University of Southern California

Viterbi School of Engineering

EE577A
VLSI System Design

Power Estimation and Optimization

References: Professor Pedram's slides, and some IEEE papers

Shahin Nazarian

Spring 2013

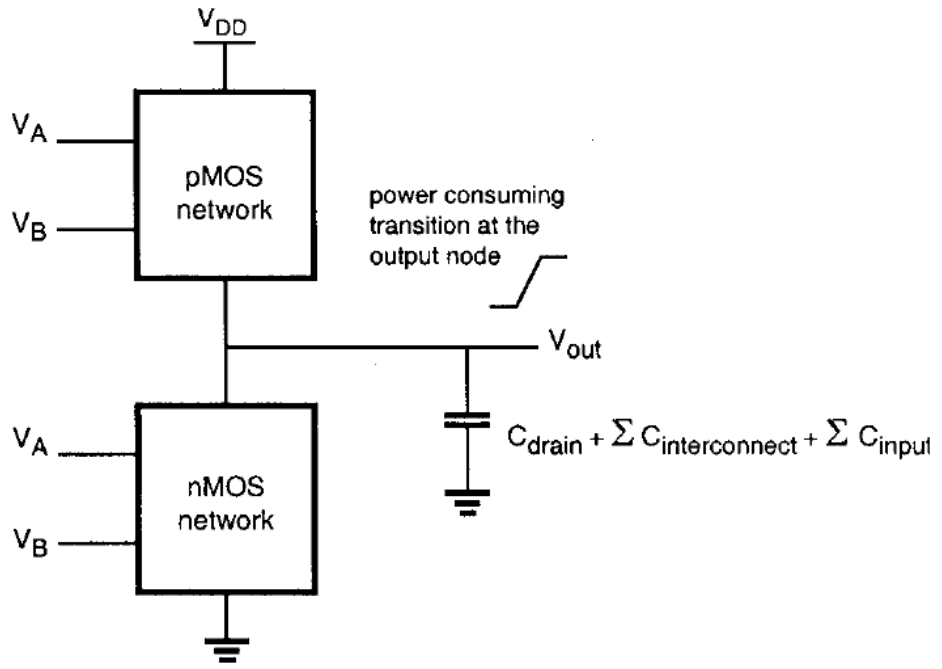
Why Power Consumption Is Important?

- In recent years, with the success and growth of personal computing device (portable desktops, multimedia products, etc), which require high-speed computation and complex functionality with low power consumption, power has increasingly been given comparable weight to area and speed
- Disposable batteries can cost > $500/watt over the life of device
- Rechargeables can cost > $50/watt over the life of device
- High performance is limited by difficulty of heat removal from chip, e.g., ~100W/chip, considering the cost of electricity as ~$5/watt over the life of device
- Every 10°C increase on operating temperature doubles failure rate for the electronic components

Background

Switching Power Consumption

- Any CMOS logic gate making an output voltage transition can be represented by its nMOS and pMOS networks, and the total load capacitance connected to its output node



Average Switching Power Consumption

- **Review**

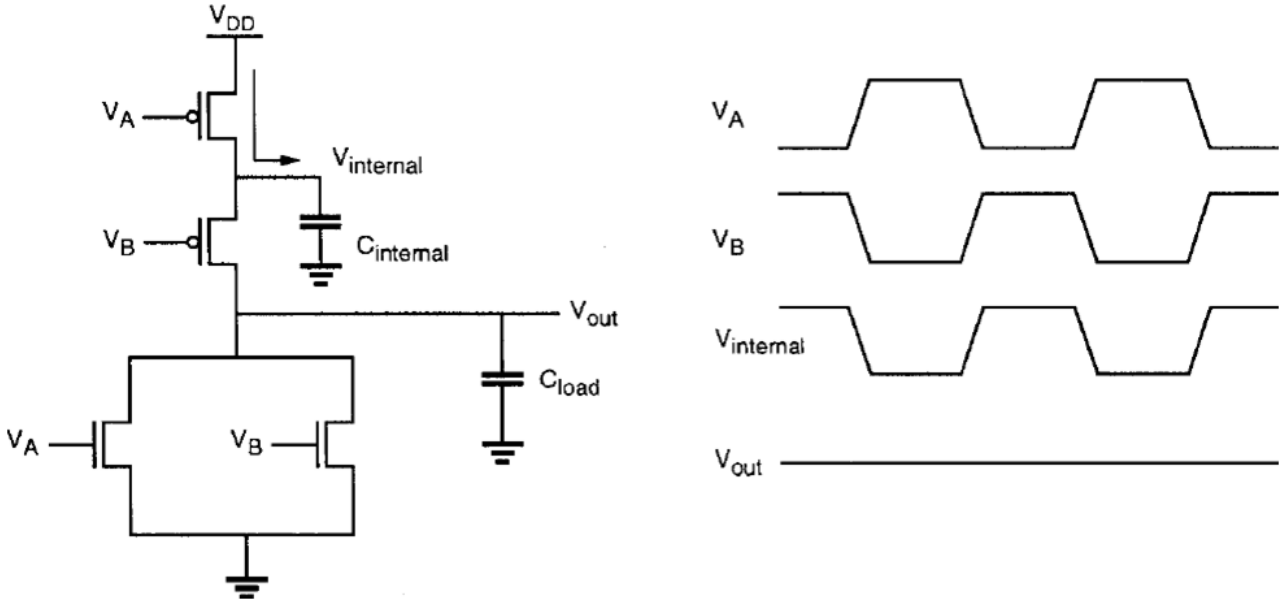
$$P_{avg} = \frac{1}{T}\left[\int_0^{T/2} V_{out}\left(-C_{load}\frac{dV_{out}}{dt}\right)dt + \int_{T/2}^{T}(V_{DD}-V_{out})\left(C_{load}\frac{dV_{out}}{dt}\right)dt\right]$$

$$P_{avg} = \frac{1}{T}C_{load}V_{DD}^{\ 2} \qquad P_{avg} = C_{load}V_{DD}^{\ 2}f_{CLK}$$

$$\boxed{P_{avg} = \frac{1}{2}C_{load}\,V_{DD}^{\ 2}\,f_{CLK}\,\beta}$$

Switching Power Consumption (Cont.)

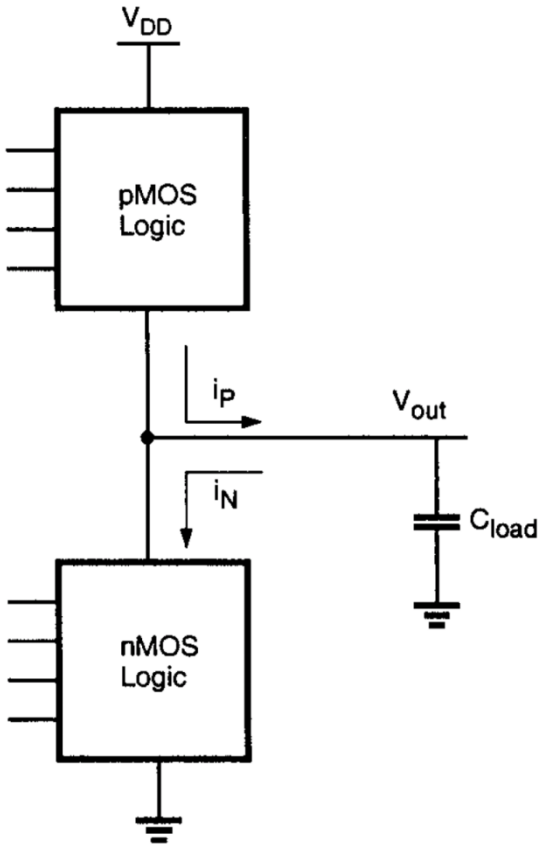
- In complex CMOS logic gates, most of internal circuit nodes also make full or partial voltage transition during switching
- Switching of the internal node in a two-input NOR gate results in dynamic power dissipation even if the output node voltage remains unchanged



Power Consumption Calculation

- Power expression of the CMOS inverter can be applied to all CMOS gates
  - General CMOS gates consist of an nMOS logic block and a pMOS block
  - Define β as the expected number of transitions (0→1 plus1→0) at the output of the gate per clock cycle time. We have:

$$P_{avg} = \frac{1}{2} C_{load} V_{DD}^{\ 2} f_{clk} \beta$$



  - Note: *prob*(0→1)=*prob*(1→0)=*p*(0)*p*(1); therefore, β = 2*p*(0)*p*(1) ; *p*(0) + *p*(1) = 1
  - Example: for 2-input NAND gate with equal probability of 0's and 1's at its inputs (random inputs,) *p*(1)=3/4 and *p*(0)=1/4;thus, β = 2(1/4)(3/4)=3/8
  - Example: for a 3-input NOR gate with random inputs, β = 2(7/8)(1/8)=7/32

Probability-based Power Calculation in a VLSI Circuit

$$P_{circuit} = \sum_{g \in gates} P_g, \quad P_g = \frac{1}{2} C_{g,load} V_{DD}^2 f \beta_g$$

where $C_{g,load}$ denotes the load capacitance of gate g,

β_g denotes the switching activity at the output of gate g

Switching Activity Calculation Algorithm for a Static CMOS Gate, g

// Assume gate g implements Boolean function, F, and that its inputs are independent

Make product terms mutually non-intersecting as you write the sum-of-products expression for F

Write each product term signal probability as the product of its input signal probabilities (notice: $s_{\bar{A}} = 1 - s_A$)

Add up signal probabilities of indvidual product terms to get the signal probability of F, s_F

Calculate the switching activity of gate g, β_g, as $2 s_F (1 - s_F)$

Example calculation for some CMOS gates -- in all cases $\beta_g = 2 s_F (1 - s_F)$:

$g : F = AB$: $s_F = s_A s_B$

$g : F = ABC$: $s_F = s_A s_B s_C$

$g : F = A + B = A\bar{B} + B$: $s_F = s_A (1 - s_B) + s_B$

$g : F = A + B + C = A\bar{B}\bar{C} + B\bar{C} + C$: $s_F = s_A (1 - s_B)(1 - s_C) + s_B (1 - s_C) + s_C$

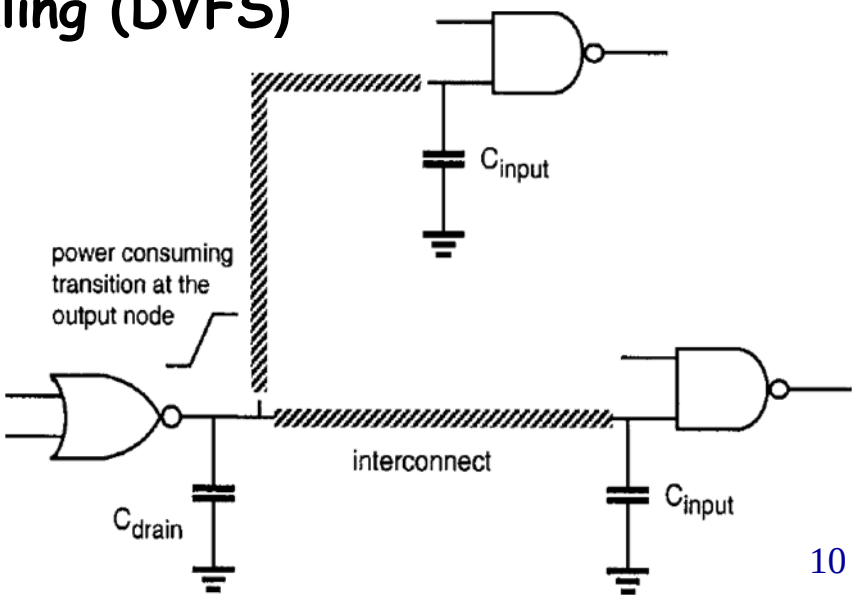
$g : F = AB + CD = AB(\bar{C} + C\bar{D}) + CD$: $s_F = s_A s_B (1 - s_C) + s_A s_B s_C (1 - s_D) + s_C s_D$

Observations about Switching Power Reduction

- There are a number of ways to reduce the switching power consumption in digital VLSI circuits:
  - Reduction of the power supply voltage
  - Reduction of the voltage swing in all nodes
  - Reduction of the switching probability (transition factor)
  - Reduction of the load capacitance
- Note that the switching power dissipation is a linear function of the clock frequency, yet simply reducing the frequency would diminish the overall system performance, and not save any energy

Dynamic Power Minimization

- Static voltage scaling (SVS)
- Trading area or latency for power
  - Pipelining
  - Parallelization
- Driving buses
  - Split buses and bus encoding
  - Low-swing bus drivers
- Clock gating
- Adiabatic circuits
- Dynamic voltage and frequency scaling (DVFS)



Optional: Voltage Scaling and Delay

- **Although the reduction of power supply voltage reduces the dynamic power dissipation, according to the alpha-power current law used for short channel devices, the delay increases**

$$delay = \frac{C_{load} V_{DD}}{k(V_{DD} - V_T)^{\alpha}} \quad 1.4 \leq \alpha \leq 2$$

Optional: Alpha-Power Law

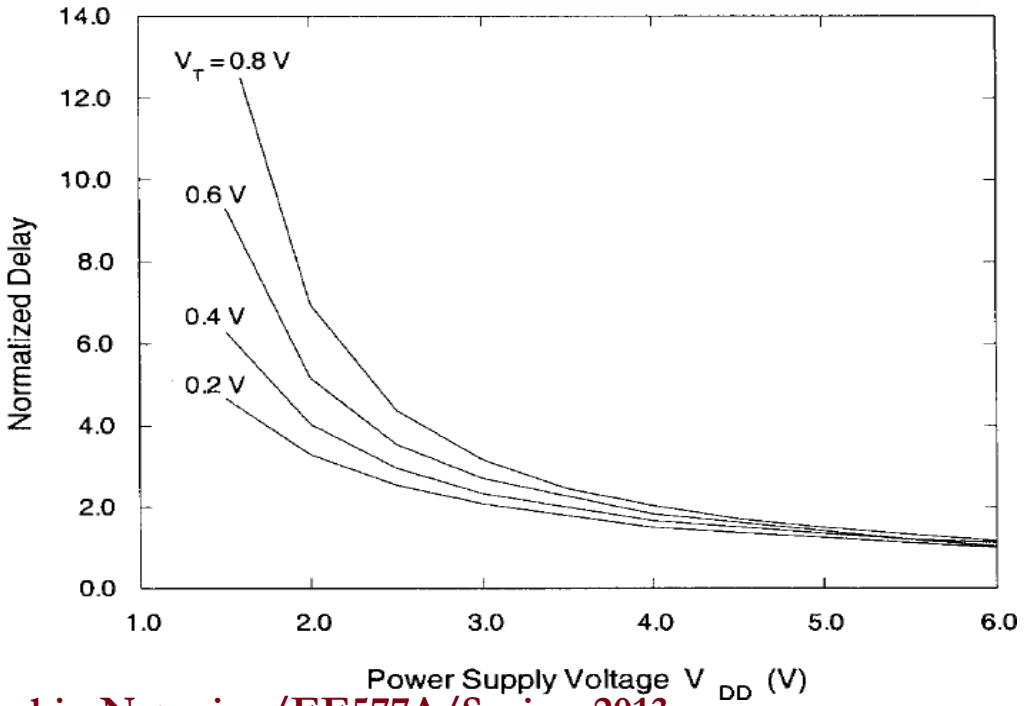
- **Variation of the normalized propagation delay of a CMOS inverter as a function of the power supply voltage V_{DD} and the threshold voltage V_{th}**
- **Alpha-Power Law:**

$$\frac{delay_{new}}{delay} = \frac{V_{DD,new}}{V_{DD}} \left(\frac{V_{DD} - V_{th}}{V_{DD,new} - V_{th}} \right)^{\alpha}$$

where $1.4 \leq \alpha \leq 2$

Short channel devices

Long channel devices



Optional: Static Voltage Scaling Review

- **Consider a scaling scenario where voltage is scaled down by S, all dimensions are scaled down by M, while doping densities are scaled up by M:**

$$Vnew = \frac{1}{s} \bullet Vold \qquad Inew = \frac{M}{S^2} \bullet Iold \qquad freq_{new} = \frac{M^2}{S} \bullet freq_{old}$$

$$Cload(new) = \frac{1}{M} \bullet Cload(old) \qquad Power_{new} = \frac{M}{S^3} \bullet Power_{old}$$

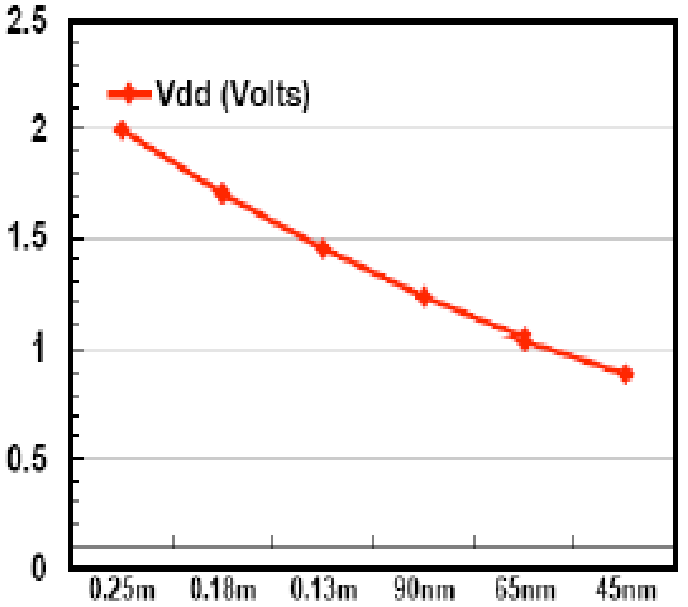
$$Energy_{new} = \frac{1}{MS^2} \bullet Energy_{old} \qquad Pow\_dens_{new} = \frac{M^3}{S^3} Pow\_dens_{old}$$

Optional: Static Voltage Scaling Example

- With S^{-1}=0.85 and M^{-1}=0.7, based on the equations on the previous page, we obtain:

$$Energy_{new} = 0.506 Energy_{old} \qquad Pow\_dens_{new} = 1.79 Pow\_dens_{old}$$

- With each generation, voltage has decreased to 0.85x, not 0.7x for constant field scaling. Thus, energy dissipation per logic gate decreases by (1- 0.85^2*0.7)=50%



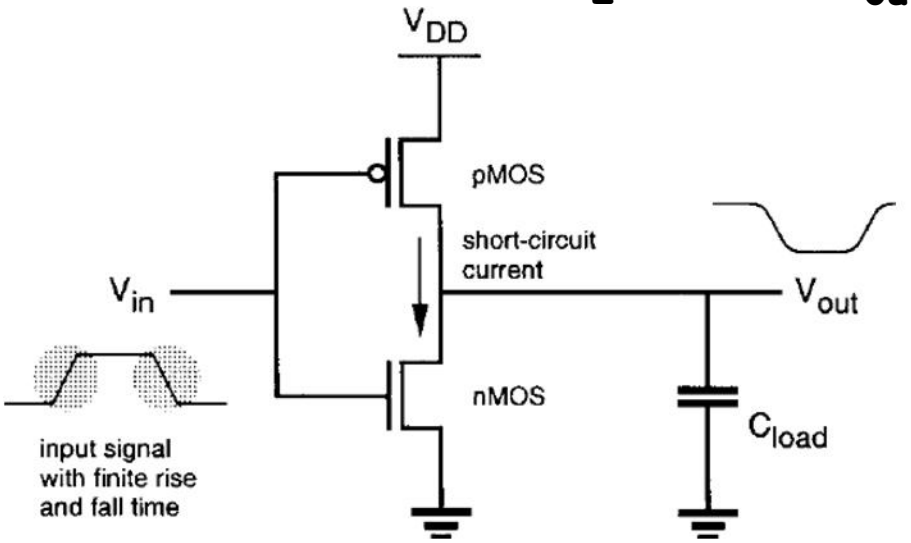
rather than by the ideal (1-$0.7^{3)}$=66% per generation

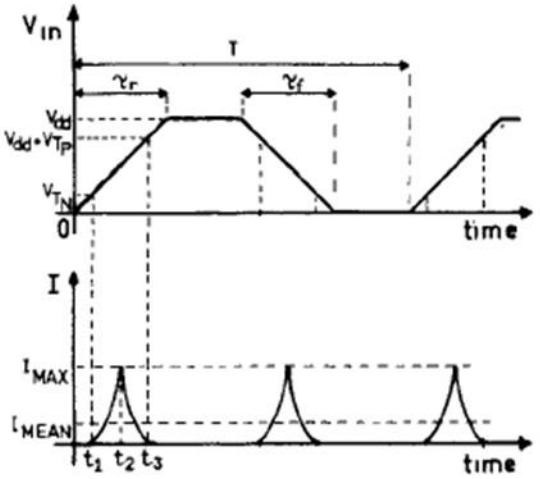
Optional: Static Voltage Scaling Example (Cont.)

- However, the number of logic gates in a chip has been increasing by 3x per generation (since the die size is increasing correspondingly), thus a net increase in the energy consumption per chip
- The power density is increasing by about 80% per generation
- Voltage can be adjusted based on different mode. For example, a laptop processor may operate at high voltage when plugged into AC adapter, but at lower voltage when using battery power

Short Circuit Power (P_{sc}) Dissipation

- Let $\tau_r = \tau_f = \tau$ denote the transition time of the input voltage, V_{in}
- Now t_1 is the time when the input voltage reaches the threshold voltage of nMOS while t_3 is the time when the input voltage reaches the threshold voltage of pMOS
- The short-circuit current flows between t_1 and t_3 and reaches it maximum at t_2 when $V_{out}=V_{dd}/2$





P_{sc} Calculation – A Model Example

- The calculation is based on the concept of an equivalent short circuit capacitance calculated under the assumption that the input and the output waveforms are linear
- For a symmetric CMOS inverter with $k_n = k_p = k$, $V_{T,n} = |V_{T,p}| = V_T$, and equal input rise and fall times, the equation is:

$$P_{sc} = \frac{1}{12} k \tau_{in} V_{DD} (V_{DD} - 2V_T)^2 \left(\frac{1}{1 + \tau_{out}/\tau_{in}} \right) f_{CLK} \beta$$

which almost reduces to Veendrick's result for $\tau_{out}/\tau_{in} = 1$

http://idc.yonsei.ac.kr/course/lvlp/papers/H.J.M.Veendrick_84.pdf

P_{SC} Calculation (Analytical Models)

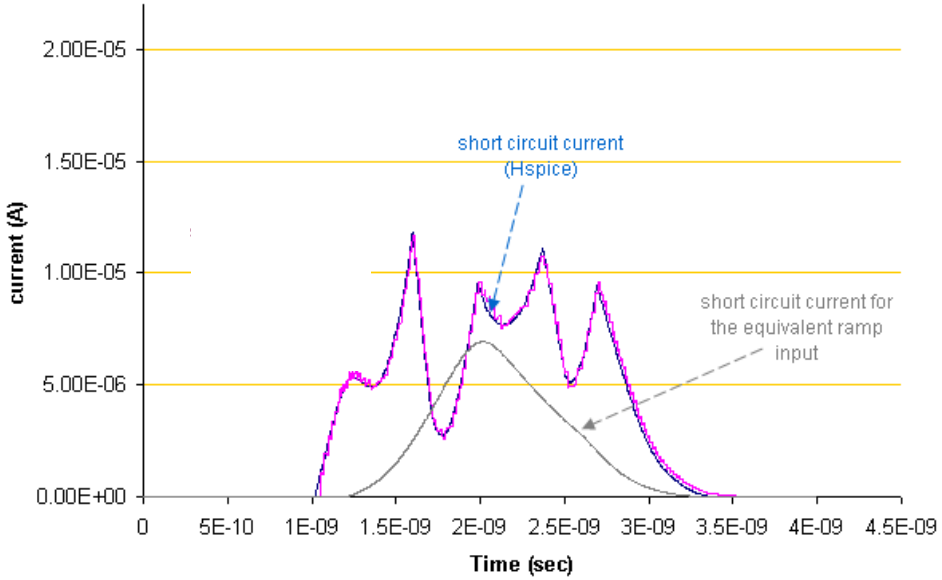
- **Closed-form analytical expressions:**
  - **For example:** $$P_{SC} = \frac{k_n \tau_{in}}{12\left(1+\frac{\tau_{out}}{\tau_{in}}\right)}(V_{DD}-2V_T)^3 V_{DD} f_{CLK} \beta$$
  - **Based on device model**
    - H. Veendrick (1984), based on Shichman–Hodges model
    - S. Vemuri et al. (1994) and K. Nose et al. (2000) based on alpha-power model
    - L. Rossello et al. (2002) based on MM9
- **Analytical models are very inaccurate because of:**
  - **Their dependence on simple and inaccurate device models**
  - **And their simplistic assumptions on device operation modes during signal transitions**

P_{SC} Calculation – Non-Analytical Models

Equivalent Ramp Input Model [Dartu, Pileggi, DAC 1996]

- Assumes a ramp signal waveform as the input

- Characterizes the logic cell current based on input signal transition time and capacitive output load, i.e., $P_{SC} = F(Slew_{in}, C_{Load-eff})$



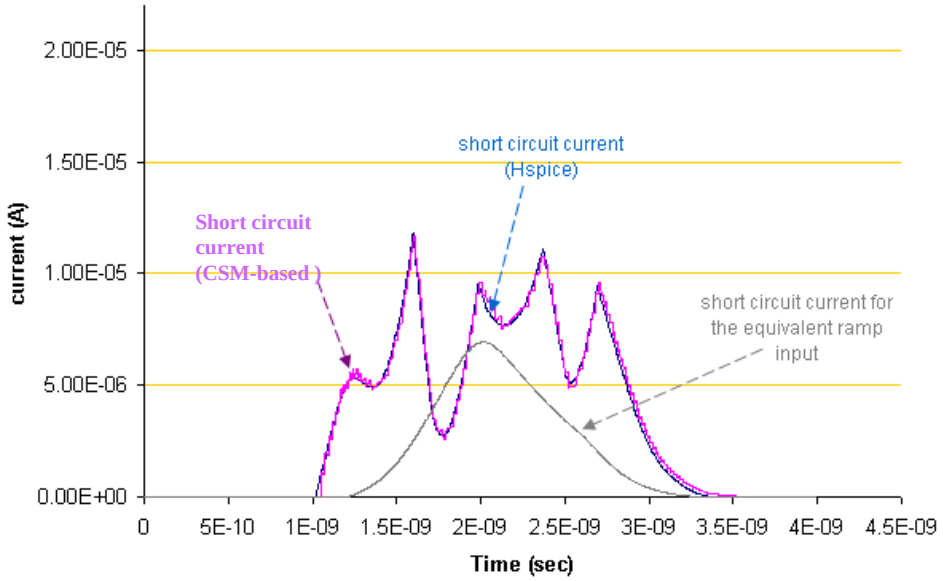
P_{SC} Calculation – Non-Analytical Models

CSM-based* Model [Nazarian, Fatemi, Pedram, TVLSI 2010]

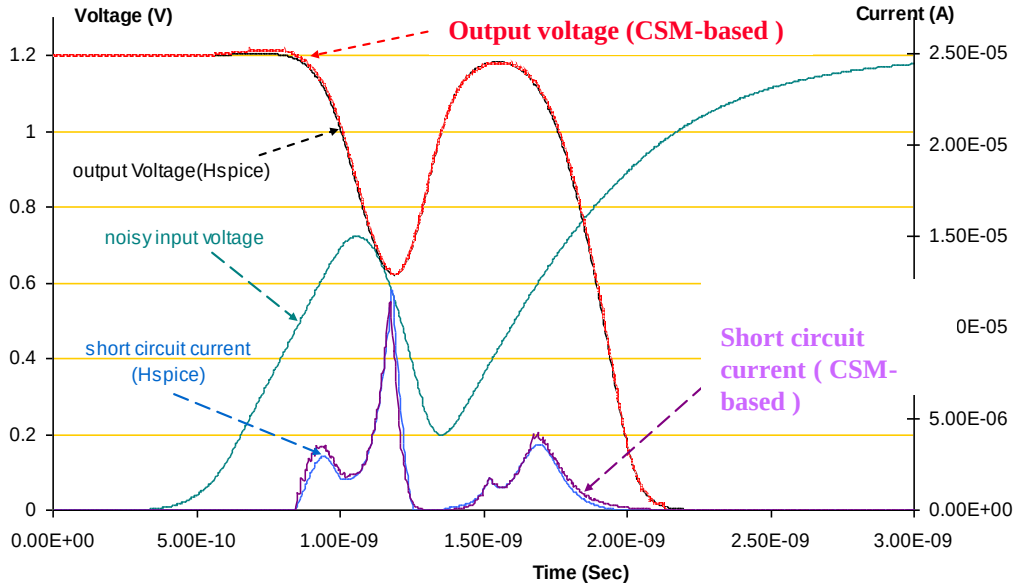
- Characterizes the logic cell based on the input and output voltage values, i.e., $P_{SC} = F(V_{in}, V_{out})$

- Can process input signal with any waveform shape

* CSM ≡ Current Source Model



CSM-based P_{SC} Calculation – Example



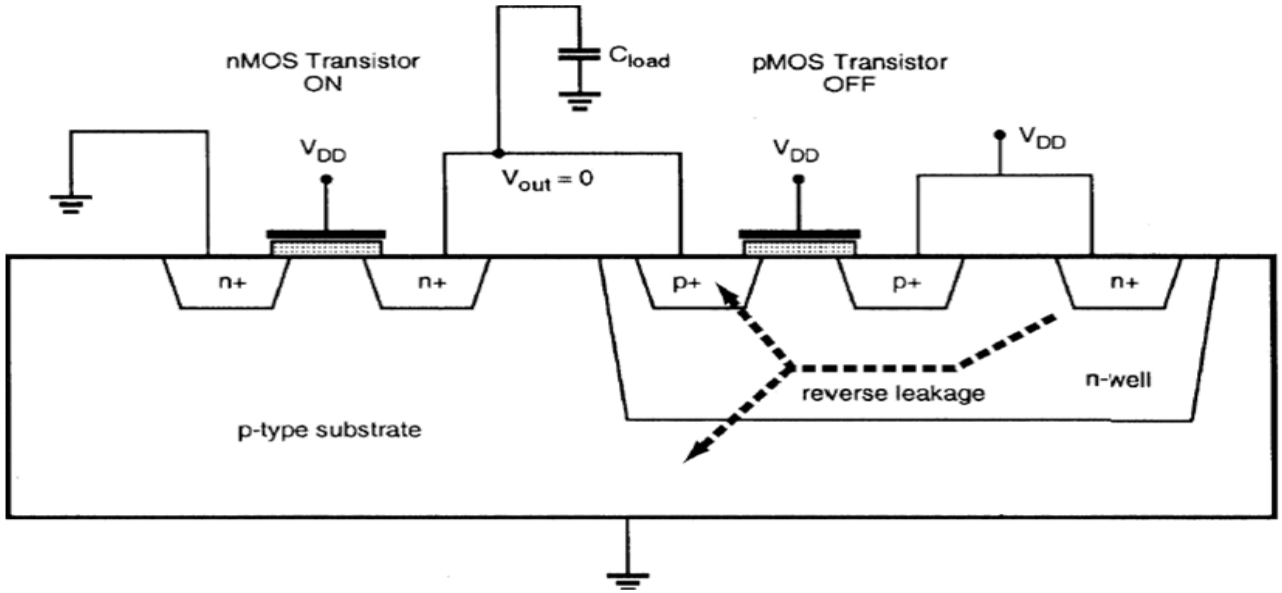
CSM-based* Model [Nazarian, Fatemi, Pedram, TVLSI 2010]

Leakage Power Dissipation

- The nMOS and pMOS transistors used in CMOS gates have non-zero reverse leakage and subthreshold current which contribute to the overall power dissipation even when there is no switching activity in a circuit
- The magnitude of the leakage current is determined mainly by the device parameters
- The reverse diode leakage occurs when the pn-junction between the drain and the bulk of a transistor is reverse-biased

Reverse Biased Junction Leakage

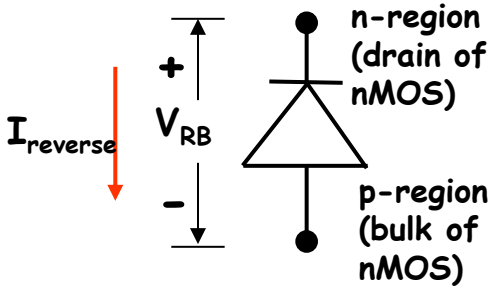
- Consider a CMOS inverter with a high input voltage
  - Although pMOS transistor is turned off, there will be a reverse potential difference of V_{DD} between its drain and the n-well



Reverse-Biased Junction Leakage (Cont.)

- The reverse leakage current of a pn-junction is expressed by

$$I_{reverse} = A \cdot J_s \cdot (1 - e^{-\frac{V_{RB}}{n\vartheta_T}})$$



  - A : the junction area
  - J_s : the *maximum reverse saturation current density* (typically 1-5 pA/mm^2)
  - n : the *emission coefficient*, usually set to 1, although can be larger depending on the type of junction
  - V_{RB} : the *reverse bias voltage* across the junction, i.e., the voltage of drain diffusion with respect to the bulk or well
  - $\upsilon_T = kT/q$ denotes the thermal voltage at absolute junction temperature, T
- $I_{reverse}$ is maximum when V_{RB} is largest, that is why we focus on the drain side and not the source side of the nMOS transistor

Subthreshold Current Leakage

- When V_{DS} is increased, the potential barrier for the electrons in the channel decreases and we have significant current even if $V_{GS} < V_{T0}$
- The channel current that flows under these conditions ($V_{GS} < V_{T0}$) is called the **sub-threshold current**. For bulk CMOS,

$$I_D(sub) \equiv I_{sub} = \frac{W}{L}\mu_e C_{ox}(n-1)\vartheta_T^2 \, e^{\frac{V_{GS}-V_T+\eta V_{DS}}{n\vartheta_T}} \left(1 - e^{\frac{-V_{DS}}{\vartheta_T}}\right)$$

$$V_{GS} = 0, V_{DS} >> \vartheta_T, \eta = 0 \Rightarrow I_{sub} = \frac{W}{L}\mu_e C_{ox}(n-1)\vartheta_T^2 \, e^{\frac{-V_T}{n\vartheta_T}} \propto \left(\frac{W}{L}\right) 10^{\frac{-V_T}{S}}$$

- The *subthreshold swing* (a.k.a. the inverse subthreshold slope), S, is equal to the voltage required to increase I_D by 10X, i.e.,
  - $n \geq 1$ is called the *body effect coefficient*
  - If $n \rightarrow 1$, $S \rightarrow 60$ mV/decade at 300 K

$$S = \left(\frac{\partial(\log_{10} I_{sub})}{\partial V_{GS}}\right)^{-1} = n\vartheta_T \ln 10 = 2.3n\frac{kT}{q}$$

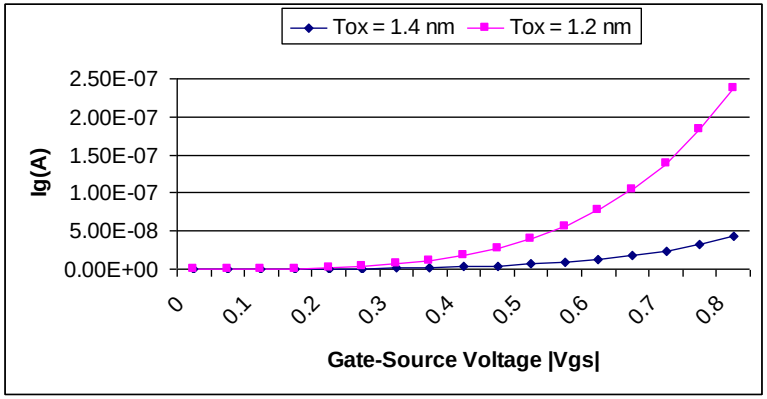
  - We want S to be small to shut off the MOSFET quickly
  - In well-designed bulk CMOS devices, S is 70 - 90 mV/decade at 300 K

Gate Leakage (Optional)

- With the advent of deep-submicron devices comes the reduction of the gate-oxide thickness. This reduction leads to a higher electric field across the oxide
  - The tunneling of electrons through the gate oxide into the substrate and from substrate to the gate becomes possible. This current is referred to as gate leakage
  - Being quantum mechanical in nature the gate leakage current is virtually temperature independent

$$I_{gate} \cong \kappa WL \left(\frac{V_{GB}}{t_{ox}} \right)^2 e^{-\alpha \frac{t_{ox}}{V_{GB}}}$$

where κ and α are fitting parameters, W is the transistor width, V_{GB} denotes the gate to bulk voltage, and t_{ox} denotes the gate oxide thickness



Source and bulk are tied together, i.e., $V_{GB}=V_{GS}$

Total Power Dissipation in CMOS VLSI Circuits

- The total power dissipation is the sum of two components: dynamic (switching plus short-circuit) and leakage (reverse biased junction, subthreshold and gate currents)

$$P_{total} = \frac{1}{2}\left(C_{load} V_{DD} + \frac{k \tau_{in}}{6\left(1 + \frac{\tau_{out}}{\tau_{in}}\right)} \left(V_{DD} - 2V_T\right)^2 \right) V_{DD} f_{CLK} \beta + V_{DD} I_{leakage}$$

$$I_{leakage} = I_{reverse} + I_{subthreshold} + I_{gate}$$

Example

Calculate the capacitive, short circuit, and subthreshold leakage components of power dissipation of a CMOS inverter with $W_p/L=2W_n/L=8$, driving an identical inverter with the following parameters:

$\mu_n = 2\mu_p = 600 cm^2/V\text{-}sec$,
$C_{ox}=2\times10^{-7}\ F/cm^2$,
$V_{T,n} = -V_{T,p} = 0.6V$, $V_{DD}=2.8V$,
$t_{in}=10ps$, $t_{out}=30ps$,
activity factor $\beta = 0.2$,
$f_{CLK}=500MHz$, die Temperature $T=85\ ^\circ C$,
Subthreshold shape parameter, $n=1.5$,
Boltzmann constant, $k=1.38\times10^{-23}J/K$, electron/hole charge, $q=1.6\times10^{-19}\ C$, and $L=0.25\mu m$

$$k'_n = \mu_n C_{ox} = 2k'_p = 2\left(\mu_p C_{ox}\right) = \left(600\frac{cm^2}{V\cdot s}\right)\left(2\times10^{-7}\frac{F}{cm^2}\right) = 120\frac{\mu A}{V^2}$$

$$k_n = k_p = \mu_n C_{ox}\frac{W_n}{L} = \mu_p C_{ox}\frac{W_p}{L} = \left(120\frac{\mu A}{V^2}\right)(4) = \left(60\frac{\mu A}{V^2}\right)(8) = 480\frac{\mu A}{V^2}$$

$$C_{load} = \left(W_n + W_p\right)LC_{ox} = 3W_nLC_{ox} = \frac{3}{2}W_pLC_{ox} = (3\times10^{-4}cm)(0.25\times10^{-4}cm)\left(2\times10^{-7}\frac{F}{cm^2}\right) = 1.5\,fF$$

$$\vartheta_T = 1.38\times10^{-23}\frac{273+85}{1.6\times10^{-19}} = 308.8\times10^{-4}V = 30.9mV$$

Note that since inverter is driving identical load: $\mu_n C_{ox}\frac{W_n}{L} = \mu_n\frac{C_{load}}{3L^2} = \left(2\mu_p\right)\frac{C_{load}}{3L^2} = \mu_p C_{ox}\frac{W_p}{L}$

$$P_{total} = \frac{1}{2}C_{load}V_{DD}{}^2 f_{CLK}\beta + \frac{k_n\tau_{in}}{12\left(1+\frac{\tau_{out}}{\tau_{in}}\right)}\left(V_{DD}-2V_T\right)^2 V_{DD} f_{CLK}\beta + \frac{\mu_n C_{load}}{3L^2}\vartheta_T^2(n-1)\,e^{\frac{-V_T}{n\vartheta_T}}V_{DD}$$

$$P_{total} = \frac{1}{2}(1.5f)(2.8V)^2(500MHz)(0.2)$$

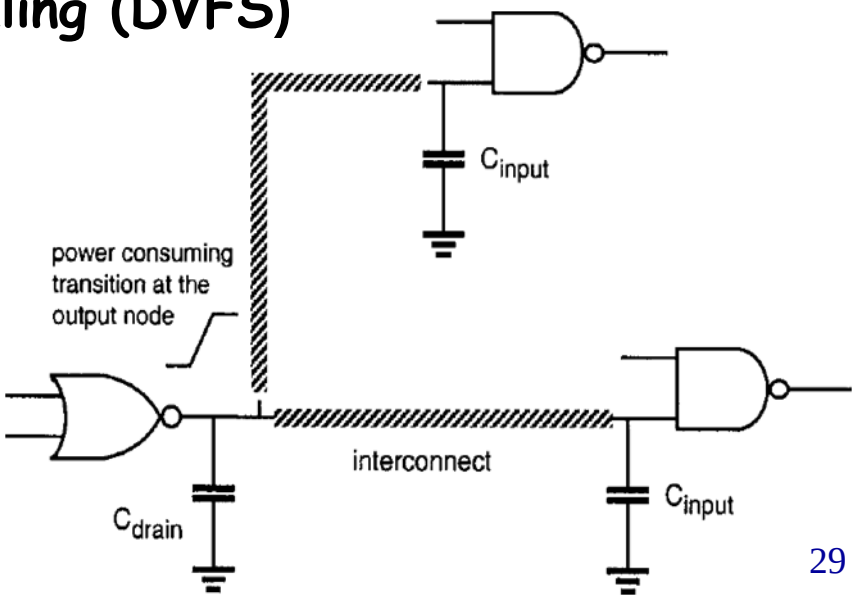
$$+\frac{\left(480\frac{\mu A}{V^2}\right)(10ps)}{12\left(1+\frac{30}{10}\right)}\left(2.8V-2\times0.6V\right)^2(2.8V)(500MHz)(0.2)$$

$$+\frac{\left(600\frac{cm^2}{V\cdot s}\right)(1.5fF)}{3\left(0.25\times10^{-4}cm\right)^2}\left(0.0309V\right)^2(0.5)e^{\frac{-0.6}{1.5\times0.0309}}(2.8V)$$

$$= 0.588\mu W + 0.0717\mu W + 1.532\,pW$$

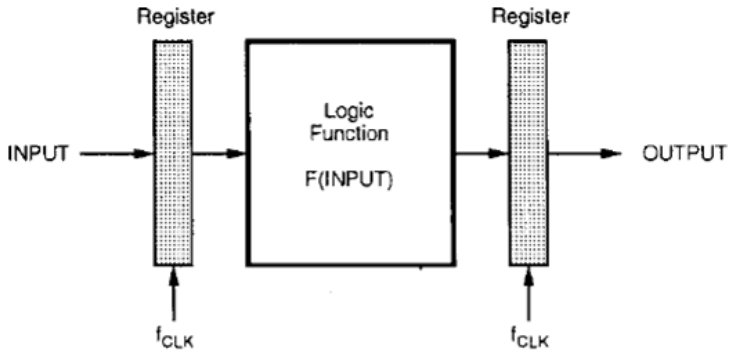
Dynamic Power Minimization

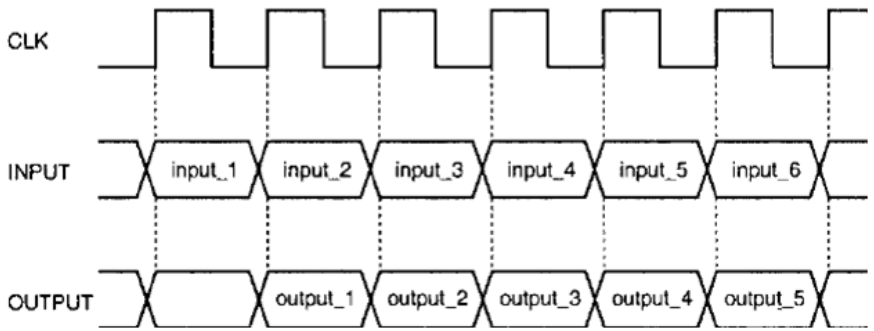
- Static voltage scaling (SVS)
- Trading area or latency for power
  - Pipelining
  - Parallelization
- Driving buses
  - Split buses and bus encoding
  - Low-swing bus drivers
- Clock gating
- Adiabatic circuits
- Dynamic voltage and frequency scaling (DVFS)



Pipelining Approach to Power Minimization

- Single-stage implementation of a logic function and its timing diagram
  - Let C_{total} be the total capacitance switched every clock cycle
  - C_{total} consists of
    - i) the capacitance switched in the input register
    - ii) the capacitance switched to implement the logic function
    - iii) the capacitance switched in the output register



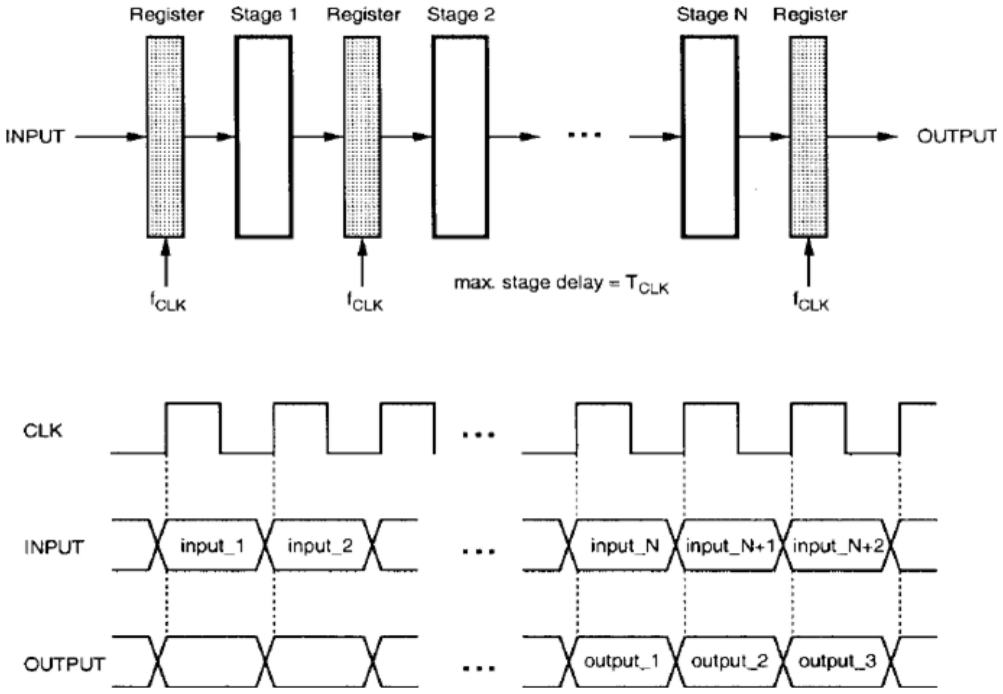


Pipelining (Cont.)

- Dynamic power consumption is

$$P_{reference} = C_{total} \cdot V_{DD}^{\ 2} \cdot f_{CLK}$$

- Consider an N-stage pipelined structure of the same function



Pipelining (Cont.)

- The logic function has been partitioned into N successive stages, and a total of (N-1) register arrays have been introduced to create the pipeline
- Suppose all stages of the partitioned function have equal delay of

$$\tau_{p(pipeline\_stage)} = \frac{\tau_{p,max(input-to-output)}}{N} = T_{CLK}$$

Then logic blocks between two successive registers can operate N-times slower while maintaining the same functional throughput. This means that the power supply voltage can be reduced to a value of

$$V_{DD,new} \cong \frac{V_{DD}}{N}$$

Pipelining (Cont.)

- The dynamic power consumption of the N-stage pipelined structure may be approximately as:

$$P_{pipeline} = [C_{total} + (N-1)C_{reg}] \cdot V_{DD,new}^{\ 2} \cdot f_{CLK}$$

where C_{reg} denotes the capacitance switched by each pipeline register

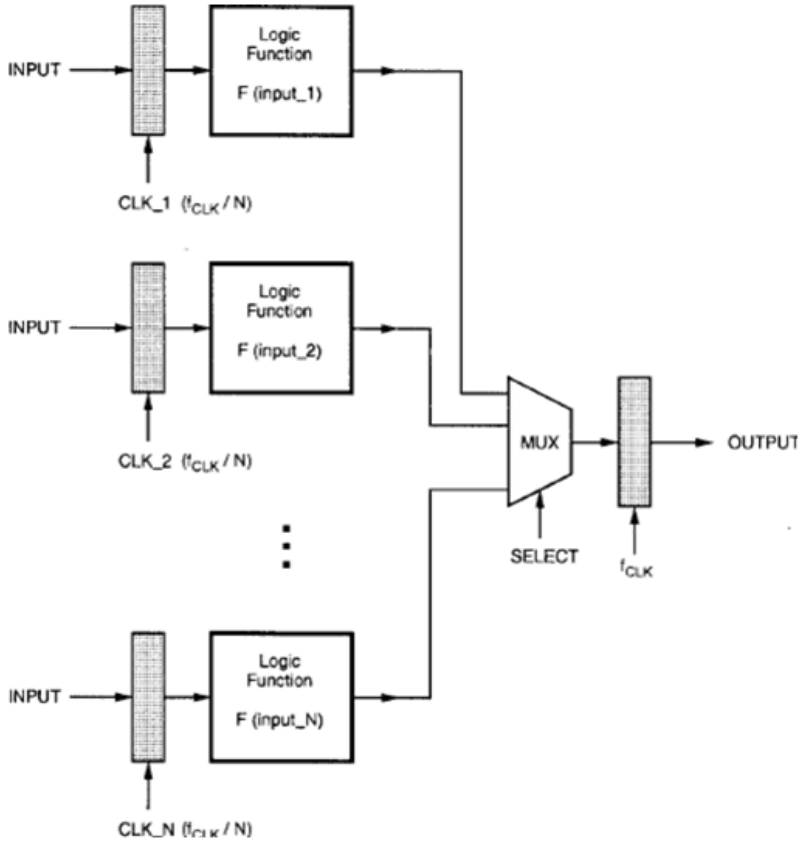
- The power reduction factor achieved in an N-stage pipeline structure is:

$$\frac{P_{pipeline}}{P_{reference}} = \left[1 + \frac{C_{reg}}{C_{total}}(N-1)\right] \cdot \frac{V_{DD,new}^{\ 2}}{V_{DD}^{\ 2}} \geq \frac{1}{N^2}$$

- The area overhead is rather small; the latency has increased from one to N cycles

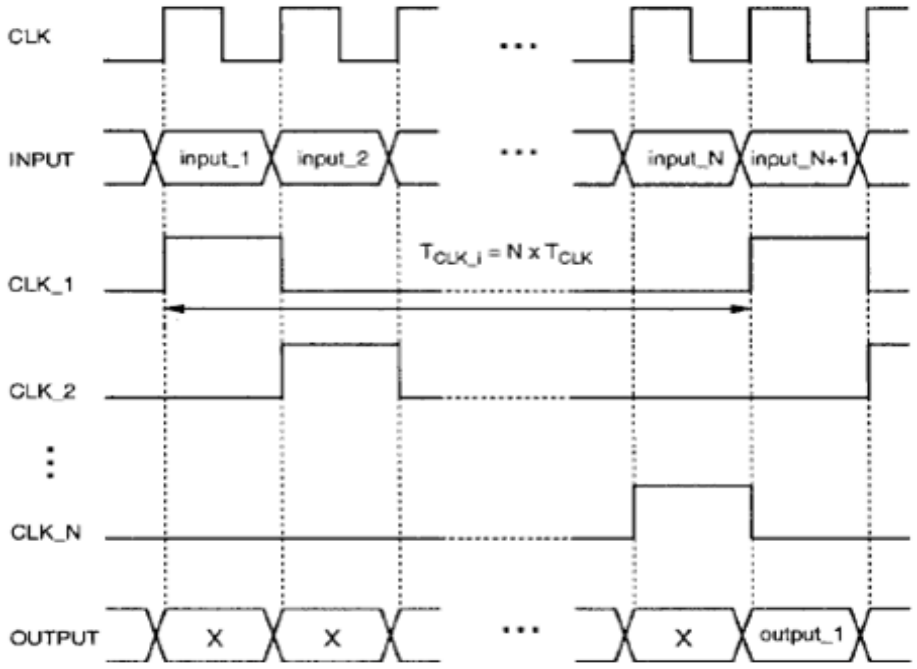
Parallel Processing Approach to Power Minimization

- Another method for trading off area for low power dissipation is to use parallelism or hardware replication
- This approach is useful when the logic function to be implemented is not suitable for pipelining
- Consider N-block parallel structure realizing the same logic function before



Parallel Processing (Cont.)

- Simplified timing diagram of the N-block parallel structure



Parallel Processing (Cont.)

- The total dynamic power dissipation of the parallel structure is:

$$P_{parallel} = N \cdot C_{total} \cdot V_{DD,new}^{\ 2} \cdot \frac{f_{CLK}}{N} + C_{MUX} \cdot V_{DD,new}^{\ 2} \cdot f_{CLK}$$

$$= \left(C_{total} + C_{MUX}\right) \cdot V_{DD,new}^{\ 2} \cdot f_{CLK}$$

- Power reduction achieved by an N-block parallel implementation is:

$$\frac{P_{parallel}}{P_{reference}} = \left(1 + \frac{C_{MUX}}{C_{total}}\right) \cdot \frac{V_{DD,new}^{\ 2}}{V_{DD}^{\ 2}} \geq \frac{1}{N^2}$$

- Two obvious overheads of this approach are the increased area and the increased latency
- The pipeline approach is more attractive because of its smaller area overhead

Driving Buses: Reduction of Switched Capacitance

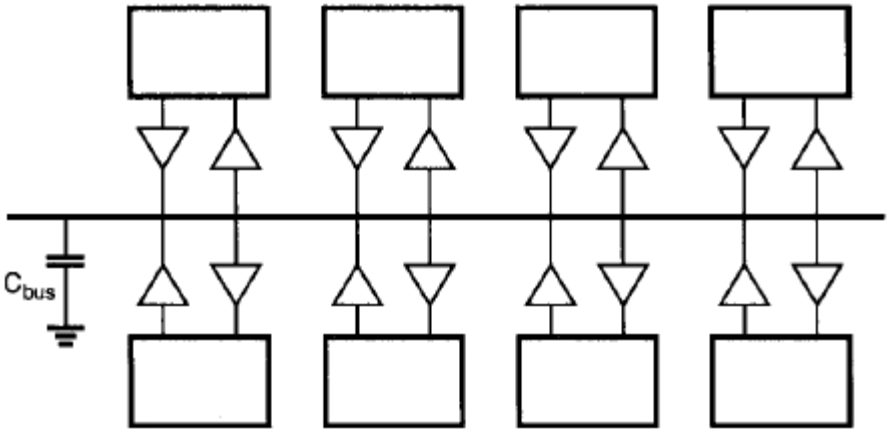
- Switching Capacitance is a key factor in Dynamic power consumption

$$P_{avg} = \frac{1}{2} C_{load} V_{DD}^{\ 2} f_{CLK} \beta$$

- At the system level, one approach to reduce the switched capacitance is to limit the use of shared resources, e.g., the driving bus
- In a system, a large number of drivers and recerivers may share the same bus, and the capacitance load of this bus would include all the capacitance that connected to it

Driving Buses: Reduction of Switched Capacitance (Cont.)

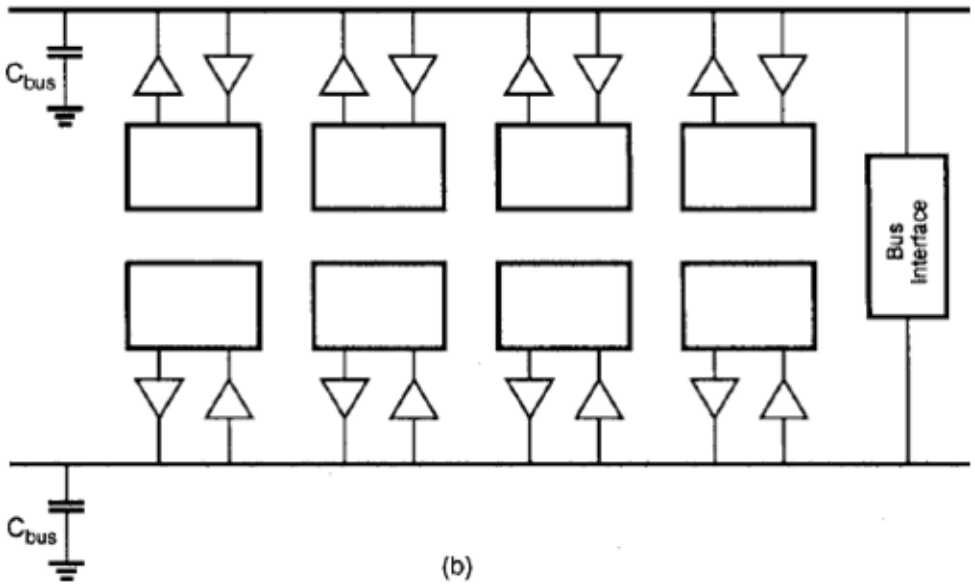
- For example, the following structure results in a large bus capacitance due to
  - -The large number of drivers and receivers sharing the same transmission medium (The bus)
  - -The parasitic capacitance of the long bus line



(a)

Bus Splitting

- To reduce the bus capacitance, the global bus structure can be partitioned into a number of smaller dedicated local buses to handle the data transmission between neighboring modules



- The bus capacitance is cut effectively into half, and a bus interface is involved to deal with the communication between buses.

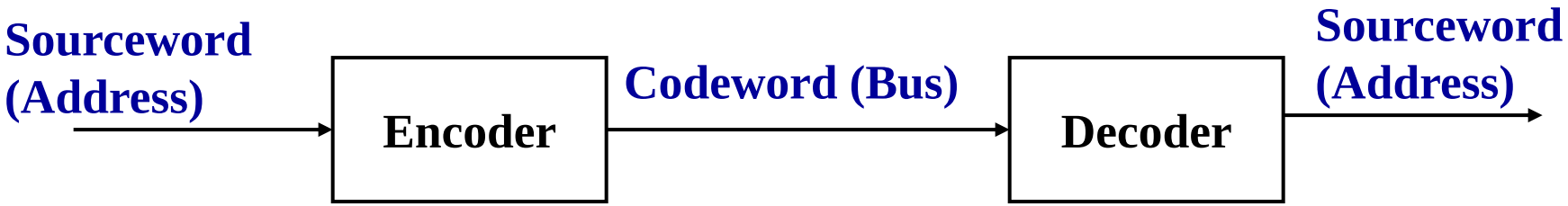
Bus Splitting Example

- Consider splitting a bus that connects 4 modules A, B, C and D into two half buses, one connecting A and B only, the other connecting C and D. The two buses are also connected thru a transfer bus with a controllable switch. The capacitance of the original bus is C_{BUS} while that of the each of the two half buses and the transfer bus is $C_{BUS}/2$. Suppose that only 10% of the data communication occurs between the (A, B) pair and the (C,D) pair. Calculate the percentage of power saving of the split bus architecture.
- Assume before and after splitting, the system work under the same clock frequency, 90% of the time only $C_{bus}/2$ would switch, and the other 10% all three buses are active so $3C_{bus}/2$ switches, base on the dynamic power consumption equation,

$$\text{Power saving} = \left(1 - \frac{\frac{C_{BUS}}{2} \times 0.9 \times V_{DD}{}^2 + \frac{3C_{BUS}}{2} \times 0.1 \times V_{DD}{}^2}{C_{BUS} \times V_{DD}{}^2}\right) * 100 = 40\%$$

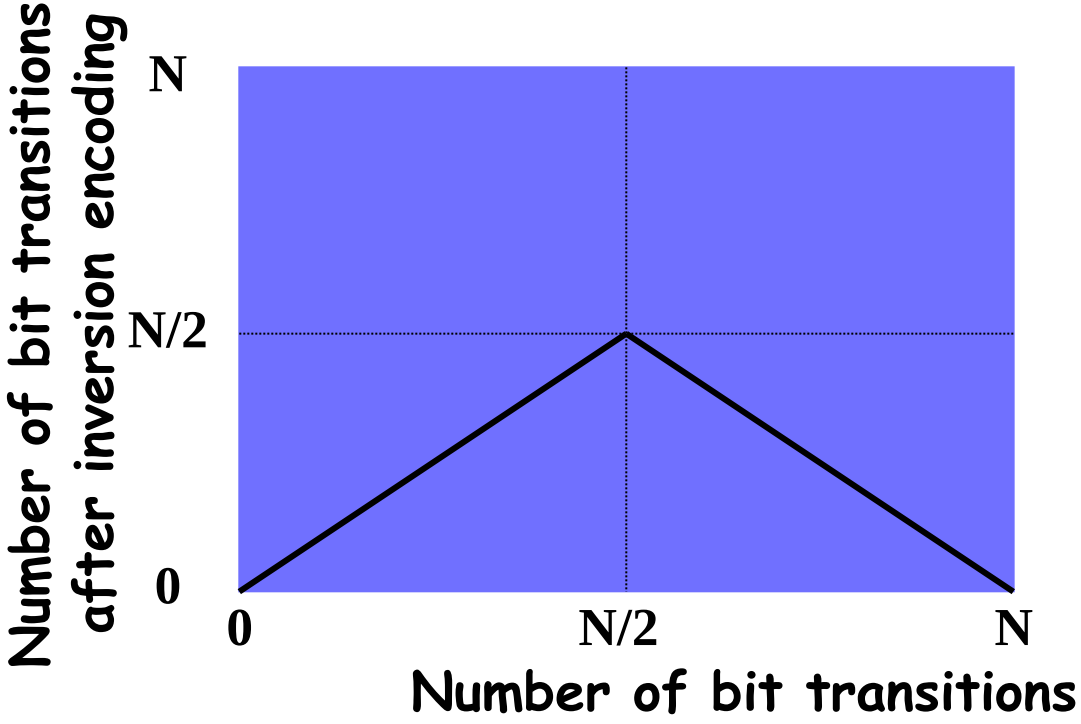
Bus Encoding

- The idea is to encode source words going on a highly capacitive bus so as to reduce switching activity on the bus



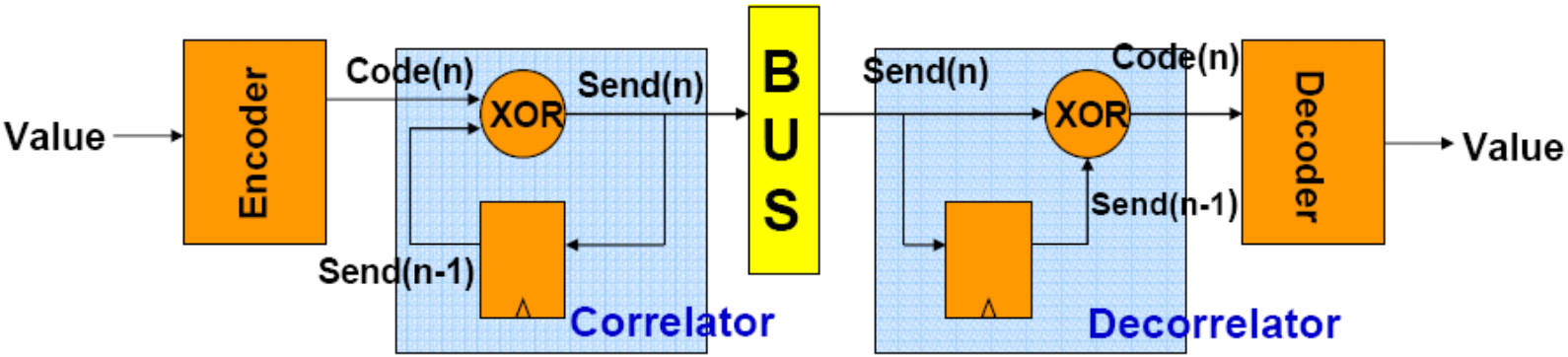
Bus Encoding Example

- For a 4-bit bus 0000 -> 1110 has three transitions. If bits of second pattern are inverted, then 0000 -> 0001 will have only one transition
  - At the receiver end, the data will be decoded back to 1110
- Bit-inversion encoding for N-bit bus:



Transition Signaling

- By XORing consecutive values on the bus switching can be reduced for high activity bus
  - for low activity bus, it may not be applicable



Send(n)=Send(n-1) XOR Code(n)
Code(n)=Send(n) XOR Send(n-1)

- The idea is by XORing, the number of back-to-back transition will be reduced

Transition Signaling Example

- Example – The first bit of the original sequence is placed on the bus – Transition signaling starts from the second bit onward

  Original sequence: 01010111010101

  activity factor=11/13

  Transition-signaled sequence: 01100101100110

  activity factor=8/13

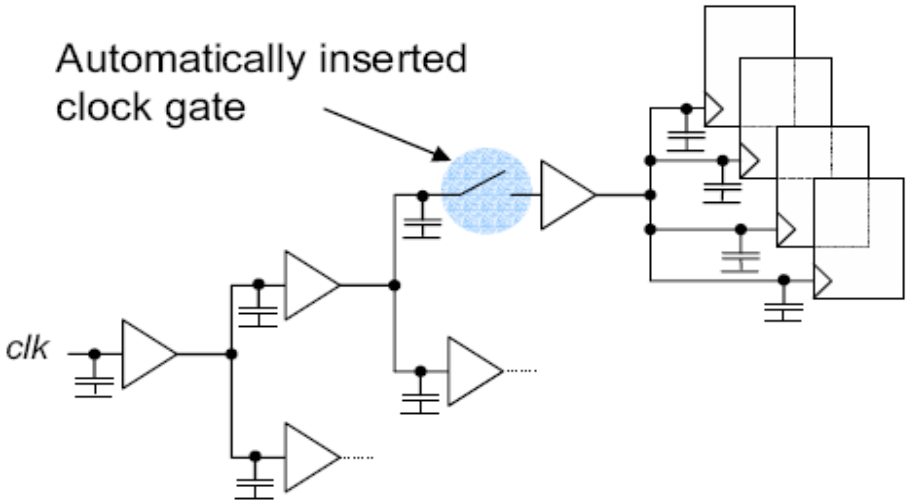
- Note:

  Transition-signaled(n) = Original(n) Xor Transition-signaled(n-1)

  Original (n) = Tansition-signaled(n-1) Xor Transition-signaled(n)

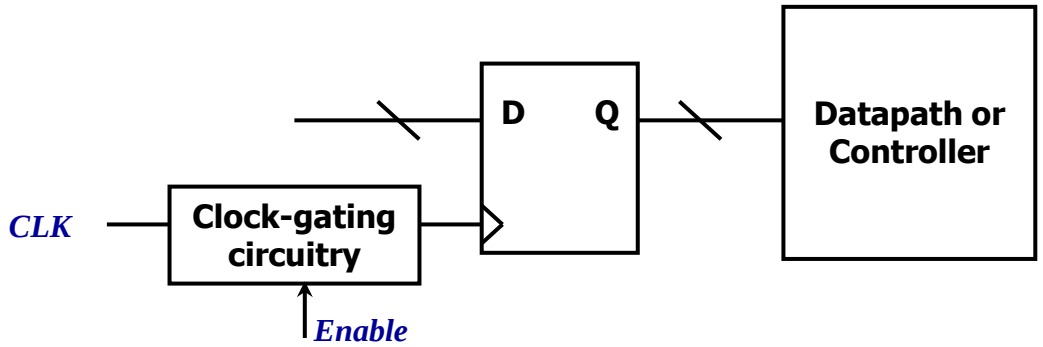
Clock Gating Principles

- Clock gating is an effective way to lower the power consumption by reducing the switching activity in CMOS logic circuits
- If certain logic blocks in a system are not immediately used during the current clock cycle, the clock signals of these blocks can be disabled to save the switching power



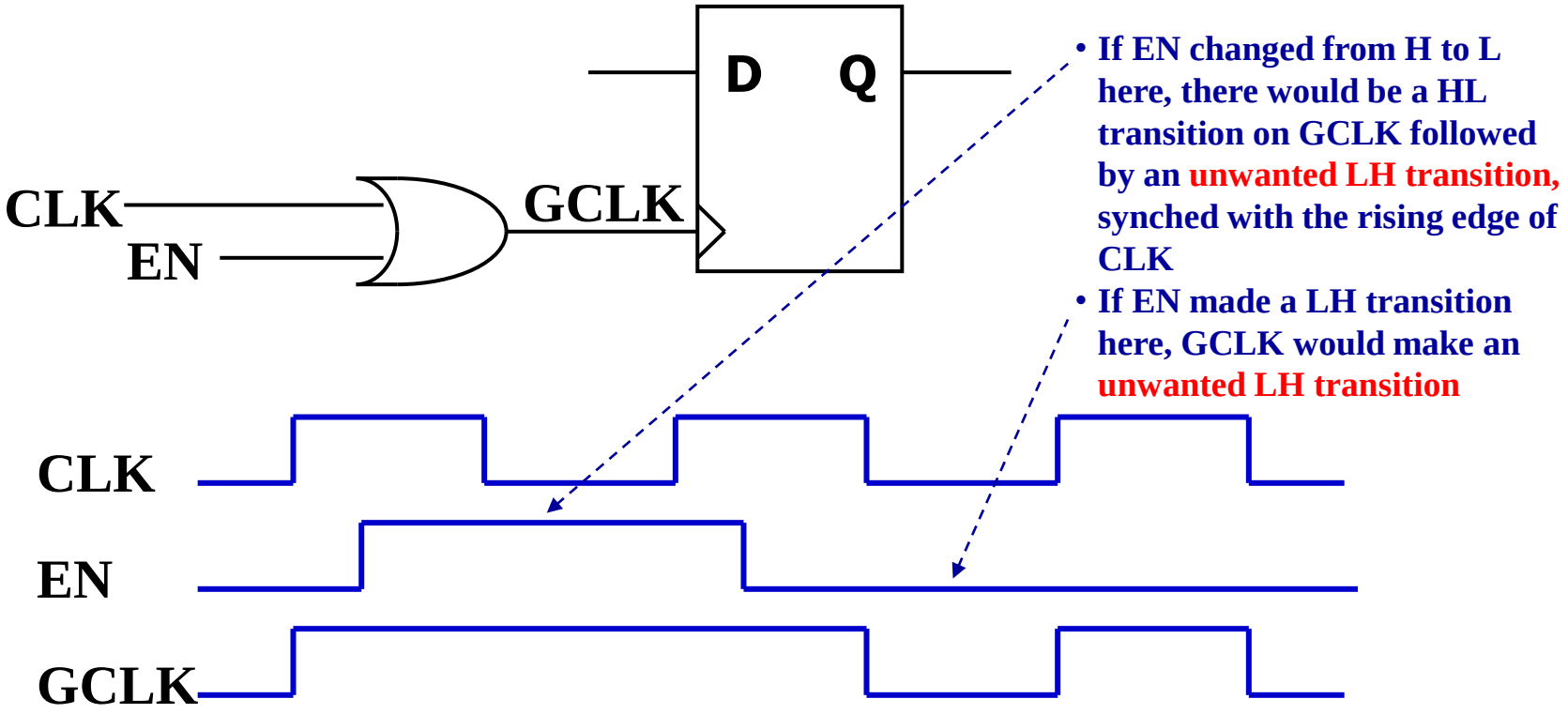
Clock is only running when required!

Clock Gating Principles (Cont.)



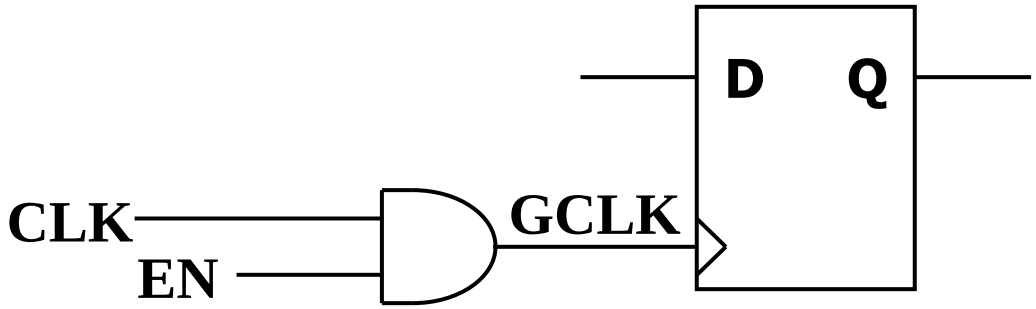
- Enable = 0 (mask the CLK) if
  - Q values remain steady compared to the previous clock cycle (Q holds its previous value)
    - Power is saved in the flip-flops
  - Or data path outputs are not needed
    - Power is saved in both the flip-flops and the data path
- Enable = 1 (do NOT mask the CLK) otherwise

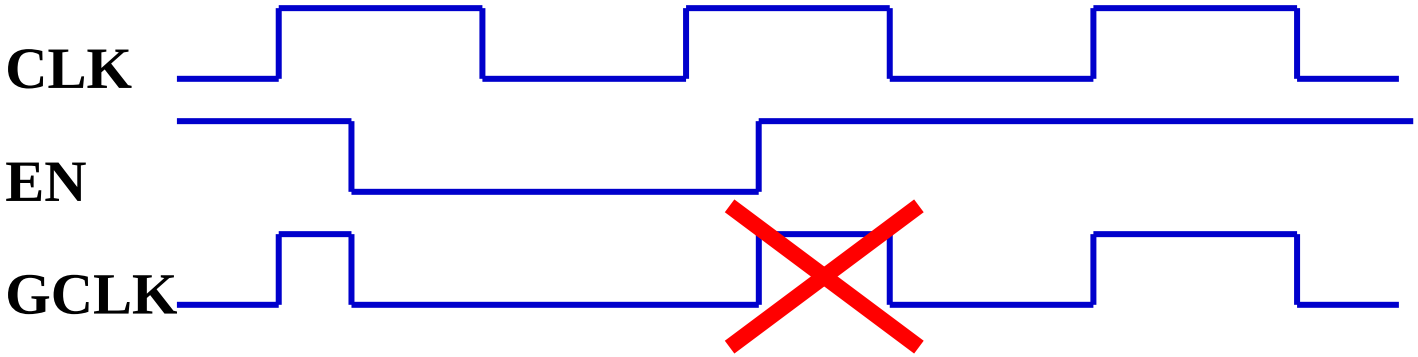
Clock Gating Using OR Logic



- Clock is enabled when EN = 0 (EN is an active low control signal)
- Necessary and Sufficient Condition: EN must stabilize within the first half of the clock cycle
- This is too restrictive a condition, and is hard to enforce in practice

Clock Gating Using AND Logic

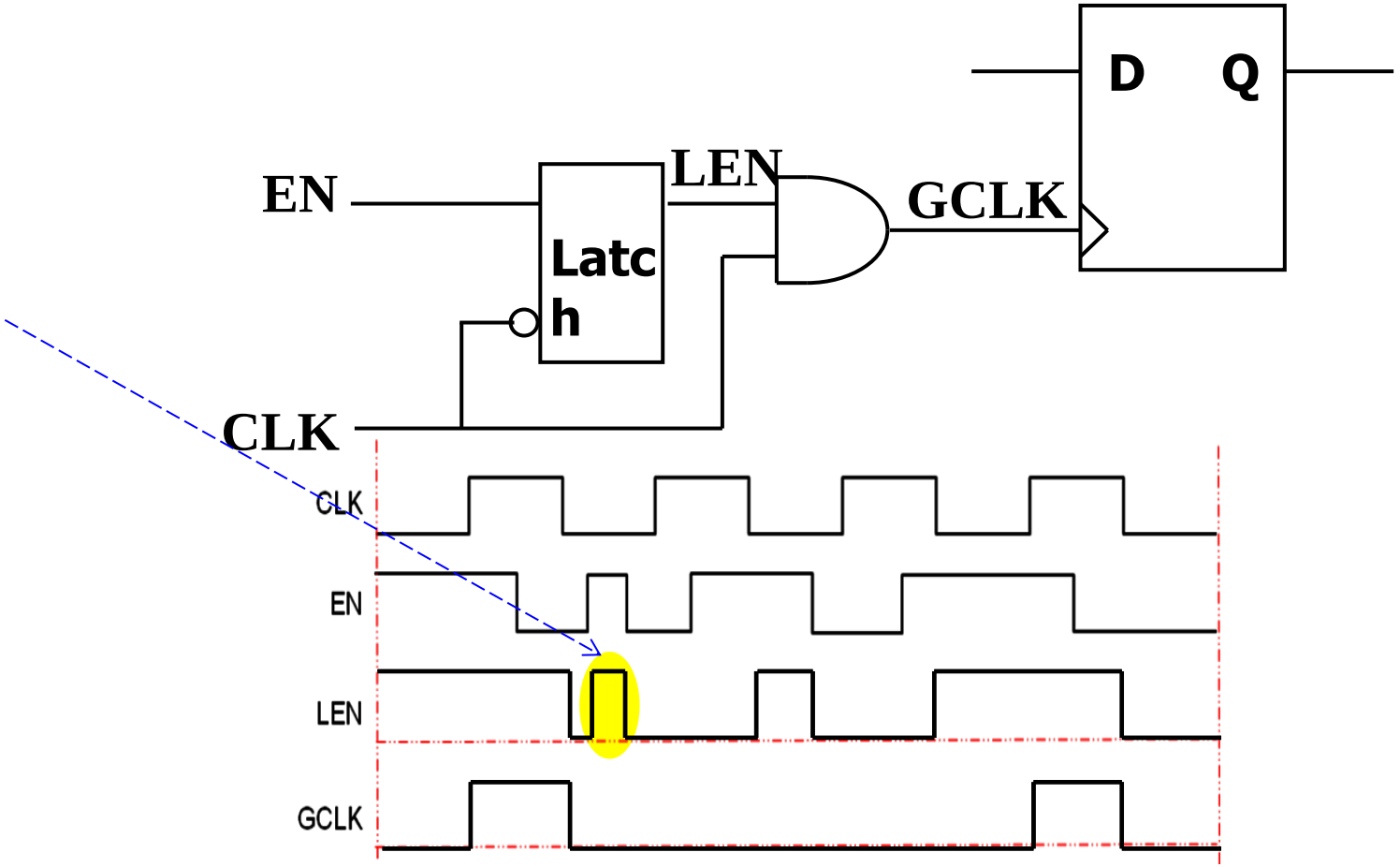




- Clock is enabled when EN = 1
- Glitch will appear even if EN stabilizes within the first half(clk = '1') of the clock cycle – this solution does not work

Clock Gating Using Negative Latch

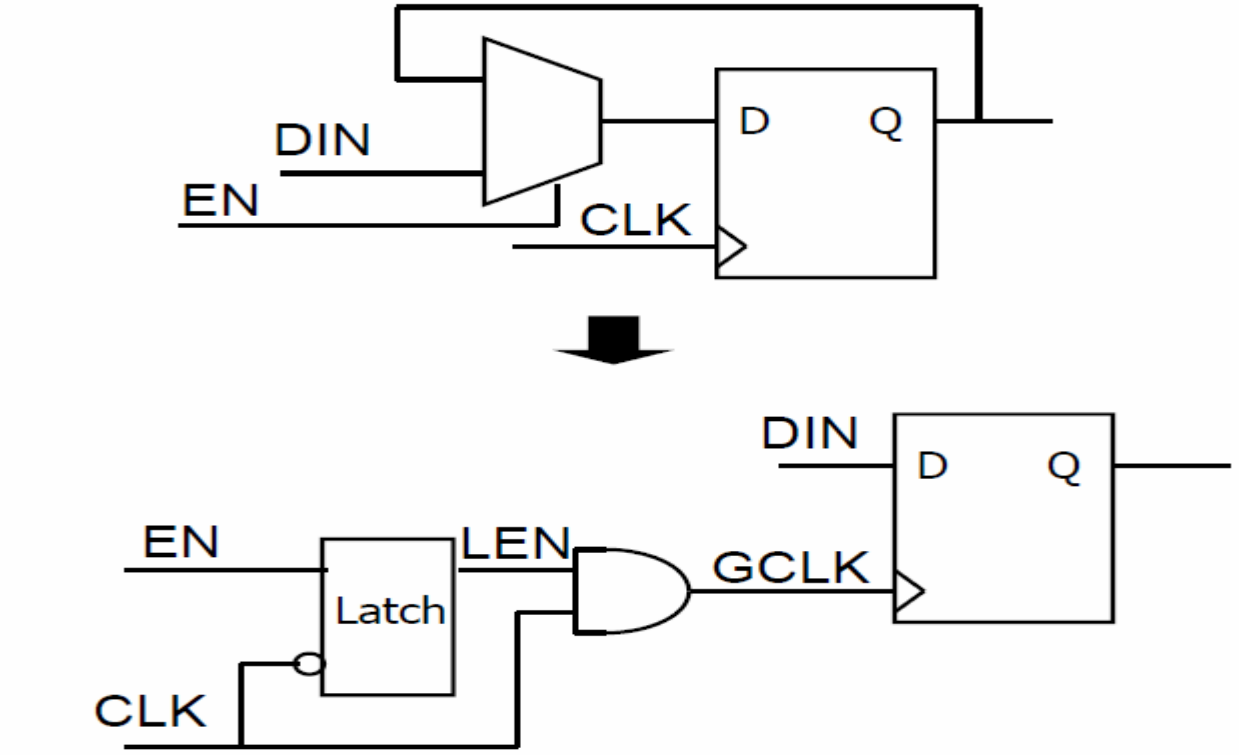
The negative latch may be replaced with a negative edge triggered D flip-flop. If so, the highlighted transition will be eliminated from LEN



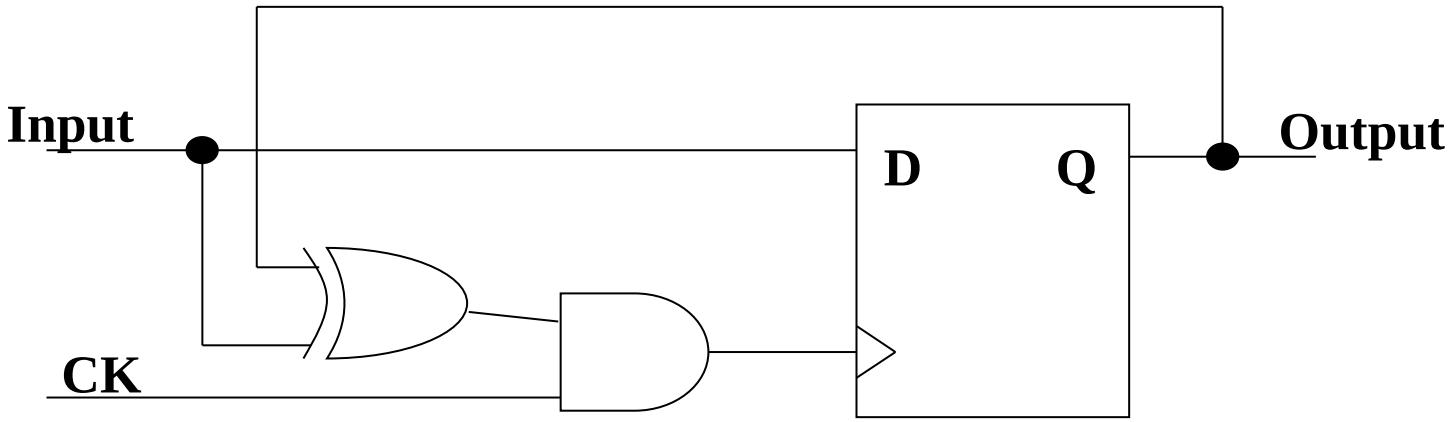
- Clock is enabled exactly if EN = 1 at the rising edge of the clock
- Any glitch on EN will be prohibited from reaching GCLK

Example Clock Gating: Register Substitution

- Clock-gating logic could be used when an RTL code such as "if (condition) out < = in" is present
- Power aware synthesis tools can identify RTL coding patterns and make the appropriate substitution



Autonomous Fine-Grain Clock-Gating

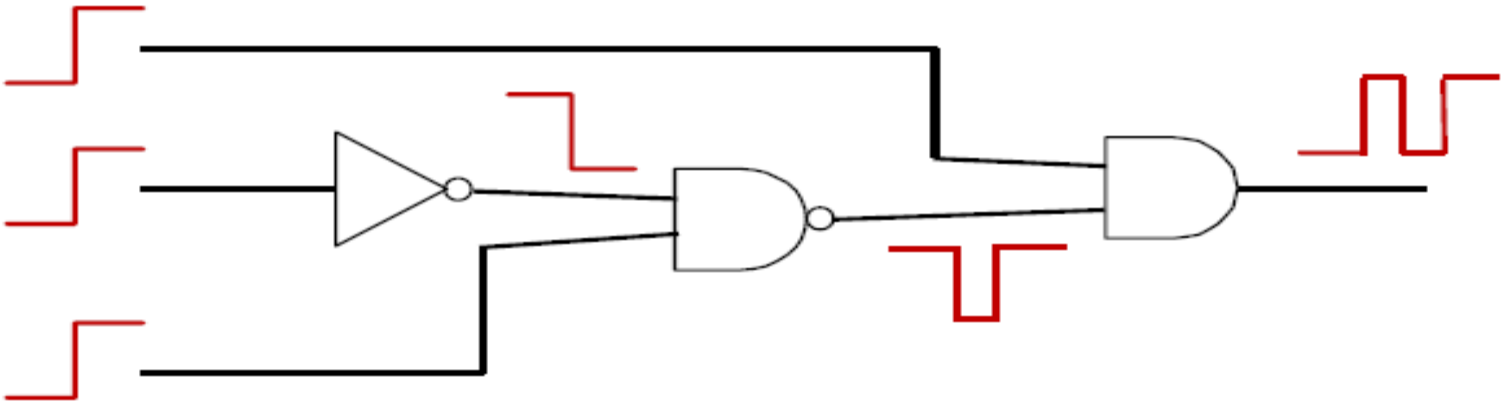


- Requires no chip-level clock enable signal
  - If next output state = present output state, then mask the clock
- The area overhead is high; there is also extra dynamic power dissipation when clock is going through. Because if the output of the DFF switches, the XOR gate may switch, thus cause switching power consumption
  - May integrate the XOR and AND gate inside the flop flop and reduce

Glitch Reduction

- The reduction of glitches is an important architecture-level measure to reduce switching activity.
- In multi-level logic circuits, the propagation delay from one logic block to the next can cause glitches, as a result of critical races or dynamic hazards. This is mainly due to a mismatch or imbalance in the path lengths in the logic network. Such a mismatch in path lengths results in a mismatch of signal timing with respect to the primary inputs

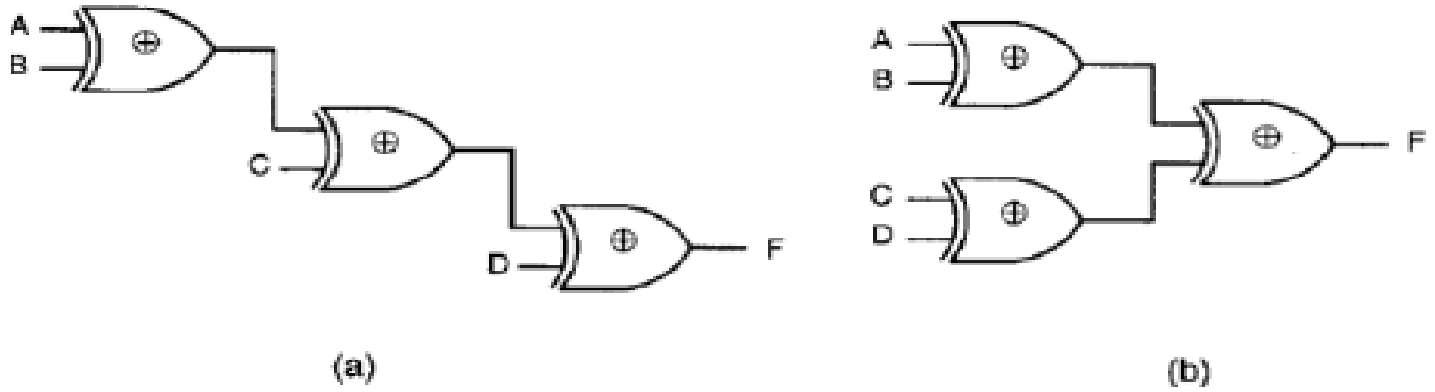
Glitch Reduction (Cont.)



- Three basic mechanisms for avoidance:
  - Use non-glicthing logic, e.g., domino
  - Add redundant logic to avoid glitching hazards
  - Adjust path delays in the design to avoid glitching hazards

Glitch Reduction (Cont.)

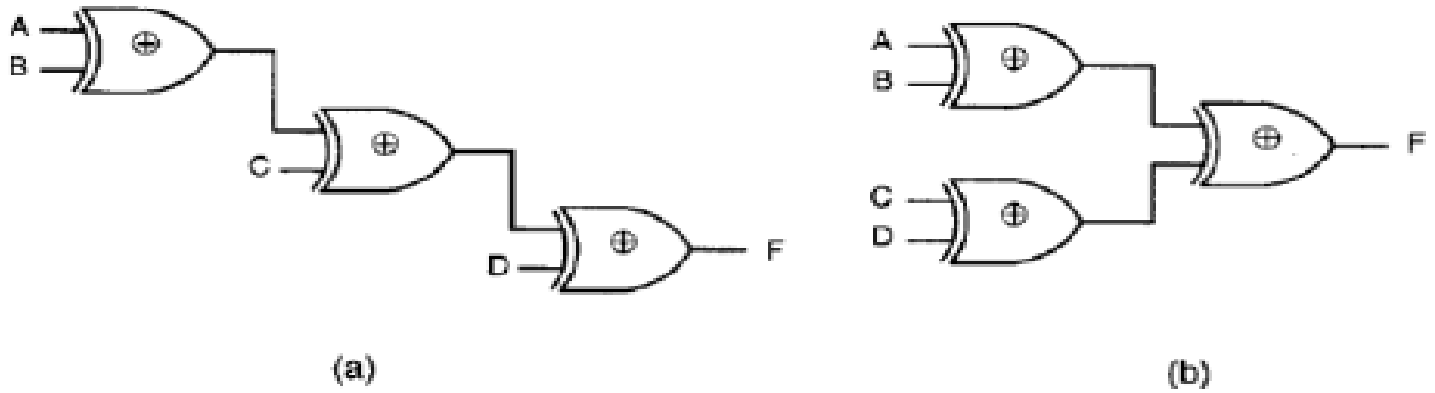
- Consider the simple parity network:



a) Implementation of a four-input parity (XOR) function using a chain structure

b) Implementation of the same function using a balanced tree structure which will reduce glitching transitions. This realization results in fewer hazards

Glitch Reduction (Cont.)



- In implementation (a), there will be glitches due to the wide disparity between the arrival times of the input signals. While in (b), on the other hand, all input arrival times are uniformly identical because the delay paths are balanced. So will reduce the glitches significantly

Adiabatic Logic Circuits

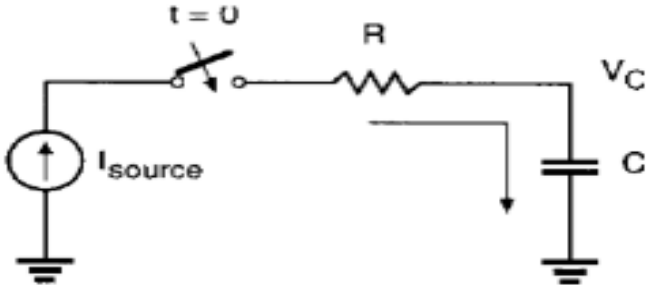
- In most of the cases, the energy drawn from the power supply is used only once before being dissipated. To increase the energy efficiency of logic circuits, recycling of the energy drawn from the power supply could be introduced
- A class of logic circuits called adiabatic logic offers the possibility of further reducing the energy dissipated during switching events, and the possibility of recycling, or reusing, some of the energy drawn from the power supply

Adiabatic Logic Circuits (Cont.)

- The term "adiabatic" is used to describe thermodynamic processes that have no energy exchange with the environment, and therefore, no energy loss in the form of dissipated heat
- Fully adiabatic operation of a circuit is an ideal condition which may only be approached asymptotically as the switching processes are slowed down
- In practice, energy dissipation associated with a charge transfer event is composed of an adiabatic component and a non-adiabatic componet, so reducing energy to zero, is not possible

Adiabatic Switching

- Constant-current source charging a load capacitance C, through a resistance R



- Assuming that the capacitance voltage Vc is zero initially, the variation of the voltage can be found as

$$V_c(t) = \frac{1}{C} \cdot I_{source} \cdot t$$

- $I_{source} = C\frac{V_c(t)}{t}$ is a constant

Adiabatic Switching (Cont.)

- The amount of energy dissipated in the resistor R from t=0 to t=T can be found as

$$E = R\int_0^T I_{source}^2 dt = R\, I_{source}^2\, T$$

- Dissipated energy during this charge-up transition can be expressed as

$$E = \frac{RC}{T} C V_{c,\max}^2$$

where Vc,max=Vc(T)

- Adiabatic logic circuits require non-standard power supplies with time-varying voltage, also called pulsed-power supplies

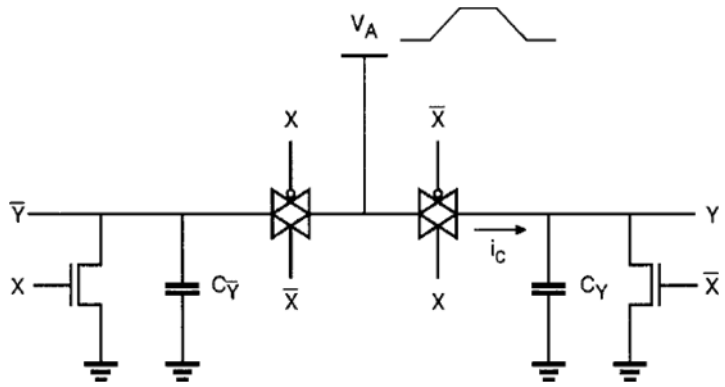
Adiabatic Logic Gates

- Adiabatic amplifier circuit which transfers the complementary input signals to it complementary outputs through CMOS transmission gates

-When the input signal X is set to a valid value, one of the two transmission gates becomes transparent.

- Next, the amplifier is energized by applying a slow voltage ramp VA, raising it from zero to VDD

- The load capacitance at one of the two complementary outputs (Y or Y) is adiabatically charged to VDD through the transmission gate while the other output (Y or Y) is kept at ground level

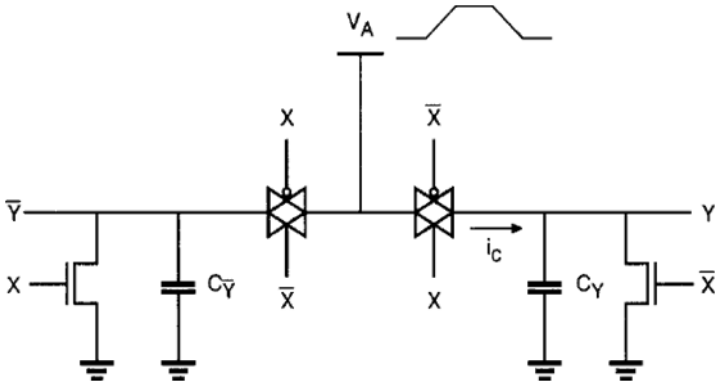


Adiabatic Logic Gates

-When the charging process is completed, the output signal pair is valid and may be used as an input to other similar circuits

- Next, the circuit is de-energized by ramping the voltage V_A back to zero

- Thus, the energy that was stored in the output load capacitance is retrieved by the power supply



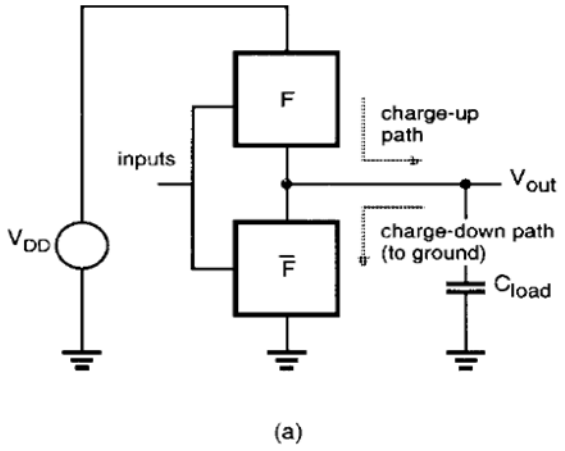
- Note that the input signal pair must be valid and stable throughout this sequence

Design of Adiabatic Logic Circuits

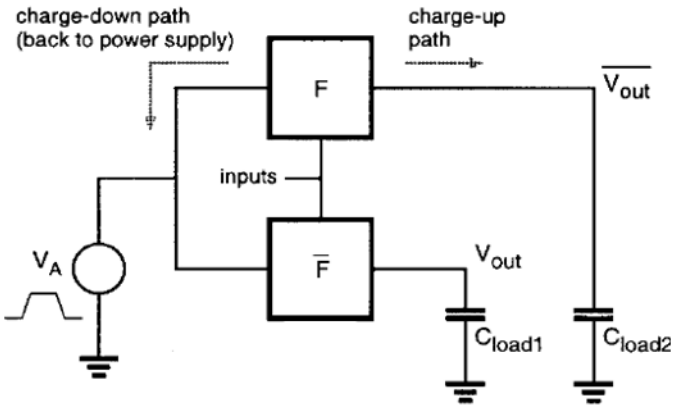
a) The general circuit topology of a conventional CMOS logic gate

b) The topology of an adiabatic logic gate implementing the same function

- Note the difference in charge-up and charge-down paths for the output capacitance



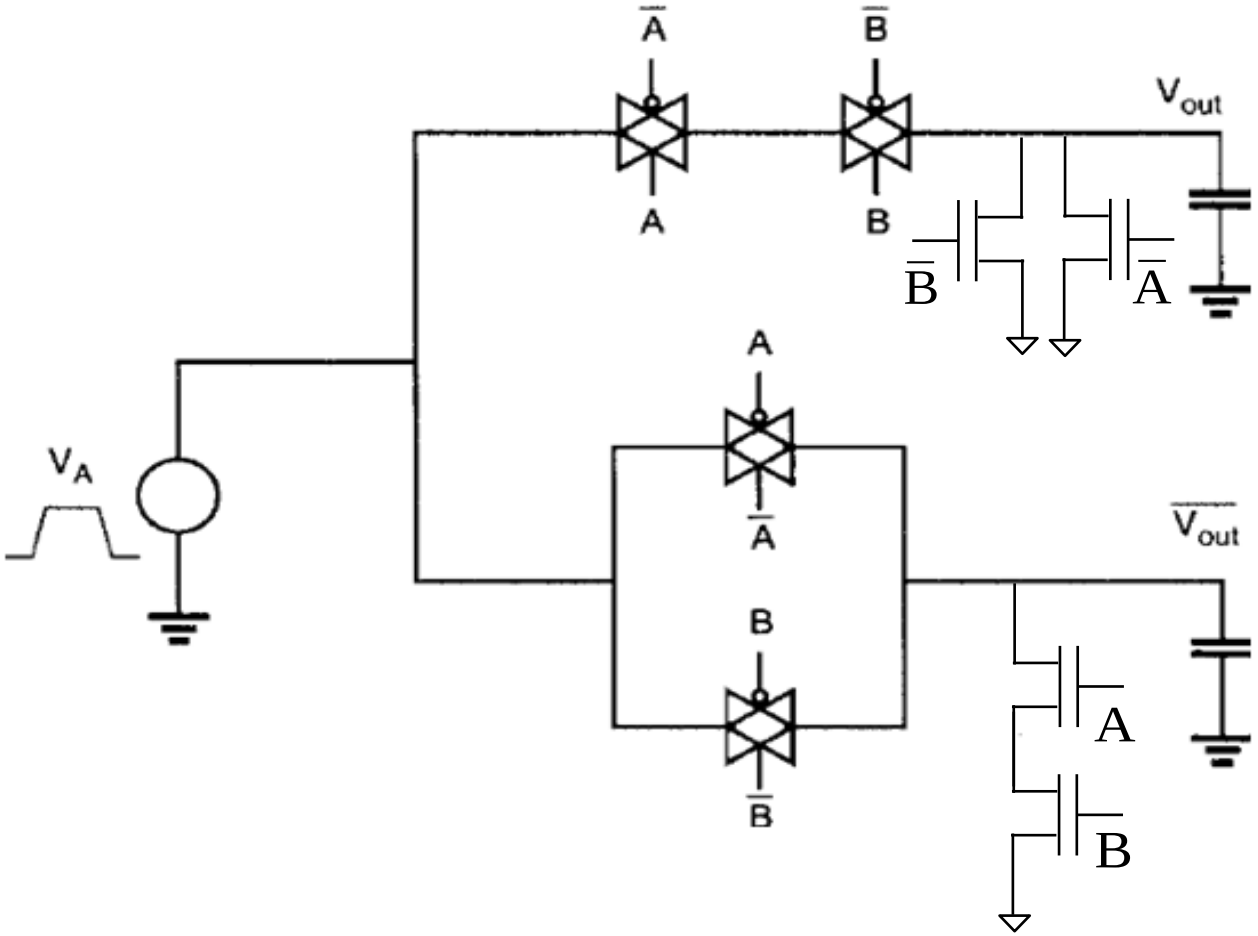
(a)



(b)

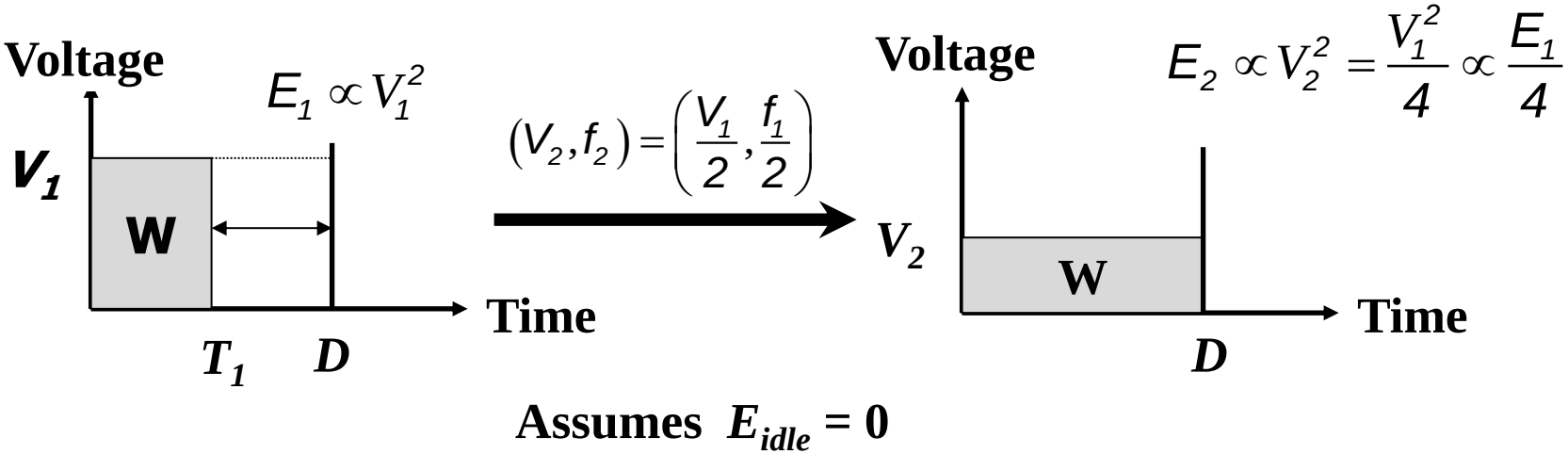
Example: Adiabatic AND/NAND Gate

- Circuit diagram of an adiabatic CMOS AND/NAND gate



Dynamic Voltage and Frequency Scaling (DVFS)

- DVFS is a method to provide variable amount of energy for a task by adaptively scaling the operating voltage and clock frequency of the chip based on its workload intensity
- Energy, E, required to run a task during $E = P \cdot T \propto V^2$
- Example: a task with workload W should be completed by a deadline, D

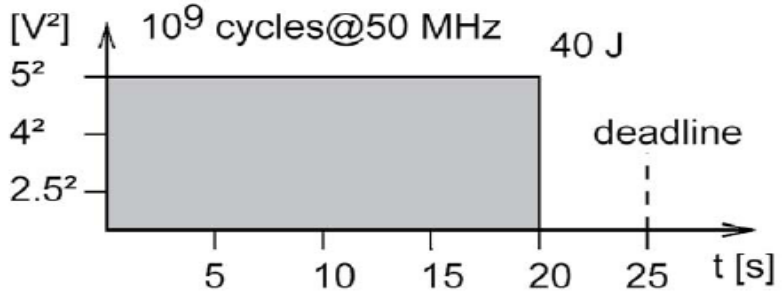


DVFS (Cont.)

V_{dd} [V]	5.0	4.0	2.5
Energy per cycle [nJ]	40	25	10
f_{max} [MHz]	50	40	25
cycle time [ns]	20	25	40

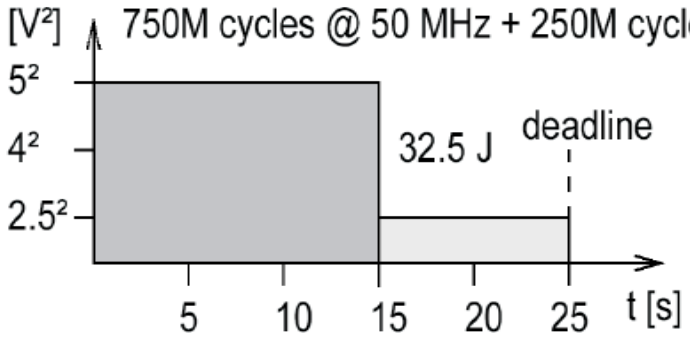
Processor with 3 voltages

a)



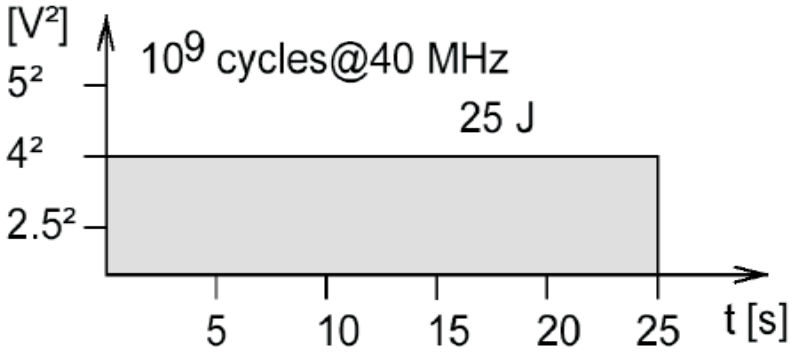
$E_a = 10^9 \times 40 \times 10^{-9} = 40$ [J]

b)



$E_b = 750\ 10^6 \times 40 \times 10^{-9}$
$+ 250\ 10^6 \times 10 \times 10^{-9}$
$= 32.5$ [J]

c)



$E_c = 10^9 \times 25 \times 10^{-9} = 25$ [J]

DVFS (Cont.)

- In the previous example:

  The deadline is 25ns. Therefore the slowest frequency the system can work at is 40MHZ

- Design (a) works at a higher frequency of

  50MHZ, and the voltage is 5v to meet the timing requirement

- Design (b) has two working clocks, 50MHZ and 25MHZ. Power is saved when working at 25MHZ

- Design (c) works at 40MHZ, which satisfies the deadline, so the lower voltage supply can be used to save power

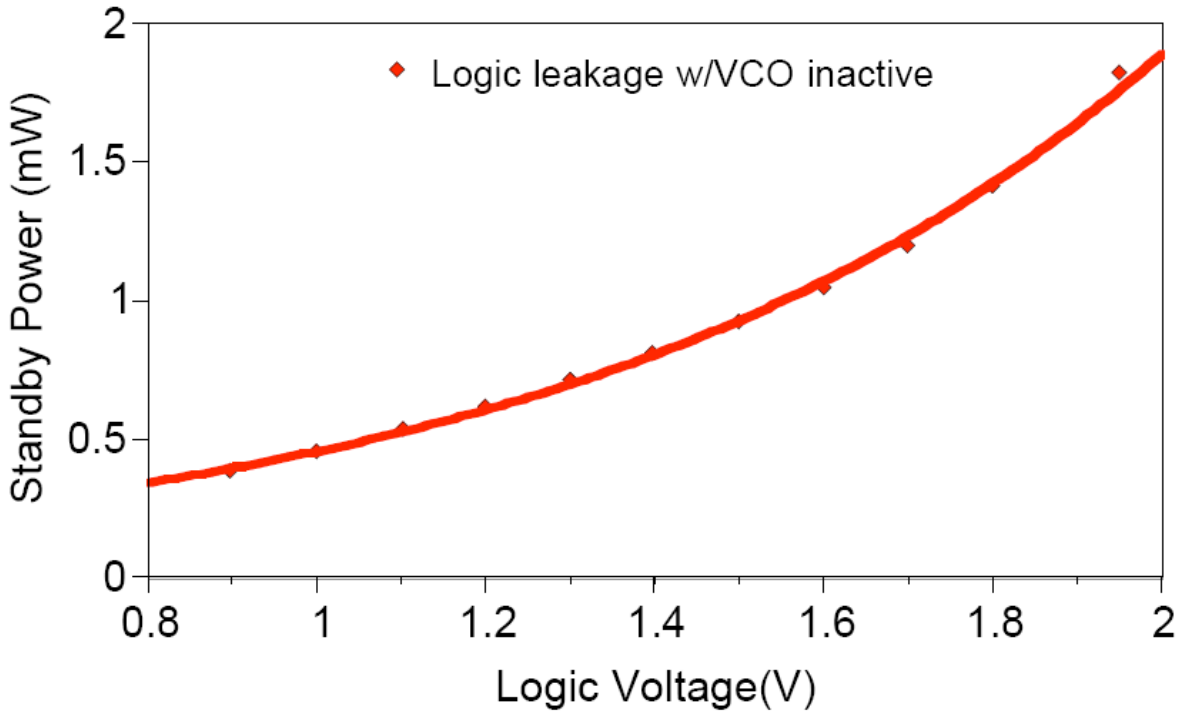
Basic Guidelines for Dynamic Power Optimization

- Do not do more than necessary
  - avoid wasteful power dissipation: clock gating
  - do not optimize for 'worst case' but for the 'current case': DVFS
  - react to the environment: DPM
- Use bus encoding, reduced swing signaling, etc.
- Use locality of reference
  - store results locally
  - avoid communication over long distances
  - avoid off-chip communications (1000 times more expensive)
- Be energy aware at all levels of your system: technological, system architecture, operating system, applications
- Do the tasks at the most energy-efficient platform/way
  - match algorithm with architecture

Leakage Power Dissipation – Background

- Gate
- Reverse-biased
- Subthreshold

Lowering and/or turning off VDD

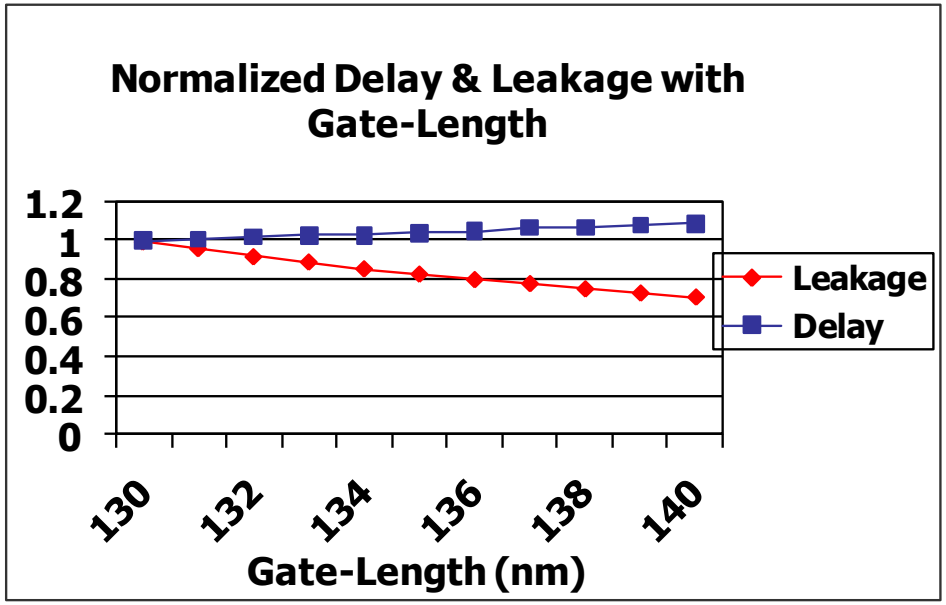


Subthreshold dominated technology

Source: Nowka, ISSCC-02

Gate Length Biasing

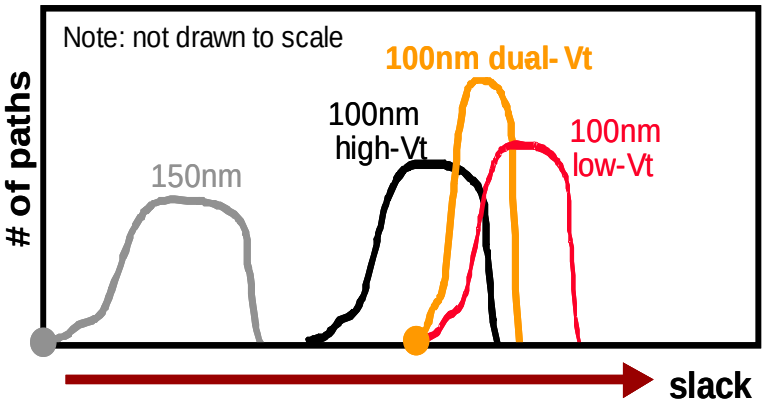
- Slightly increase (bias) the gate-length (line width) of devices
  - Slightly increases delay
  - Significantly reduces leakage
  - Bias only the non-critical devices

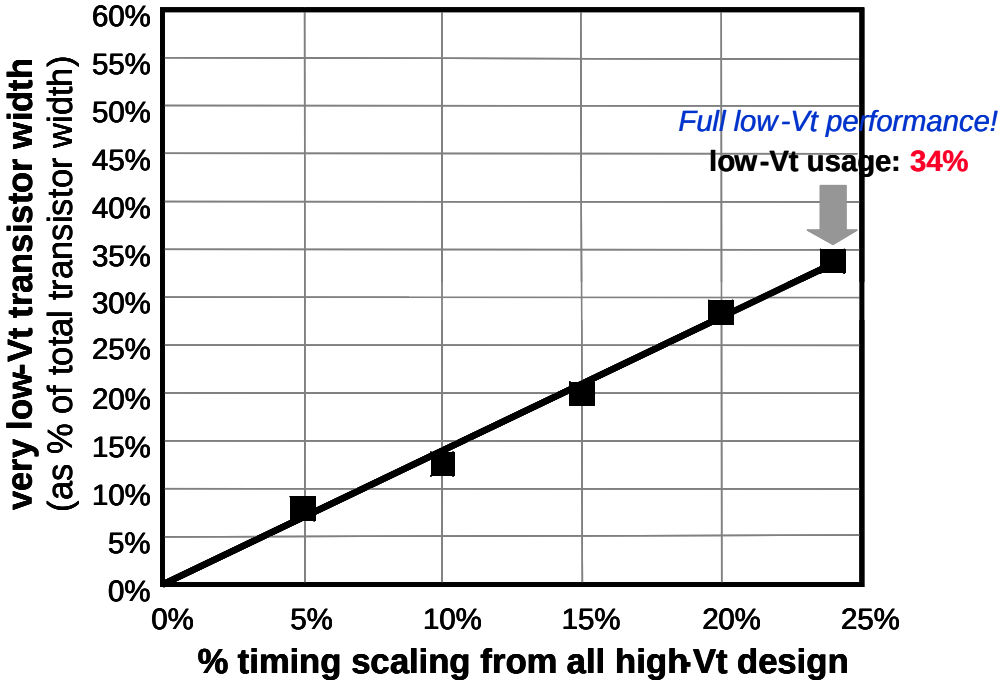


- Advantages:
  - Reduces runtime leakage and leakage variability
  - Can work in conjunction with V_{th} assignment → Gives finer control over delay-leakage tradeoff
  - Post-layout technique, no additional masks required
- 15-40% leakage and 30-60% leakage variability reduction for 90nm with dual-V_{th} assignment

Dual-V_{th} Design for Leakage Control

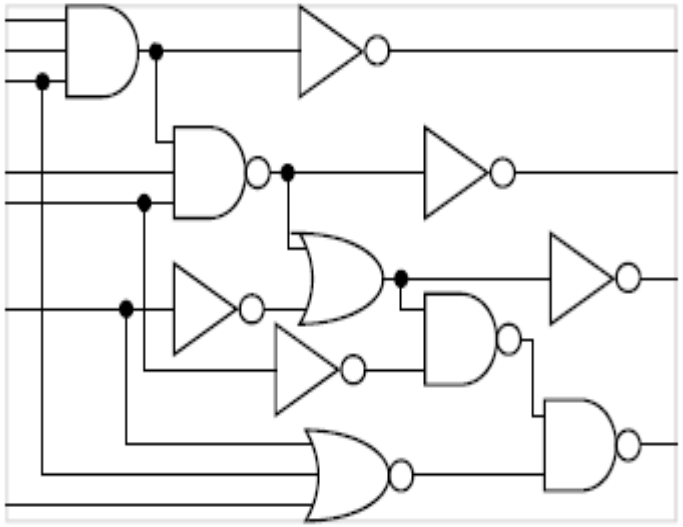
- Use low-V_{th} devices in CMOS gates that lie on timing critical paths of the circuit to maintain circuit speed while employing high-V+ devices in CMOS gates off the critical paths so as to minimize the subthreshold leakage





Dual-V_{th} Optimization Example

- The following circuit is designed in 65nm CMOS technology using low threshold transistors. Each gate has a delay of 5ps and a leakage current of 10nA. Given that a gate with high threshold transistors has a delay of 12ps and leakage of 1nA, optimally design the circuit with dual-threshold gates to minimize the leakage current without increasing the critical path delay.

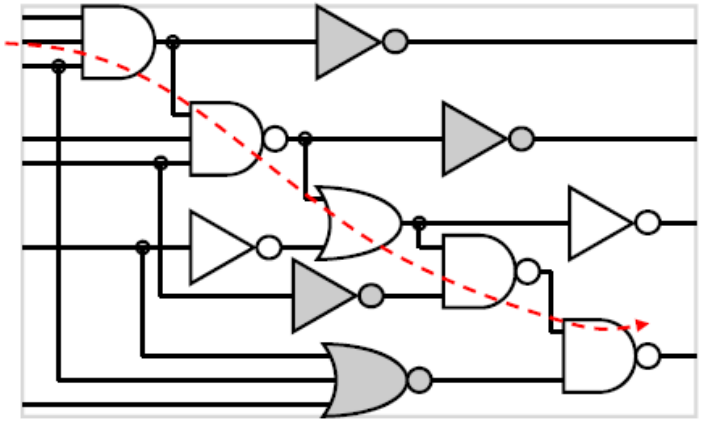


(a) What is the percentage reduction in leakage power?

(b) What will the leakage power reduction be if a 30% increase in the critical path delay is allowed?

Dual-V_{th} Optimization Example (Cont.)

- Part (a): Three critical paths are from the first, second and third inputs to the last output, shown by a dashed line arrow. Each has five gates and a delay of 25ps. None of the five gates on the critical path (red arrow) can be assigned a high threshold
- Also, the two inverters that are on four-gate paths cannot be assigned high threshold because then the delay of those paths will become 27ps. The remaining three inverters and the NOR gate can be assigned high threshold. These gates are shaded grey in the circuit
- The reduction in leakage power = $1 - (4 \times 1 + 7 \times 10)/(11 \times 10)$ $= 32.73\%$
- Critical path delay = 25ps

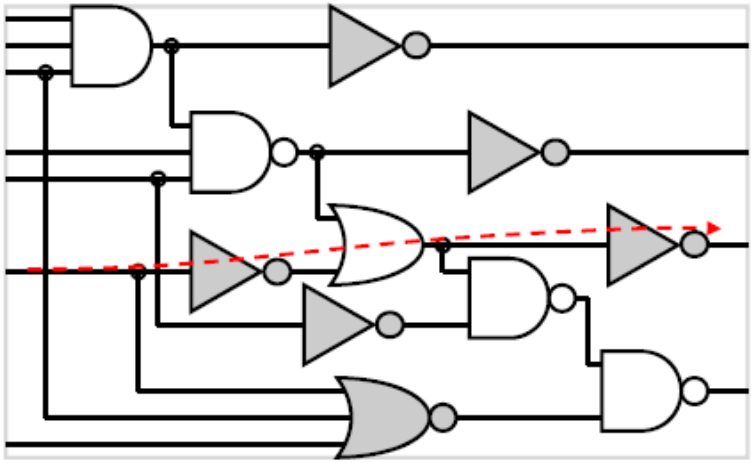


Dual-V_{th} Optimization Example (Cont.)

- Part (b): Several solutions are possible
- Notice that any 3-gate path can have 2 high threshold gates. Four and five gate paths can have only one high threshold gate. One solution is shown in the figure below where six high threshold gates are shown with shading and the critical path is shown by a dashed red line arrow. The reduction in leakage power =

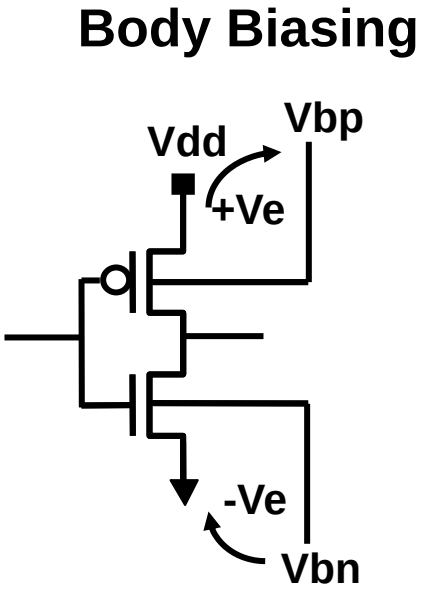
$$1 - (6 \times 1 + 5 \times 10)/(11 \times 10) = 49.09\%$$

- Critical path delay = 29ps

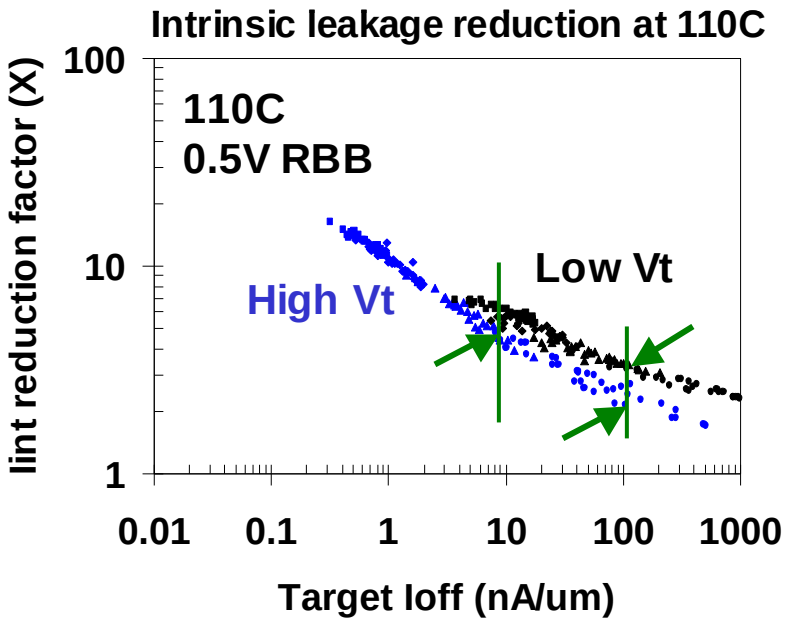


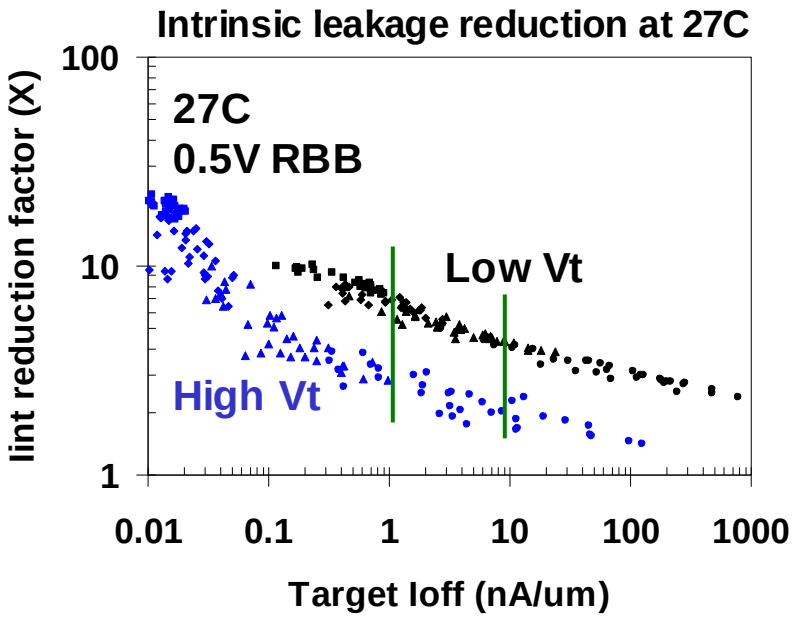
Reverse and Forward Body Biasing

- Reverse Body Biasing (RBB)
  - No Bias = Low V_{th} (Active mode)
  - Apply Reverse Bias = High V_{th} (Sleep Mode)
- Forward Body Biasing (FBB)
  - No bias = High V_{th} (Sleep Mode)
  - Apply Forward Bias = Low V_{th} (Active mode)
  - It is crucial to limit FBB to ensure that the source-bulk pn-junction remains in cut-off when FBB is applied



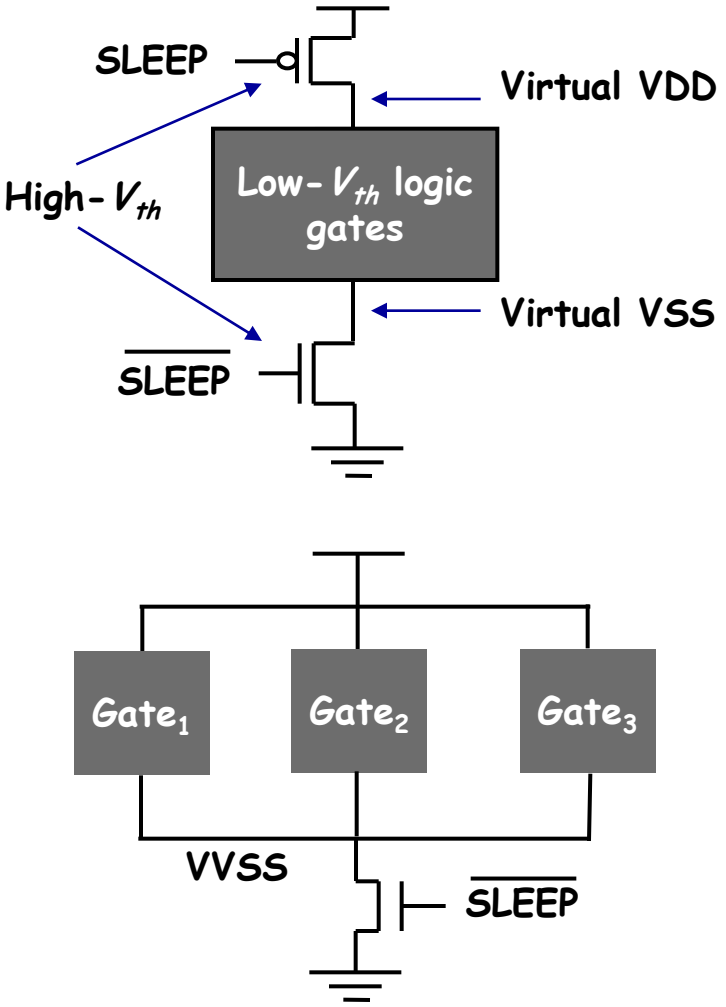
Reverse Body Bias (RBB)



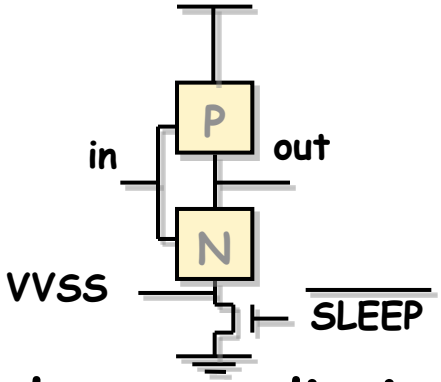


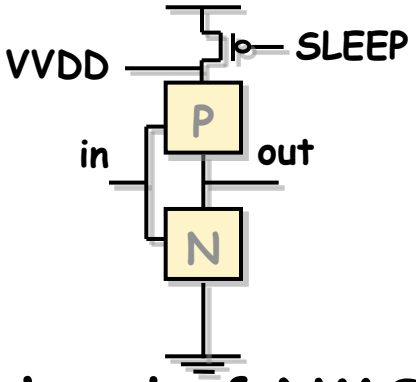
Power Gating (Multi-Threshold CMOS)

- High-V_{th} power switches are connected to low-V_{th} logic gates
  - Achieves high performance due to low-V_{th} logic gates
  - Reduces leakage power dramatically due to the series-connected high-V_{th} power switch
- Typically only a header or a footer sleep transistor is used, not both
- A single sleep transistor may be shared along several logic gates

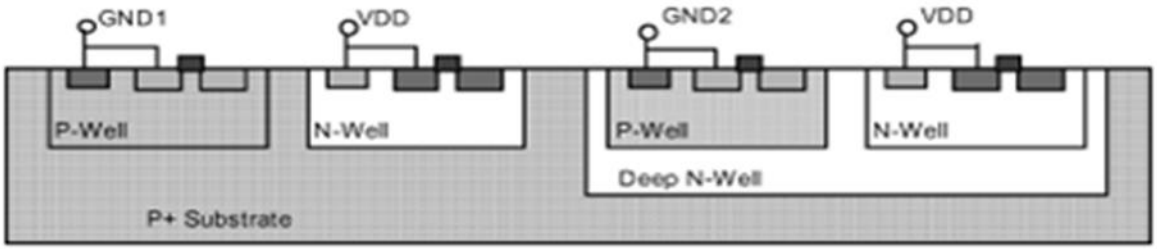


Header vs. Footer Switches





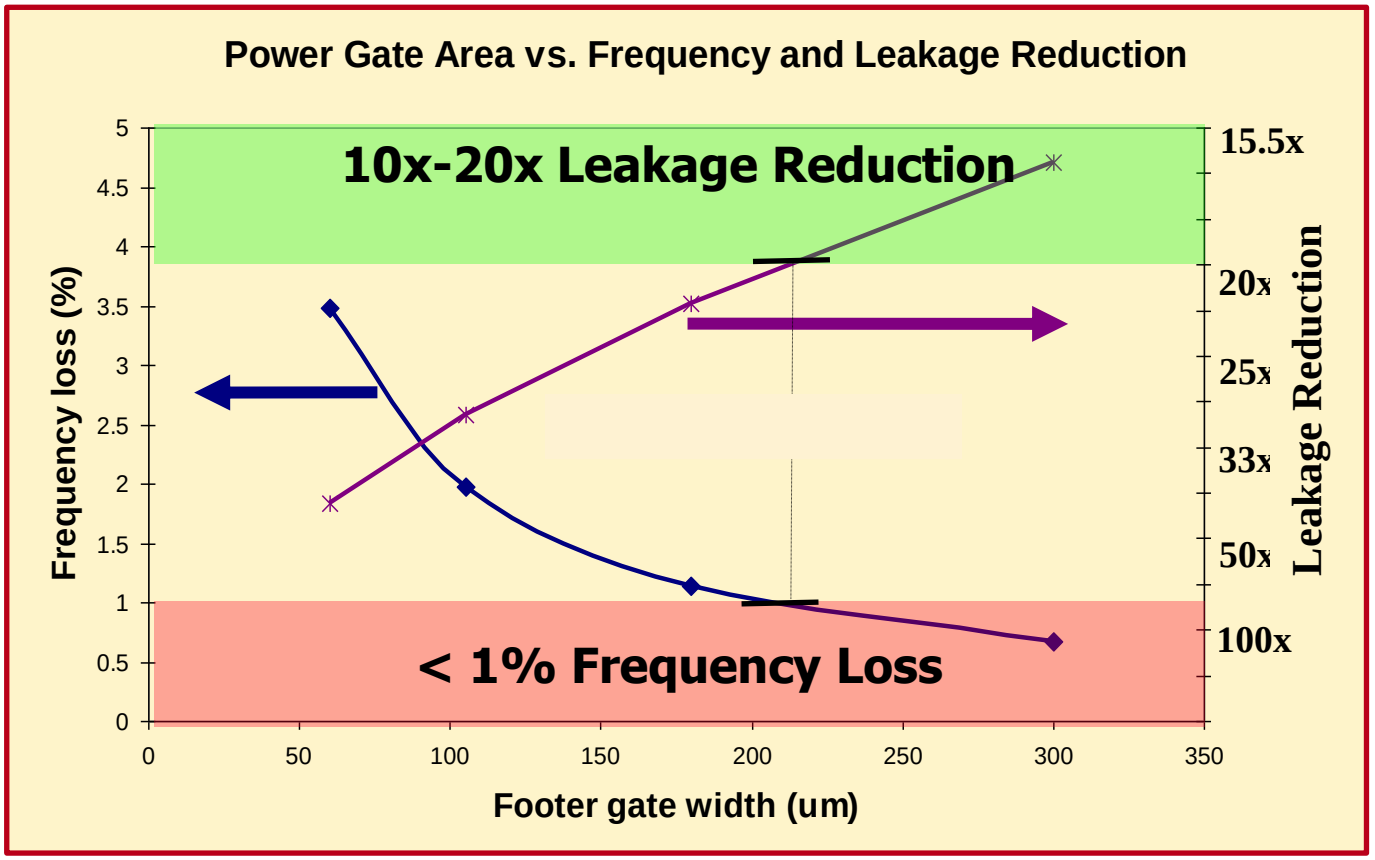
- Area and power dissipation overhead of NMOS footer transistors are lower due to higher mobility of electrons
- PMOS header transistors are more compatible with two-well bulk CMOS process where a high-performance NMOS transistor realized in the substrate is desirable



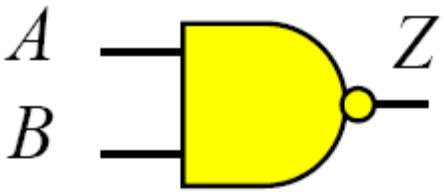
Triple well structure provides an isolating N layer between the local P-well and the P-substrate.

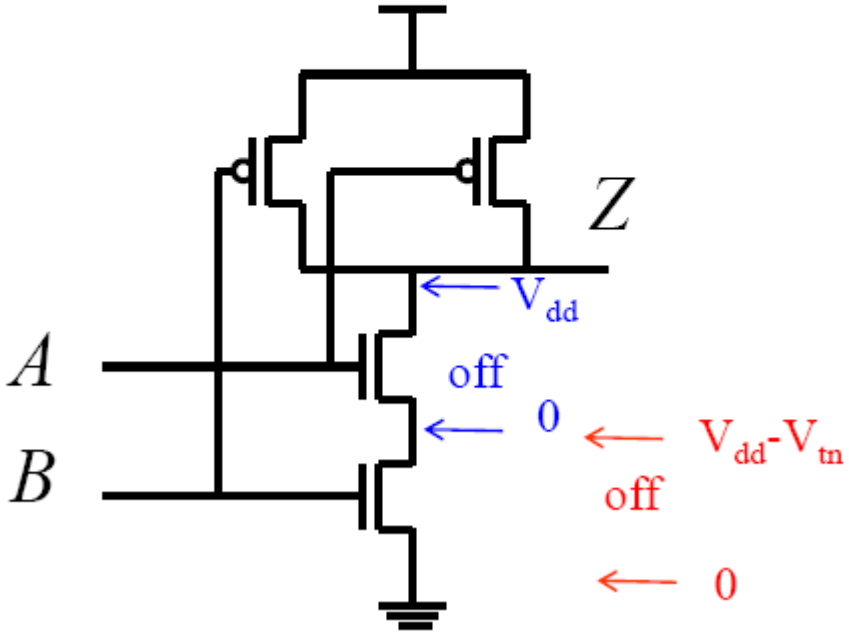
Footer Gate Width Selection

A 2-stage pipelined 40-bit ALU (IBM)



State Dependence of the Leakage Current: NAND2





TSMC CLN65LP (tcbn65gplustc)

TC, 1.0V, 25 °C

Cell: ND2D0, Cell area = 1.44λ^2

Pin input capacitance = 0.8 fF

2.4fF input cap, 4ps input trans. time:

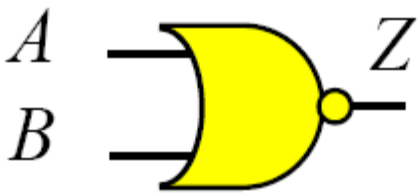
τ_{pHL}=26.3ps , Tfall=38.5ps

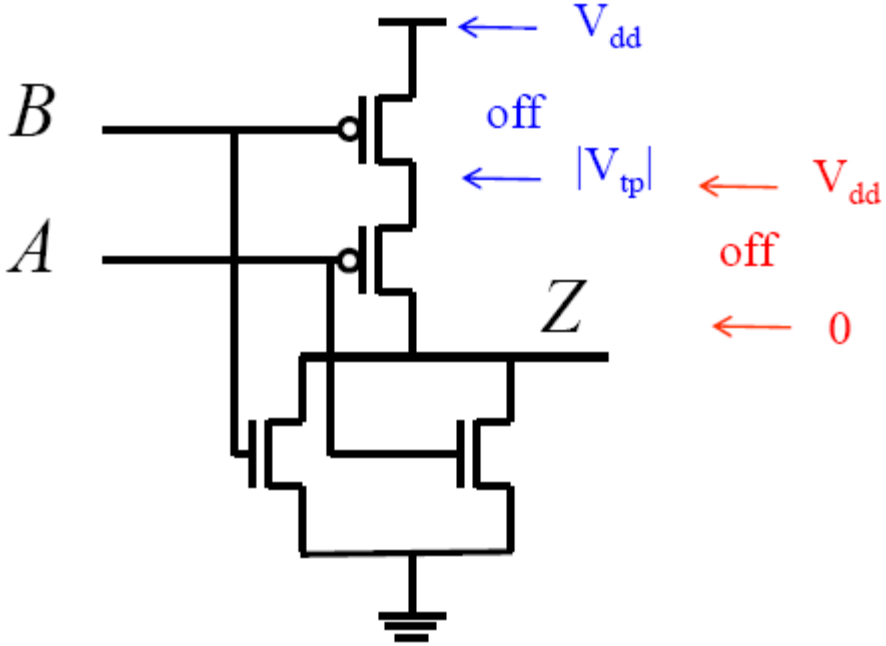
τ_{pLH}=22.2ps , Trise=34.5ps

$E_{SC,fall}$ = 0.1 fJ, $E_{SC,rise}$ = 1.1 fJ

A B	Leakage
0 0	2.281 nW
0 1	4.029 nW
1 0	2.185 nW
1 1	7.713 nW

State Dependence of the Leakage Current: NOR2





TSMC CLN65LP (tcbn65gplustc)

TC, 1.0V, 25 °C

Cell: ND2D0, Cell area = 1.44λ^2

Pin input capacitance = 0.8 fF

2.4fF input cap, 4ps input trans. time:

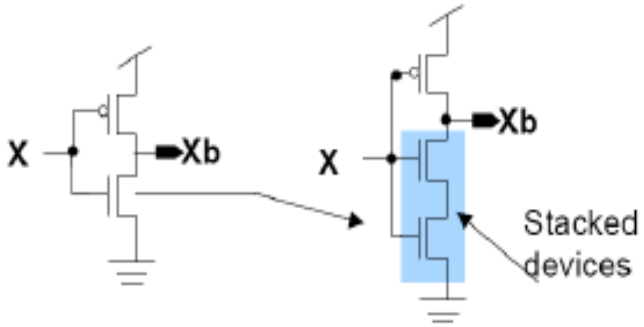
t_{pHL}=14.8ps , Tfall=21.0ps

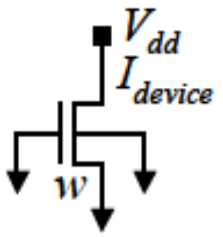
t_{pLH}=42.9ps , Trise=68.0ps

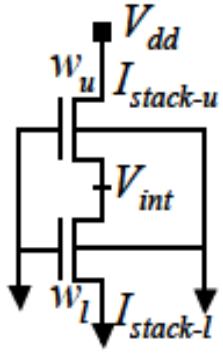
$E_{SC,fall}$ = 0.0 fJ, $E_{SC,rise}$ = 1.1 fJ

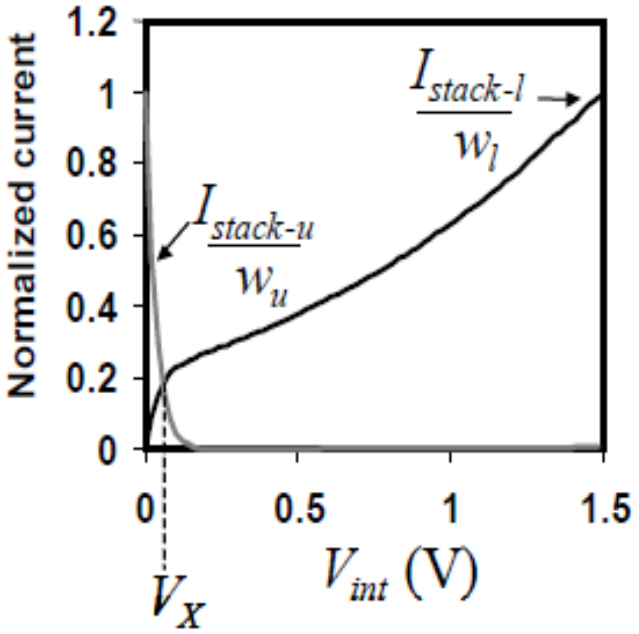
A B	Leakage
0 0	5.533 nW
0 1	2.882 nW
1 0	4.704 nW
1 1	2.924 nW

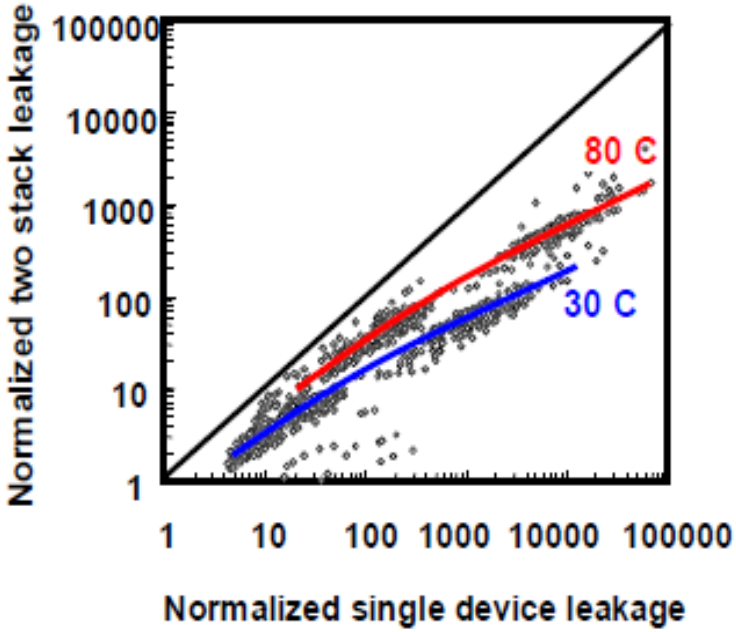
Stacked Transistors











Source: De-2004

Basic Guidelines for Leakage Optimization

- Exploit advances in CMOS technology and manufacturing processes
  - high-k dielectric and metal gates for the gate stack
  - low-k inter-layer dielectric for metal interconnect
- Go to sleep as soon as possible
  - avoid standby power dissipation in active mode: power gating
  - react to the environment: DPM
- Operate at minimum possible voltage and maximum threshold voltages
  - lower supply voltage results in lower active mode leakage
  - higher threshold voltages result in lower leakage in active and sleep modes