
CSE477

VLSI Digital Circuits

Fall 2002

Lecture 12&13: Designing for Low Power

Mary Jane Irwin (www.cse.psu.edu/~mji)
www.cse.psu.edu/~cg477

[Adapted from Rabaey's *Digital Integrated Circuits*, ©2002, J. Rabaey et al.]

Review: Designing Fast CMOS Gates

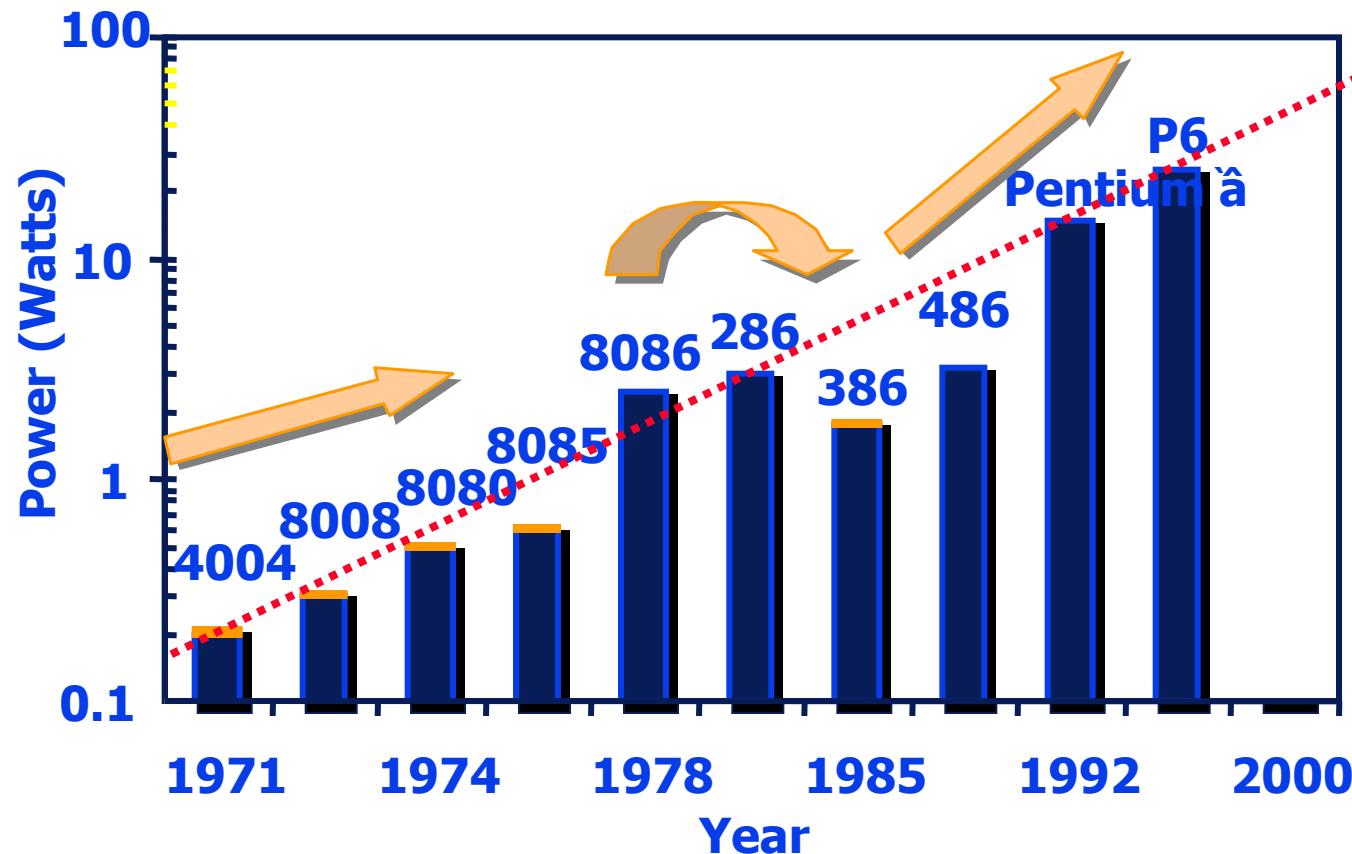
- ❑ Transistor sizing
- ❑ Progressive transistor sizing
 - ❑ fet closest to the output is smallest of series fets
- ❑ Transistor ordering
 - ❑ put latest arriving signal closest to the output
- ❑ Logic structure reordering
 - ❑ replace large fan-in gates with smaller fan-in gate network
- ❑ Logical effort
- ❑ Buffer (inverter) insertion
 - ❑ separate large fan-in from large C_L with buffers
 - ❑ uses buffers so there are no more than four TGs in series

Why Power Matters

- ❑ Packaging costs
- ❑ Power supply rail design
- ❑ Chip and system cooling costs
- ❑ Noise immunity and system reliability
- ❑ Battery life (in portable systems)
- ❑ Environmental concerns
 - ❑ Office equipment accounted for 5% of total US commercial energy usage in 1993
 - ❑ *Energy Star* compliant systems

Why worry about power? -- Power Dissipation

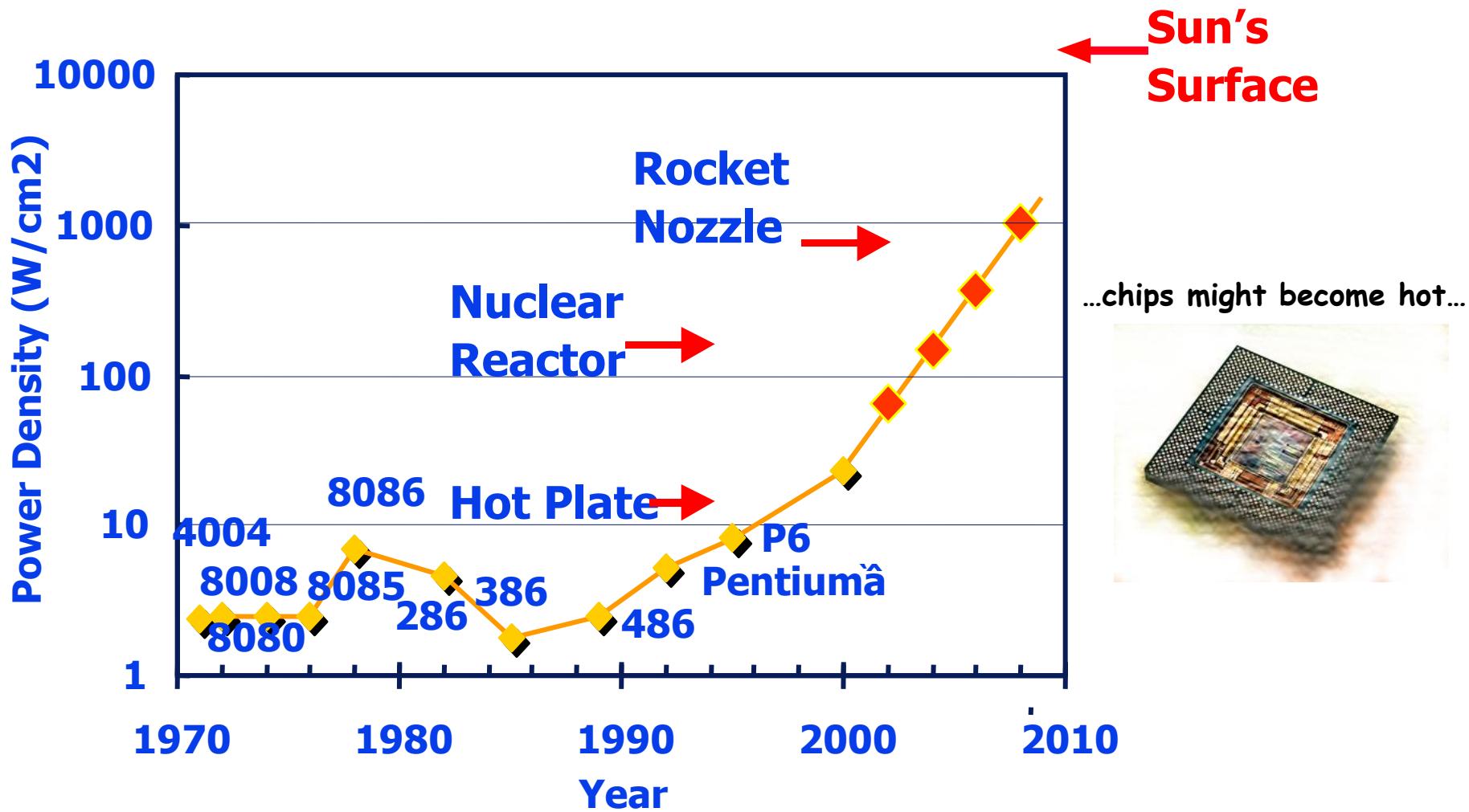
Lead microprocessors power continues to increase



Power delivery and dissipation will be prohibitive

Source: Borkar, De Intel®

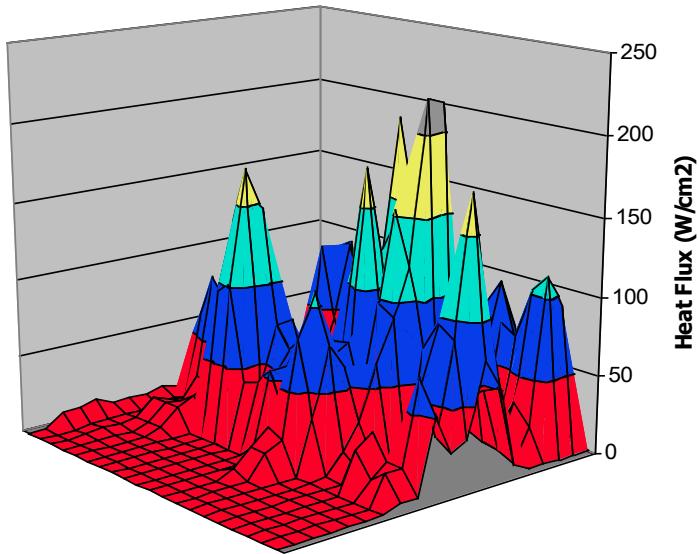
Why worry about power? -- Chip Power Density



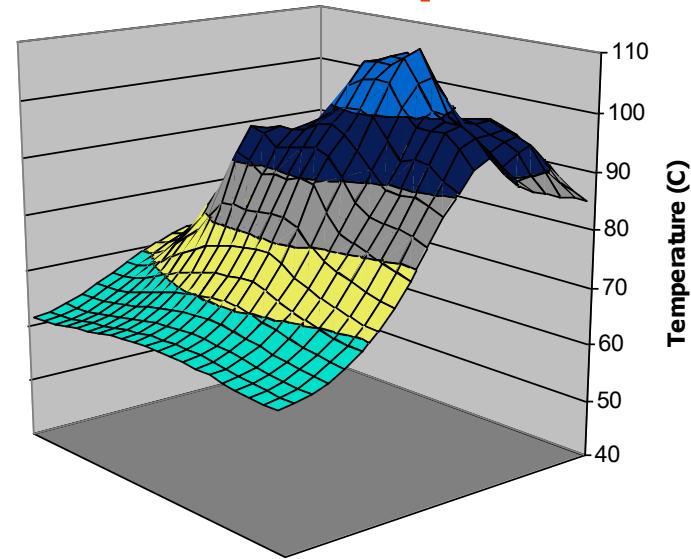
Source: Borkar, De Intel®

Chip Power Density Distribution

Power Map



On-Die Temperature



- ❑ Power density is not uniformly distributed across the chip
- ❑ Silicon is not a good heat conductor
- ❑ Max junction temperature is determined by hot-spots
 - ❑ Impact on packaging, w.r.t. cooling

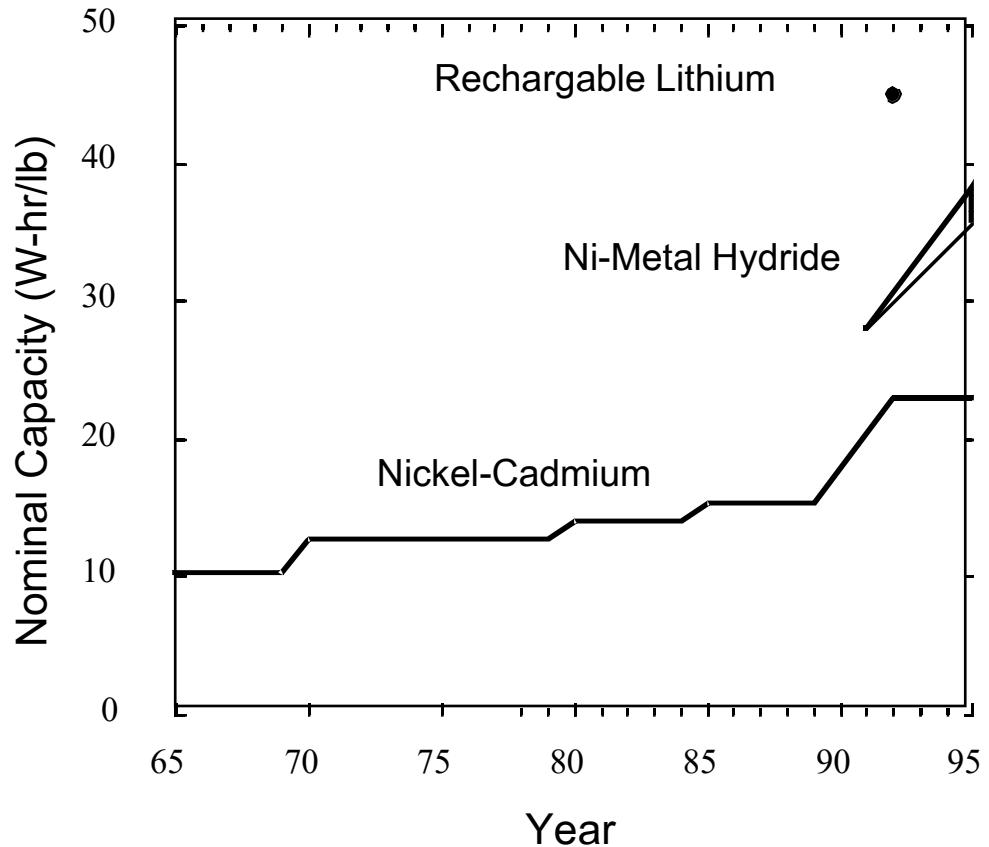
Problem Illustration

***What happens
when the
CPU cooler is
removed?***



www.tomshardware.de
www.tomshardware.com

Why worry about power ? -- Battery Size/Weight



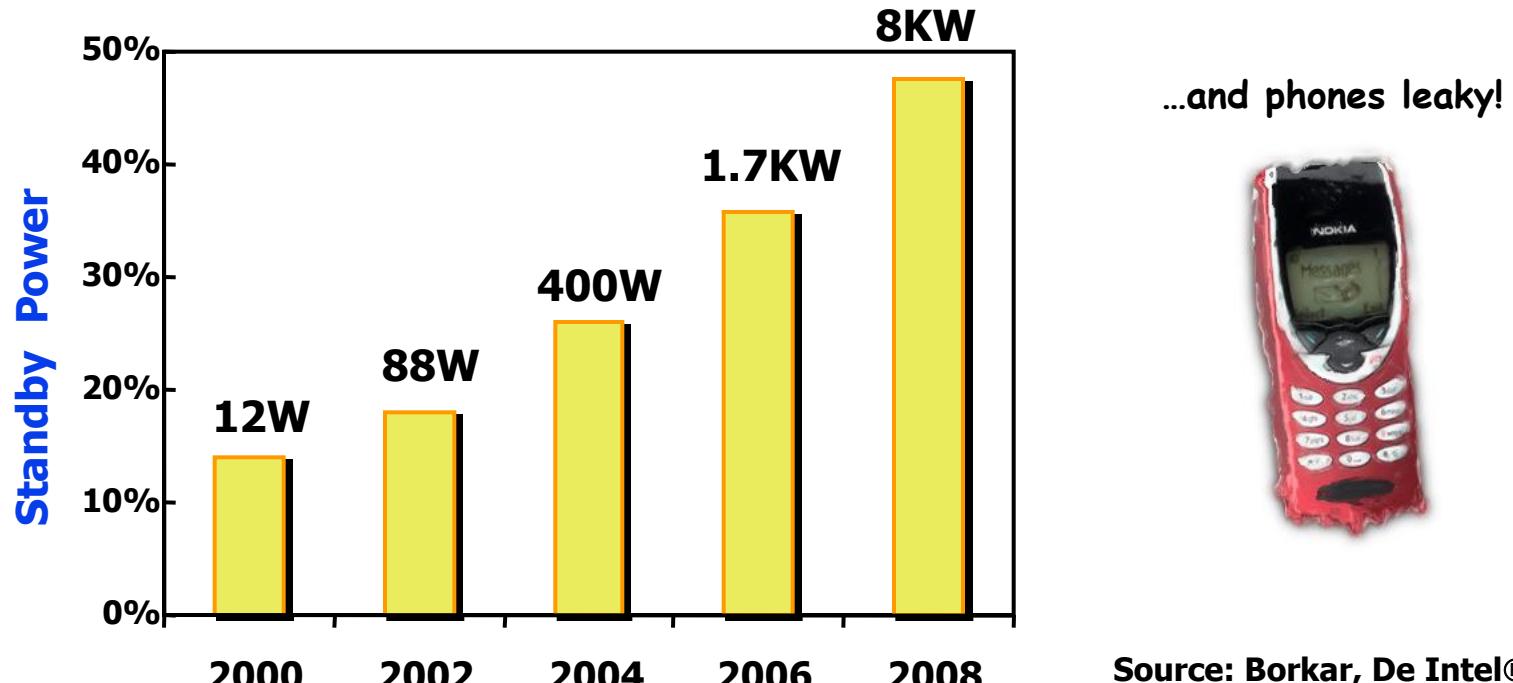
Expected battery lifetime increase
over the next 5 years: **30 to 40%**

From Rabaey, 1995

Why worry about power? -- Standby Power

Year	2002	2005	2008	2011	2014
Power supply V_{dd} (V)	1.5	1.2	0.9	0.7	0.6
Threshold V_T (V)	0.4	0.4	0.35	0.3	0.25

- Drain leakage will increase as V_T decreases to maintain noise margins and meet frequency demands, leading to excessive **battery draining standby** power consumption.

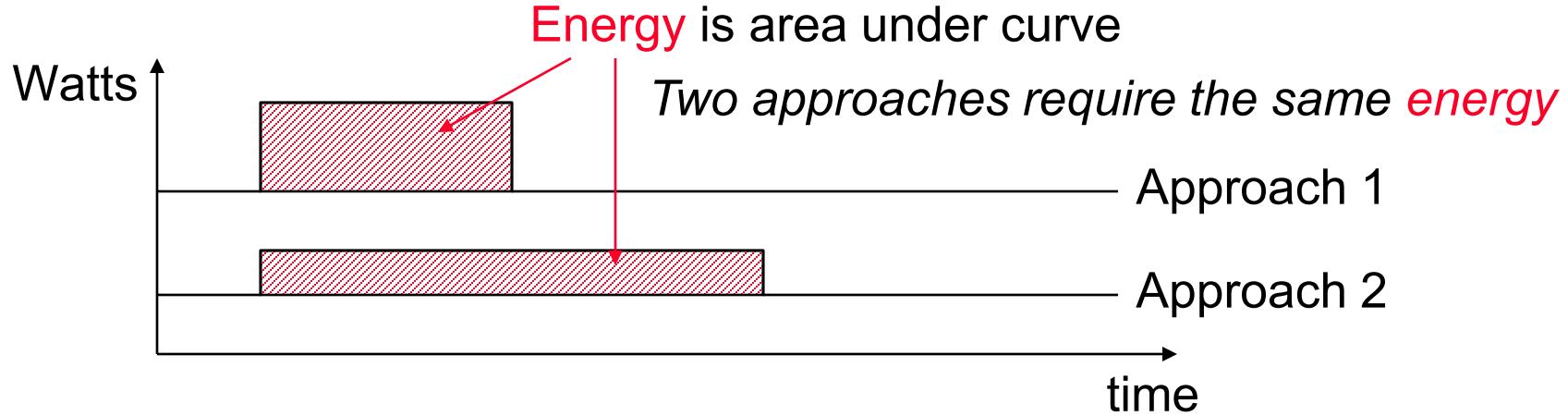
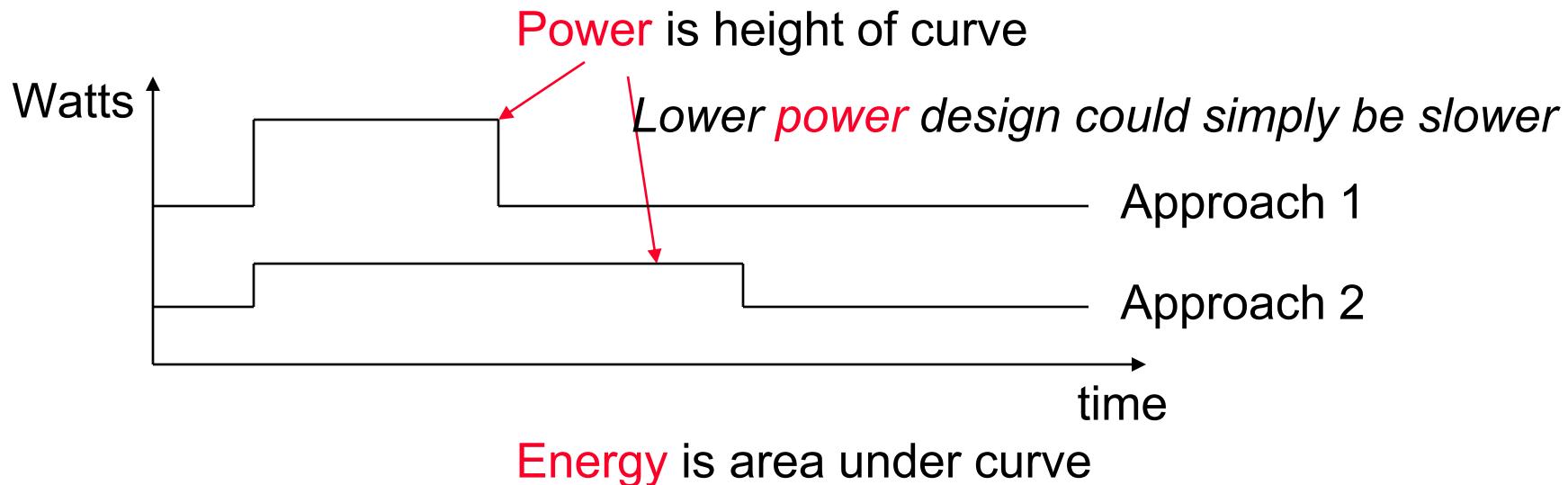


Source: Borkar, De Intel®

Power and Energy Figures of Merit

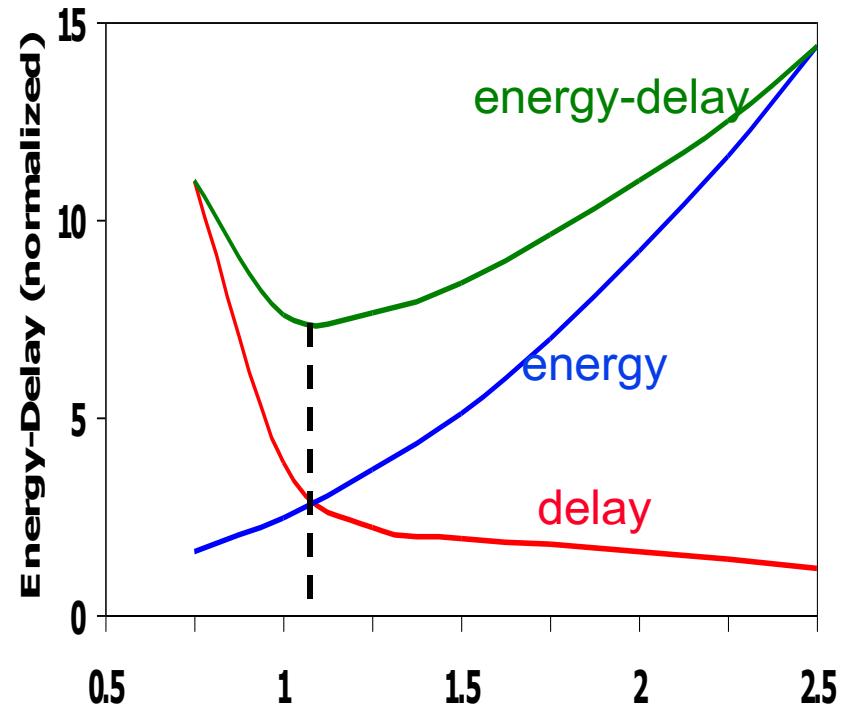
- Power consumption in Watts
 - determines battery life in hours
- Peak power
 - determines power ground wiring designs
 - sets packaging limits
 - impacts signal noise margin and reliability analysis
- Energy efficiency in Joules
 - rate at which power is consumed over time
- Energy = power * delay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

Power versus Energy



PDP and EDP

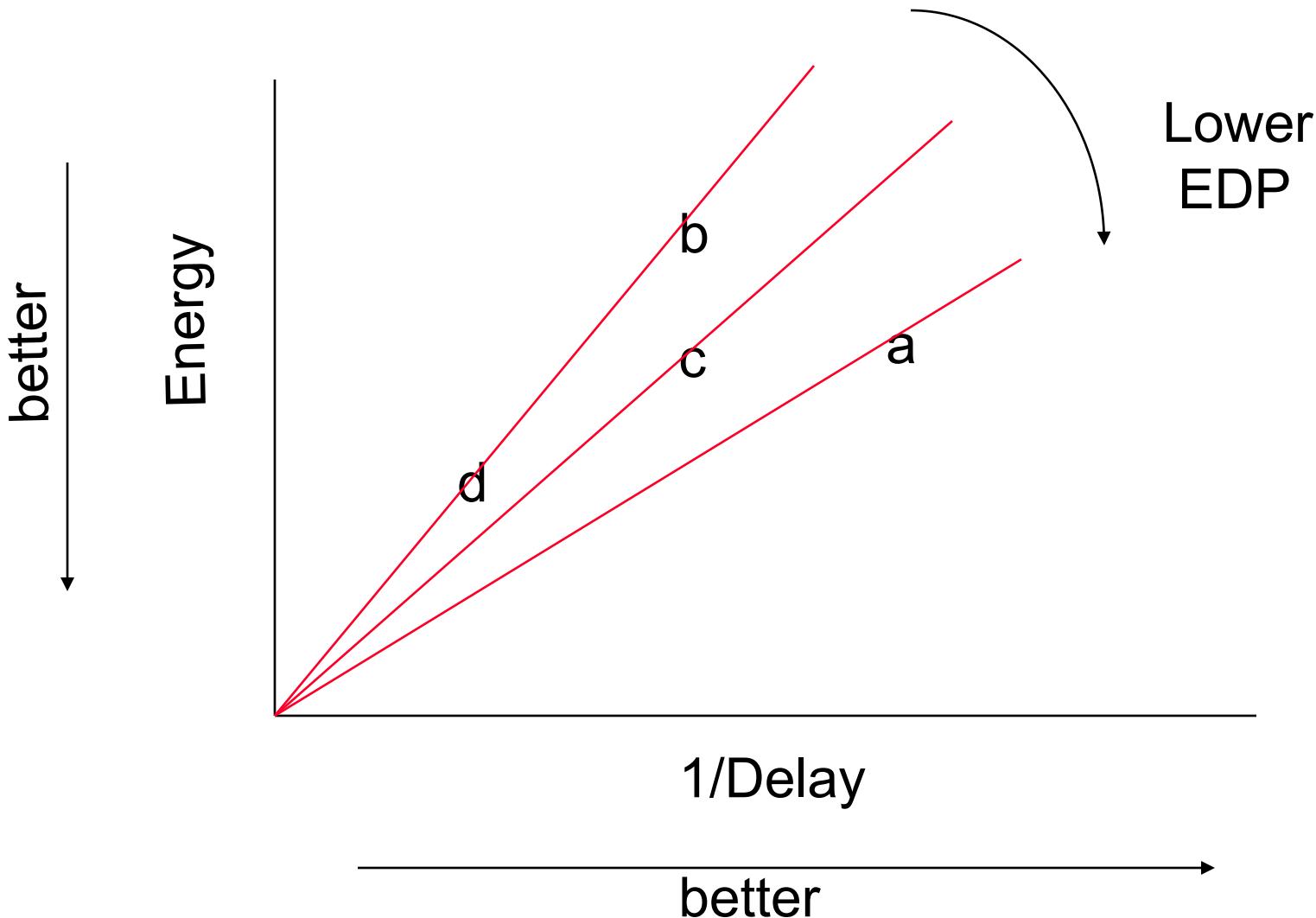
- Power-delay product (PDP) = $P_{av} * t_p = (C_L V_{DD}^2)/2$
 - PDP is the average **energy** consumed per switching event (Watts * sec = Joule)
 - **lower** power design could simply be a **slower** design
- Energy-delay product (EDP) = $PDP * t_p = P_{av} * t_p^2$
 - EDP is the average **energy** consumed multiplied by the computation time required
 - takes into account that one can **trade** increased delay for lower energy/operation (e.g., via supply voltage scaling that increases delay, but decreases energy consumption)



- allows one to understand tradeoffs better

Understanding Tradeoffs

- ❑ Which design is the “best” (fastest, coolest, both) ?



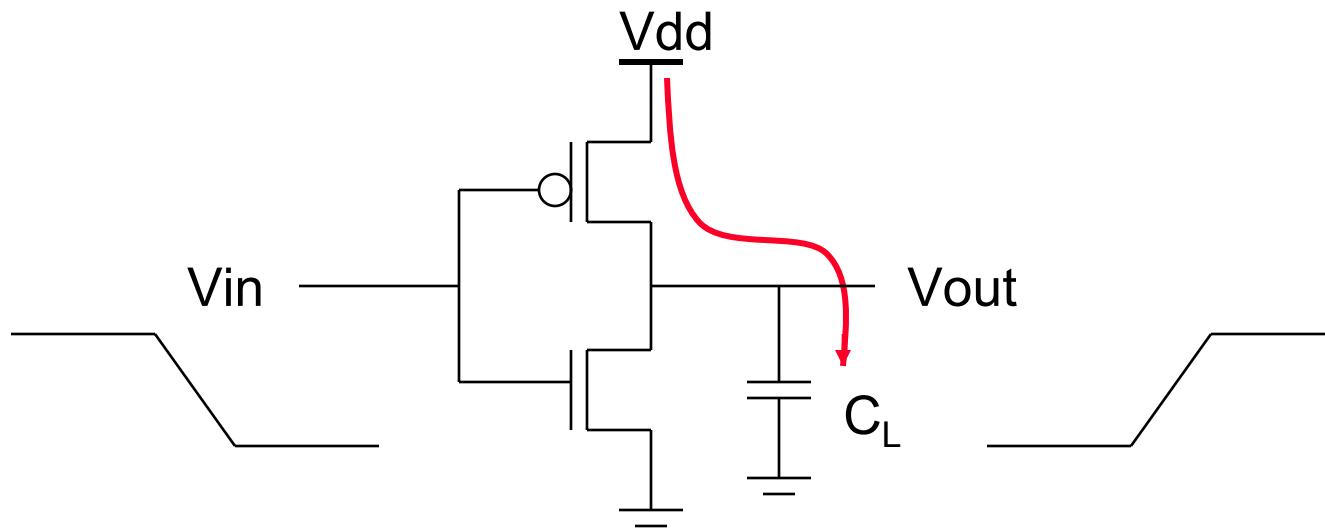
CMOS Energy & Power Equations

$$E = C_L V_{DD}^2 P_{0 \rightarrow 1} + t_{sc} V_{DD} I_{peak} P_{0 \rightarrow 1} + V_{DD} I_{leakage}$$

$$f_{0 \rightarrow 1} = P_{0 \rightarrow 1} * f_{clock}$$

$$P = \begin{array}{l} C_L V_{DD}^2 f_{0 \rightarrow 1} \\ \text{Dynamic power} \end{array} + \begin{array}{l} t_{sc} V_{DD} I_{peak} f_{0 \rightarrow 1} \\ \text{Short-circuit power} \end{array} + \begin{array}{l} V_{DD} I_{leakage} \\ \text{Leakage power} \end{array}$$

Dynamic Power Consumption



$$\text{Energy/transition} = C_L * V_{DD}^2 * P_{0 \rightarrow 1}$$

$$P_{dyn} = \text{Energy/transition} * f = C_L * V_{DD}^2 * (P_{0 \rightarrow 1} * f)$$

$$P_{dyn} = C_{EFF} * V_{DD}^2 * f \quad \text{where } C_{EFF} = P_{0 \rightarrow 1} C_L$$

Not a function of transistor sizes!
Data dependent - a function of switching activity!

Lowering Dynamic Power

Capacitance:

Function of fan-out,
wire length, transistor
sizes

Supply Voltage:
Has been dropping
with successive
generations

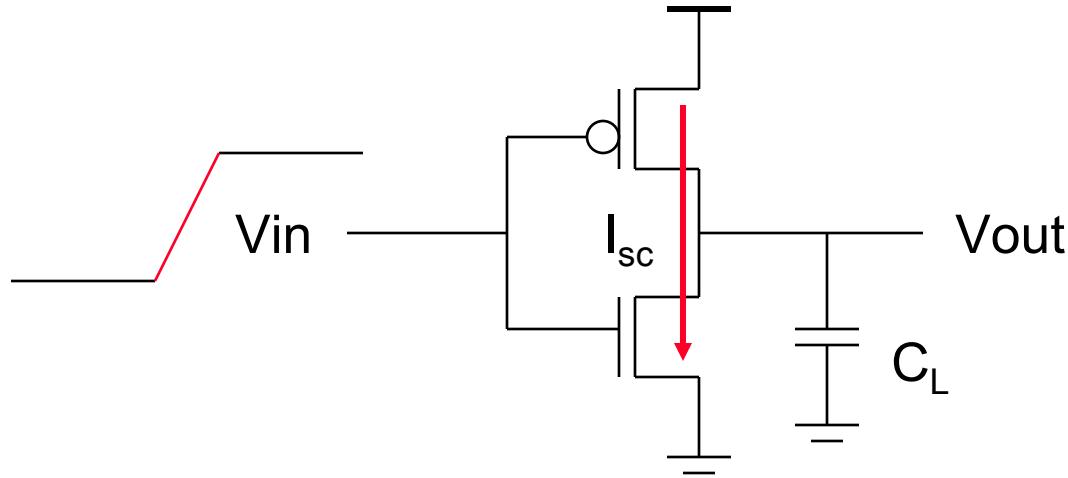
$$P_{\text{dyn}} = C_L V_{\text{DD}}^2 P_{0 \rightarrow 1} f$$

Activity factor:

How often, on average,
do wires switch?

Clock frequency:
Increasing...

Short Circuit Power Consumption



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.

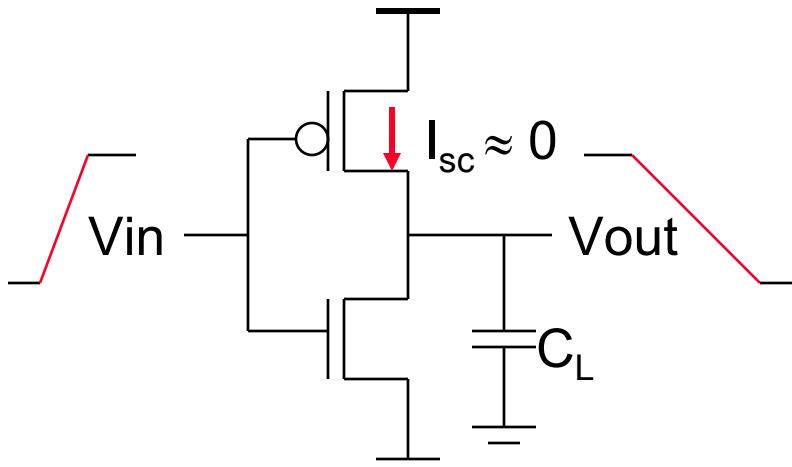
Short Circuit Currents Determinates

$$E_{sc} = t_{sc} V_{DD} I_{peak} P_{0 \rightarrow 1}$$

$$P_{sc} = t_{sc} V_{DD} I_{peak} f_{0 \rightarrow 1}$$

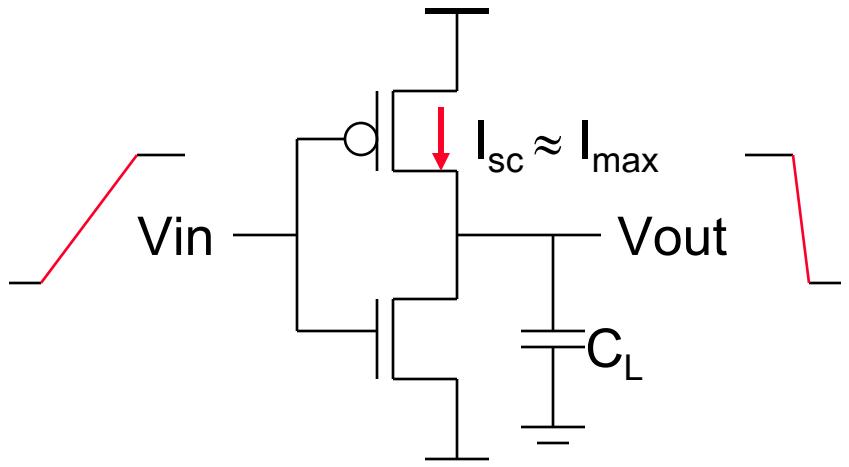
- Duration and slope of the input signal, t_{sc}
- I_{peak} determined by
 - the saturation current of the P and N transistors which depend on their **sizes**, process technology, temperature, etc.
 - strong function of the ratio between input and output slopes
 - a function of C_L

Impact of C_L on P_{sc}



Large capacitive load

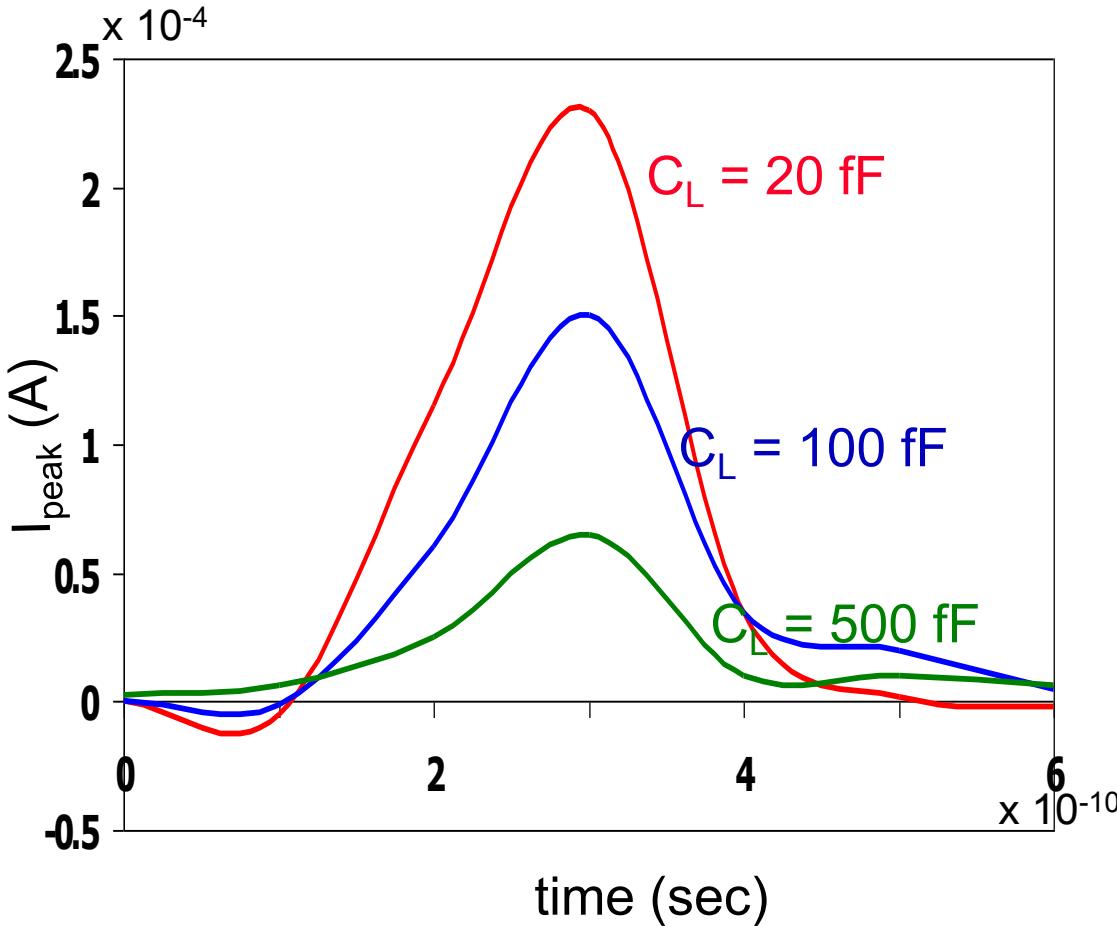
Output fall time significantly larger than input rise time.



Small capacitive load

Output fall time substantially smaller than the input rise time.

I_{peak} as a Function of C_L

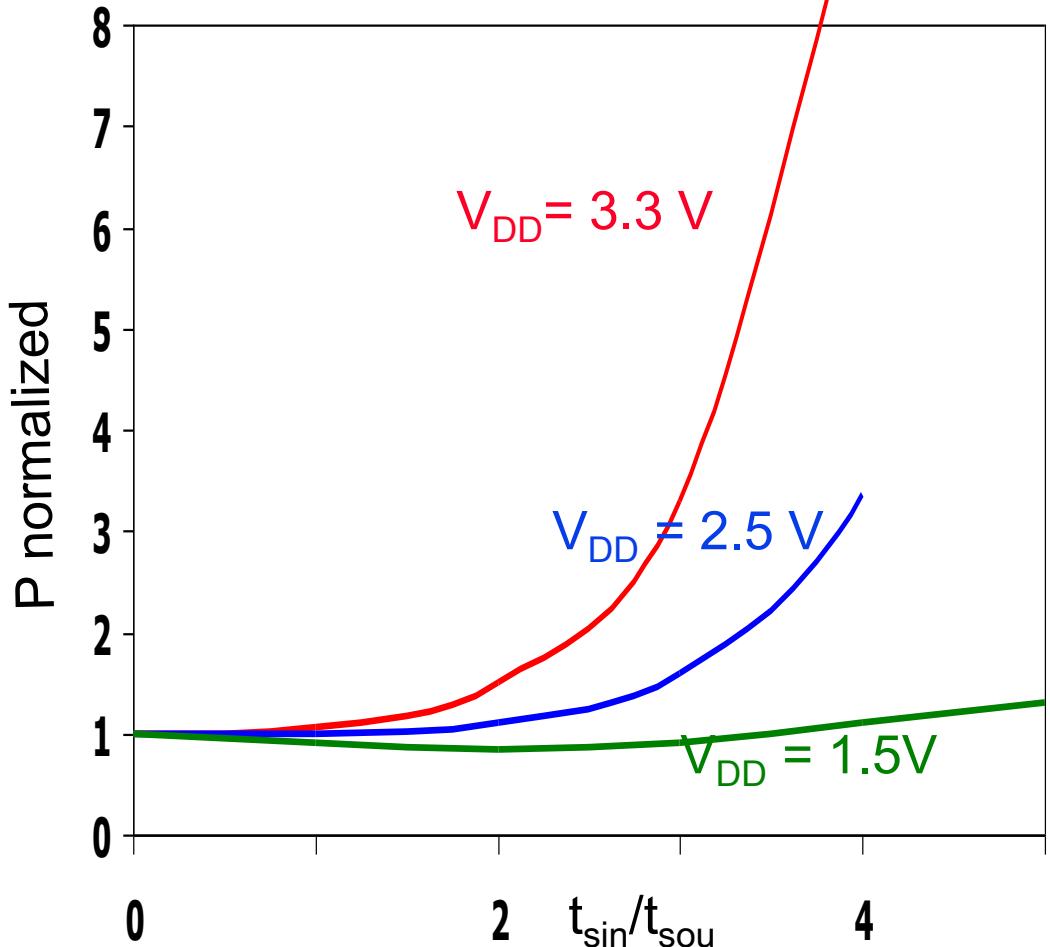


500 psec input slope

When load capacitance is small, I_{peak} is large.

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

P_{sc} as a Function of Rise/Fall Times



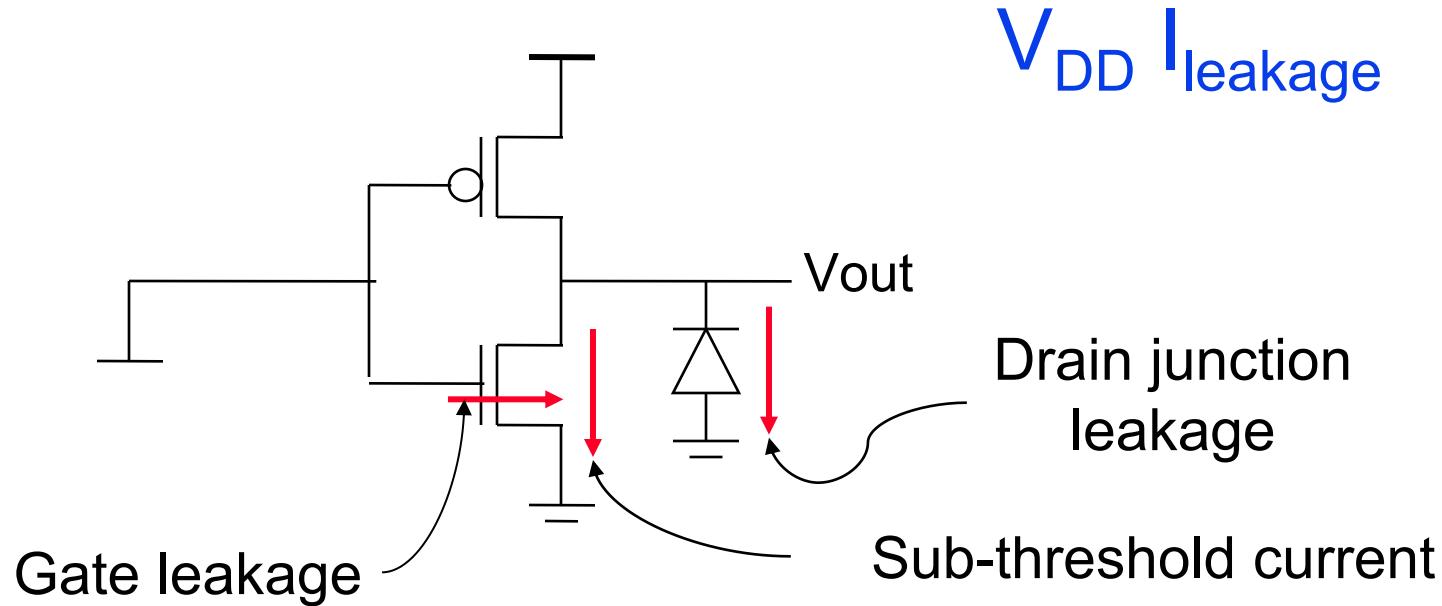
When load capacitance is small ($t_{\text{sin}}/t_{\text{sou}} > 2$ for $V_{DD} > 2\text{V}$) the power is dominated by P_{sc}

If $V_{DD} < V_{Tn} + |V_{Tp}|$ then P_{sc} is eliminated since both devices are never on at the same time.

$$\begin{aligned}W/L_p &= 1.125 \mu\text{m}/0.25 \mu\text{m} \\W/L_n &= 0.375 \mu\text{m}/0.25 \mu\text{m} \\C_L &= 30 \text{ fF}\end{aligned}$$

normalized wrt zero input
rise-time dissipation

Leakage (Static) Power Consumption

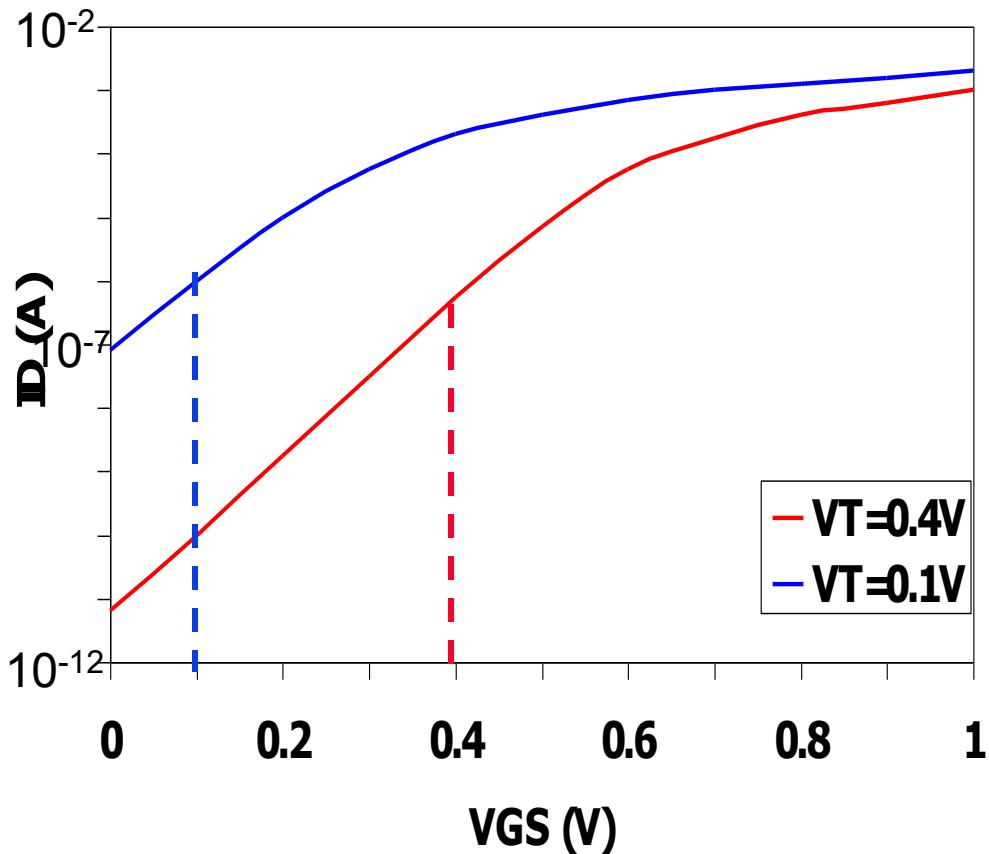


Sub-threshold current is the dominant factor.

All increase **exponentially** with temperature!

Leakage as a Function of V_T

- Continued scaling of supply voltage and the subsequent scaling of threshold voltage will make subthreshold conduction a dominate component of power dissipation.



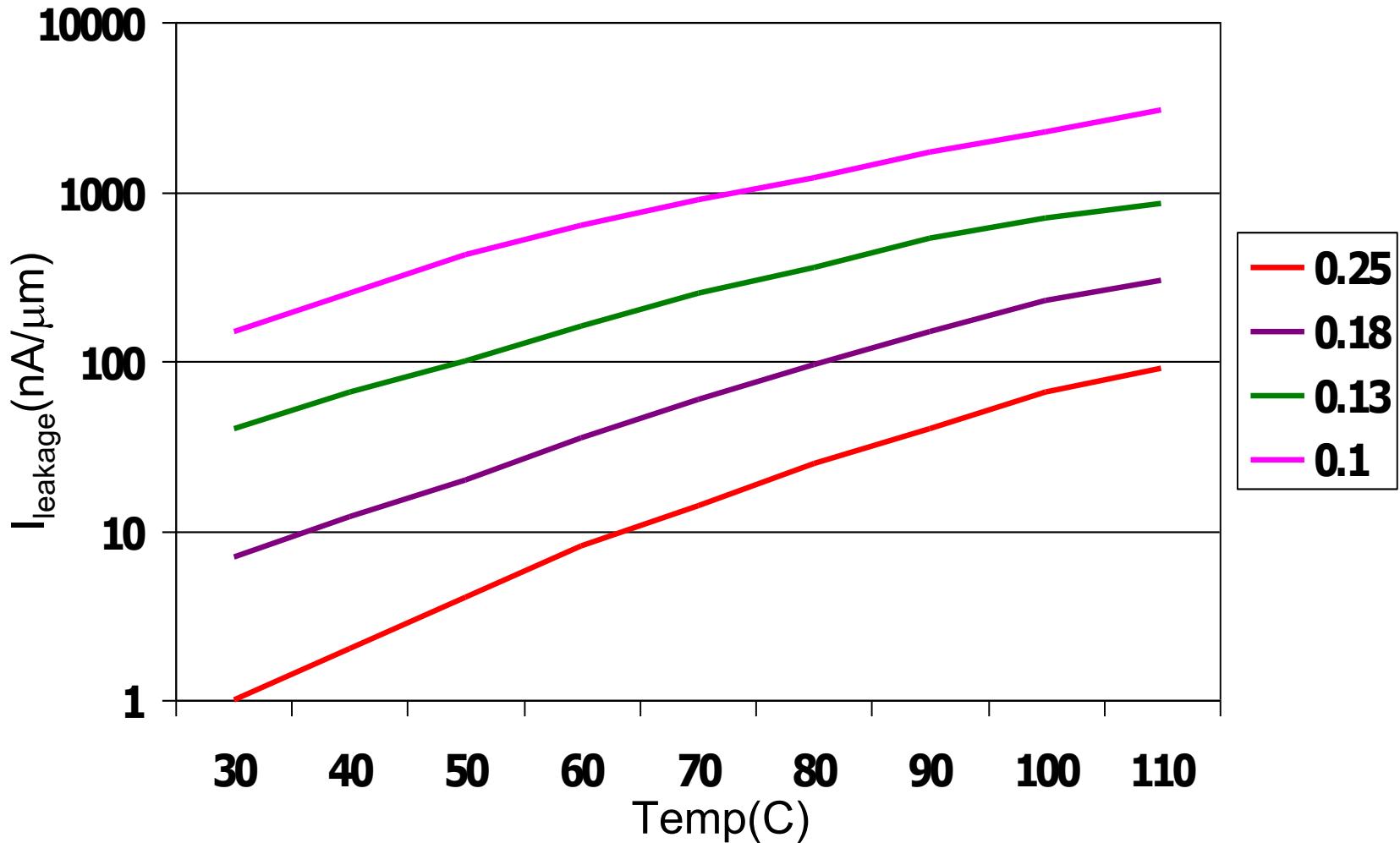
- An 90mV/decade V_T roll-off - so each 255mV increase in V_T gives 3 orders of magnitude reduction in leakage (but adversely affects performance)

TSMC Processes Leakage and V_T

	CL018 G	CL018 LP	CL018 ULP	CL018 HS	CL015 HS	CL013 HS
V_{dd}	1.8 V	1.8 V	1.8 V	2 V	1.5 V	1.2 V
T_{ox} (effective)	42 fF	42 fF	42 fF	42 fF	29 fF	24 fF
L_{gate}	0.16 μm	0.16 μm	0.18 μm	0.13 μm	0.11 μm	0.08 μm
I_{DSat} (n/p) ($\mu\text{A}/\mu\text{m}$)	600/260	500/180	320/130	780/360	860/370	920/400
I_{off} (leakage) ($\text{pA}/\mu\text{m}$)	20	1.60	0.15	300	1,800	13,000
V_{Tn}	0.42 V	0.63 V	0.73 V	0.40 V	0.29 V	0.25 V
FET Perf. (GHz)	30	22	14	43	52	80

From MPR, 2000

Exponential Increase in Leakage Currents



From De, 1999

Review: Energy & Power Equations

$$E = C_L V_{DD}^2 P_{0 \rightarrow 1} + t_{sc} V_{DD} I_{peak} P_{0 \rightarrow 1} + V_{DD} I_{leakage}$$

$$f_{0 \rightarrow 1} = P_{0 \rightarrow 1} * f_{clock}$$

$$P = C_L V_{DD}^2 f_{0 \rightarrow 1} + t_{sc} V_{DD} I_{peak} f_{0 \rightarrow 1} + V_{DD} I_{leakage}$$

Dynamic power (~90% today and decreasing relatively)

Short-circuit power (~8% today and decreasing absolutely)

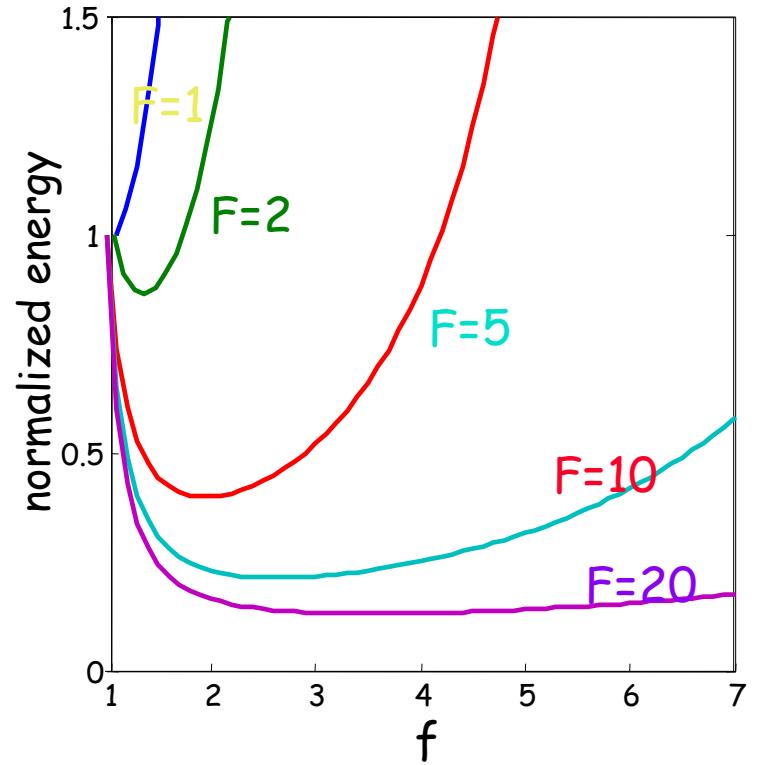
Leakage power (~2% today and increasing)

Power and Energy Design Space

	Constant Throughput/Latency	Variable Throughput/Latency	
Energy	Design Time	Non-active Modules	Run Time
Active	Logic Design Reduced V_{dd} Sizing Multi- V_{dd}	Clock Gating	DFS, DVS (Dynamic Freq, Voltage Scaling)
Leakage	+ Multi- V_T	Sleep Transistors Multi- V_{dd} Variable V_T	+ Variable V_T

Dynamic Power as a Function of Device Size

- ❑ Device sizing affects dynamic energy consumption
 - gain is largest for networks with large overall effective fan-outs ($F = C_L/C_{g,1}$)
- ❑ The optimal gate sizing factor (f) for dynamic energy is smaller than the one for performance, especially for large F 's
 - e.g., for $F=20$,
 $f_{opt}(\text{energy}) = 3.53$ while
 $f_{opt}(\text{performance}) = 4.47$
- ❑ If energy is a concern avoid oversizing beyond the optimal



From Nikolic, UCB

Dynamic Power Consumption is Data Dependent

- ❑ Switching activity, $P_{0 \rightarrow 1}$, has two components
 - ❑ A static component – function of the logic topology
 - ❑ A dynamic component – function of the timing behavior (glitching)

2-input NOR Gate

A	B	Out
0	0	1
0	1	0
1	0	0
1	1	0

Static transition probability

$$\begin{aligned}P_{0 \rightarrow 1} &= P_{\text{out}=0} \times P_{\text{out}=1} \\&= P_0 \times (1-P_0)\end{aligned}$$

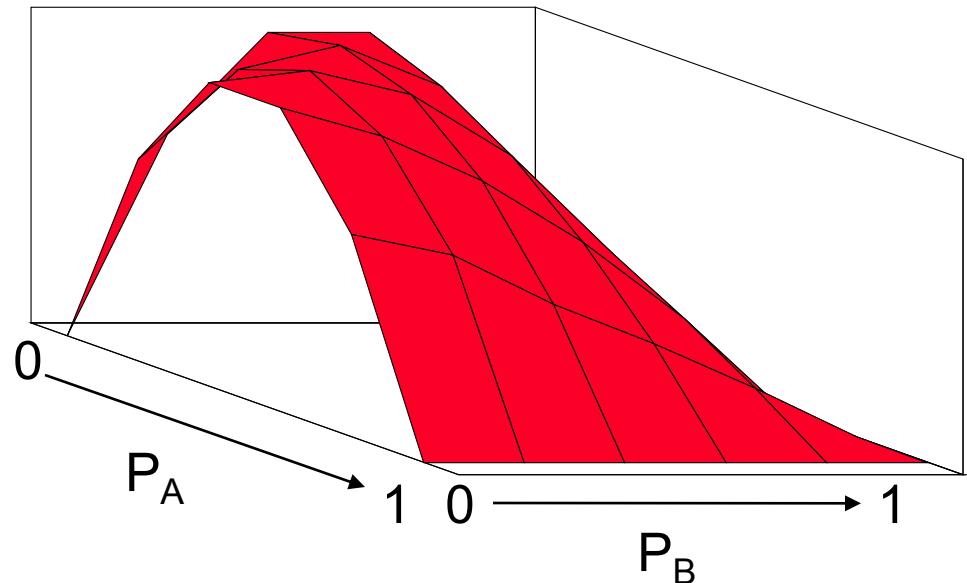
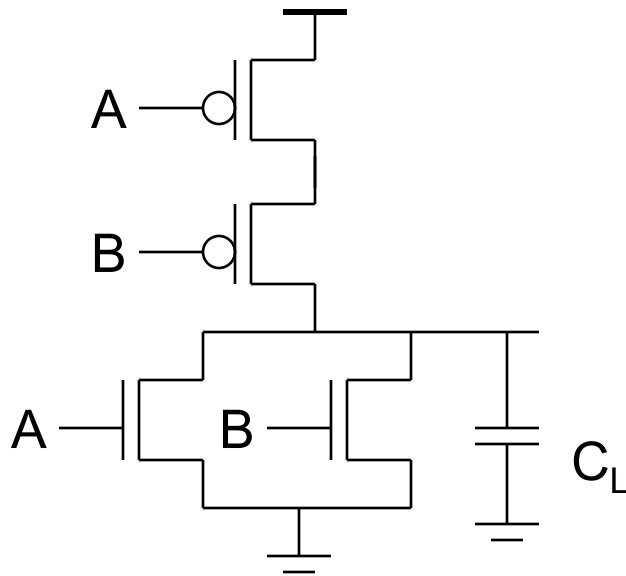
With input signal probabilities

$$\begin{aligned}P_{A=1} &= 1/2 \\P_{B=1} &= 1/2\end{aligned}$$

NOR static transition probability
 $= 3/4 \times 1/4 = 3/16$

NOR Gate Transition Probabilities

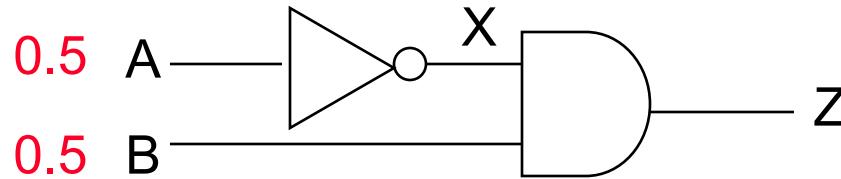
- Switching activity is a strong function of the input signal statistics
 - P_A and P_B are the probabilities that inputs A and B are one



$$P_{0 \rightarrow 1} = P_0 \times P_1 = (1 - (1 - P_A)(1 - P_B)) (1 - P_A)(1 - P_B)$$

Transition Probabilities for Some Basic Gates

	$P_{0 \rightarrow 1} = P_{\text{out}=0} \times P_{\text{out}=1}$
NOR	$(1 - (1 - P_A)(1 - P_B)) \times (1 - P_A)(1 - P_B)$
OR	$(1 - P_A)(1 - P_B) \times (1 - (1 - P_A)(1 - P_B))$
NAND	$P_A P_B \times (1 - P_A P_B)$
AND	$(1 - P_A P_B) \times P_A P_B$
XOR	$(1 - (P_A + P_B - 2P_A P_B)) \times (P_A + P_B - 2P_A P_B)$



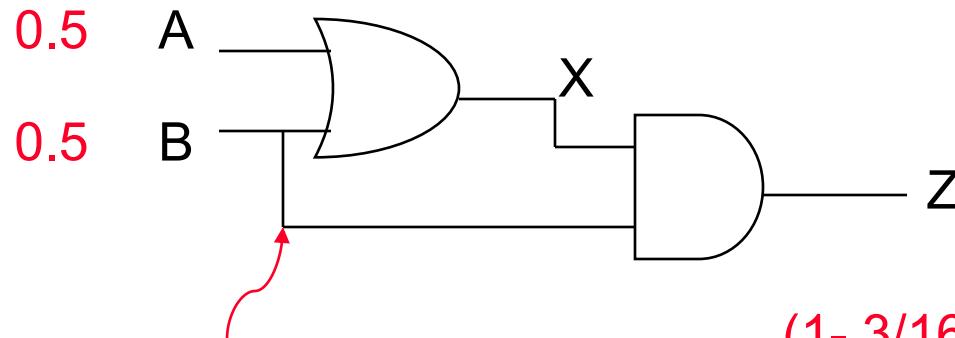
$$\begin{aligned}\text{For } X: P_{0 \rightarrow 1} &= P_0 \times P_1 = (1 - P_A) P_A \\ &= 0.5 \times 0.5 = 0.25\end{aligned}$$

$$\begin{aligned}\text{For } Z: P_{0 \rightarrow 1} &= P_0 \times P_1 = (1 - P_X P_B) P_X P_B \\ &= (1 - (0.5 \times 0.5)) \times (0.5 \times 0.5) = 3/16\end{aligned}$$

Inter-signal Correlations

- Determining switching activity is complicated by the fact that signals exhibit correlation in space and time
 - reconvergent fan-out

$$(1-0.5)(1-0.5) \times (1-(1-0.5)(1-0.5)) = 3/16$$



Reconvergent

$$(1 - 3/16 \times 0.5) \times (3/16 \times 0.5) = 0.085$$

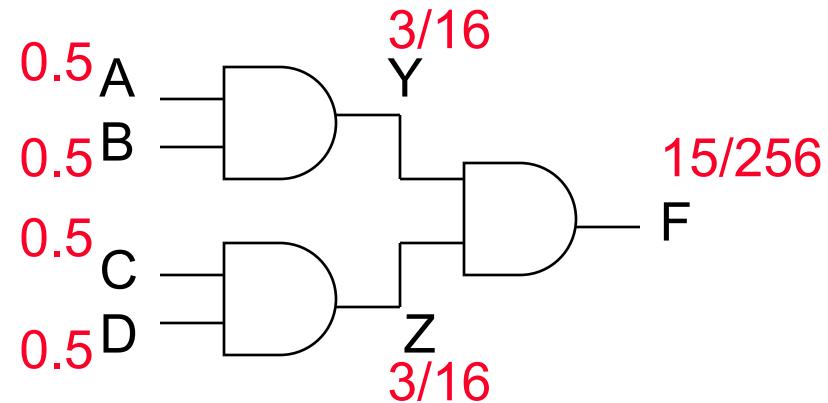
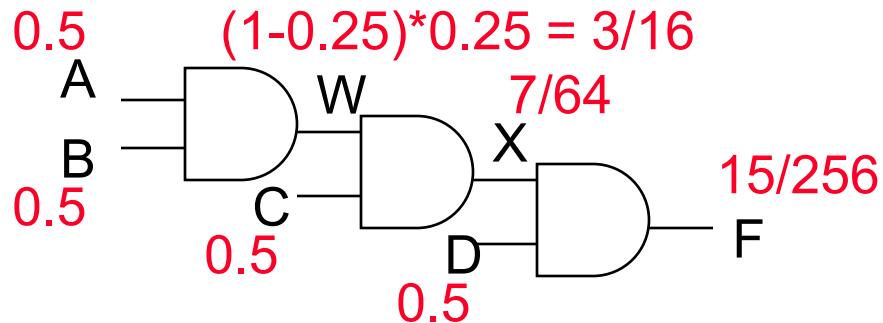
$$P(Z=1) = P(B=1) \& P(A=1 | B=1)$$

- Have to use conditional probabilities

Logic Restructuring

- Logic restructuring: changing the topology of a logic network to reduce transitions

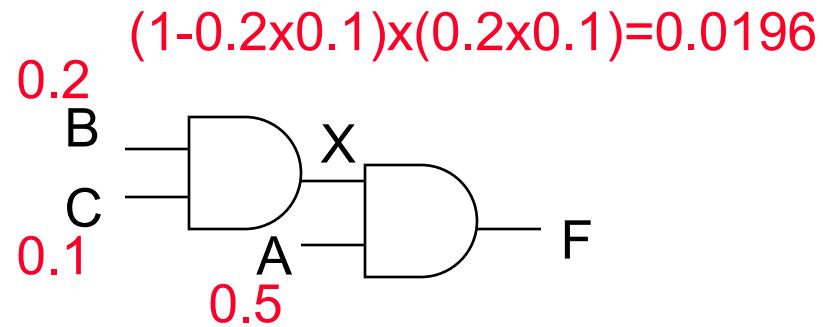
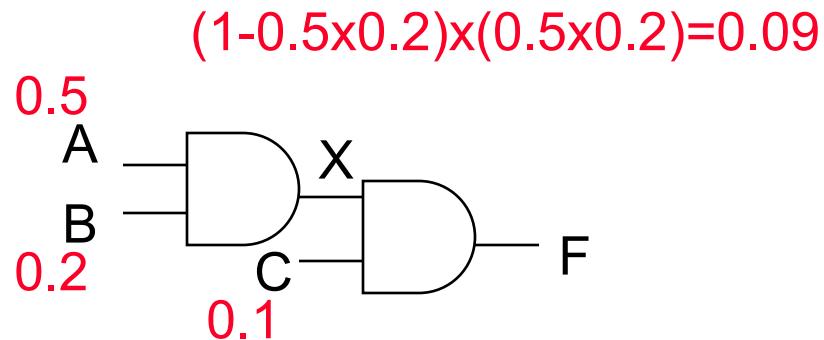
$$\text{AND: } P_{0 \rightarrow 1} = P_0 \times P_1 = (1 - P_A P_B) \times P_A P_B$$



Chain implementation has a lower overall switching activity than the tree implementation for random inputs

Ignores glitching effects

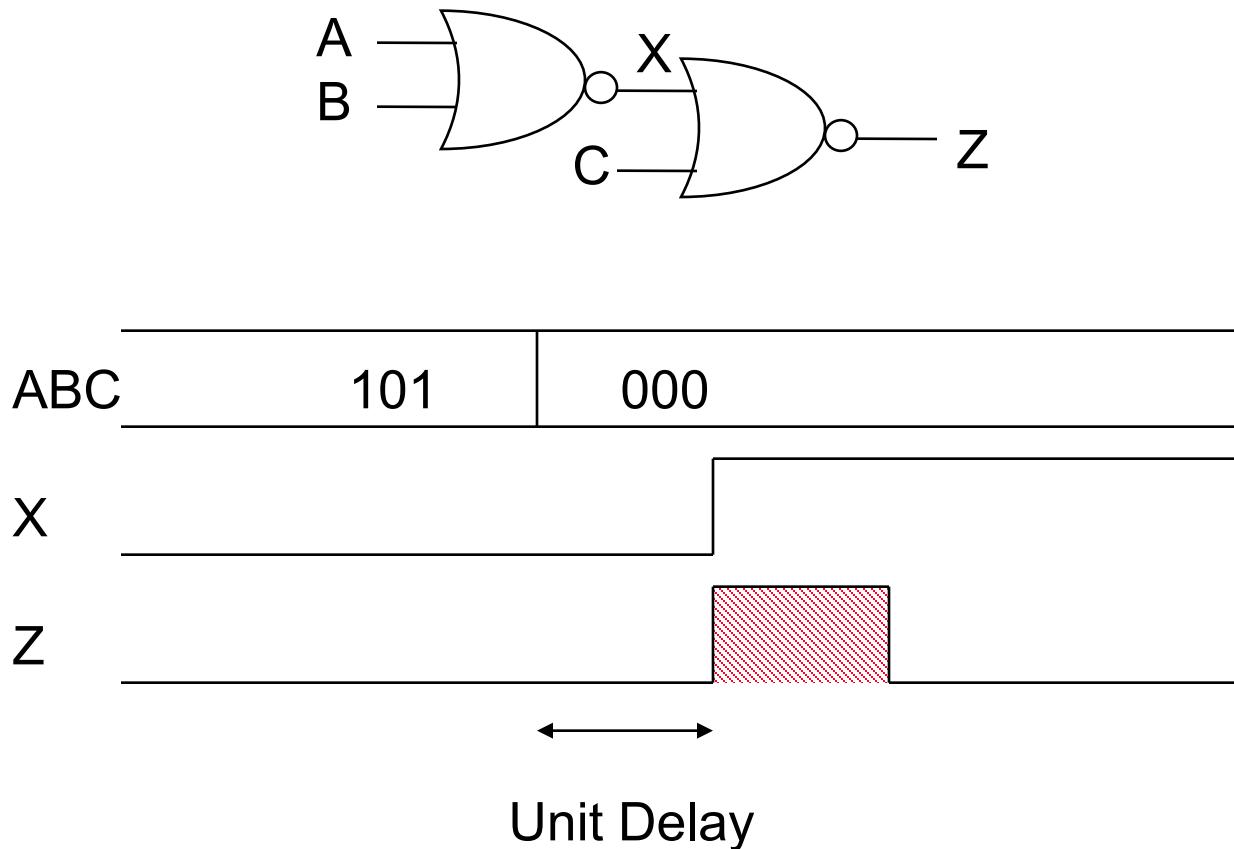
Input Ordering



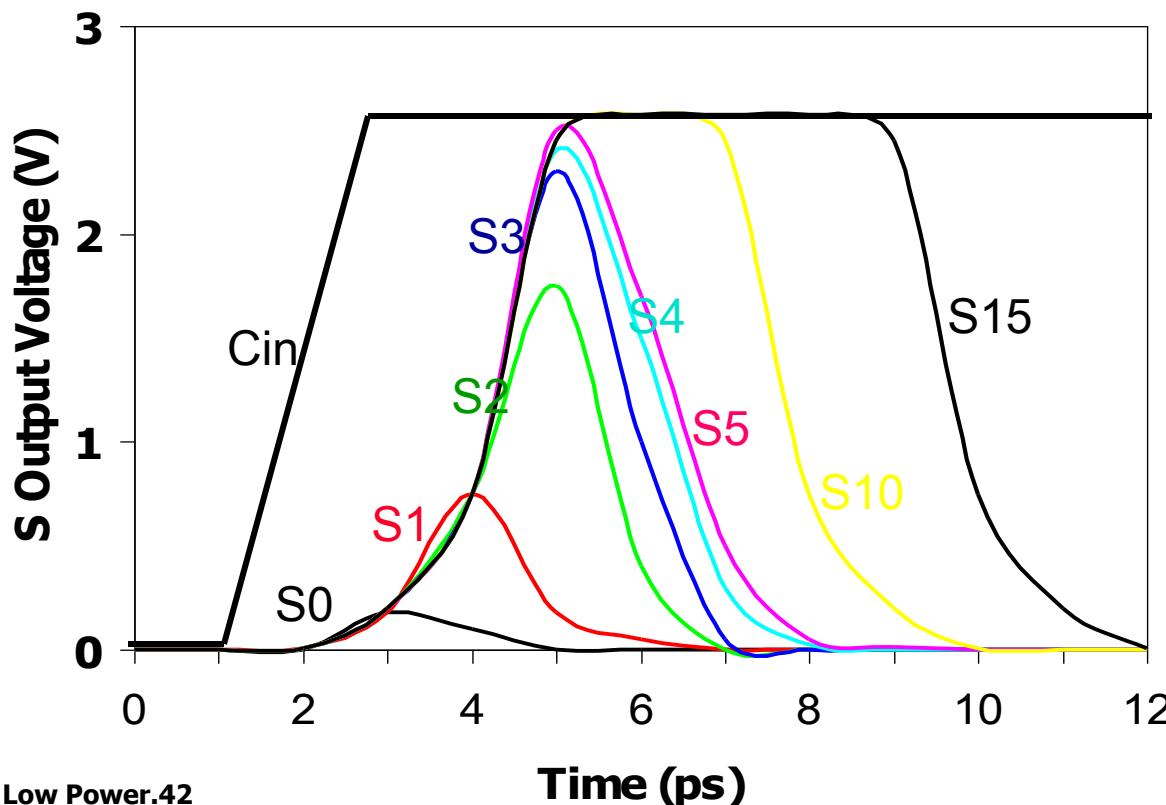
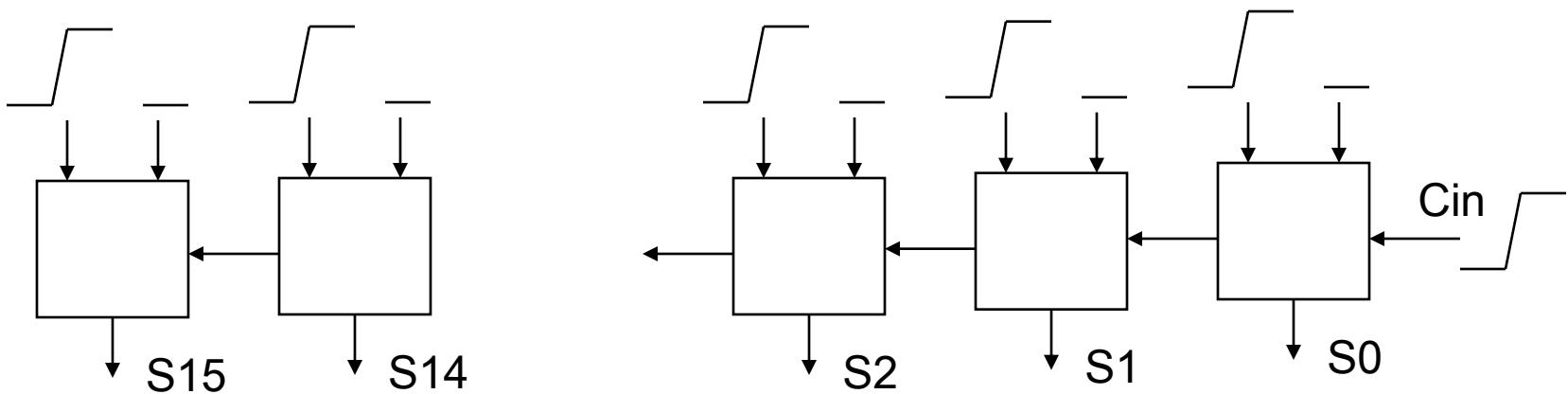
Beneficial to postpone the introduction of signals with a **high** transition rate (signals with signal probability close to 0.5)

Glitching in Static CMOS Networks

- ❑ Gates have a nonzero propagation delay resulting in spurious transitions or **glitches** (dynamic hazards)
 - glitch: node exhibits multiple transitions in a single cycle before settling to the correct logic value

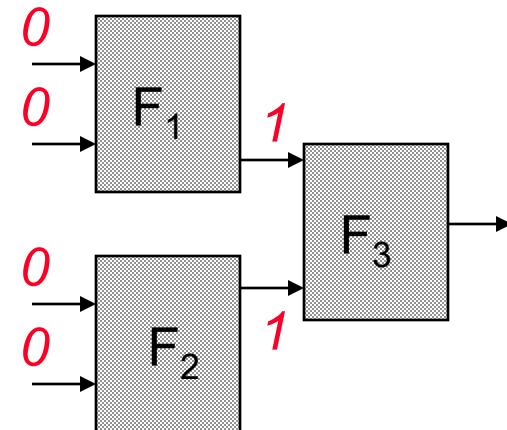
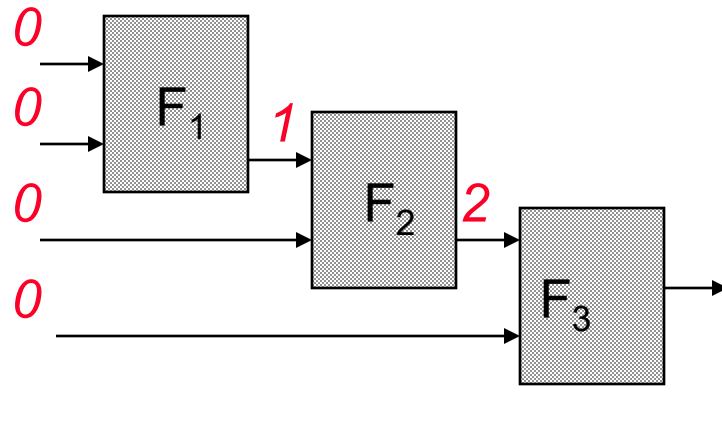


Glitching in an RCA



Balanced Delay Paths to Reduce Glitching

- Glitching is due to a mismatch in the path lengths in the logic network; if all input signals of a gate change simultaneously, no glitching occurs



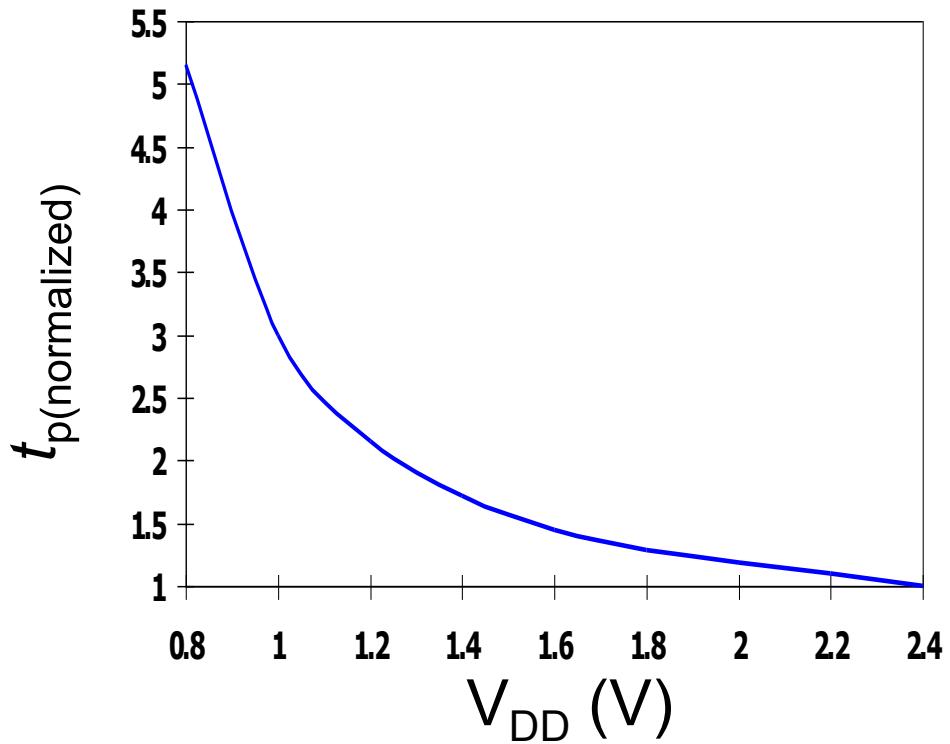
So equalize the lengths of timing paths through logic

Power and Energy Design Space

	Constant Throughput/Latency	Variable Throughput/Latency	
Energy	Design Time	Non-active Modules	Run Time
Active	Logic Design Reduced V_{dd} Sizing Multi-V_{dd}	Clock Gating	DFS, DVS (Dynamic Freq, Voltage Scaling)
Leakage	+ Multi- V_T	Sleep Transistors Multi- V_{dd} Variable V_T	+ Variable V_T

Dynamic Power as a Function of V_{DD}

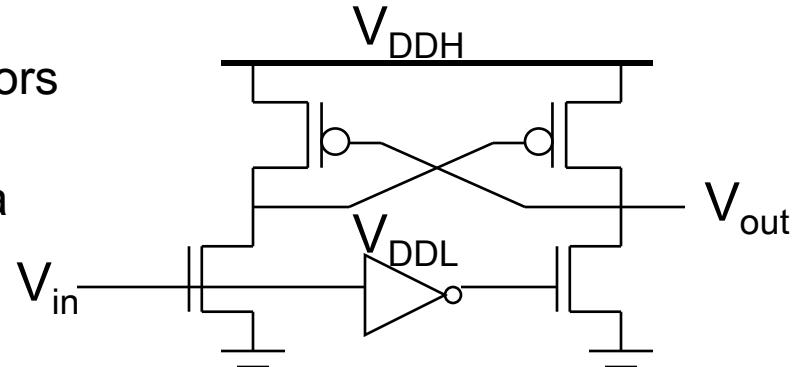
- Decreasing the V_{DD} **decreases** dynamic energy consumption (quadratically)
- But, **increases** gate delay (decreases performance)



- Determine the critical path(s) at **design time** and use high V_{DD} for the transistors on those paths for speed. Use a lower V_{DD} on the other gates, especially those that drive large capacitances (as this yields the largest energy benefits).

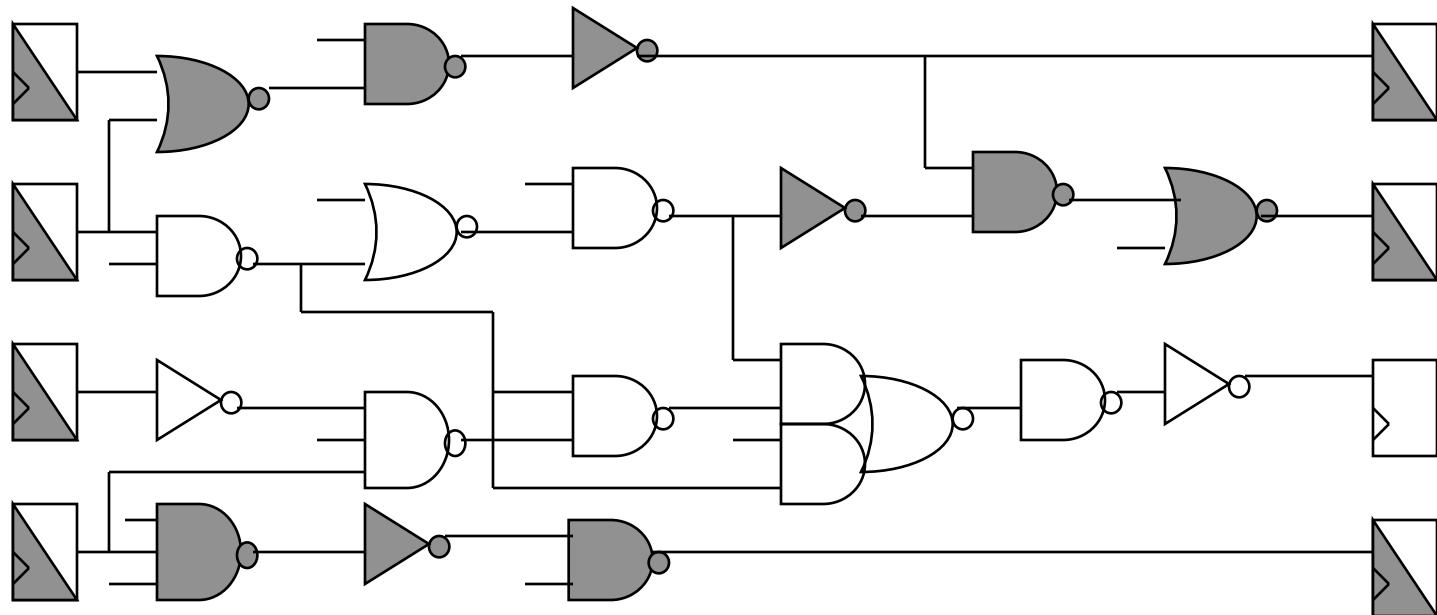
Multiple V_{DD} Considerations

- ❑ How many V_{DD} ? – Two is becoming common
 - ❑ Many chips already have two supplies (one for core and one for I/O)
- ❑ When combining multiple supplies, **level converters** are required whenever a module at the lower supply drives a gate at the higher supply (step-up)
 - ❑ If a gate supplied with V_{DDL} drives a gate at V_{DDH} , the PMOS never turns off
 - The cross-coupled PMOS transistors do the level conversion
 - The NMOS transistor operate on a reduced supply
 - ❑ Level converters are not needed for a step-down change in voltage
 - ❑ Overhead of level converters can be mitigated by doing conversions at register boundaries and embedding the level conversion inside the flipflop (see Figure 11.47)



Dual-Supply Inside a Logic Block

- ❑ Minimum energy consumption is achieved if **all** logic paths are critical (have the same delay)
- ❑ Clustered voltage-scaling
 - ❑ Each path starts with V_{DDH} and switches to V_{DDL} (gray logic gates) when delay **slack** is available
 - ❑ Level conversion is done in the flipflops at the end of the paths



Power and Energy Design Space

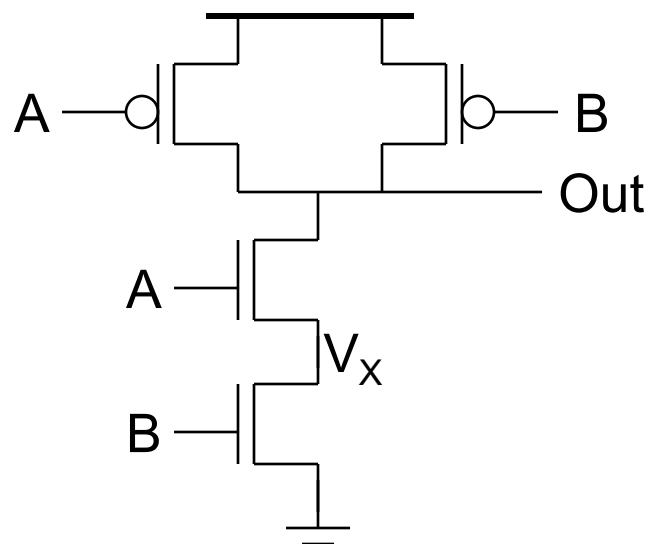
	Constant Throughput/Latency	Variable Throughput/Latency	
Energy	Design Time	Non-active Modules	
Active	Logic Design Reduced V_{dd} Sizing Multi- V_{dd}	Clock Gating	DFS, DVS (Dynamic Freq, Voltage Scaling)
Leakage	+ Multi- V_T	Sleep Transistors Multi- V_{dd} Variable V_T	+ Variable V_T

Stack Effect

- Leakage is a function of the circuit topology and the value of the inputs

$$V_T = V_{T0} + \gamma(\sqrt{|-2\phi_F|} + V_{SB}) - \sqrt{|-2\phi_F|}$$

where V_{T0} is the threshold voltage at $V_{SB} = 0$; V_{SB} is the source-bulk (substrate) voltage; γ is the **body-effect coefficient**



A	B	V_x	I_{SUB}
0	0	$V_T \ln(1+n)$	$V_{GS}=V_{BS}=-V_x$
0	1	0	$V_{GS}=V_{BS}=0$
1	0	$V_{DD}-V_T$	$V_{GS}=V_{BS}=0$
1	1	0	$V_{SG}=V_{SB}=0$

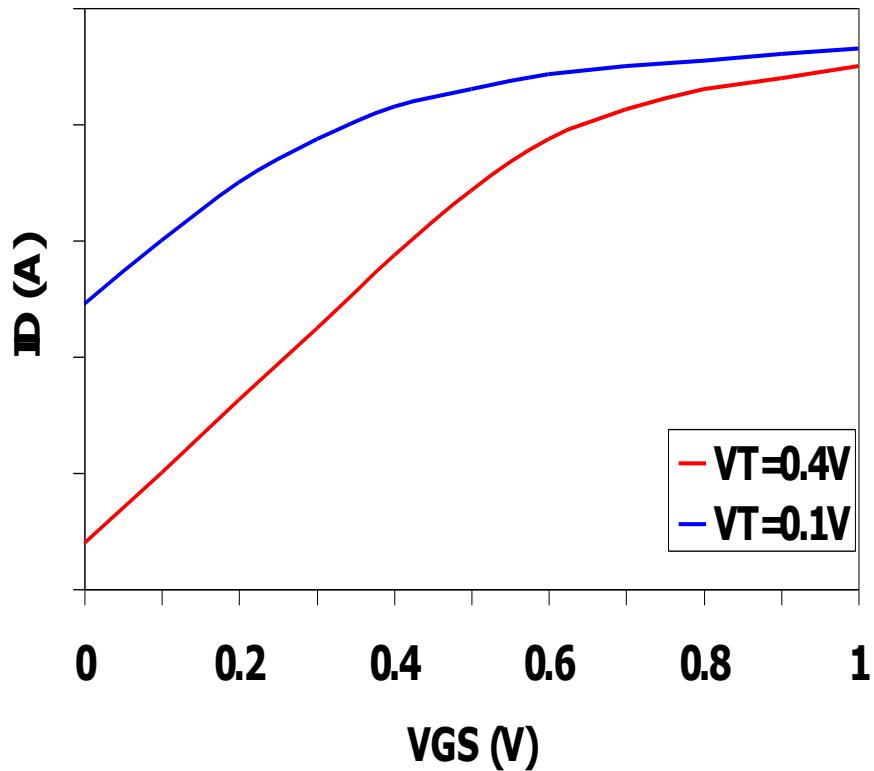
- Leakage is least when $A = B = 0$
- Leakage reduction due to stacked transistors is called the **stack effect**

Short Channel Factors and Stack Effect

- In short-channel devices, the subthreshold leakage current depends on V_{GS} , V_{BS} **and** V_{DS} . The V_T of a short-channel device decreases with increasing V_{DS} due to **DIBL** (drain-induced barrier loading).
 - Typical values for DIBL are 20 to 150mV change in V_T per voltage change in V_{DS} so the stack effect is even more significant for short-channel devices.
 - V_X reduces the drain-source voltage of the top nfet, increasing its V_T and lowering its leakage
- For our 0.25 micron technology, V_X settles to ~ 100 mV in steady state so $V_{BS} = -100$ mV and $V_{DS} = V_{DD} - 100$ mV which is 20 times smaller than the leakage of a device with $V_{BS} = 0$ mV and $V_{DS} = V_{DD}$

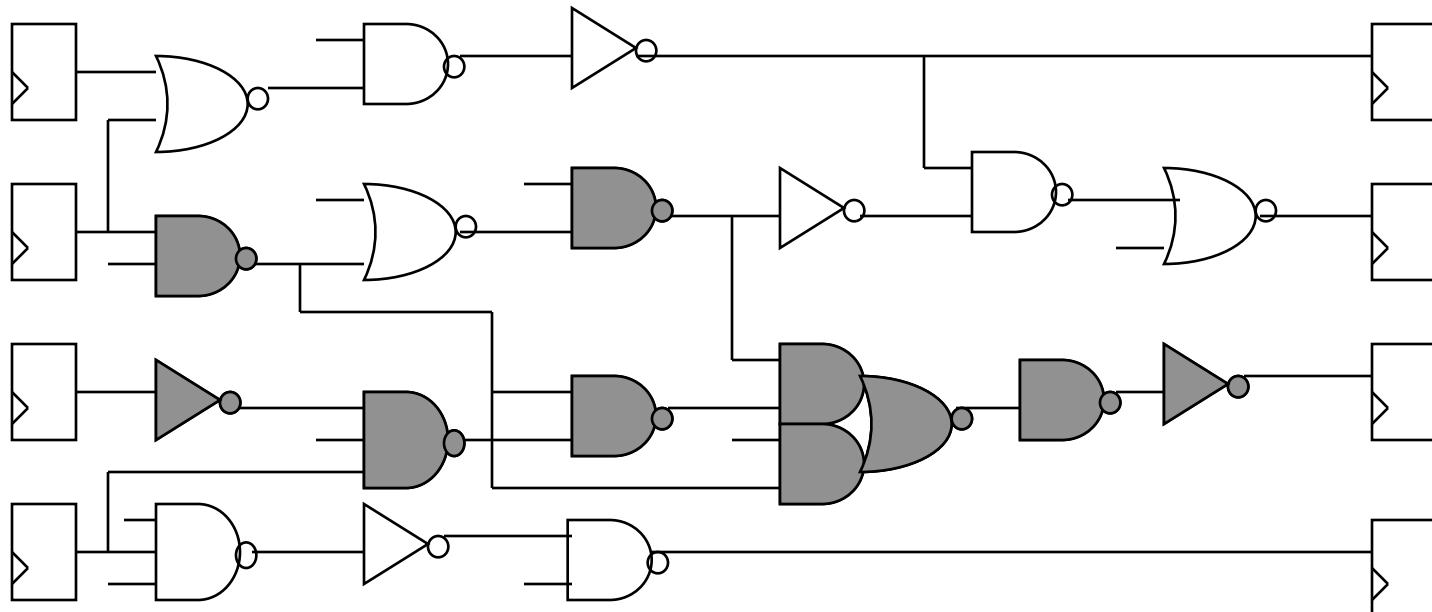
Leakage as a Function of Design Time V_T

- ❑ Reducing the V_T **increases** the sub-threshold leakage current (exponentially)
 - 90mV reduction in V_T increases leakage by an order of magnitude
- ❑ But, reducing V_T **decreases** gate delay (increases performance)
- ❑ Determine the critical path(s) at **design time** and use low V_T devices on the transistors on those paths for speed. Use a high V_T on the other logic for leakage control.
 - A careful assignment of V_T 's can reduce the leakage by as much as 80%



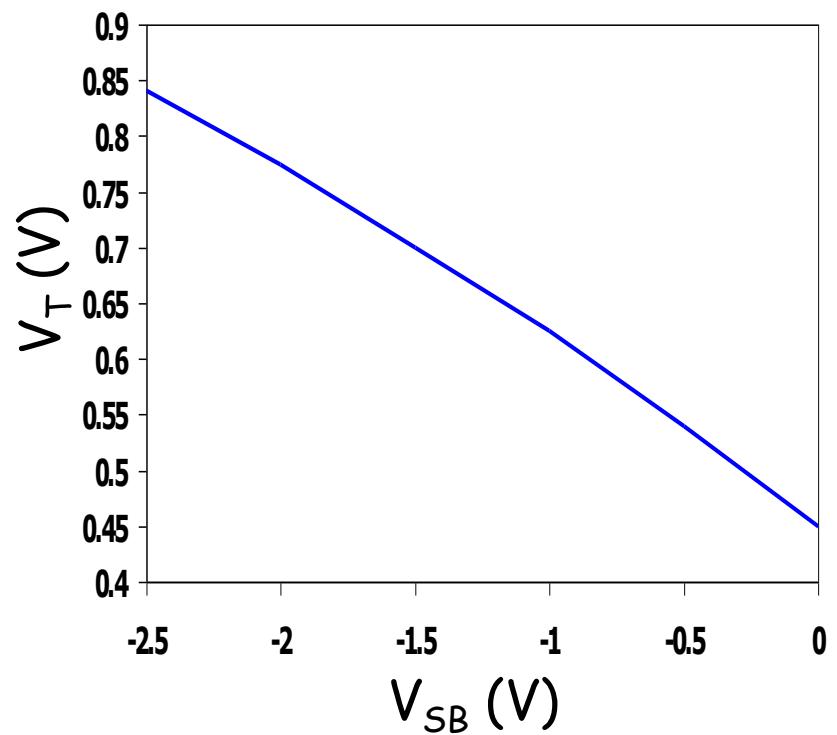
Dual-Thresholds Inside a Logic Block

- ❑ Minimum energy consumption is achieved if **all** logic paths are critical (have the same delay)
- ❑ Use lower threshold on timing-critical paths
 - ❑ Assignment can be done on a per gate or transistor basis; no clustering of the logic is needed
 - ❑ No level converters are needed



Variable V_T (ABB) at Run Time

- ❑ $V_T = V_{T0} + \gamma(\sqrt{|-2\phi_F|} + V_{SB}) - \sqrt{|-2\phi_F|}$
- ❑ For an n-channel device, the substrate is normally tied to ground ($V_{SB} = 0$)
- ❑ A negative bias on V_{SB} causes V_T to increase
- ❑ Adjusting the substrate bias at **run time** is called **adaptive body-biasing** (ABB)
 - ❑ Requires a dual well fab process



Next Lecture and

Reminders

❑ Next lecture (after midterm)

- ❑ Dynamic logic
 - Reading assignment – Rabaey, et al, 6.3

❑ Reminders

- ❑ HW3 due Oct 10th (hand in to TA)
- ❑ Class cancelled on Oct 10th as make up for evening midterm
- ❑ Class cancelled on Oct 15th due to fall break
- ❑ I will be out of town Oct 10th through Oct 15th and Oct 18th through Oct 23rd, so office hours during those periods are cancelled
- ❑ There will be a guest lecturer on Oct 22nd
- ❑ Evening midterm exam scheduled
 - Wednesday, October 16th from 8:15 to 10:15pm in 260 Willard
 - Only one midterm conflict filed for so far