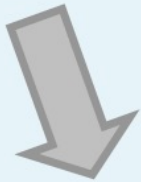


ECE1388 VLSI Design Methodology

Lecture 7: Design for Low Power

Problems of Power Dissipation



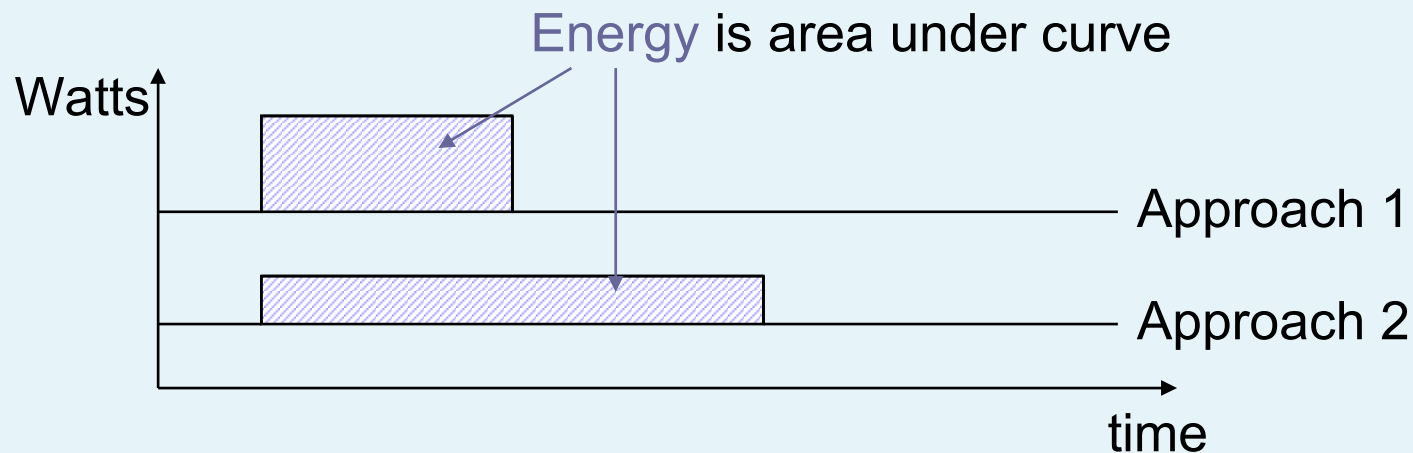
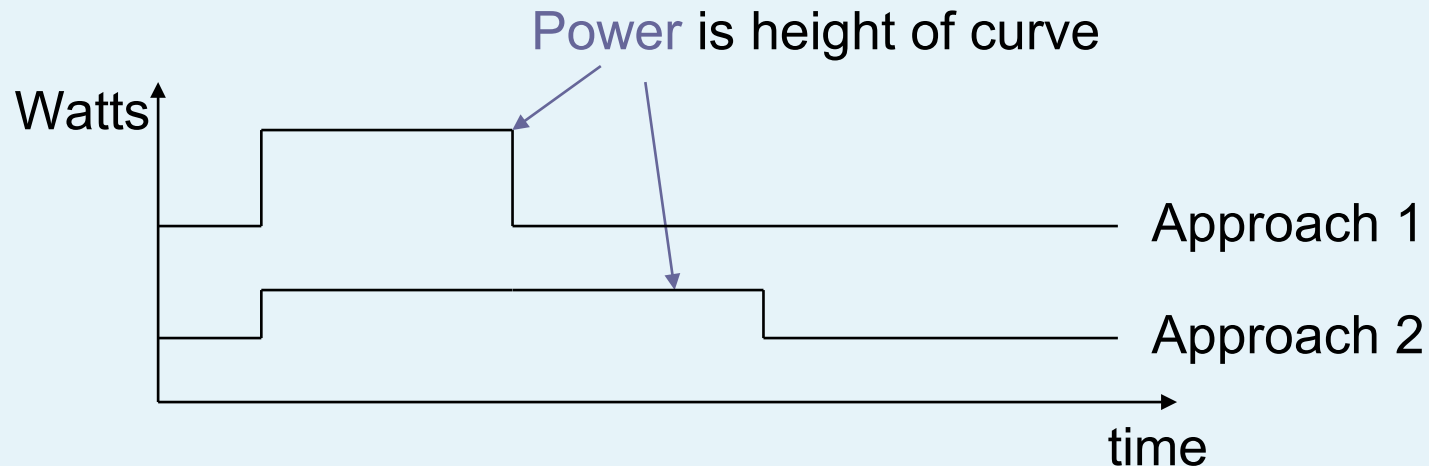
- Continuously increasing performance demands
- ➔ Increasing power dissipation of technical devices
- ➔ Today: power dissipation is a main problem
- High **Power dissipation** leads to:



- ☹ Reduced time of operation
- ☹ Higher weight (batteries)
- ☹ Reduced mobility

- ☹ High efforts for cooling
- ☹ Increasing operational costs
- ☹ Reduced reliability

Recap: Energy and Power



$$\text{Energy} = \text{Power} * \text{time for calculation} = \text{Power} * \text{Delay}^6$$

Power in Circuit Elements

$$P_{VDD}(t) = I_{DD}(t)V_{DD}$$



$$P_R(t) = \frac{V_R^2(t)}{R} = I_R^2(t)R$$



$$\begin{aligned} E_C &= \int_0^{\infty} I(t)V(t)dt = \int_0^{\infty} C \frac{dV}{dt} V(t)dt \\ &= C \int_0^{V_C} V(t)dV = \frac{1}{2} CV_C^2 \end{aligned}$$

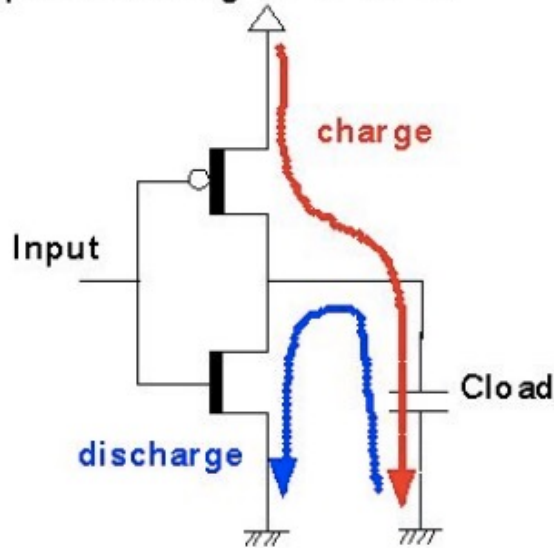


Types of Power Dissipation

Dynamic Power

($\approx 40 - 70\%$ today and decreasing relatively)

Input switching to '1' or '0'

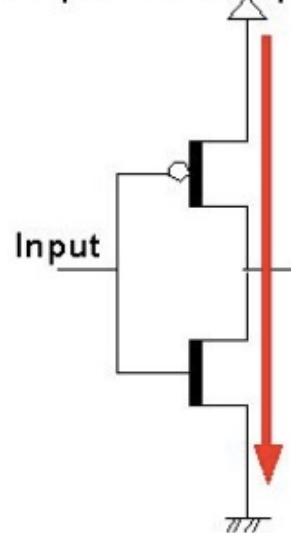


(a) Capacitive Current

Short-circuit power

($\approx 10\%$ today and decreasing absolutely)

$V_{th} < \text{Input} < V_{DD} - |V_{thp}|$

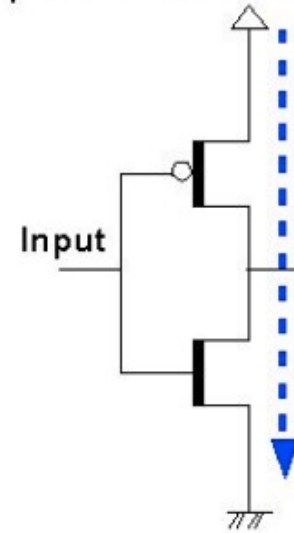


(b) Short Circuit Current

Static power

($\approx 20 - 50\%$ today and increasing)

Input : '1' or '0' steady state



(c) Static Leakage Current

Adapted from Low-power Digital Design lecture notes by Rana Ian Owen

Power Dissipation Sources

□ $P_{total} = P_{dynamic} + P_{short-circuit} + P_{static}$

□ Dynamic power:

– Switching load capacitances

$$P_{\text{switching}} = \alpha C V_{DD}^2 f$$

□ Short-circuit current (Vdd-to-GND on input edges)

□ Static power:

$$P_{static} = (I_{sub} + I_{gate} + I_{junct} + I_{contention}) V_{DD}$$

– Subthreshold leakage

– Gate leakage

– Junction leakage

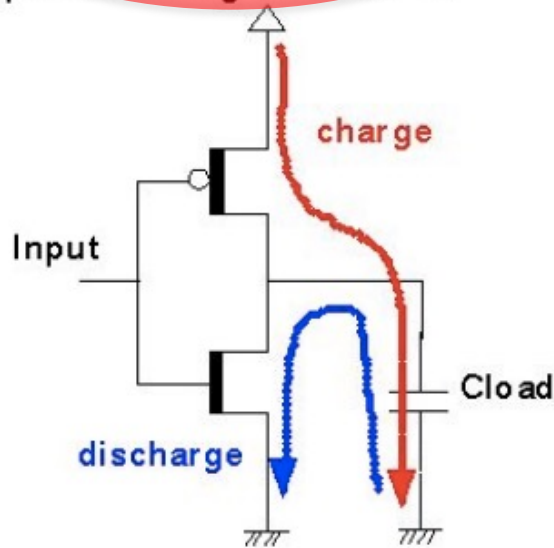
– Contention current

Types of Power Dissipation

Dynamic Power

($\approx 40 - 70\%$ today and decreasing relatively)

Input switching to '1' or '0'

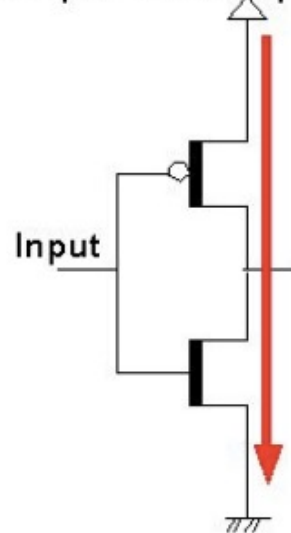


(a) Capacitive Current

Short-circuit power

($\approx 10\%$ today and decreasing absolutely)

$V_{th} < \text{Input} < V_{DD} - |V_{thp}|$

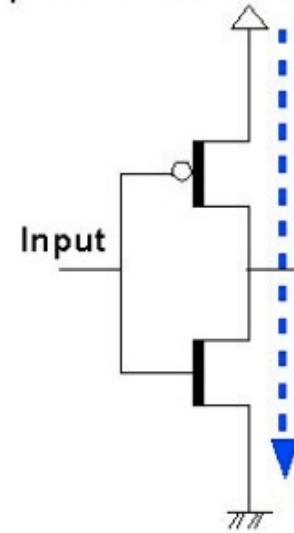


(b) Short Circuit Current

Static power

($\approx 20 - 50\%$ today and increasing)

Input : '1' or '0' steady state

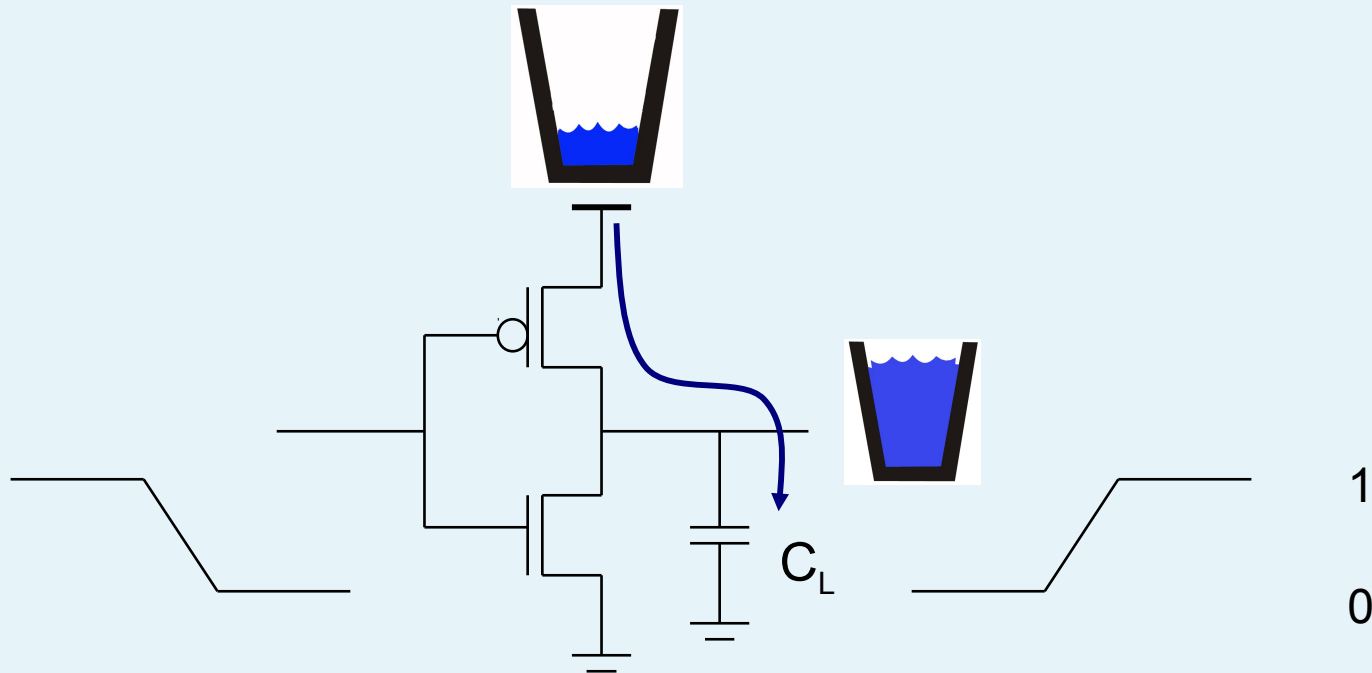


(c) Static Leakage Current

Adapted from Low-power Digital Design lecture notes by Rana Ian Owen

Dynamic Power Dissipation

- Voltage (Volt, V)
 - Current (Ampere, A)
 - Energy
- ↔
- Water pressure (bar)
Water quantity per second (liter/s)
Amount of Water



Energy consumption is proportional to capacitive load!⁴

Charging a Capacitor

- When the gate output rises
 - *Recall*: Energy stored in capacitor is

$$E_C = \frac{1}{2} C_L V_{DD}^2$$

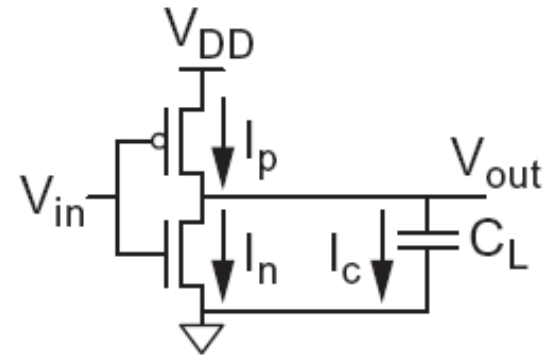
- But energy drawn from the supply is

$$E_{VDD} = \int_0^{\infty} I(t) V_{DD} dt = \int_0^{\infty} C_L \frac{dV}{dt} V_{DD} dt$$

$$= C_L V_{DD} \int_0^{V_{DD}} dV = C_L V_{DD}^2$$

- Half the energy from V_{DD} is dissipated in the pMOS transistor as heat, the other half stored in capacitor

- When the gate output falls
 - Energy in capacitor is dumped to GND
 - Dissipated as heat in the nMOS transistor



Dynamic Power

- ❑ Energy spent per cycle is $C_L V_{DD}^2$
- ❑ Suppose the system clock frequency = f
 - i.e., f cycles are performed per second
 - the switching power at frequency f is $CV_{DD}^2 f$
- ❑ Activity factor
 - Let $f_{sw} = \alpha f$, where α = activity factor
 - If the signal is a clock, $\alpha = 1$
 - If the signal switches once per cycle, $\alpha = 1/2$
- ❑ Dynamic power:

$$P_{\text{switching}} = \alpha C V_{DD}^2 f$$

Dynamic Power Reduction

□ $P_{\text{switching}} = \alpha C V_{DD}^2 f$

□ Try to minimize:

- Activity factor
- Capacitance
- **Supply voltage**
- Frequency

} Discussed next,
In that order

Activity Factor Estimation

- ❑ Let $P_i = \text{Prob}(\text{node } i = 1)$
 - $\bar{P}_i = 1 - P_i$ - probability node is 0
- ❑ $\alpha_i = \bar{P}_i * P_i$ - probability of a transition ($0 \rightarrow 1$ or $1 \rightarrow 0$)
- ❑ Completely random data has $P = 0.5$ and $\alpha = 0.25$

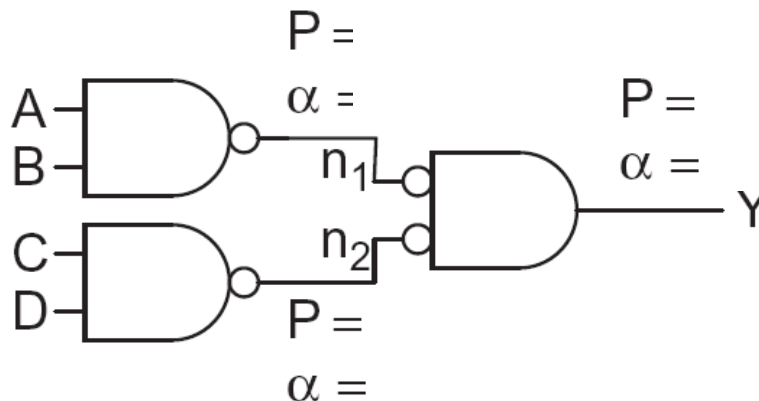
- ❑ Data is typically not completely random
- ❑ Data propagating through ANDs and ORs has lower activity factor
 - Depends on design, but typically $\alpha \approx 0.1$

Switching Probability

Gate	P_Y
AND2	$P_A P_B$
AND3	$P_A P_B P_C$
OR2	$1 - \bar{P}_A \bar{P}_B$
NAND2	$1 - P_A P_B$
NOR2	$\bar{P}_A \bar{P}_B$
XOR2	$P_A \bar{P}_B + \bar{P}_A P_B$

Example

- ❑ A 4-input AND is built out of two levels of gates
 - NAND ($1 - P_A P_B$) + NOR ($\bar{P}_A \bar{P}_B$)
- ❑ Estimate the activity factor at each node
 - if the inputs have $P = 0.5$ ($\alpha_i = \bar{P}_i * P_i$)

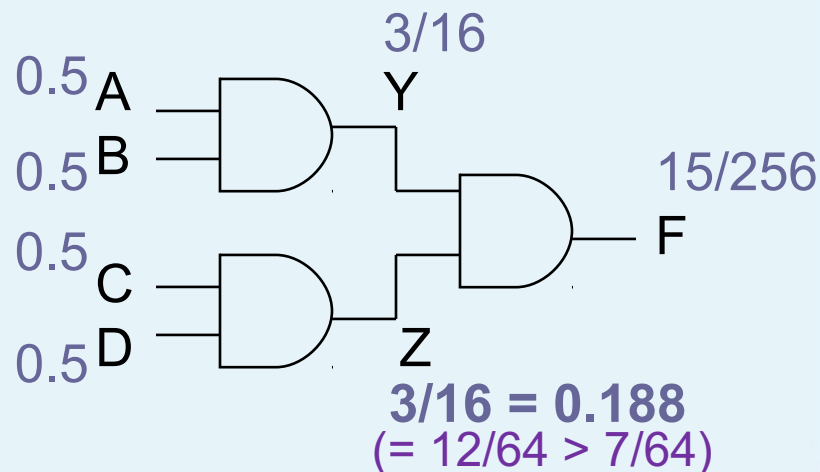
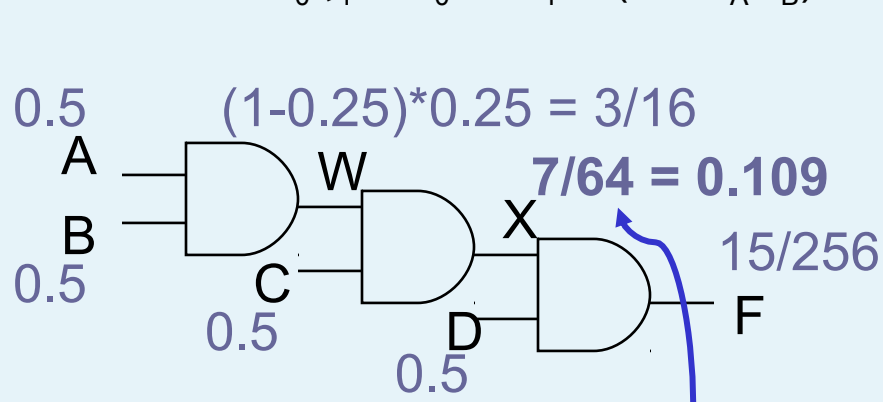


Logic Restructuring

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

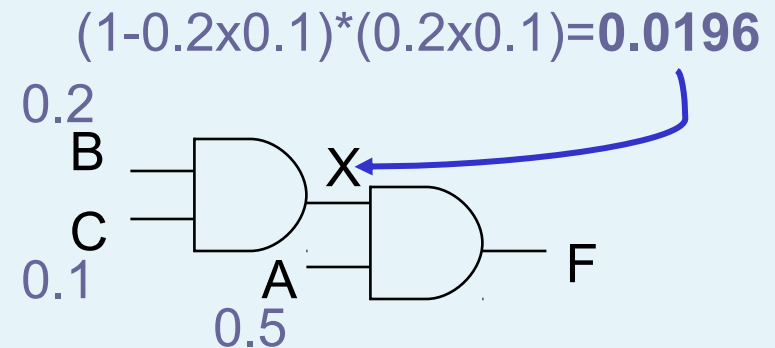
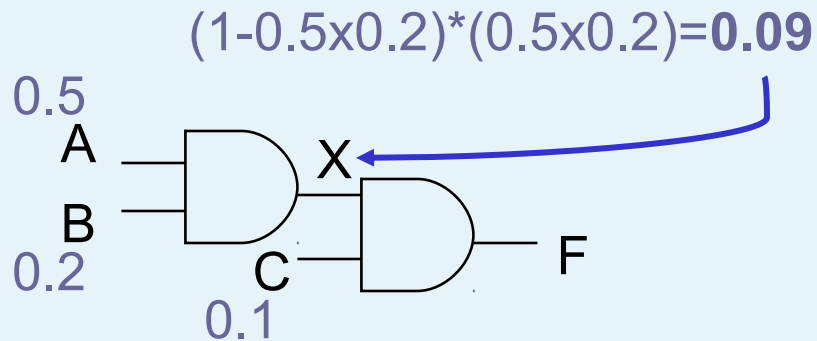
Gate	P_Y
AND2	$P_A P_B$

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



- Chain implementation has a lower overall switching activity than tree implementation for random inputs
- BUT: Ignores glitching effects (discussed later)

Input Ordering

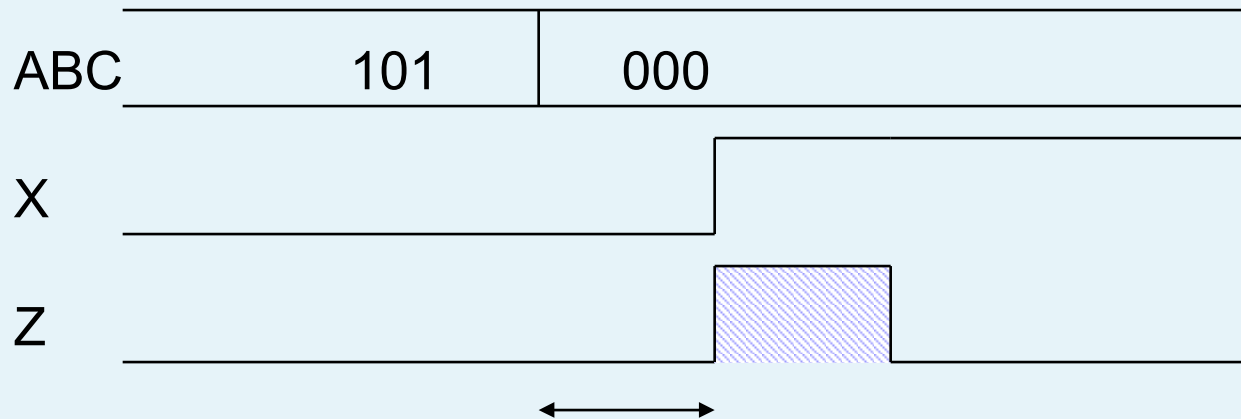
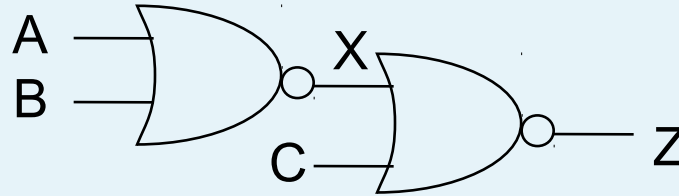


$$\text{AND: } P_{0 \rightarrow 1} = (1 - P_A P_B) * P_A P_B$$



Beneficial: postponing introduction of signals with a high transition rate (signals with signal probability close to 0.5)

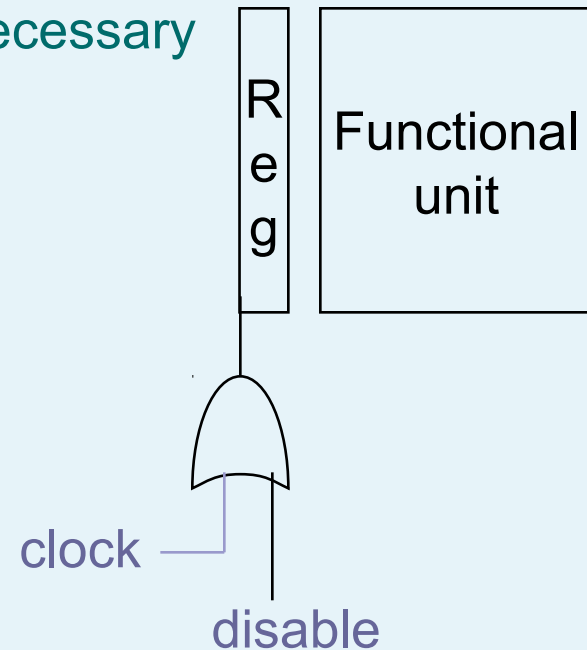
Glitching



Unit Delay

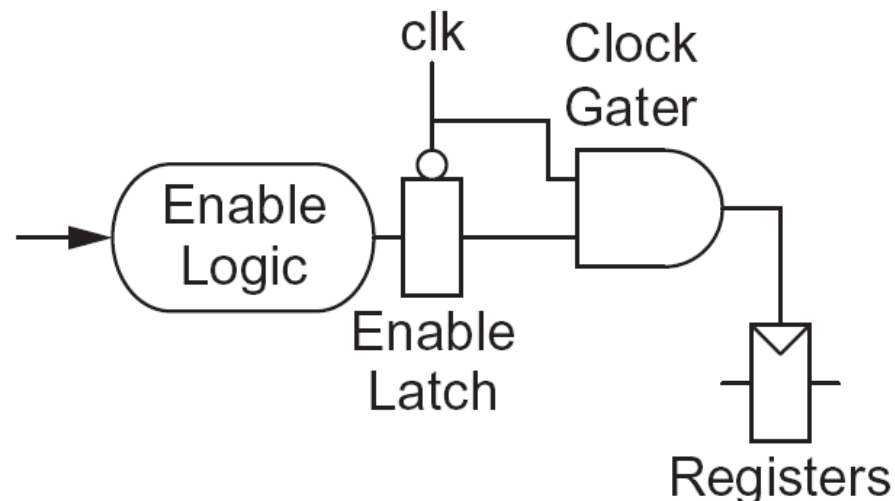
Clock Gating

- Most popular method for power reduction of clock signals and functional units
- Gate off clock to idle functional units
- Logic for generation of **disable** signal necessary
 - ➡ Higher complexity of control logic
 - ➡ Higher power consumption
 - ➡ Signals timing critical for avoiding of clock glitches at OR gate output
 - ➡ Additional gate delay on clock signal



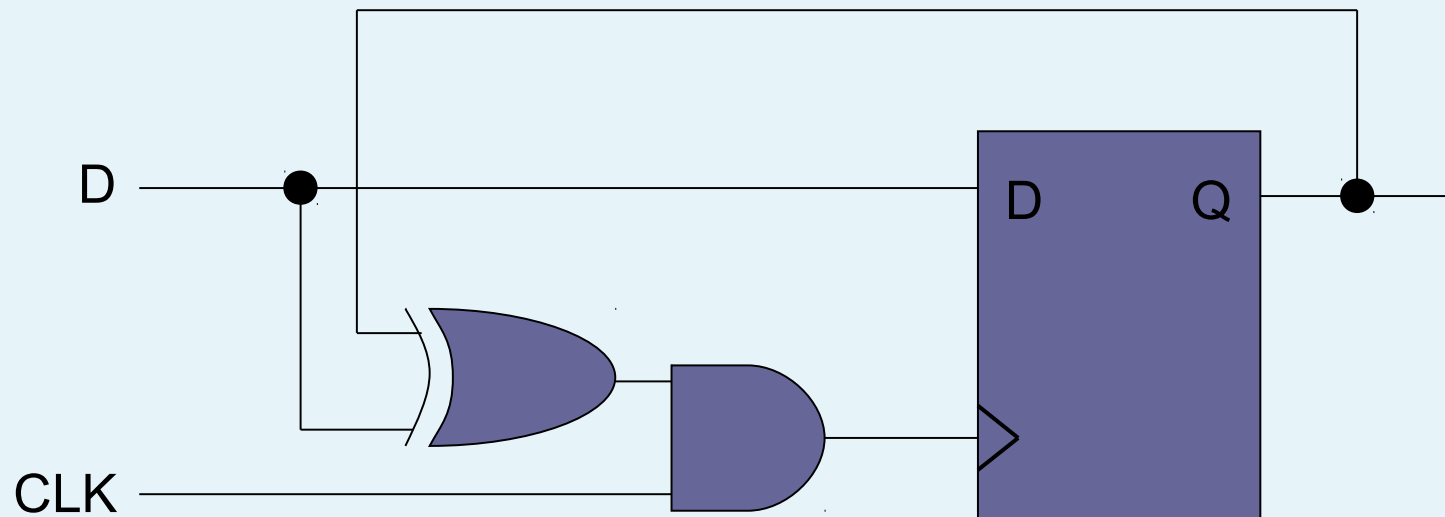
Clock Gating cont'd

- ❑ The best way to reduce the activity is to turn off the clock to registers in unused blocks
 - Saves clock activity ($\alpha = 1$)
 - Eliminates all switching activity in the block
 - Requires determining if block will be used



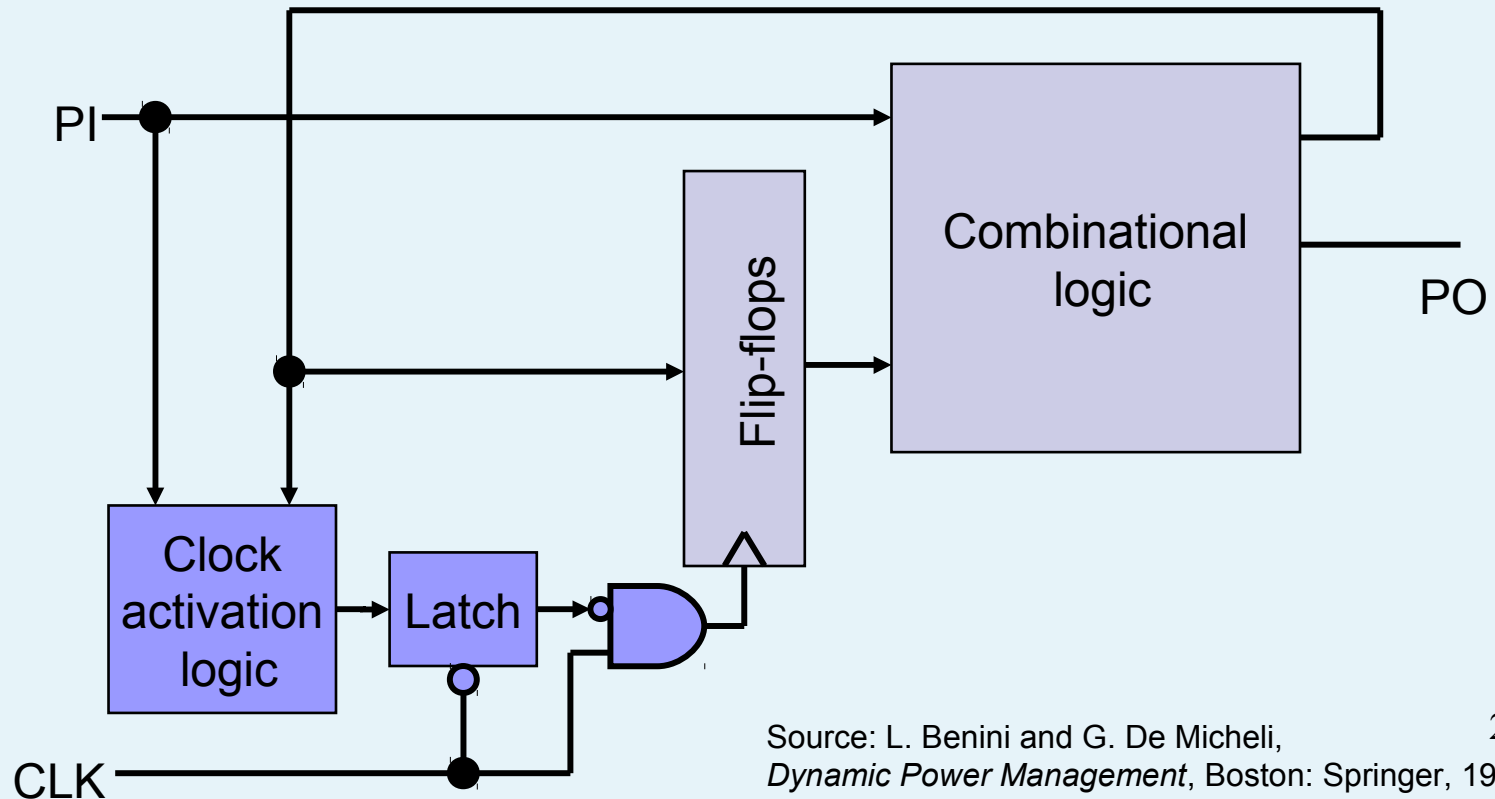
Clock Gating cont'd

- Clock-Gating in Low-Power Flip-Flop



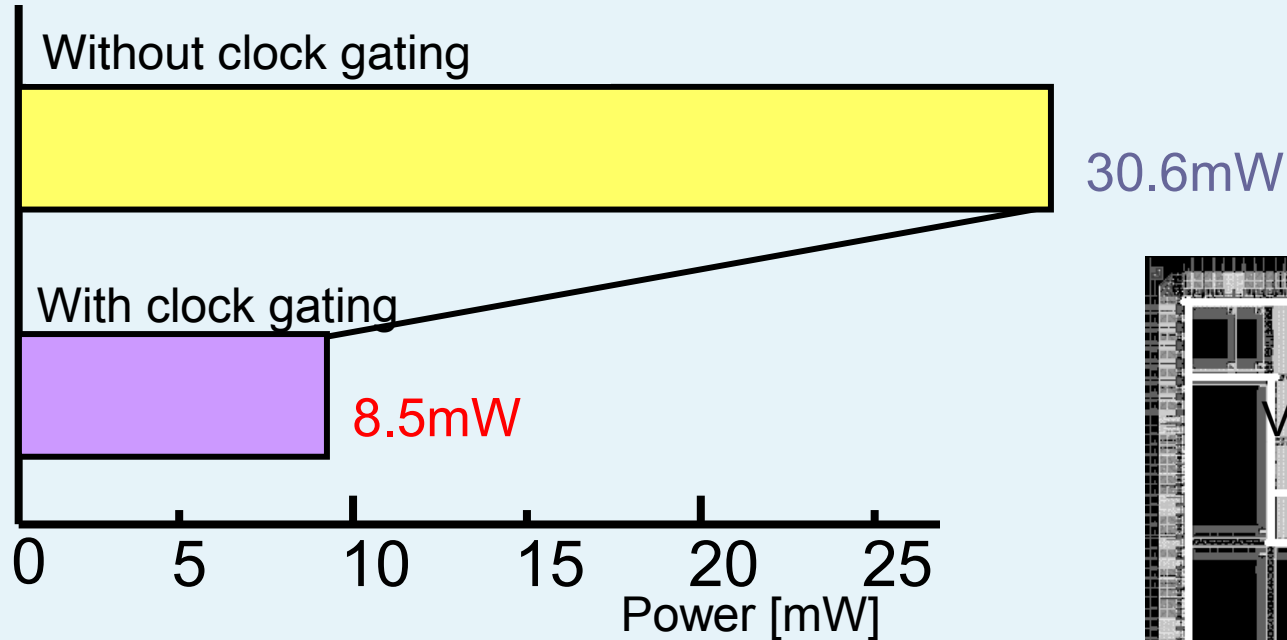
Clock Gating cont'd

- Clock gating over consideration of state in Finite-State-Machines (FSM)

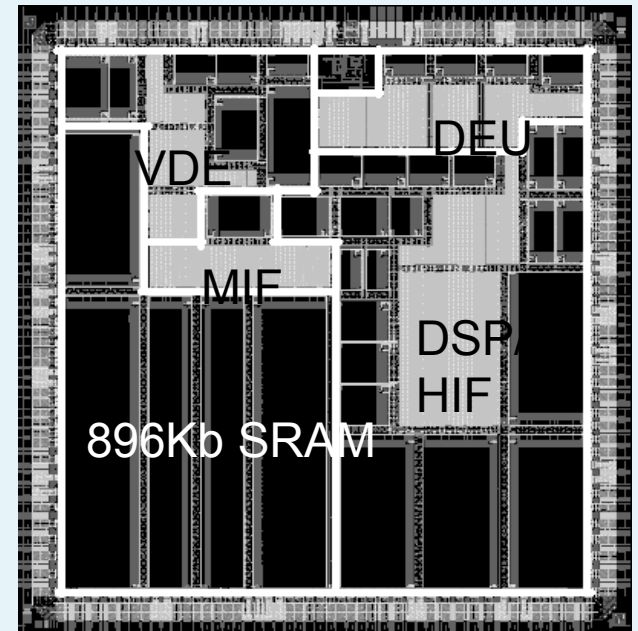


Source: L. Benini and G. De Micheli,
Dynamic Power Management, Boston: Springer, 1998.

Clock Gating: Example



- 90% of FlipFlops clock-gated
- 70% power reduction by clock-gating



MPEG4 decoder 24

Source: M. Ohashi, Matsushita, 2002

Adapted from Low-power Digital Design lecture notes by Rana Ian Owen

Capacitance

- ❑ Gate capacitance
 - Fewer stages of logic
 - Small gate sizes
- ❑ Wire capacitance
 - Good floorplanning to keep communicating blocks close to each other
 - Drive long wires with inverters or buffers rather than complex gates

Dynamic Power and Device Size

■ Device Sizing (= changing gate width)

- Affects input capacitance C_{in}
- Affects load capacitance C_{load}
- Affects dynamic power consumption P_{dyn}

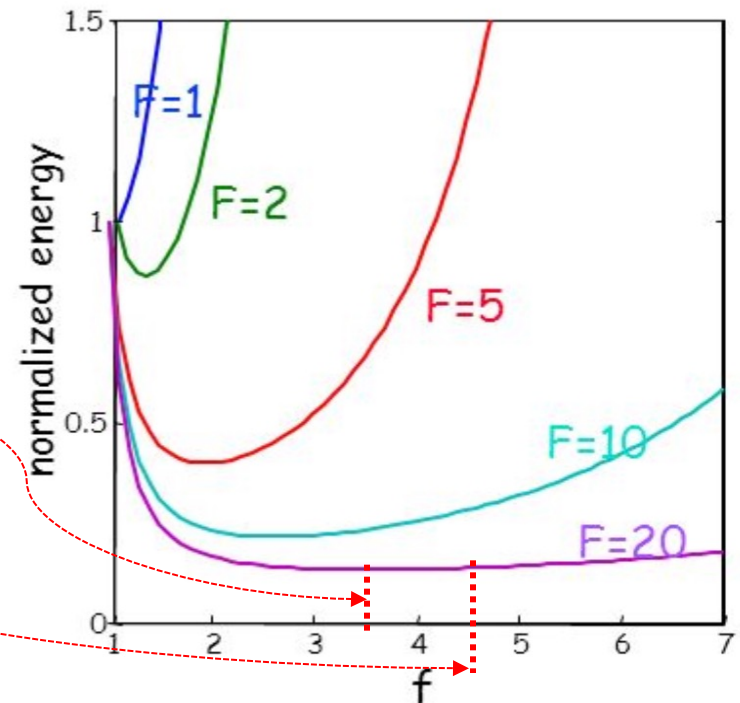
Recall: for best performance (i.e., delay), the best stage effort is: $\hat{f} = g_i h_i = F^{\frac{1}{N}}$

- 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
 $f_{opt}(\text{performance}) = 4.47 (= \sqrt{20})$

(Assuming two stages with minimum-size first stage)

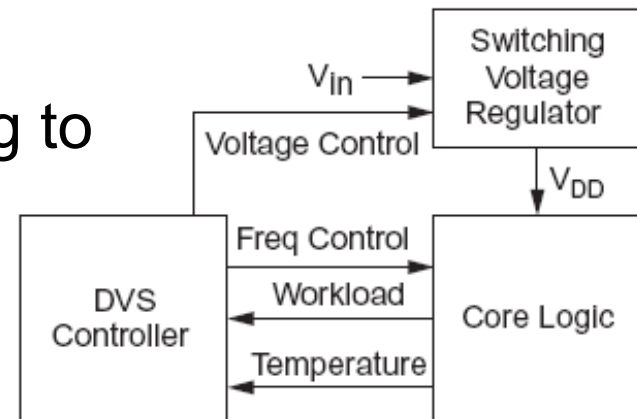
- If energy is a concern avoid oversizing beyond the optimal



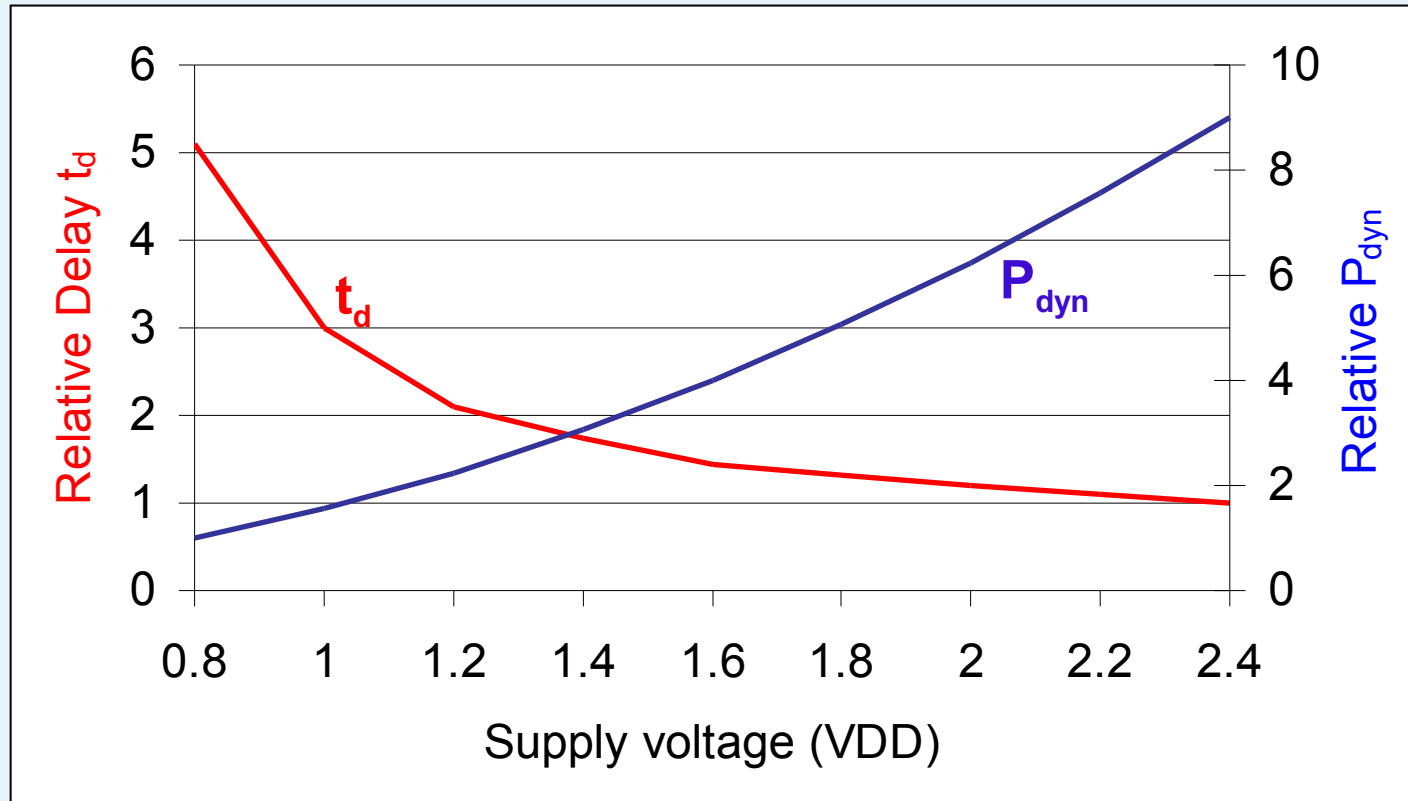
Adapted from Nikolic, UCB

Voltage / Frequency

- ❑ Run each block at the lowest possible voltage and frequency that meets performance requirements
- ❑ Voltage Domains
 - Provide separate supplies to different blocks
- ❑ Dynamic Voltage Scaling
 - Adjust V_{DD} and f according to workload



V_{DD} versus Delay and Power



Dynamic Power can be traded by delay

Multiple V_{DD}

- Main ideas:
 - Use of different supply voltages within the same design
 - High V_{DD} for critical parts (high performance needed)
 - Low V_{DD} for non-critical parts (only low performance demands)
- At design phase:
 - Determine critical path(s) (see upper next slide)
 - High V_{DD} for gates on those paths
 - Lower V_{DD} on the other gates (in non-critical paths)
 - For low V_{DD} : prefer gates that drive large capacitances (yields the largest energy benefits)
- Usually two different V_{DD} (but more are possible)

Multiple V_{DD} cont'd

- Level converters:

- Necessary, when module at lower supply drives gate at higher supply (step-up)

- If gate supplied with V_{DDL} drives a gate supplied with V_{DDH}

- ➔ then PMOS never turns off

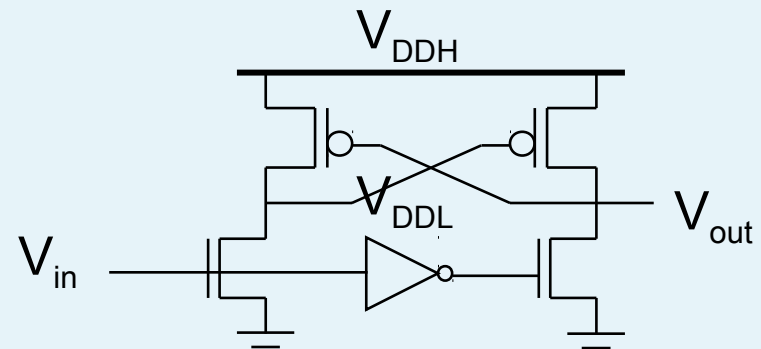
- Possible implementation:

- Cross-coupled PMOS transistors
 - NMOS transistor operate on reduced supply

- No need of level converters for step-down change in voltage

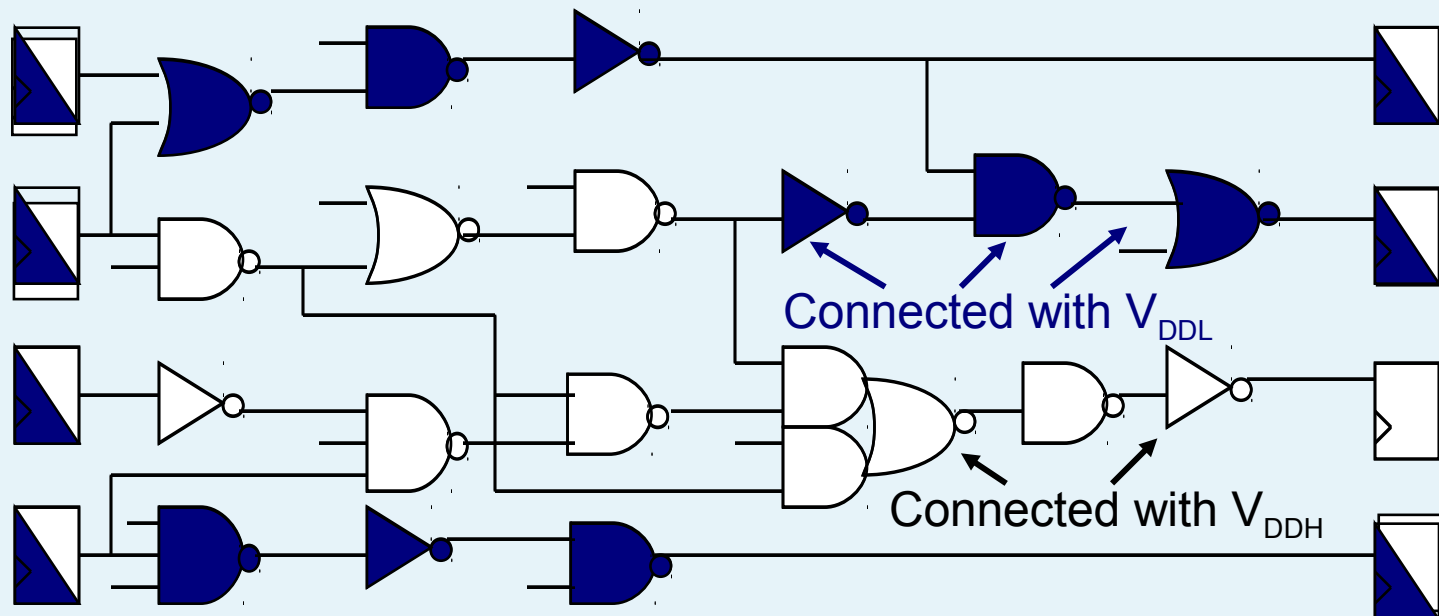
- Reducing of overhead:

- Conversions at register boundaries
 - Embedding of inside flipflop



Multiple V_{DD} in Data Paths

- Minimum energy consumption when **all** logic paths are critical (same delay)
- Possible Algorithm: clustered voltage-scaling
 - Each path starts with V_{DDH} and switches to V_{DDL} (blue gates) when **slack** is available
 - Level conversion in flipflops at end of paths

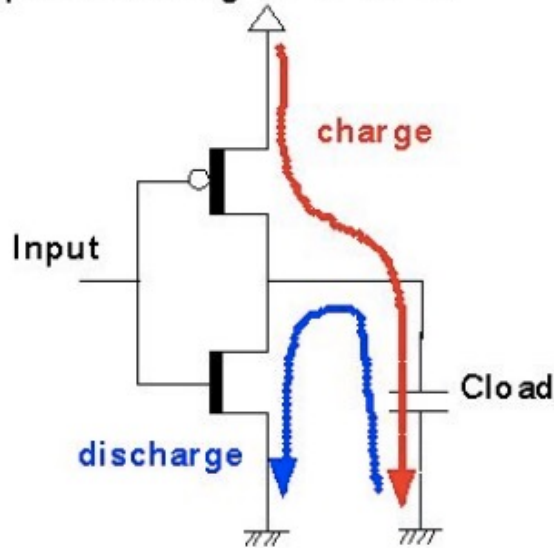


Types of Power Dissipation

Dynamic Power

($\approx 40 - 70\%$ today and decreasing relatively)

Input switching to '1' or '0'

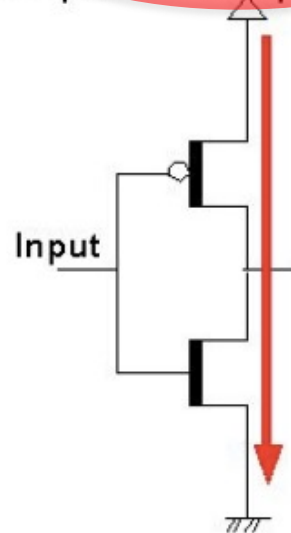


(a) Capacitive Current

Short-circuit power

($\approx 10\%$ today and decreasing absolutely)

$V_{DD} > \text{Input} > V_{DD} - |V_{thp}|$

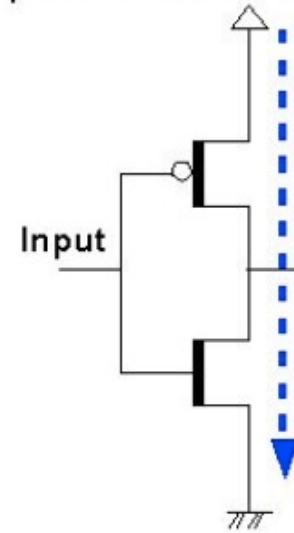


(b) Short Circuit Current

Static power

($\approx 20 - 50\%$ today and increasing)

Input : '1' or '0' steady state

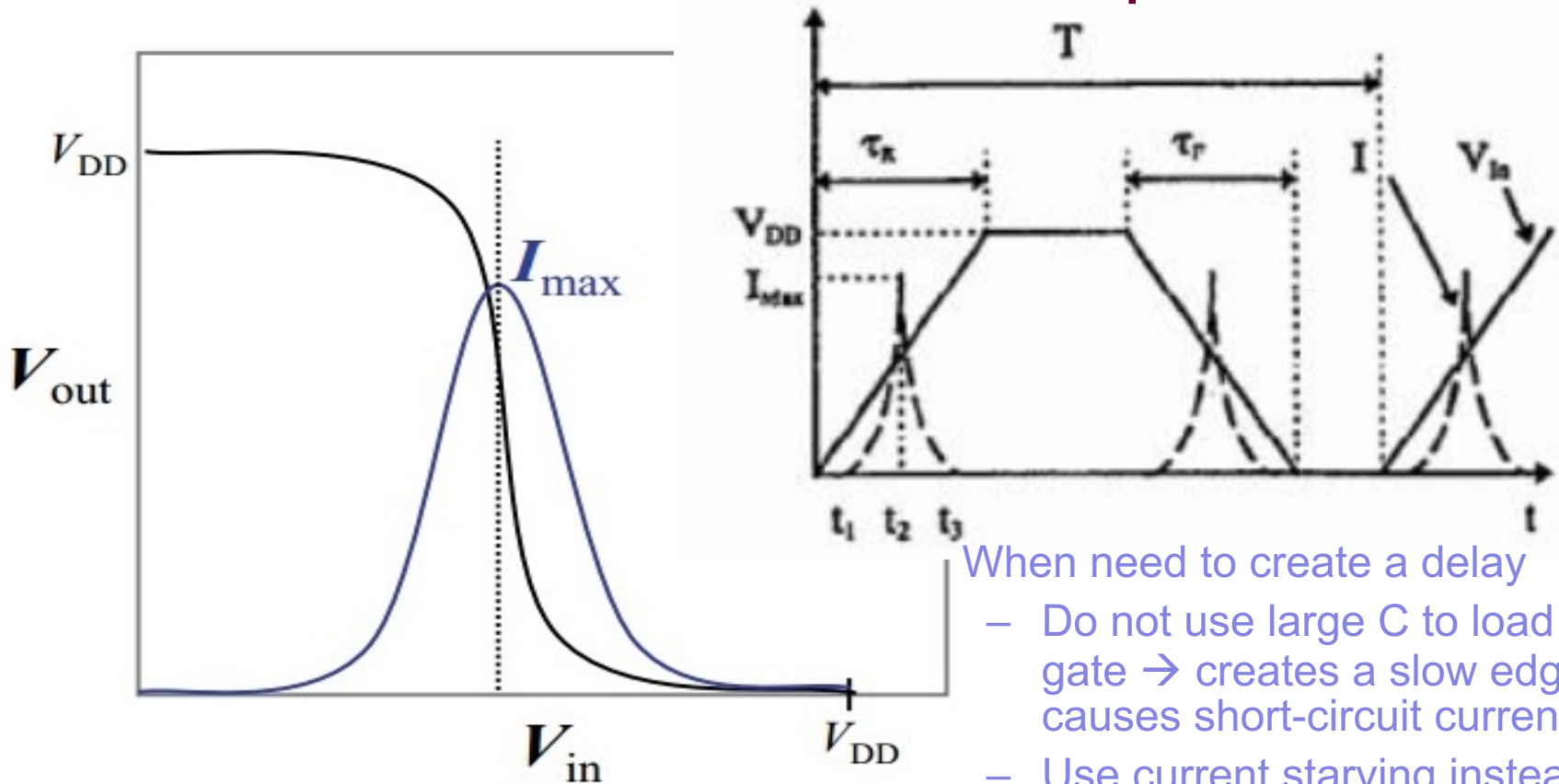


(c) Static Leakage Current

Short Circuit Current

- ❑ When transistors switch, both nMOS and pMOS networks may be momentarily ON at once
- ❑ Leads to a blip of “short circuit” current.
- ❑ $< 10\%$ of dynamic power if rise/fall times are comparable for input and output

Short-circuit Power Dissipation



- When need to create a delay
- Do not use large C to load a gate \rightarrow creates a slow edge, causes short-circuit current
 - Use current starving instead

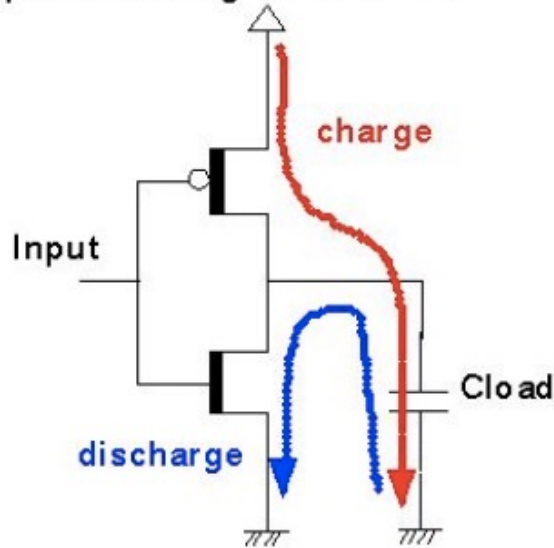
- Note that its peak value is when both NMOS and PMOS are in saturation.
- Ideally, there should be no short circuit current but there is a transition zone in the Inverter characteristic curve when both NMOS and PMOS are on.
- When $V_{in} = 0.5 V_{DD}$, NMOS is not completely ON and PMOS is not completely OFF, both are in saturation region, and hence peak value of short circuit current

Types of Power Dissipation

Dynamic Power

($\approx 40 - 70\%$ today and decreasing relatively)

Input switching to '1' or '0'

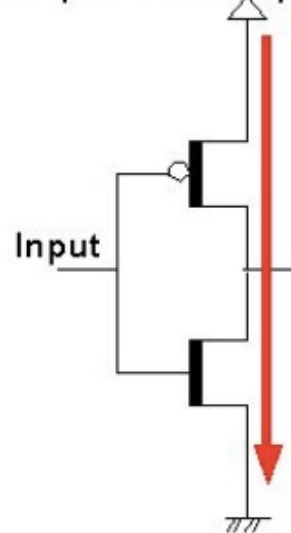


(a) Capacitive Current

Short-circuit power

($\approx 10\%$ today and decreasing absolutely)

$V_{th} < \text{Input} < V_{DD} - |V_{thp}|$

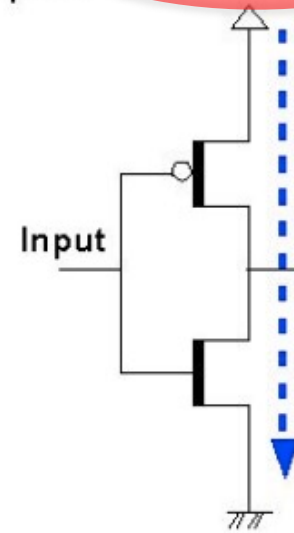


(b) Short Circuit Current

Static power

($\approx 20 - 50\%$ today and increasing)

Input : '1' or '0' steady state



(c) Static Leakage Current

Static Power

- ❑ Static power is consumed even when chip is quiescent.
 - Leakage draws power from nominally OFF devices
 - Ratioed circuits burn power in fight between ON transistors

Recall: Subthreshold Leakage

- Subthreshold leakage exponential with V_{gs}

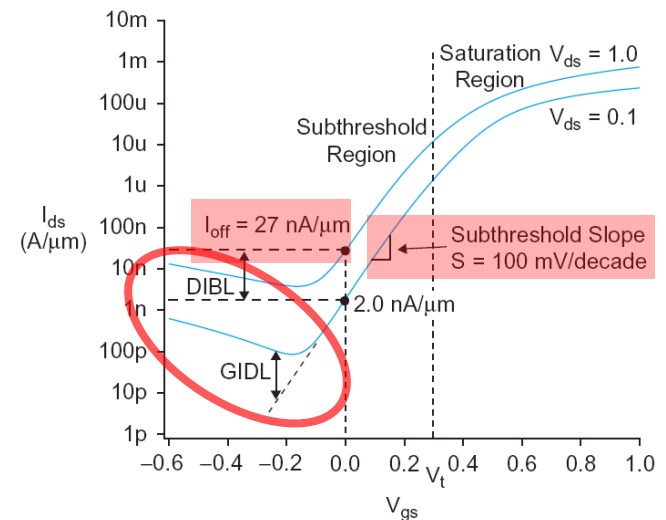
$$I_{ds} = I_{ds0} e^{\frac{V_{gs} - V_{t0} + \eta V_{ds} - k_{\gamma} V_{sb}}{n v_T}} \left(1 - e^{\frac{-V_{ds}}{v_T}} \right)$$

- n is process dependent
 - typically 1.3-1.7
- Rewrite relative to I_{off} on log scale

$$I_{ds} = I_{off} 10^{\frac{V_{gs} + \eta(V_{ds} - V_{dd}) - k_{\gamma} V_{sb}}{S}} \left(1 - e^{\frac{-V_{ds}}{v_t}} \right)$$

$$S = \left[\frac{d(\log_{10} I_{ds})}{dV_{gs}} \right]^{-1} = n v_T \ln 10$$

- $S \approx 100 \text{ mV/decade}$ @ room temperature



DIBL: Drain-Induced Barrier Lowering
GIDL: Gate-Induced Drain Leakage

Subthreshold Leakage

- For $V_{ds} > 50$ mV

$$I_{sub} \approx I_{off} 10^{\frac{V_{gs} + \eta(V_{ds} - V_{DD}) - k_{\gamma} V_{sb}}{S}}$$

- I_{off} = leakage at $V_{gs} = 0$, $V_{ds} = V_{DD}$

Typical values in 65 nm

$$I_{off} = 100 \text{ nA}/\mu\text{m} @ V_t = 0.3 \text{ V}$$

$$I_{off} = 10 \text{ nA}/\mu\text{m} @ V_t = 0.4 \text{ V}$$

$$I_{off} = 1 \text{ nA}/\mu\text{m} @ V_t = 0.5 \text{ V}$$

$$\eta = 0.1$$

$$k_{\gamma} = 0.1$$

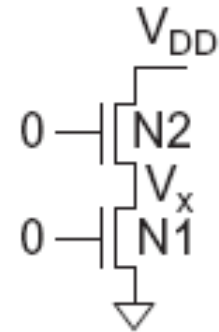
$$S = 100 \text{ mV/decade}$$

- To reduce leakage:

- Increase V_t : *multiple* V_t
 - Use low V_t only in critical circuits
- Decrease V_b
 - *Reverse body bias* in sleep
 - Or forward body bias in active mode

Stack Effect

- Series OFF transistors have less leakage
 - $V_x > 0$, so N2 has negative V_{gs}



$$I_{sub} = \underbrace{I_{off} 10^{\frac{\eta(V_x - V_{DD})}{S}}}_{N1} = \underbrace{I_{off} 10^{\frac{-V_x + \eta((V_{DD} - V_x) - V_{DD}) - k_\gamma V_x}{S}}}_{N2}$$

$$V_x = \frac{\eta V_{DD}}{1 + 2\eta + k_\gamma}$$

$$I_{sub} = I_{off} 10^{\frac{-\eta V_{DD} \left(\frac{1 + \eta + k_\gamma}{1 + 2\eta + k_\gamma} \right)}{S}} \approx I_{off} 10^{\frac{-\eta V_{DD}}{S}}$$

- **Leakage through 2-stack reduces ~10x**
- Leakage through 3-stack reduces further

Leakage Control

- ❑ Leakage and delay trade off
 - Aim for low leakage in sleep and low delay in active mode
- ❑ To reduce leakage:
 - Increase V_t : *multiple V_t*
 - Use low V_t only in critical circuits
 - Increase V_s : *stack effect*
 - *Input vector control* in sleep
 - Decrease V_b
 - *Reverse body bias* in sleep
 - Or forward body bias in active mode

Gate Leakage

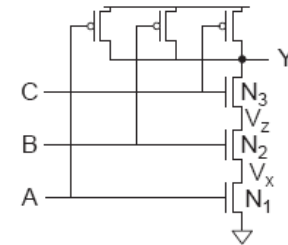
- ❑ Extremely strong function of t_{ox} and V_{gs}
 - Negligible for older processes
 - Approaches subthreshold leakage at 65 nm and beyond in some processes
- ❑ An order of magnitude less for pMOS than nMOS
- ❑ Control leakage in the process using $t_{ox} > 10.5 \text{ \AA}$
 - High-k gate dielectrics help
 - Some processes provide multiple t_{ox}
 - e.g. thicker oxide for 3.3 V I/O transistors
- ❑ Control leakage in circuits by limiting V_{DD}

NAND3 Leakage Example

□ 100 nm process

$$I_{gn} = 6.3 \text{ nA} \quad I_{gp} = 0$$

$$I_{offn} = 5.63 \text{ nA} \quad I_{offp} = 9.3 \text{ nA}$$



Input State (ABC)	I_{sub}	I_{gate}	I_{total}	V_x	V_z
000	0.4	0	0.4	stack effect	stack effect
001	0.7	0	0.7	stack effect	$V_{DD} - V_t$
010	0.7	1.3	2.0	intermediate	intermediate
011	3.8	0	3.8	$V_{DD} - V_t$	$V_{DD} - V_t$
100	0.7	6.3	7.0	0	stack effect
101	3.8	6.3	10.1	0	$V_{DD} - V_t$
110	5.6	12.6	18.2	0	0
111	28	18.9	46.9	0	0

ORDER OF MAGNITUDE DIFFERENCE
DUE TO STACKING OF NMOS

HIGH GATE-TO-SOURCE (VDD-TO-GND)
LEAKAGE FOR NMOS (NOT PMOS)

Data from [Lee03]

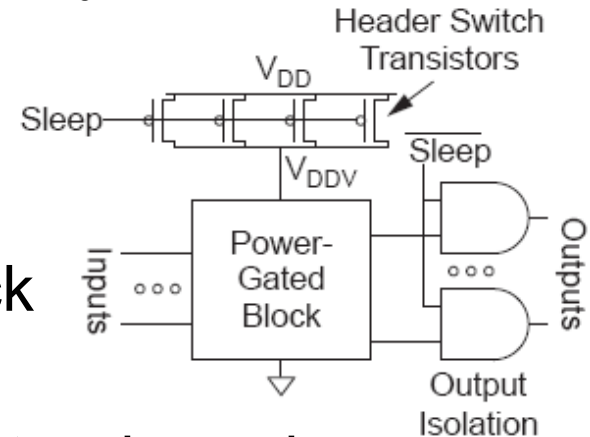
Junction Leakage

- ❑ From reverse-biased p-n junctions
 - Between diffusion and substrate or well
- ❑ Ordinary diode leakage is negligible
- ❑ Band-to-band tunneling (BTBT) can be significant
 - Especially in high- V_t transistors where other leakage is small
 - Worst at $V_{db} = V_{DD}$
- ❑ Gate-induced drain leakage (GIDL) exacerbates
 - Worst for $V_{gd} = -V_{DD}$ (or more negative)

Power Gating

- ❑ Turn OFF power to blocks when they are idle to save leakage

- Use virtual V_{DD} (V_{DDV})
- Gate outputs to prevent invalid logic levels to next block



- ❑ Voltage drop across sleep transistor degrades performance during normal operation
 - Size the transistor wide enough to minimize impact
- ❑ Switching wide sleep transistor costs dynamic power
 - Only justified when circuit sleeps long enough