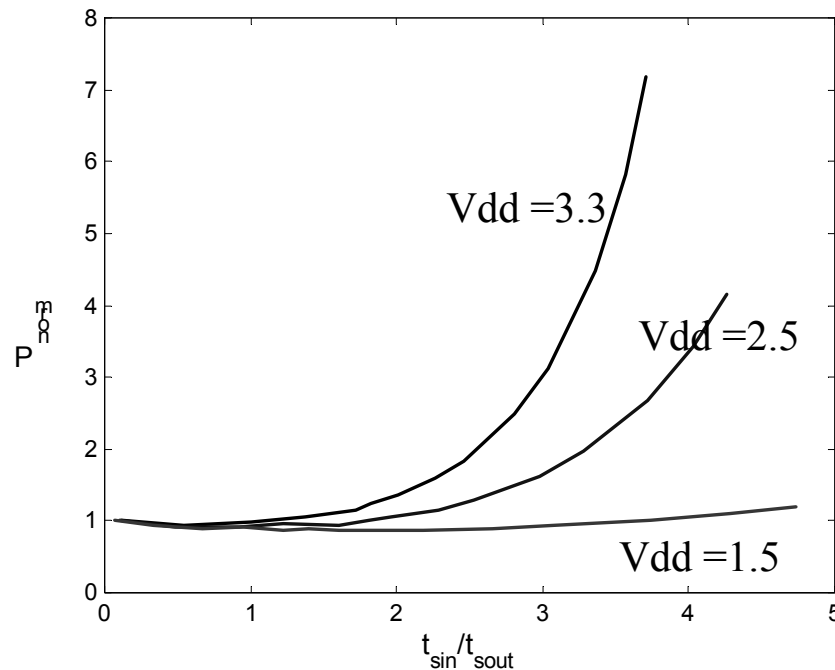


# How to keep Short-Circuit Currents Low?



Short circuit current goes to zero if  $t_{fall} \gg t_{rise}$ ,  
but can't do this for cascade logic, so ...

# Minimizing Short-Circuit Power



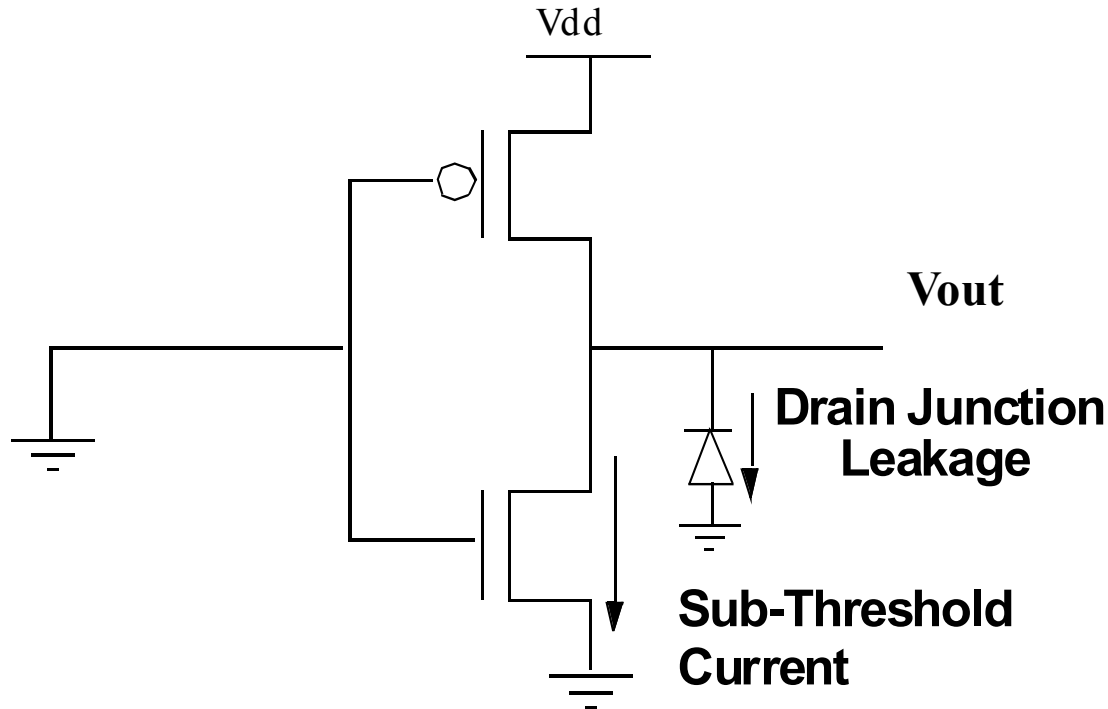
- Keep the input and output rise/fall times the same  
( $< 10\%$  of Total Consumption)

from [Veendrick84]

(*IEEE Journal of Solid-State Circuits*, August 1984)

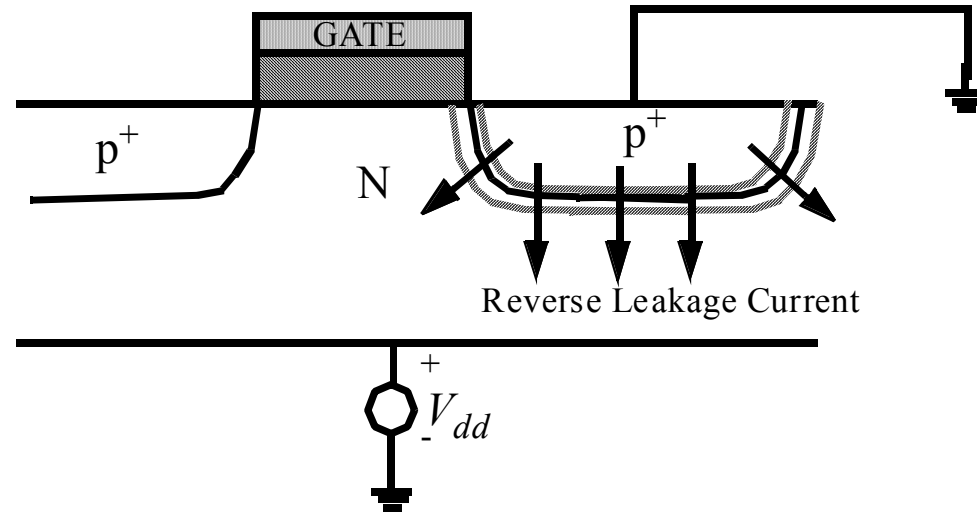
- If  $V_{dd} < V_{tn} + |V_{tp}|$  then short-circuit power can be *eliminated*!

# Leakage



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

# Reverse-Biased Diode Leakage

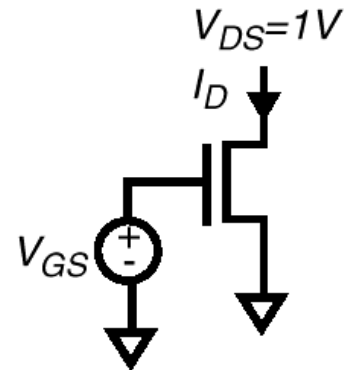
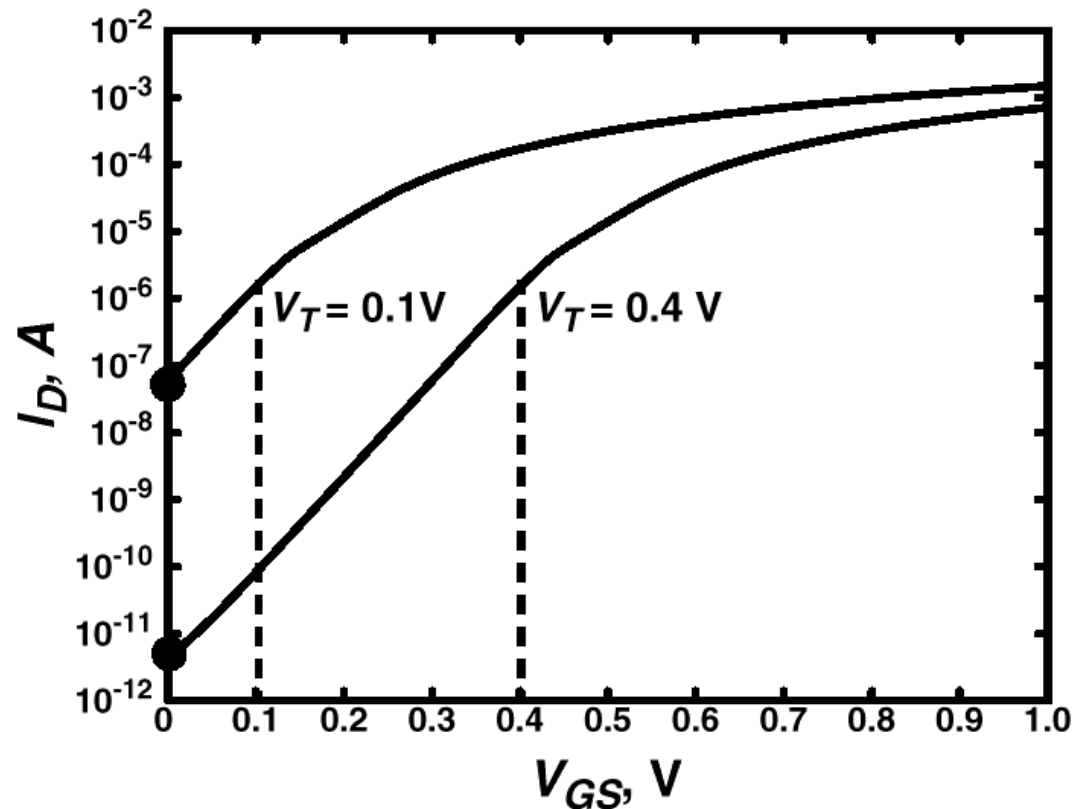


$$I_{DL} = J_S \times A$$

$J_S = 10\text{-}100 \text{ pA}/\mu\text{m}^2$  at 25 deg C for 0.25 $\mu\text{m}$  CMOS  
 $J_S$  doubles for every 9 deg C!

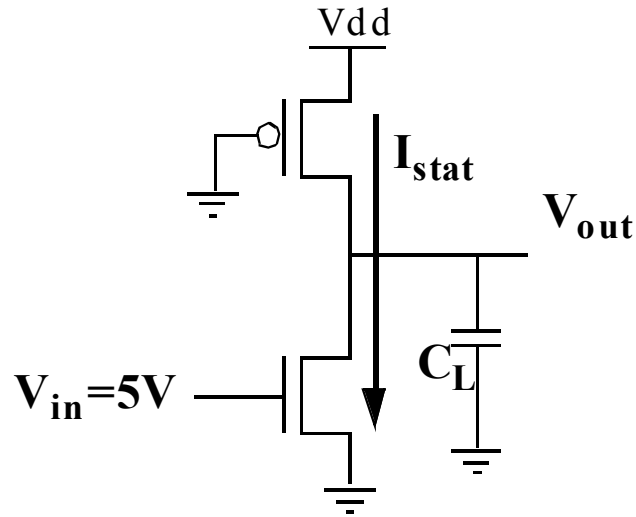


# Subthreshold Leakage Component



- Leakage control is critical for low-voltage operation

# ***Static Power Consumption***



$$P_{\text{stat}} = P_{(\text{In}=1)} \cdot V_{\text{dd}} \cdot I_{\text{stat}}$$

Wasted energy ...

Should be avoided in almost all cases,  
but could help reducing energy in others (e.g. sense amps)

# ***Principles for Power Reduction***

- Prime choice: Reduce voltage!
  - Recent years have seen an acceleration in supply voltage reduction
  - Design at very low voltages still open question (0.6 ... 0.9 V by 2010!)
- Reduce switching activity
- Reduce physical capacitance
  - Device Sizing: for  $F=20$ 
    - $f_{opt}(\text{energy})=3.53$ ,  $f_{opt}(\text{performance})=4.47$



# ***Impact of Technology Scaling***

# ***Goals of Technology Scaling***

- ❑ Make things cheaper:
  - Want to sell more functions (transistors) per chip for the same money
  - Build same products cheaper, sell the same part for less money
  - Price of a transistor has to be reduced
- ❑ But also want to be faster, smaller, lower power

# ***Technology Scaling***

- ❑ Goals of scaling the dimensions by 30%:
  - Reduce gate delay by 30% (increase operating frequency by 43%)
  - Double transistor density
  - Reduce energy per transition by 65% (50% power savings @ 43% increase in frequency)
- ❑ Die size used to increase by 14% per generation
- ❑ Technology generation spans 2-3 years

# Technology Generations

Table 2. Time overlap of semiconductor technology generations.																	
95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12
350 nm	1	2	3	4	5												
-2	-1	250 nm	1	2	3	4	5										
-4	-3	-2	-1	100 nm	1	2	3	4	5								
-6	-5	-4	-3	-2	-1	150 nm	1	2	3	4	5						
-8	-7	-6	-5	-4	-3	-2	-1	180 nm	1	2	3	4	5				
-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	100 nm	1	2	3	4	5	
			-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	70 nm	1	2	3
						-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	50 nm

# ***Technology Evolution (2000 data)***

International Technology Roadmap for Semiconductors

<b>Year of Introduction</b>	<b>1999</b>	<b>2000</b>	<b>2001</b>	<b>2004</b>	<b>2008</b>	<b>2011</b>	<b>2014</b>
<b>Technology node [nm]</b>	<b>180</b>		<b>130</b>	<b>90</b>	<b>60</b>	<b>40</b>	<b>30</b>
<b>Supply [V]</b>	<b>1.5-1.8</b>	<b>1.5-1.8</b>	<b>1.2-1.5</b>	<b>0.9-1.2</b>	<b>0.6-0.9</b>	<b>0.5-0.6</b>	<b>0.3-0.6</b>
<b>Wiring levels</b>	<b>6-7</b>	<b>6-7</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>9-10</b>	<b>10</b>
<b>Max frequency [GHz], Local-Global</b>	<b>1.2</b>	<b>1.6-1.4</b>	<b>2.1-1.6</b>	<b>3.5-2</b>	<b>7.1-2.5</b>	<b>11-3</b>	<b>14.9-3.6</b>
<b>Max <math>\mu</math>P power [W]</b>	<b>90</b>	<b>106</b>	<b>130</b>	<b>160</b>	<b>171</b>	<b>177</b>	<b>186</b>
<b>Bat. power [W]</b>	<b>1.4</b>	<b>1.7</b>	<b>2.0</b>	<b>2.4</b>	<b>2.1</b>	<b>2.3</b>	<b>2.5</b>

Node years: 2007/65nm, 2010/45nm, 2013/33nm, 2016/23nm



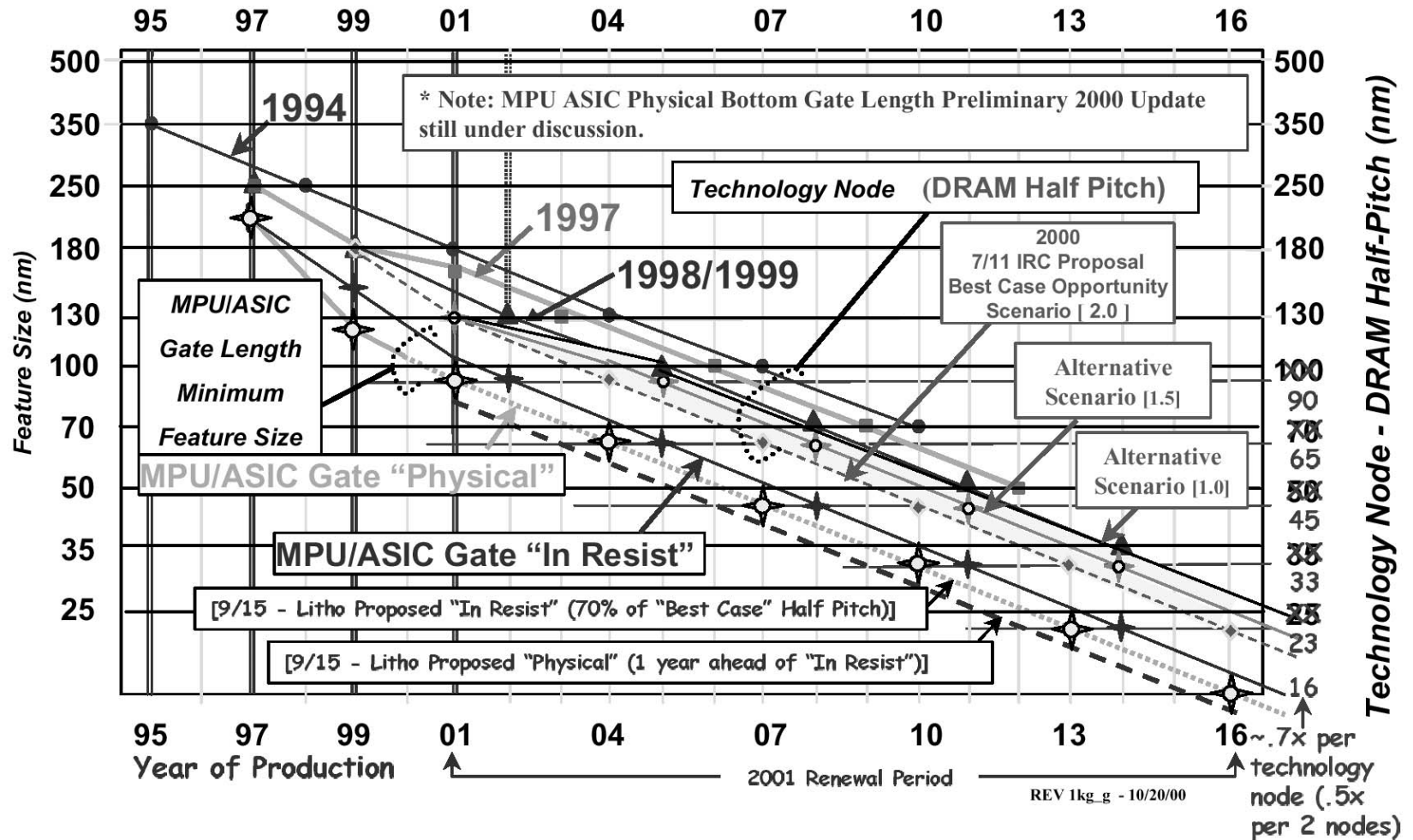
# Technology Evolution (1999)

Year of Introduction	1994	1997	2000	2003	2006	2009
Channel length ( $\mu\text{m}$ )	0.4	0.3	0.25	0.18	0.13	0.1
Gate oxide (nm)	12	7	6	4.5	4	4
$V_{DD}$ (V)	3.3	2.2	2.2	1.5	1.5	1.5
$V_T$ (V)	0.7	0.7	0.7	0.6	0.6	0.6
NMOS $I_{Dsat}$ (mA/ $\mu\text{m}$ ) (@ $V_{GS} = V_{DD}$ )	0.35	0.27	0.31	0.21	0.29	0.33
PMOS $I_{Dsat}$ (mA/ $\mu\text{m}$ ) (@ $V_{GS} = V_{DD}$ )	0.16	0.11	0.14	0.09	0.13	0.16

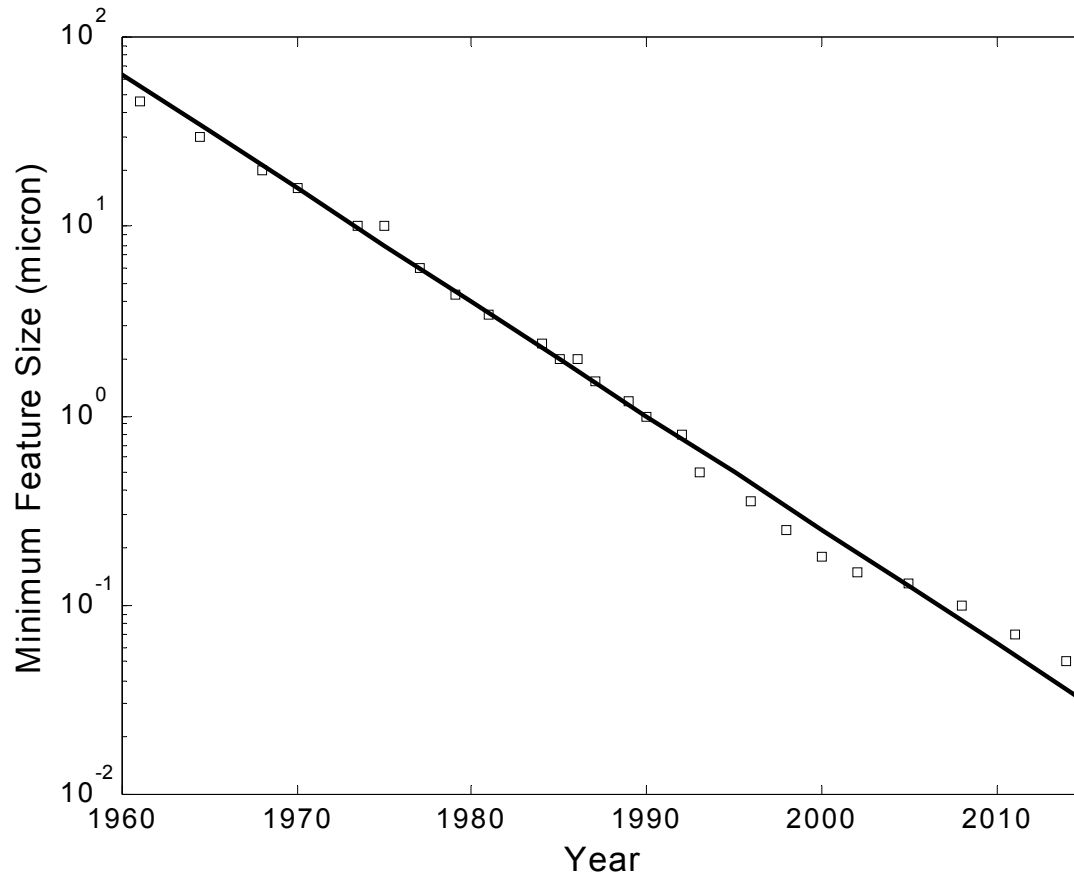
# ITRS Technology Roadmap

## Acceleration Continues

(Including MPU/ASIC "Physical Gate Length" Proposal)

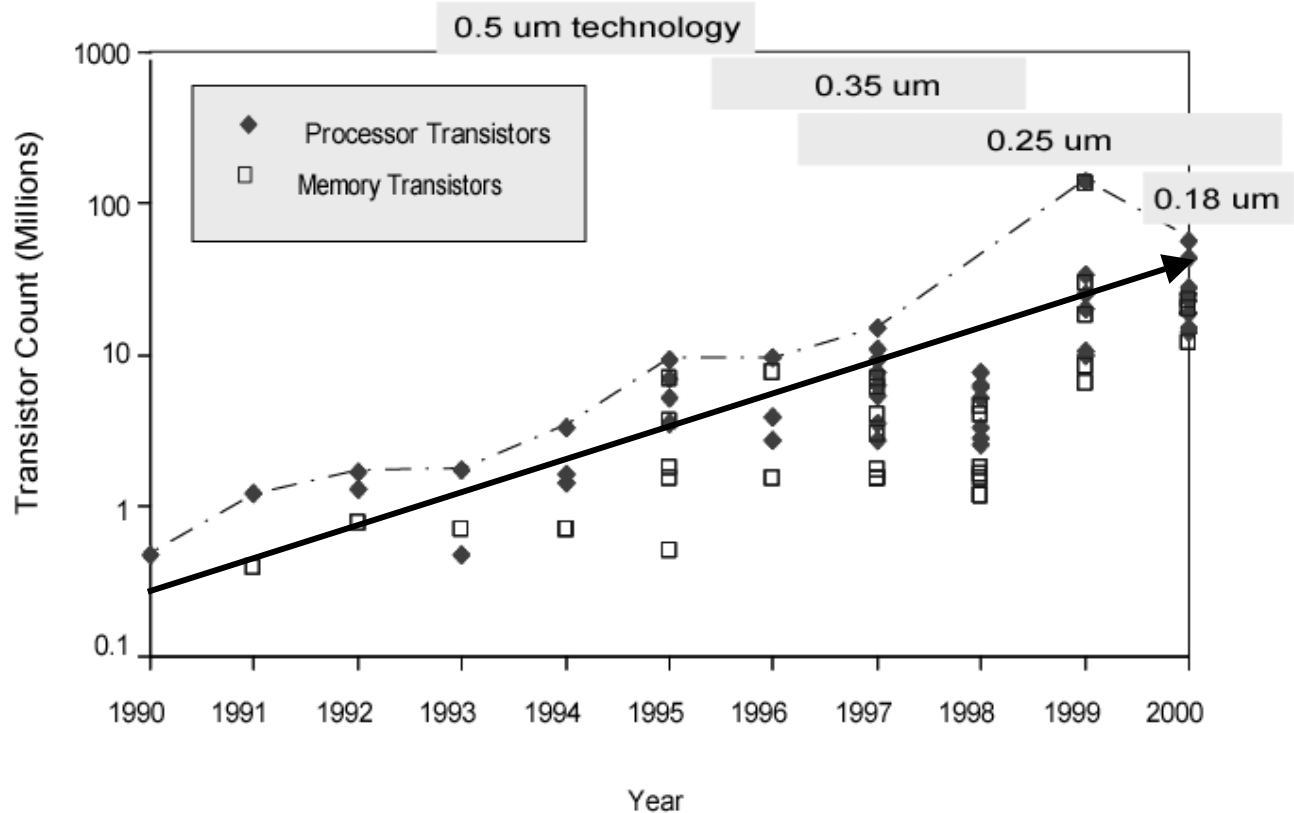


# Technology Scaling (1)



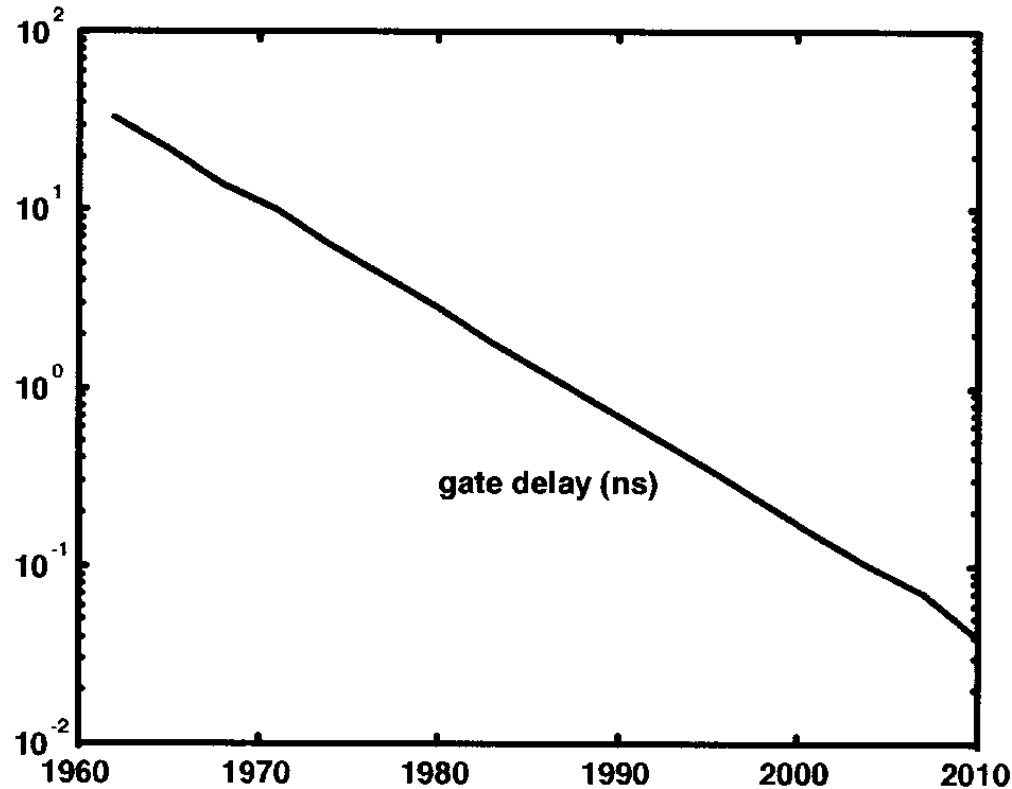
***Minimum Feature Size***

# Technology Scaling (2)



***Number of components per chip***

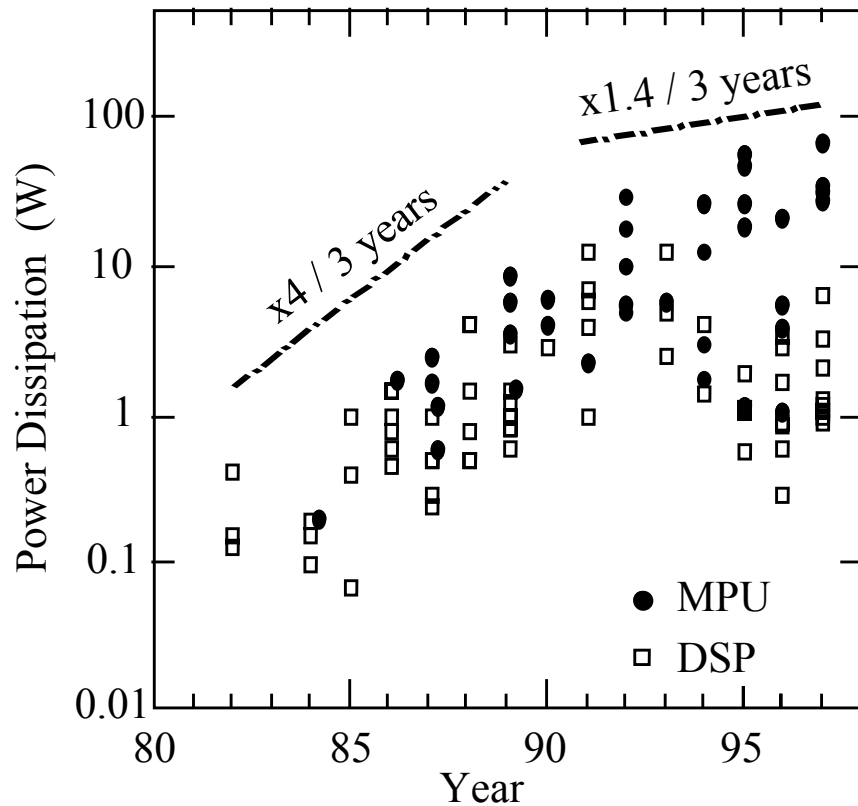
# Technology Scaling (3)



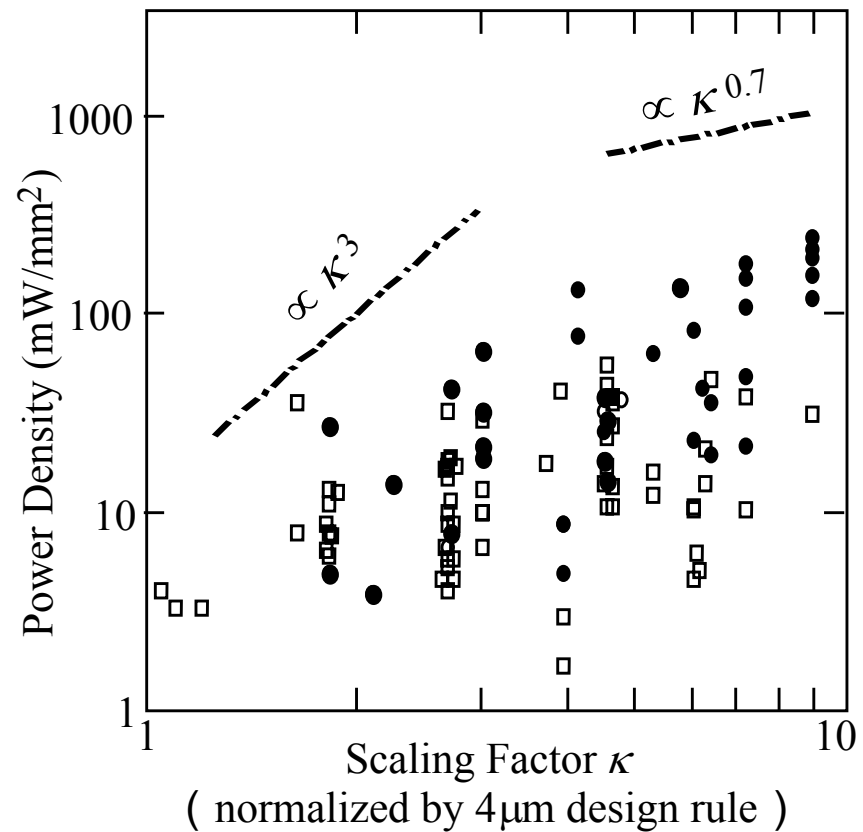
$t_p$  decreases by 13%/year  
50% every 5 years!

***Propagation Delay***

# Technology Scaling (4)



(a) Power dissipation vs. year.



(b) Power density vs. scaling factor.

From Kuroda

# ***Technology Scaling Models***

- **Full Scaling (Constant Electrical Field)**

ideal model — dimensions and voltage scale together by the same factor  $S$

- **Fixed Voltage Scaling**

most common model until recently —  
only dimensions scale, voltages remain constant

- **General Scaling**

most realistic for today's situation —  
voltages and dimensions scale with different factors

# Scaling Relationships for Long Channel Devices

Parameter	Relation	Full Scaling	General Scaling	Fixed Voltage Scaling
$W, L, t_{ox}$		$1/S$	$1/S$	$1/S$
$V_{DD}, V_T$		$1/S$	$1/U$	$1$
$N_{SUB}$	$V/W_{depl}^2$	$S$	$S^2/U$	$S^2$
Area/Device	$WL$	$1/S^2$	$1/S^2$	$1/S^2$
$C_{ox}$	$1/t_{ox}$	$S$	$S$	$S$
$C_L$	$C_{ox}WL$	$1/S$	$1/S$	$1/S$
$k_n, k_p$	$C_{ox}W/L$	$S$	$S$	$S$
$I_{av}$	$k_{n,p} V^2$	$1/S$	$S/U^2$	$S$
$t_p$ (intrinsic)	$C_L V / I_{av}$	$1/S$	$U/S^2$	$1/S^2$
$P_{av}$	$C_L V^2 / t_p$	$1/S^2$	$S/U^3$	$S$
PDP	$C_L V^2$	$1/S^3$	$1/SU^2$	$1/S$

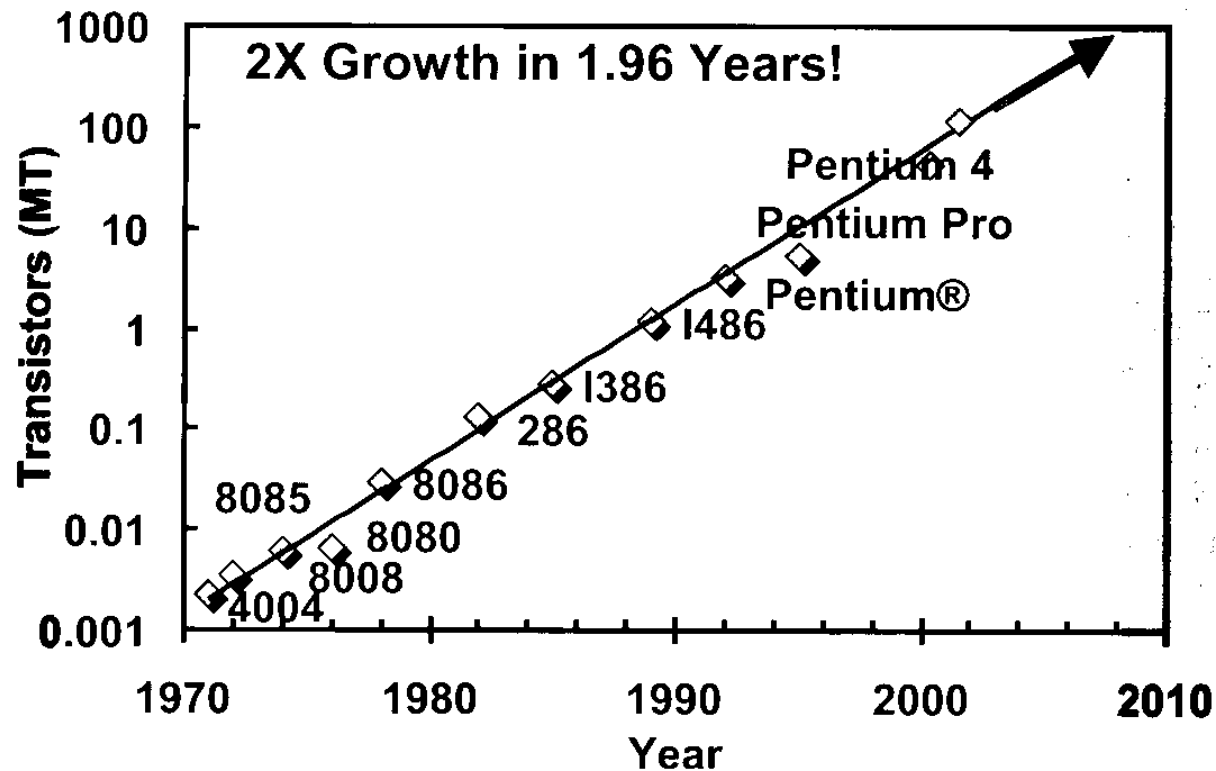


# Transistor Scaling

## (velocity-saturated devices)

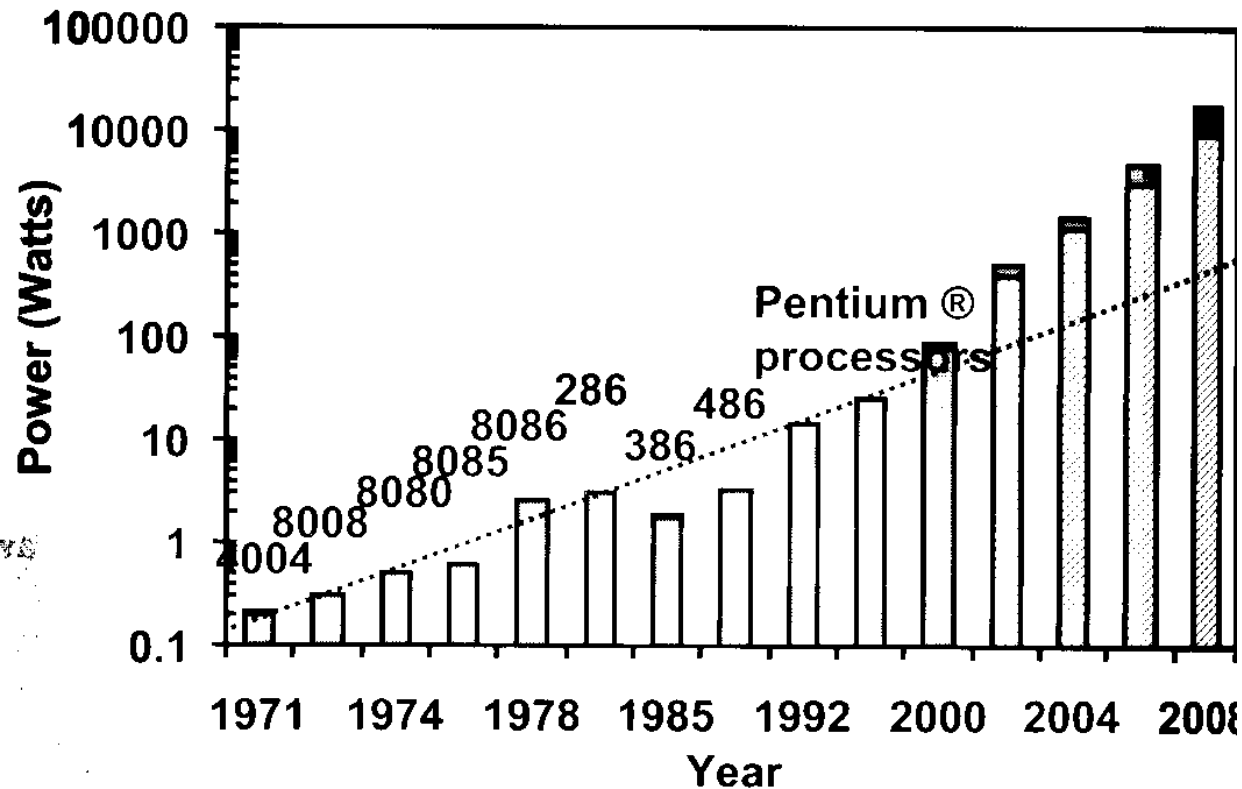
Parameter	Relation	Full Scaling	General Scaling	Fixed-Voltage Scaling
$W, L, t_{ox}$		$1/S$	$1/S$	$1/S$
$V_{DD}, V_T$		$1/S$	$1/U$	1
$N_{SUB}$	$V/W_{depl}^2$	$S$	$S^2/U$	$S^2$
Area/Device	$WL$	$1/S^2$	$1/S^2$	$1/S^2$
$C_{ox}$	$1/t_{ox}$	$S$	$S$	$S$
$C_{gate}$	$C_{ox}WL$	$1/S$	$1/S$	$1/S$
$k_n, k_p$	$C_{ox}W/L$	$S$	$S$	$S$
Current Density	$I_{sat}/Area$	$S$	$S^2/U$	$S^2$
$R_{on}$	$V/I_{sat}$	1	1	1
$P$	$I_{sat}V$	$1/S^2$	$1/U^2$	1

# $\mu$ Processor Scaling



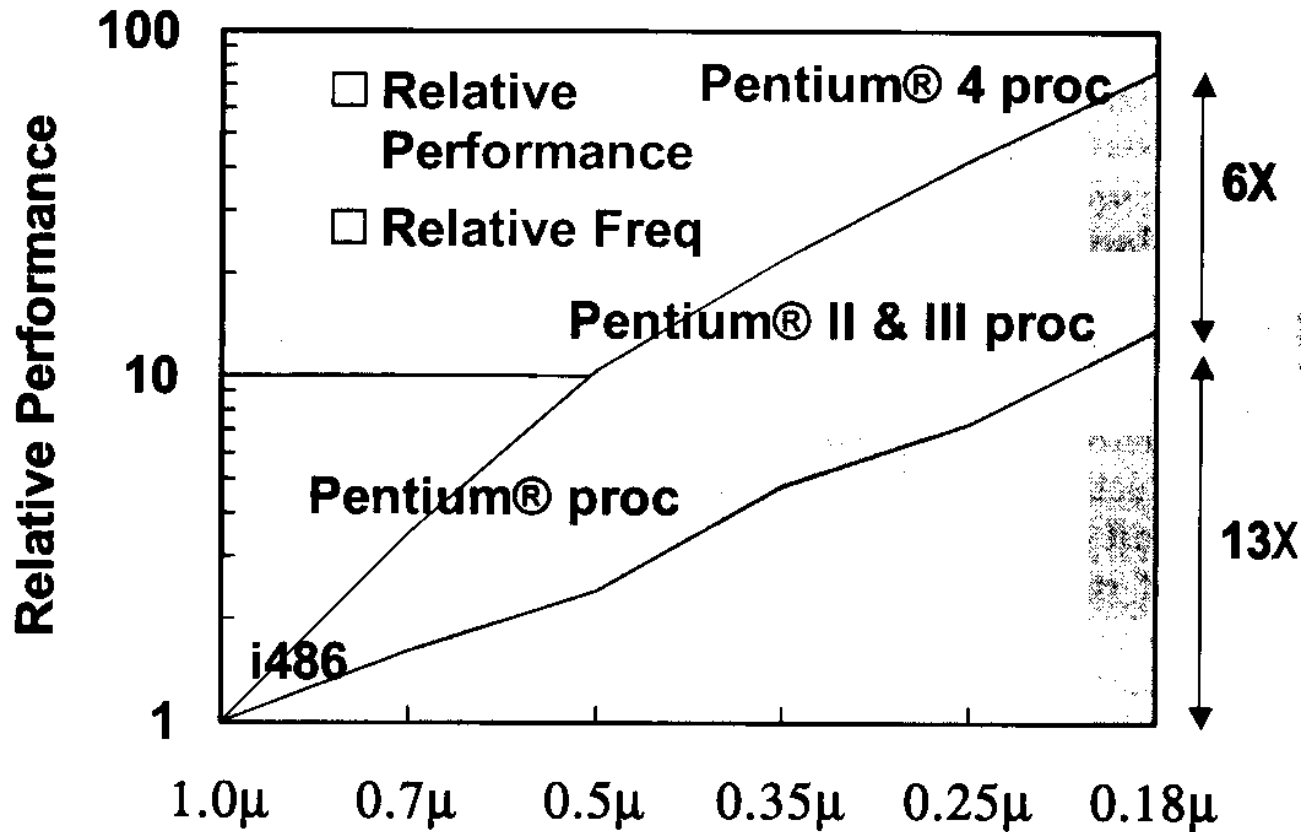
P.Gelsinger:  $\mu$ Processors for the New Millenium, ISSCC 2001

# $\mu$ Processor Power



P.Gelsinger:  $\mu$ Processors for the New Millenium, ISSCC 2001

# $\mu$ Processor Performance



P.Gelsinger:  $\mu$ Processors for the New Millenium, ISSCC 2001

# 2010 Outlook

- ❑ Performance 2X/16 months
  - 1 TIP (terra instructions/s)
  - 30 GHz clock
- ❑ Size
  - No of transistors: 2 Billion
  - Die: 40\*40 mm
- ❑ Power
  - 10kW!!
  - Leakage: 1/3 active Power

P.Gelsinger:  $\mu$ Processors for the New Millenium, ISSCC 2001

# ***Some interesting questions***

- ❑ What will cause this model to break?
- ❑ When will it break?
- ❑ Will the model gradually slow down?
  - Power and power density
  - Leakage
  - Process Variation