High-Level Synthesis and Low Power Design

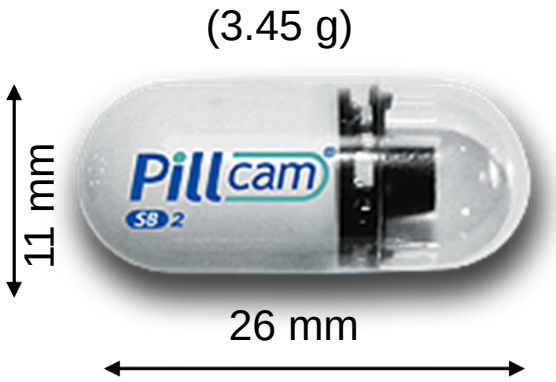
- *CMOS Digital Integrated Circuits - Analysis and Design* (3rd Edition)
  Sung-Mo (Steve) Kang and Yusuf Leblebici
  McGraw Hill

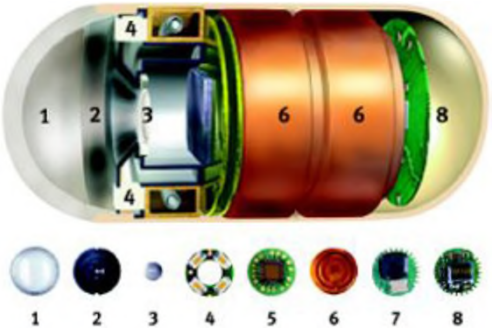
- *Digital Integrated Circuits*
  Jan M Rabaey
  Prentice Hall

- *Application-Specific Integrated Circuits*
  Michael J S Smith
  Addison-Wesley

Why Low Power Design?

- Portability:
  - Size and weight of battery pack: Heavy battery pack is not practical and restricts amount of battery power that can be loaded at any one time.
  - Recharging Interval: frequent recharging poses inconvenience and reduce user's satisfaction in using the product.
  - Battery capacity has only increased by a factor of two to four in the last 30 years or so, while computational power of digital ICs has increased by more than four orders of magnitude.



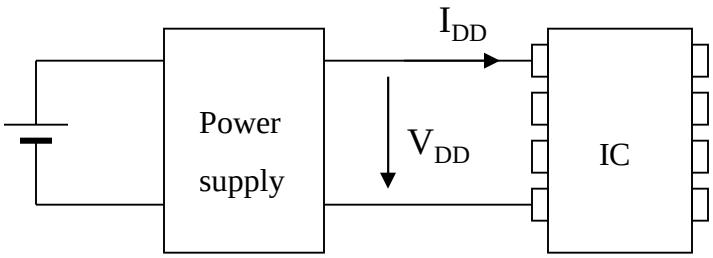


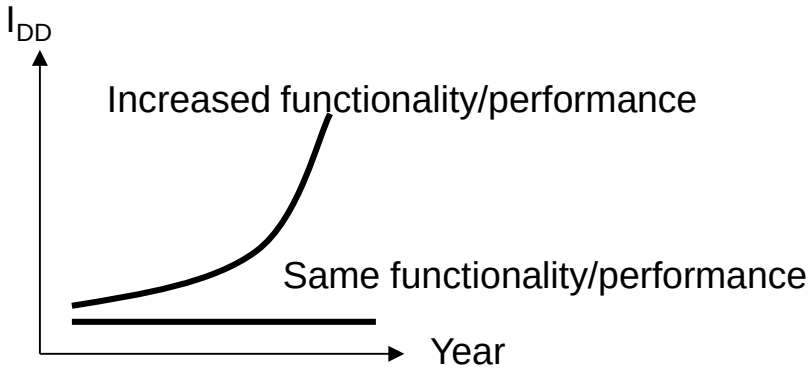
Why Low Power Design?

- Reliability and Manufacturing Cost:
  - Rapidly increasing integration density, clock frequency and computational power of μP results in rising power dissipation.
  - Power consumption of μPs has increased almost linearly with area-frequency product over the years, e.g., DEC21164, die area = 3 cm^2 and runs on 300 MHz clock dissipates 50W of power.
  - Expensive packaging and cooling techniques are needed as insufficient cooling leads to high operating temperature, which tends to exacerbate several Si failure mechanisms.
- Global Awareness of Environmental Concerns:
  - Global warming and environmental pollution are directly or indirectly associated with the increasing percentage of electrical energy usage for computing and communication in modern workplace.

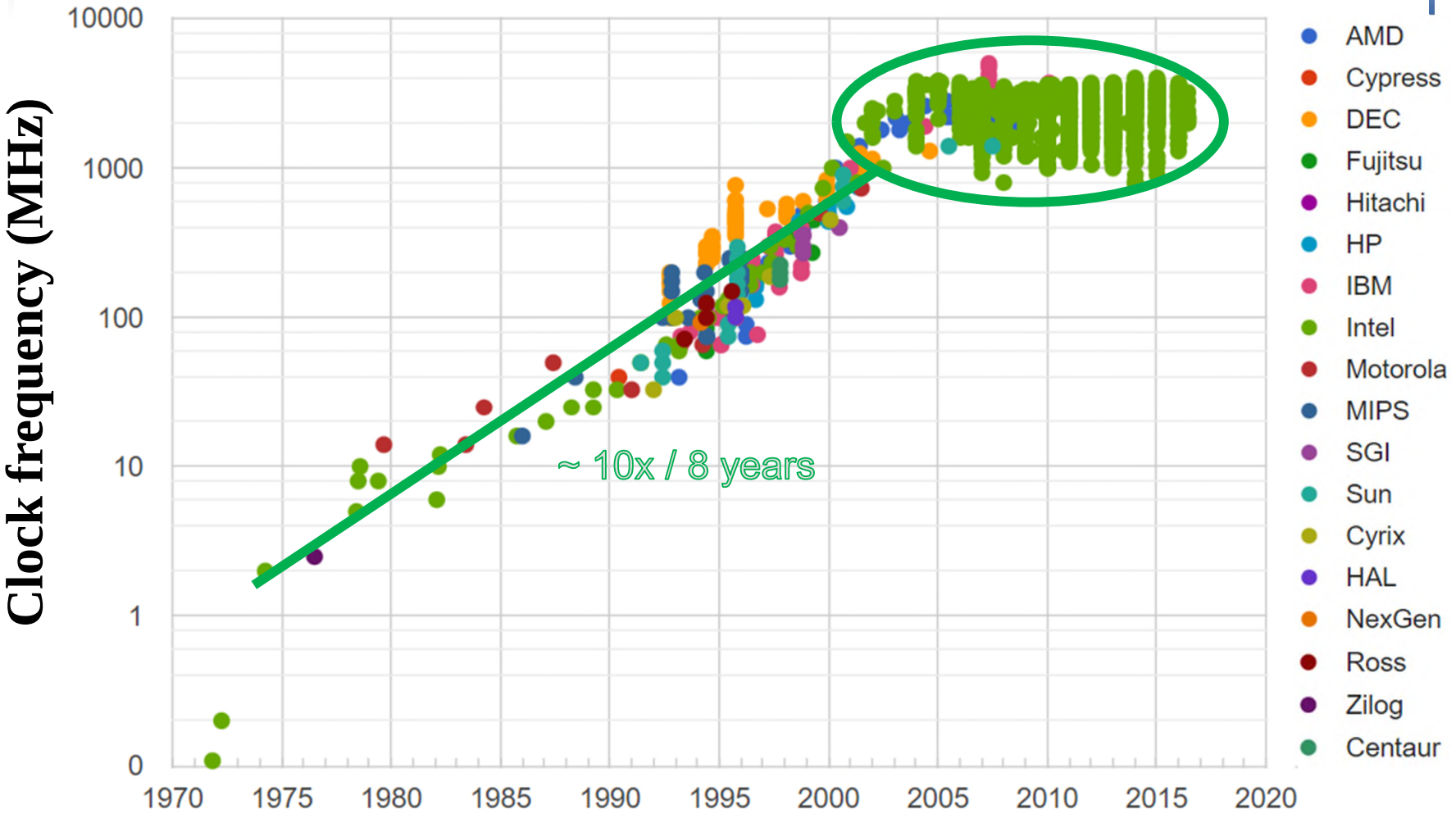
Challenges?

- Reduce the weight of battery and extend battery life before recharging.
- Conflicting objectives. Performance improvement, versatile functionalities, multimedia applications and high-density integration causes a drastic increase in power dissipation due to increased operating frequency and increased chip size.
- Battery-less applications like sensor nodes, smart card and endorscopy capsules may rely on weak and limited harvested power.
- Heat dissipation: a problem of not only handheld devices and mobile terminals, but also high-performance desktop systems and data center.
- Production yield and reliability.



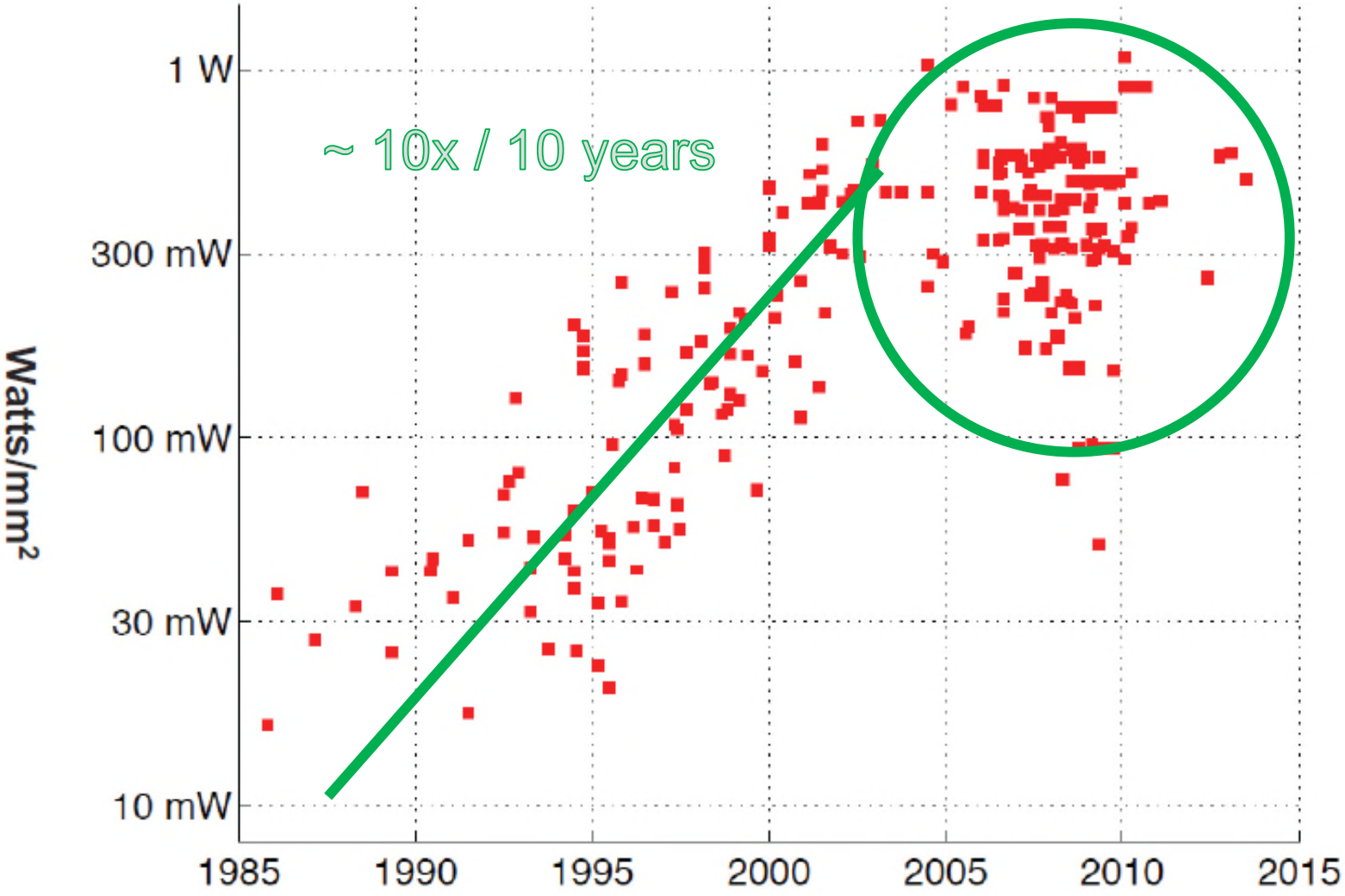


Performance Improvement



http://cpudb.stanford.edu

Power Density Limit



http://cpudb.stanford.edu

Power Consumption in CMOS Digital Circuits

$$P = P_{\text{dynamic}} + P_{\text{static}} = P_{\text{switching}} + P_{\text{sc}} + P_{\text{leakage}}$$

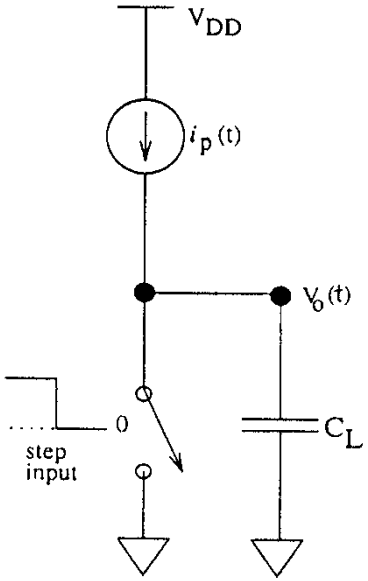
- There are three major source of power dissipation compoents associated with a CMOS logic gate.
  - Static power caused by the leakage current I_{leak} and other static current I_{static} due to the value of the input voltage;
  - Dynamic power caused by charging and discharging node capacitance;
  - Dynamic power caused by the short-circuit current I_{sc} during the switching transient.

- In earlier CMOS design, dynamic power is the main source of power dissipation, but leakage power increases as technology scales and becomes as significant today.

Switching Power Dissipation

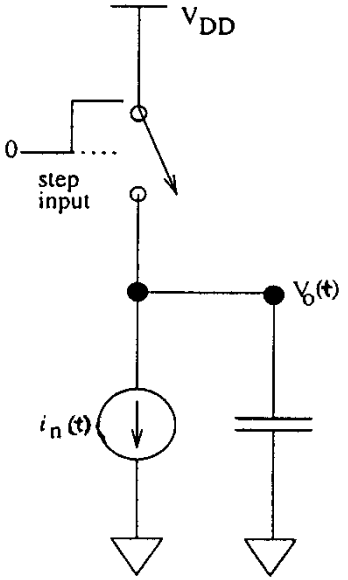
- Average switching power P_{sw} required to charge and discharge a capacitance C_L at a switching frequency $f = 1/T$ is given by:

$$P_{sw} = \frac{1}{T}\int_0^T i_o(t)\,v_o(t)\,dt$$



- During charging phase, the output current is given by:

$$i_o = i_p = C_L\frac{dv_o}{dt}$$



- During the discharging phase, the output current is given by:

$$i_o = i_n = -C_L\frac{dv_o}{dt}$$

Switching Power Dissipation

- Finally, P_{sw} is given by:

$$P_{sw} = \frac{1}{T}\int_0^{V_{DD}} C_L v_o dv_o - \frac{1}{T}\int_{V_{DD}}^0 C_L v_o dv_o = \frac{C_L V_{DD}^2}{T} = C_L V_{DD}^2 f$$

- In general, switching power of a gate *i* is a function of power supply (V_{DD}), voltage swing (V_i), operating frequency (f) and total load and parasitic capacitance (C_i)

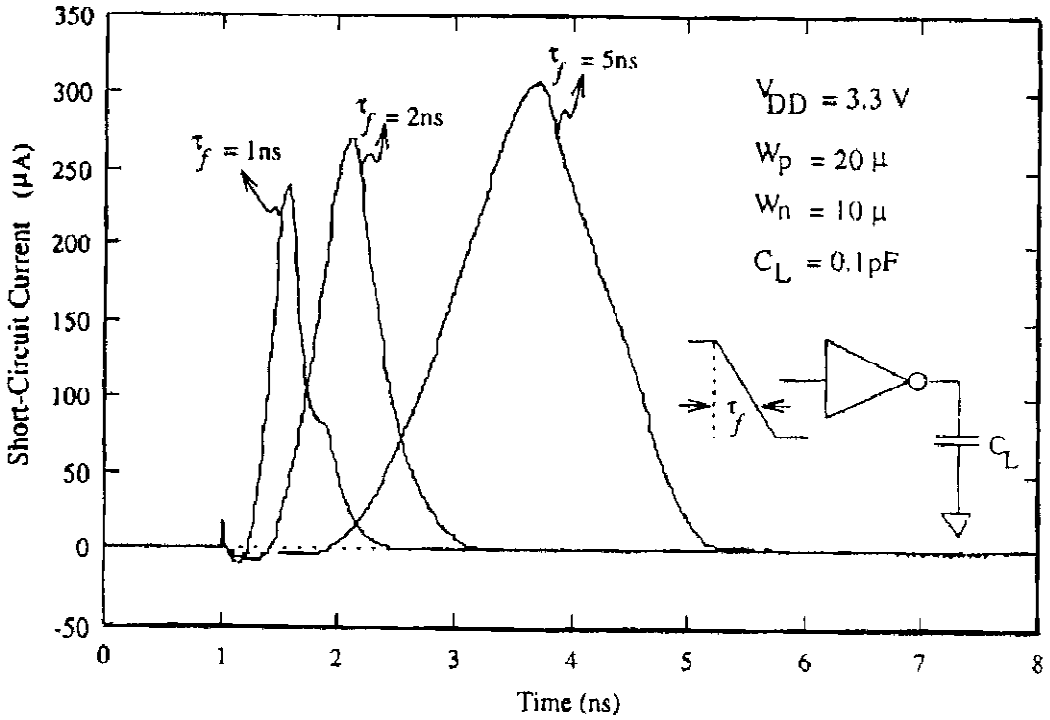
$$P_{sw} = \alpha_i \times C_i \times V_i \times V_{DD} \times f$$

  As not all nodes switch at the full rate of f, the activity factor α of a node is the probability that the node changes its state from 1 to 0 or vice versa in a clock cycle.

- Reducing V_{DD} leads to quadratic power savings.

Short-Circuit Power Dissipation

- If a CMOS logic gate is driven with input voltage waveform with finite rise and fall times, both nMOS and pMOS in the circuits may conduct simultaneously for a short amount of time during switching, forming a direct current path between V_{DD} and ground.



$$P_{sc} = I_{avg} V_{DD}$$

If $C_L = 0, \tau_r = \tau_f, k_n = k_p, V_{TP} = -V_{TN}$,

$$I_{avg} = \frac{2}{T}\left\{\int_{t_1}^{t_2} i(t)\,dt + \int_{t_2}^{t_3} i(t)\,dt\right\}$$

$$= \frac{4}{T}\int_{t_1}^{t_2} i(t)\,dt = \frac{4}{T}\int_{t_1}^{t_2} \frac{k}{2}\left(v_i(t) - V_T\right)^2 dt$$

$$v_i(t) = \frac{V_{DD}}{\tau} t,\ t_1 = \frac{V_{DD}}{\tau} t \text{ and } t_2 = \frac{\tau}{2}$$

Short-circuit current as a function of the input slope

Symmetric CMOS Inverter Short-Circuit Power

- Consider a symmetric CMOS inverter with $k_n = k_p = k$ and $V_{T,n} = |V_{T,p}| = V_T$, with very small capacitive load. If the input waveform has equal rise and fall times, $\tau_{rise} = \tau_{fall} = \tau$, the time averaged short-circuit current is:

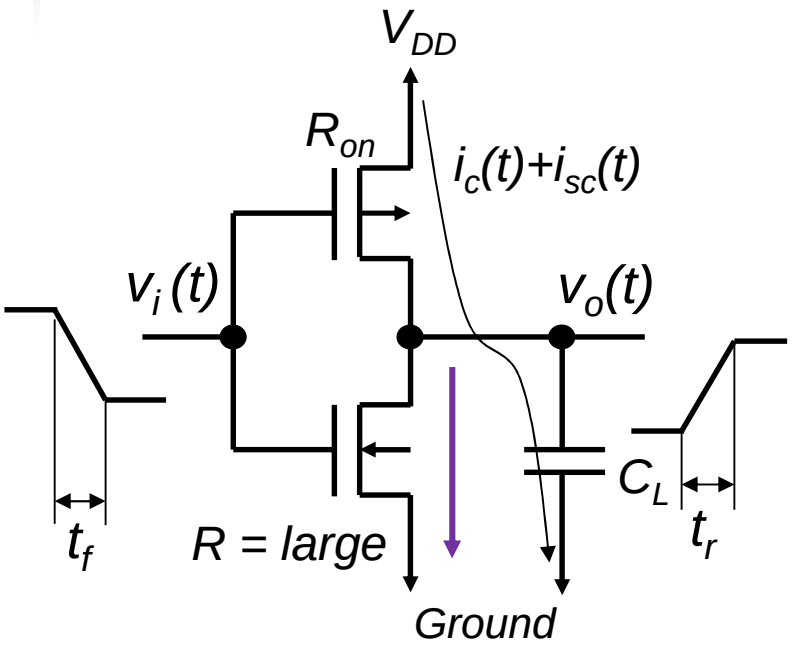
$$I_{sc} = \frac{1}{12} \cdot \frac{k \cdot \tau \cdot f}{V_{DD}} \left(V_{DD} - 2V_T \right)^3$$

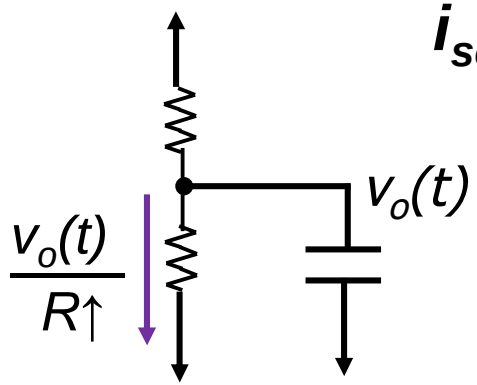
- The short-circuit power dissipation becomes:

$$P_{sc} = \frac{1}{12} \cdot k \cdot \tau \cdot f \cdot \left(V_{DD} - 2V_T \right)^3$$

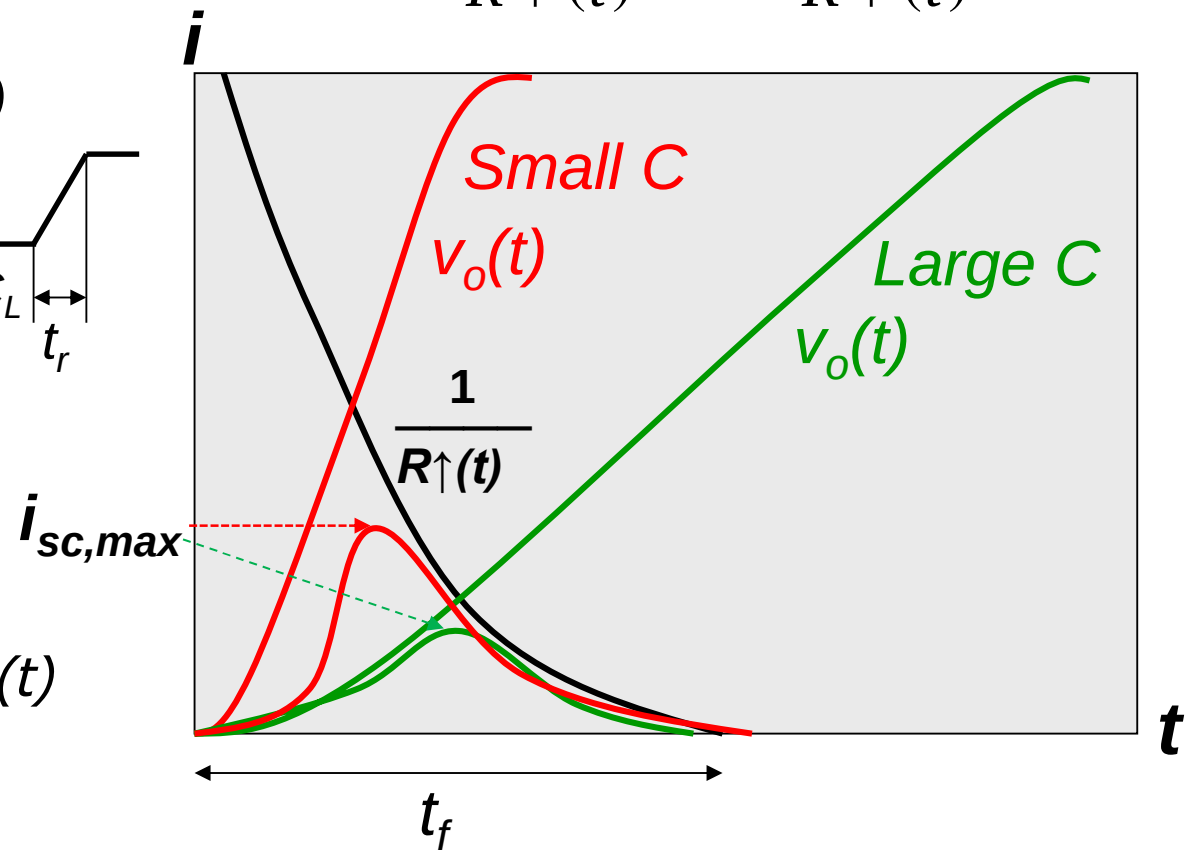
- Short circuit power increases with increase of input rise and fall time.
- For a given input rise and fall time, short circuit power decreases as output capacitance increases.
- The short-circuit power dissipation can be reduced if the output transition time is larger than the input voltage transition time.

$I_{sc,max}$, Capacitance and rise time





$$i_{sc}(t) = \frac{v_o(t)}{R\uparrow(t)} = \frac{V_{DD}\left(1 - e^{-\frac{t}{R\downarrow(t)C}}\right)}{R\uparrow(t)}$$



Transition Probability of Static CMOS Gates

P_0: Probability of output logic at logic 0;

P_1: Probability of output logic at logic 1, $P_1 = 1 - P_0$.

$$P_{0\to1} = P_0 \cdot P_1 = \left(\frac{n_0}{2^n}\right) \cdot \left(\frac{2^n - n_0}{2^n}\right)$$

n_0:number of 0s in the o/p of truth table.

NAND2

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

Out = 1, P_1 = ¾;

Out = 0; P_0 = ¼

$$P_{0\to1} = \frac{1}{4} \times \frac{3}{4} = \frac{3}{16}$$

$$P_{1\to0} = \frac{3}{4} \times \frac{1}{4} = \frac{3}{16}$$

Switching $P_{0\to1} = P_{1\to0}$

$$P_{0\to0} = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$$

$$P_{1\to1} = \frac{3}{4} \times \frac{3}{4} = \frac{9}{16}$$

Relative Switching Power

- Let α_i be the switching activity at the output node *i*

$$\alpha_i = P_{0\to1} + P_{1\to0}$$

- Let C_i be the capacitive load at the output node *i* and *f* be the frequency of change at the inputs of the gate, then the switching power is

$$P_{sw} = \alpha_i C_i V^2 f$$

- The relative switching power at node *i* is given by

$$P_{rsw} = \alpha_i C_i$$

- For *N* nodes, the relative switching power is given by

$$P_{rsw} = \sum_{i=1}^{N} \alpha_i C_i$$

Example: Relative Switching Power

NOR2

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

$$P_{0\to1} = \frac{1}{4}\times\frac{3}{4} = \frac{3}{16}$$
$$P_{1\to0} = \frac{3}{4}\times\frac{1}{4} = \frac{3}{16}$$

Switching

$$P_{0\to0} = \frac{3}{4}\times\frac{3}{4} = \frac{9}{16}$$
$$P_{1\to1} = \frac{1}{4}\times\frac{1}{4} = \frac{1}{16}$$

$$P_{rsw} = \frac{6}{16}C_L$$

XOR2

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

$$P_{0\to1} = \frac{2}{4}\times\frac{2}{4} = \frac{4}{16}$$
$$P_{1\to0} = \frac{2}{4}\times\frac{2}{4} = \frac{4}{16}$$

Switching

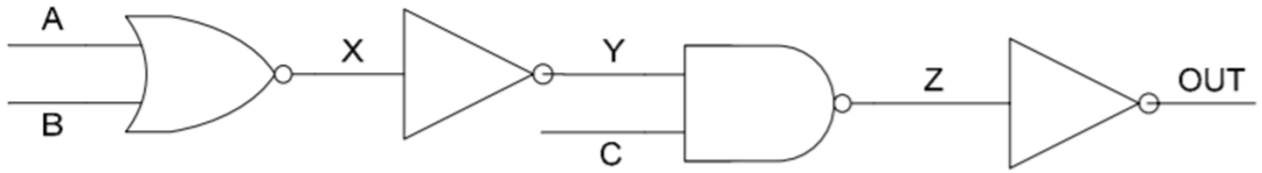
$$P_{0\to0} = \frac{2}{4}\times\frac{2}{4} = \frac{4}{16}$$
$$P_{1\to1} = \frac{2}{4}\times\frac{2}{4} = \frac{4}{16}$$

$$P_{rsw} = \frac{8}{16}C_L$$

Example: Switching Probability Calculation

Given $P(A=1) = P(B=1) = P(C=1) = 0.5$,

calculate $P_{X:0\to1}$, $P_{Y:0\to1}$, $P_{Z:0\to1}$, and $P_{OUT:0\to1}$.



$P_{X:0\to1} = P(X=0) \times P(X=1) = ¾ \times ¼ = 3/16.$

$P_{Y:0\to1} = P_{X:1\to0} = P(X=1) \times P(X=0) = ¼ \times ¾ = 3/16.$

Since Y is logically $A + B$, $P(Y=1) = ¾$.

Since $P(C=1) = 0.5$, $P(Z=0) = ¾ \times ½ = 3/8$.

$P_{Z:0\to1} = P(Z=0) \times P(Z=1) = 3/8 \times 5/8 = 15/64.$

$P_{OUT:0\to1} = P(Z=1) \times P(Z=0) = 5/8 \times 3/8 = 15/64$

Example: Switching Activity of Row Decoder

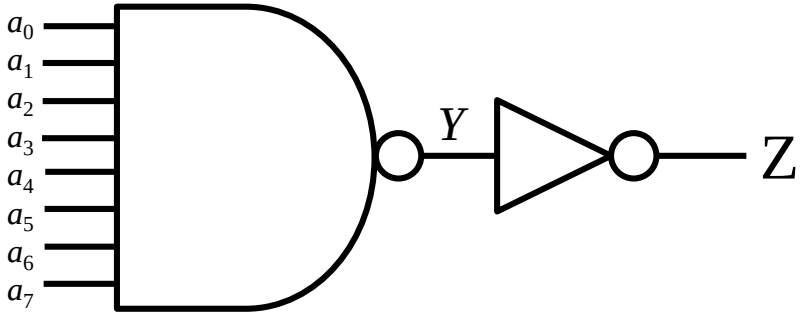
An 8-input AND function is used as a row decoder of an SRAM.

Which of the following schemes has lower relative switching power:

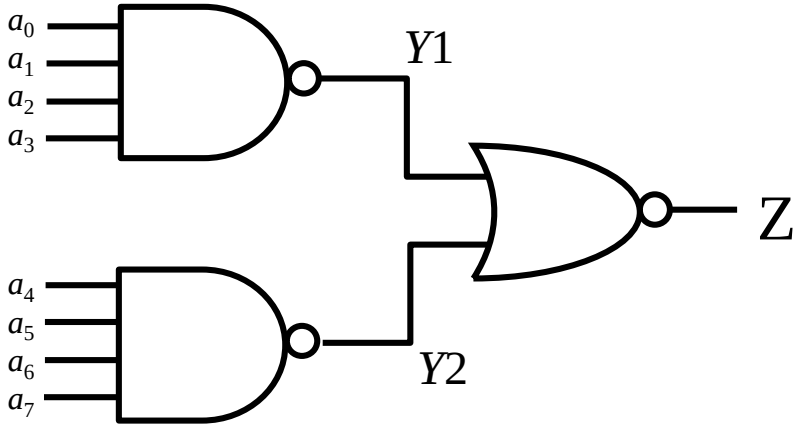
Scheme A: $Z = \overline{a_0 a_1 a_2 a_3 a_4 a_5 a_6 a_7}$

Scheme B: $Z = \overline{a_0 a_1 a_2 a_3} + \overline{a_4 a_5 a_6 a_7}$

if the capacitive loading at the internal nodes are 1 unit and at the output is 5 units



Scheme A



Scheme B

Example: Switching Activity of Row Decoder

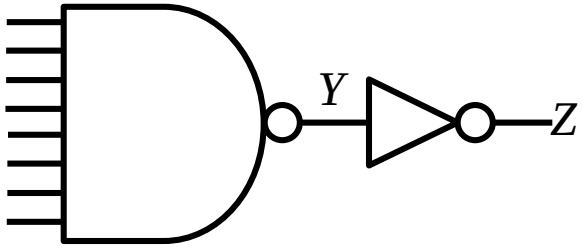
Scheme A:

$P_1(Y) = 255/256$, $P_0(Y) = 1/256$, $P_{0\to1}(Y) = P_{1\to0}(Y) = 255/65536$

$P_1(Z) = 1/256$, $P_0(Z) = 255/256$, $P_{0\to1}(Z) = P_{1\to0}(Z) = 255/65536$

$\alpha_Y = P_{0\to1}(Y) + P_{1\to0}(Y) = 510/65536.$

$\alpha_Z = P_{0\to1}(Z) + P_{1\to0}(Z) = 510/65536.$



Total switching activity = $\alpha_Y + \alpha_Z = 0.01556$

Relative switching power = $\alpha_Y C_Y + \alpha_Z C_Z$

= (510/65536)×1 + (510/65536)×5

= 0.04669

Example: Switching Activity of Row Decoder

Scheme B:

$P_1(Y1) = P_1(Y2) = 15/16$, $P_0(Y1) = P_0(Y2) = 1/16$,

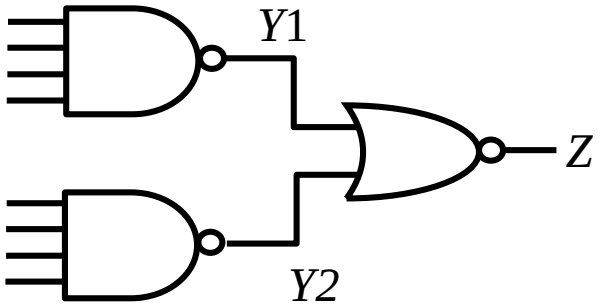
$P_{0\to1}(Y1) = P_{1\to0}(Y1) = 15/256$,

$P_{0\to1}(Y2) = P_{1\to0}(Y2) = 15/256$,

$P_1(Z) = 1/256$, $P_0(Z) = 255/256$,

$P_{0\to1}(Z) = P_{1\to0}(Z) = 255/65536$

$\alpha_{Y1} = \alpha_{Y2} = 30/256$ and $\alpha_Z = 510/65536$.



Total switching activity = $\alpha_{Y1} + \alpha_{Y2} + \alpha_Z = 0.24216$

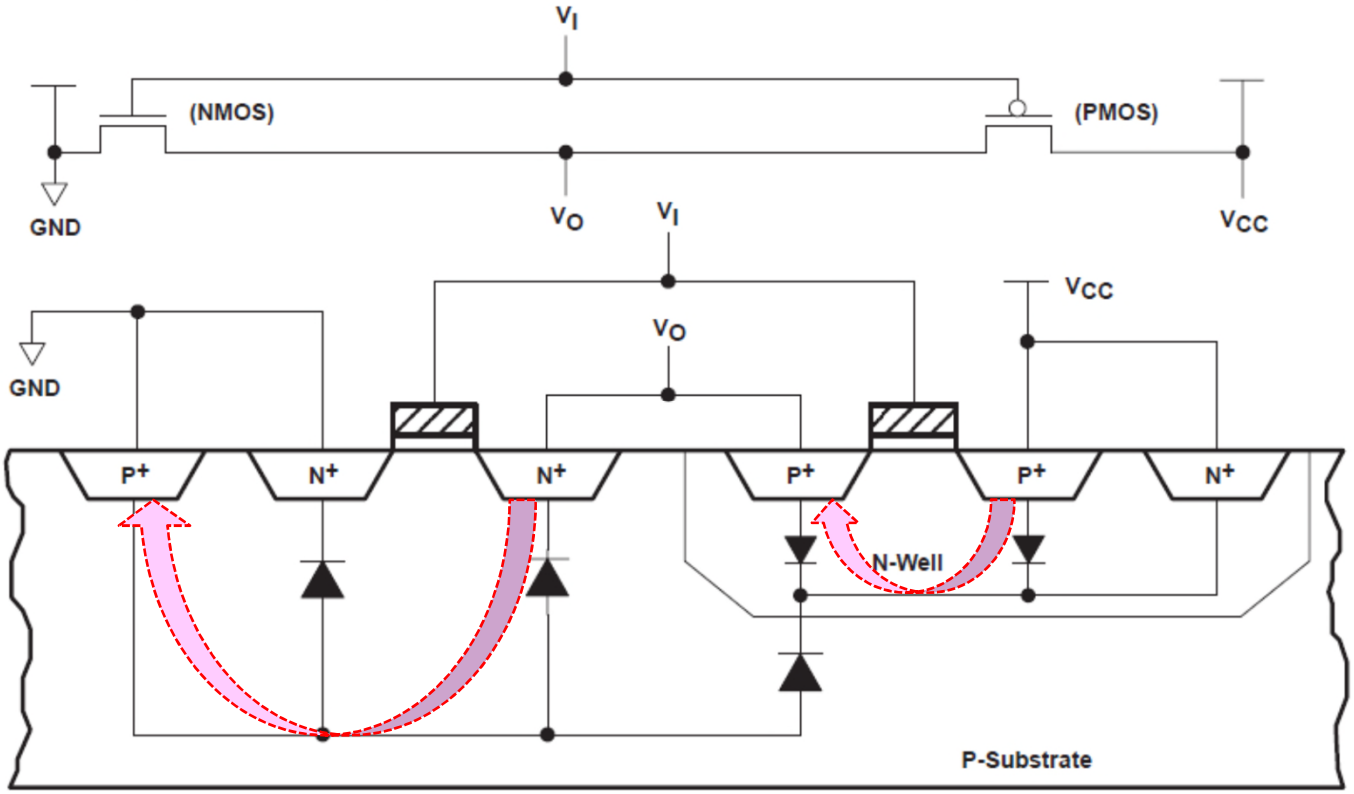
Relative switching power = $\alpha_{Y1}C_{Y1} + \alpha_{Y2}C_{Y2}\ \alpha_ZC_Z$

= (30/256)×1 + (30/256)×1 + (510/65536)×5 = 0.27328

Scheme A is better than Scheme B since it has both lower switching activity and relative switching power.

Static Power – Reverse Leakage Current

- The first main leakage current component in a MOSFET is the reverse diode leakage. It occurs when the p-n junction between the drain and the bulk of the transistor is reversed biased.
- CMOS inverter example:



Static Power – Reverse Leakage Current

- The reverse leakage current of a pn-junction (diode) is given by

$$I_d = I_S \left(\exp\left(\frac{qV_d}{nkT} \right) - 1 \right)$$

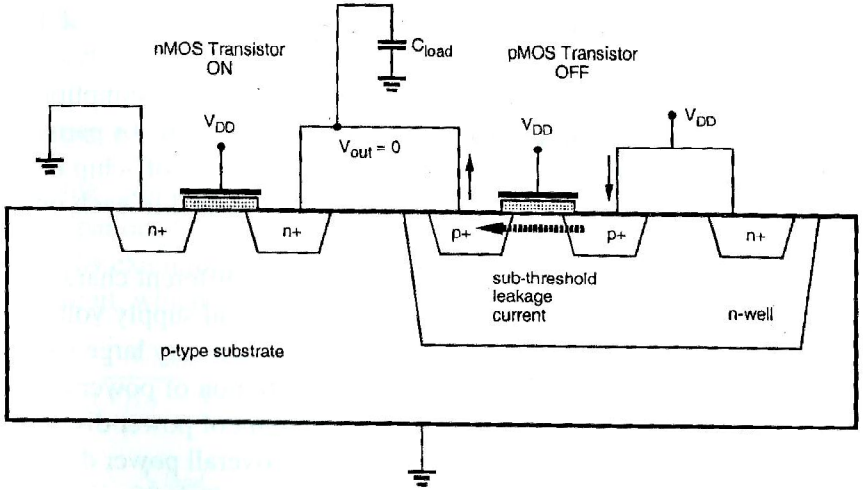
where n is the emission coefficient of the diode (~1), V_d is the bias voltage across the junction, q is the electronic charge, T is the temperature and k is the Boltzmann constant, and $I_s = A \times J_S$ is the reverse saturation current. A is the junction area and J_s is the reverse saturation current density, typically 1-5 pA/μm^2.

$$q = 1.60217663 \times 10^{-19}\,\mathrm{C}$$

$$k = 1.380649 \times 10^{-23}\ \mathrm{m^2\ kg\ s^{-2}\ K^{-1}}$$

Static Power – Subthreshold Current

- The second component is the subthreshold current. It is due to carrier diffusion between the source and the drain regions of the transistor in weak inversion.
- The behavior of an MOS in subthreshold region is similar to a bipolar. The subthreshold current has an exponential dependence on gate voltage.
- The subthreshold current is significant when the gate-to-source voltage is less than but close to the threshold voltage of the device.
- Subthreshold leakage current can occur even when there is no switching activity.

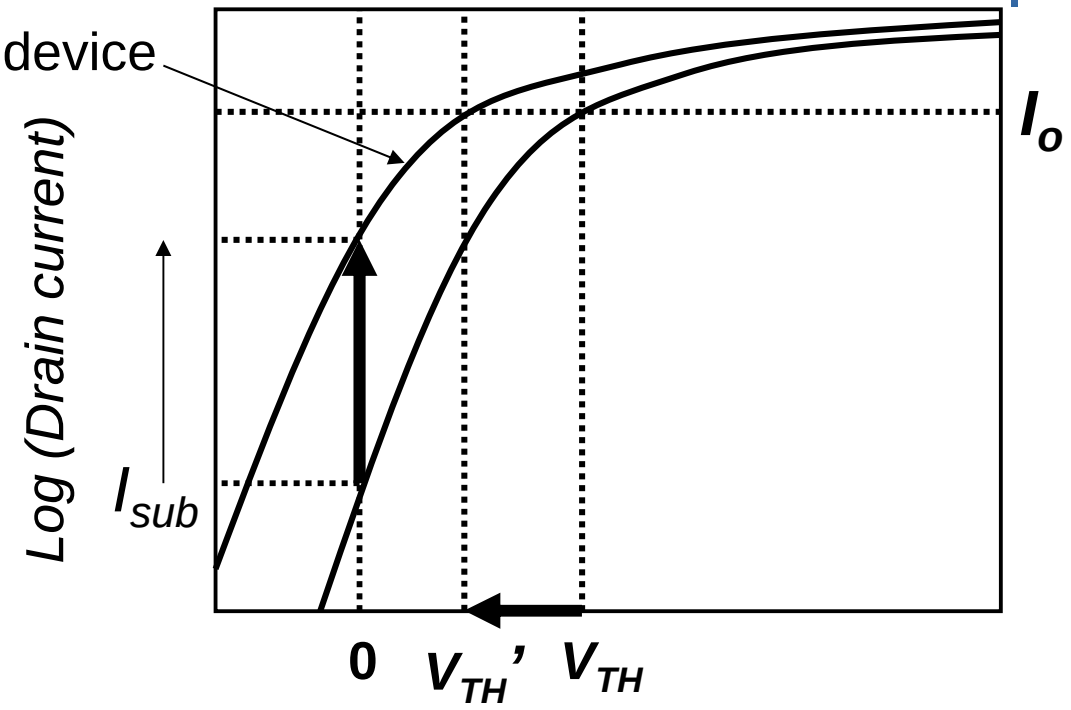


Static Power – Subthreshold Current

For $0 < V_{in} < V_T$,

$$I_{DS} = I_O \frac{W_{eff}}{W_O} 10^{\frac{V_{in} - V_t}{S}}$$

Subthreshold swing S is the gate voltage swing required to reduce the drain current by one decade.

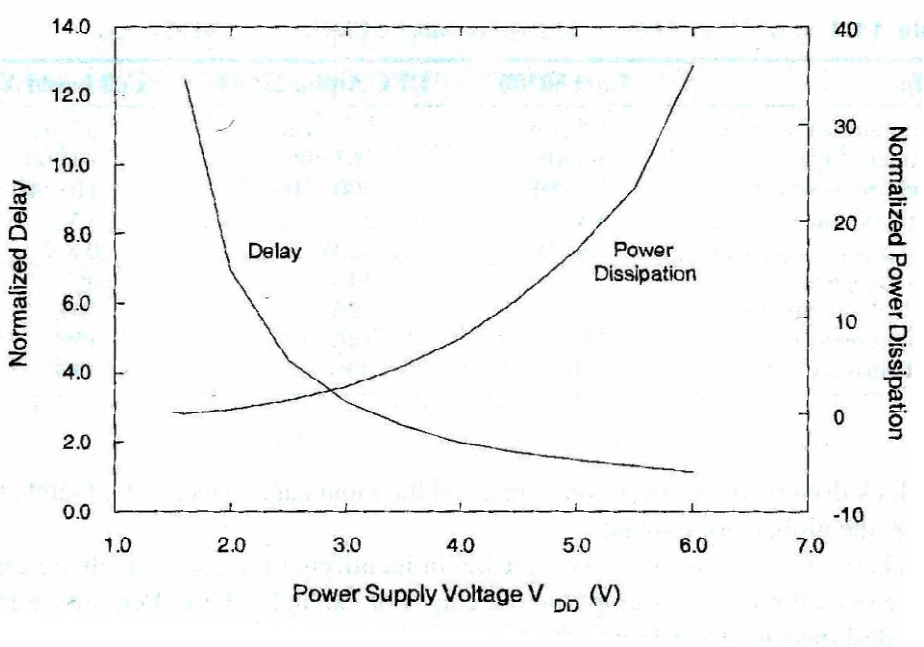


One measure to limit the subthreshold current is to avoid very low threshold voltage, with some speed penalty, so that V_{GS} of the nMOS remains safely below $V_{T,n}$ when the input is logic zero, and $|V_{GS}|$ of pMOS remains safely above $|V_{T,p}|$ when the input is logic one.

Supply Voltage Scaling

Reducing supply voltage leads to quadratic reduction of power consumption if switching frequency remains constant.

$V_{TN} = |V_{TP}| = 0.8$ V



Reduction of supply voltage:

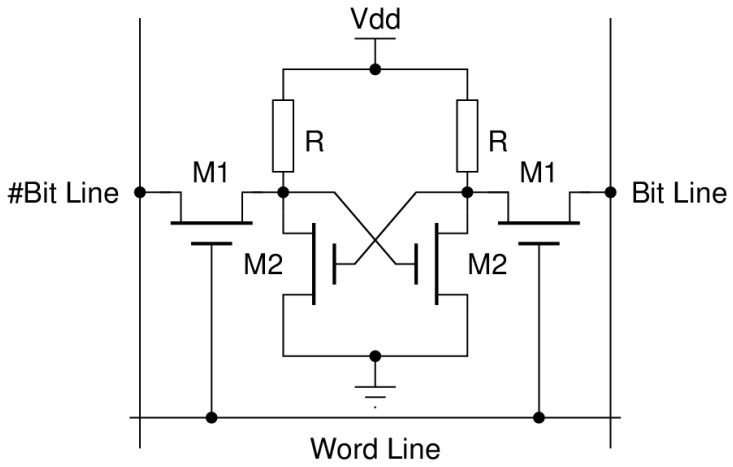
i. the drivability of MOSFETs will decrease,
ii. signals will become smaller, and
iii. the threshold voltage variations will become more prominent.

- Lowering the supply voltage increases gate delay.
- The increase of the gate delay time is serious when the operating voltage is reduced to 2 V or less, even with scaling down the device dimensions.

Example: High resistive load memory cell

A 1 Mb 4T SRAM chip is fabricated using a 180-nm CMOS with a threshold voltage of 0.5 V. If the total standby current of the memory chip is less than 20 μA, calculate the value of the pull-up resistive load if the power supply voltage of the memory chip is 1.8 V.

If the total standby current of the memory chip and the calculated pull-up high resistive load must remain the same, is it possible to increase the density of the memory chip to 4-Mb?



Example: High resistive load memory cell

1 Mb = 2^{20} bits = 1,048,576 bits

Leakage current/cell = 20 μA/2^{20} = 19.07 *p*A/bit.

Pull-up resistive load R of each cell

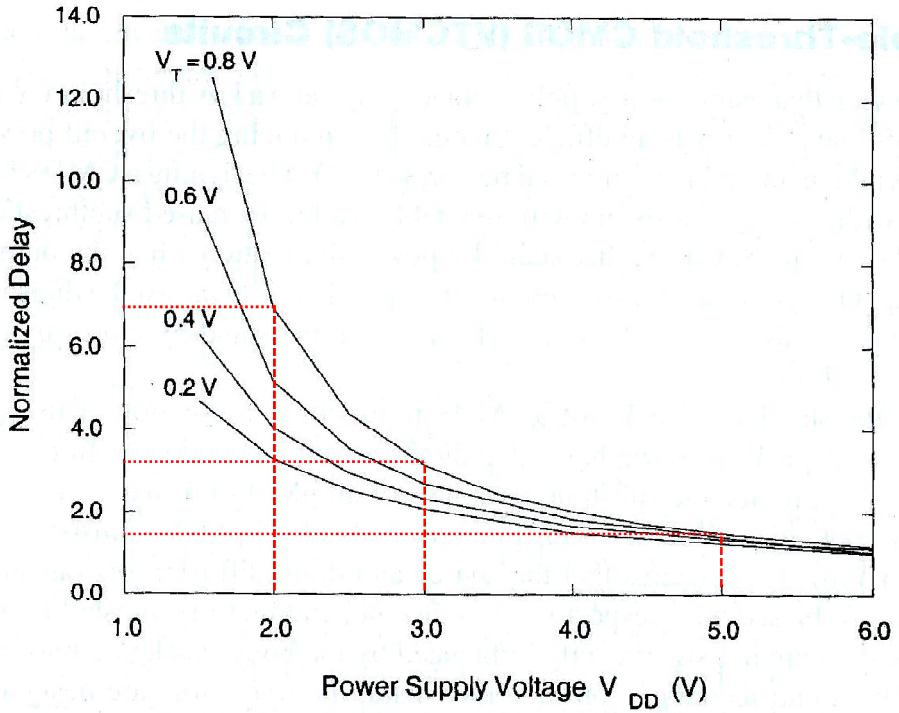
= *V*/*I* = 1.8 V/ 19.07 *p*A = 94.39 GΩ

If the memory array is increased to 4-Mb, the power supply must be scaled down to 1.8 V/4 = 0.45 V in order to keep the total leakage current below 20 μA.

The SRAM will fail to operate properly as the supply voltage 0.45 V is lower than the threshold voltage 0.5 V of the device.

Threshold Voltage Scaling

- The negative of reducing supply voltage on delay can be compensated for, if the threshold voltage is scaled down accordingly.
- When scaled linearly, reduced threshold voltages allow circuit to produce the same speed-performance at a lower V_{DD}.

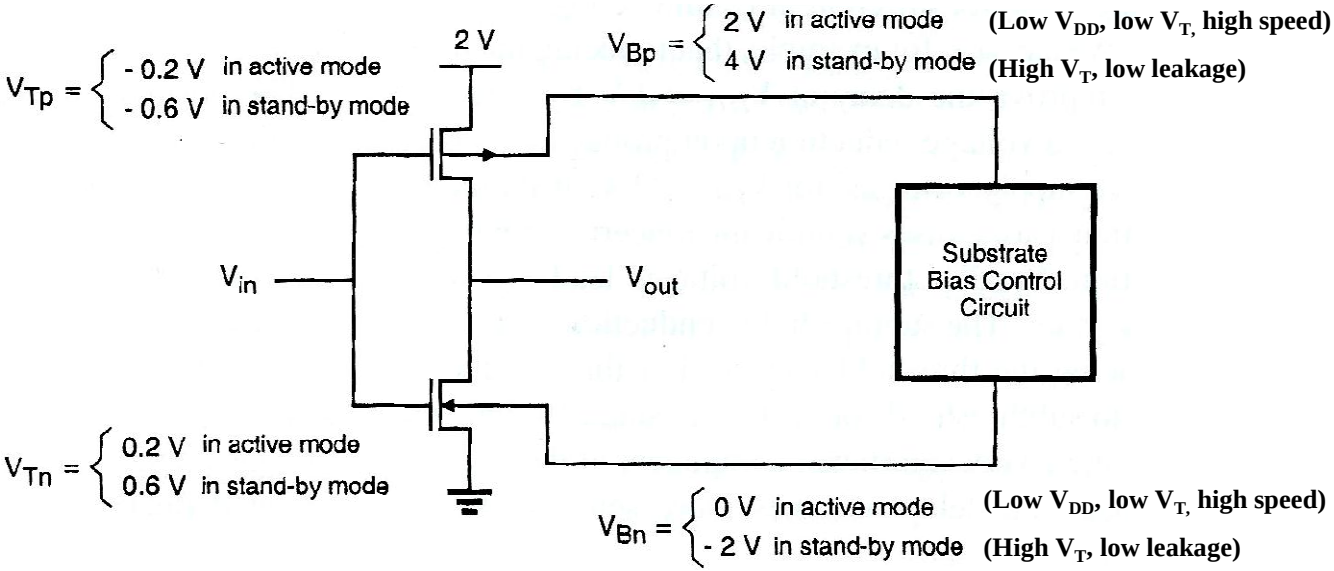


Limitations of Threshold Voltage Scaling

- Threshold voltage may not be scaled to the same extent as supply voltage.
- Reduction of threshold voltage sharply increase the MOSFET cut-off current and degrades its ON/OFF ratio.
- Reduction of threshold voltage increases the transient power dissipation due to the switching transient current.
- Smaller threshold voltages lead to smaller noise margins.
- For threshold voltages smaller than 0.2V, leakage due to subthreshold conduction in stand-by may become a very significant component of overall power consumption.
- Propagation delay becomes more sensitive to process-related variations of threshold voltage.
- Thus a compromise needs to be found for the V_T/V_{DD} ratio in order to achieve both low-power and high-speed operation.

Variable-Threshold CMOS (VTCMOS) Circuits

- V_T of an MOS transistor is a function of source-to-substrate voltage V_{SB}.
- In conventional CMOS circuits, the substrate terminals of all nMOS and pMOS transistors are connected to ground and V_{DD}, respectively to ensure that the source and drain diffusion regions are always reverse-biased with respect to the substrate.
- In VTCMOS circuit, the transistors are designed inherently with a low V_T, and the substrate bias voltages of nMOS and pMOS transistors are generated by a variable substrate bias control circuit.



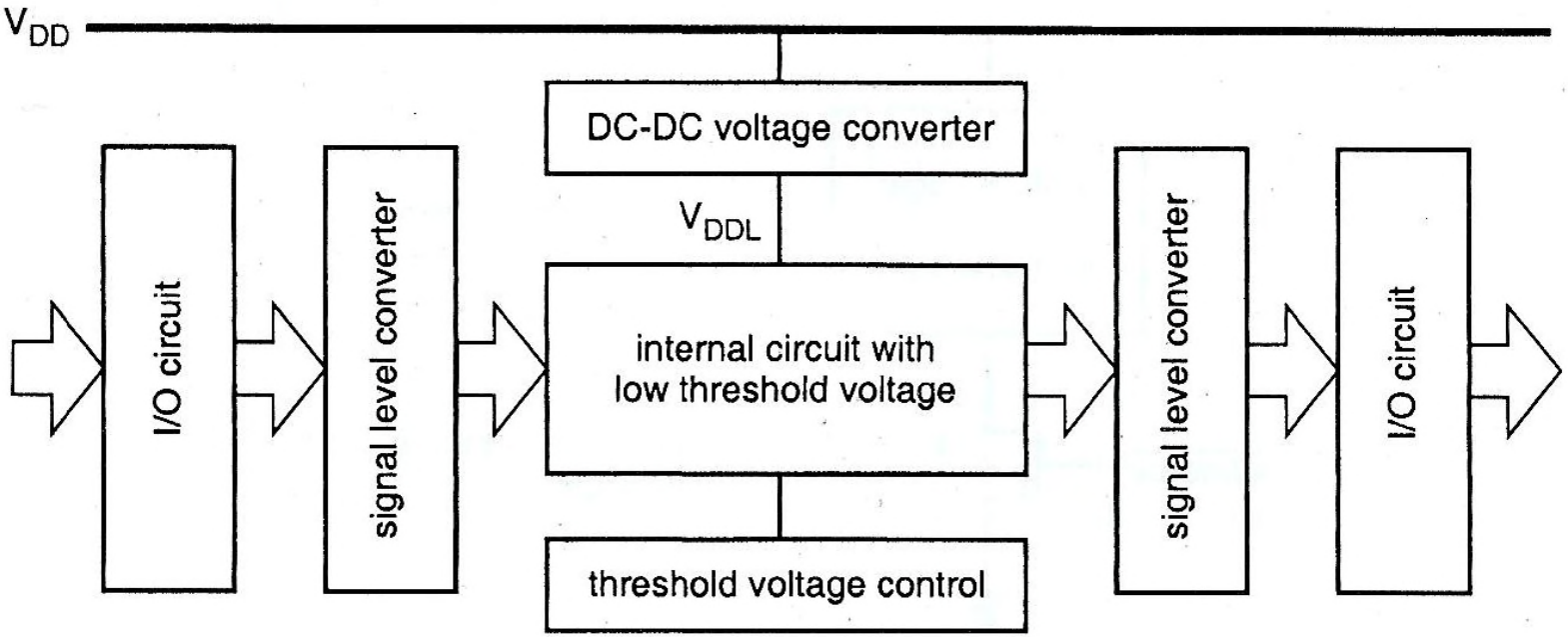
Variable-Threshold CMOS (VTCMOS) Circuits

- In active mode, the substrate bias voltages of nMOS and pMOS transistors are $V_{Bn} = 0$ and $V_{Bp} = V_{DD}$, respectively. The circuit operates with low V_{DD} (low power dissipation) and low V_T (high switching speed).
- In stand-by mode, the substrate bias control circuit generates a lower substrate bias voltage for the nMOS and a higher substrate bias voltage for the pMOS. Due to the backgate-bias effect, the magnitudes of V_{Tn} and V_{Tp} both increase and reduce the leakage power dissipation significantly.

- This technique requires twin-well or triple-well CMOS technology in order to apply different substrate bias voltages to different parts of the chip.
- Separate power pins may be required if the substrate bias voltage levels are not generated on-chip.
- Area overhead of the substrate bias control circuitry is usually neglible compared to the overall chip area.

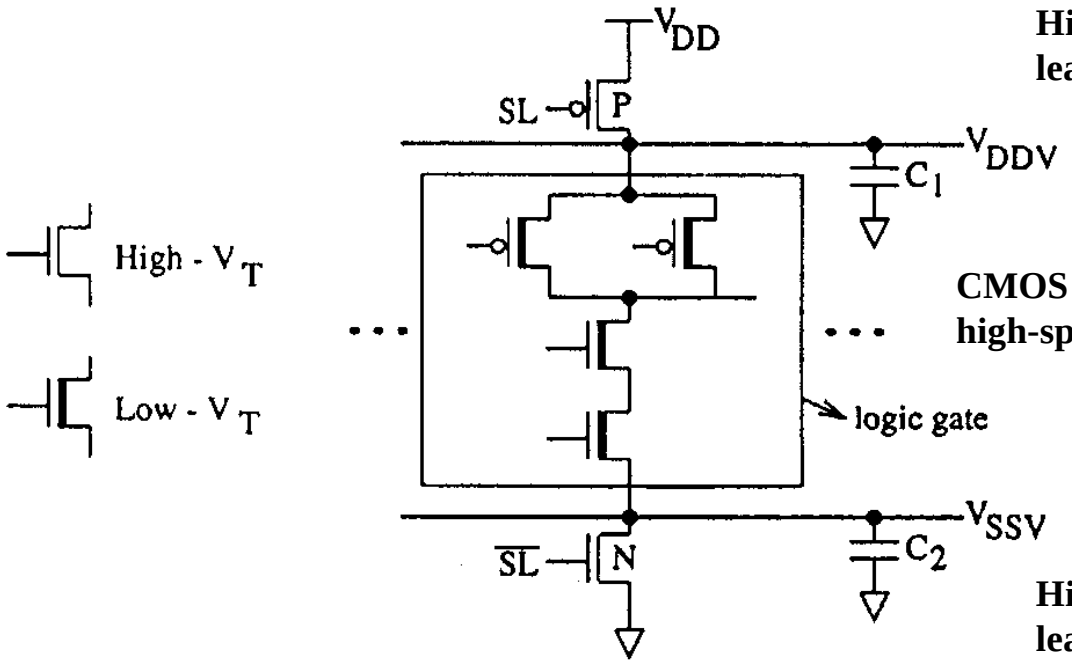
Typical Low-Power Chip with VTCMOS

- I/O circuits operate with high external supply voltage V_{DD} to increase noise margin and drivability with peripheral devices.
- DC-DC converter generate low internal voltage V_{DDL}.
- Level converters reduce voltage swing of incoming input signal and increase voltage swing of outgoing output signal of the internal low-voltage circuitry with VTCMOS technique.



Multiple-Threshold CMOS (MTCMOS) Circuits

- Low-V_T transistors are used to design logic gates where switching speed is essential. High-V_T transistors are used to effectively isolate the logic gates in stand-by and to prevent leakage dissipation.
- Signal SL is used to switch the gate in active or sleep (stand-by) mode.
- The virtual supply lines V_{DDV} and V_{SSV} are common for many gates.



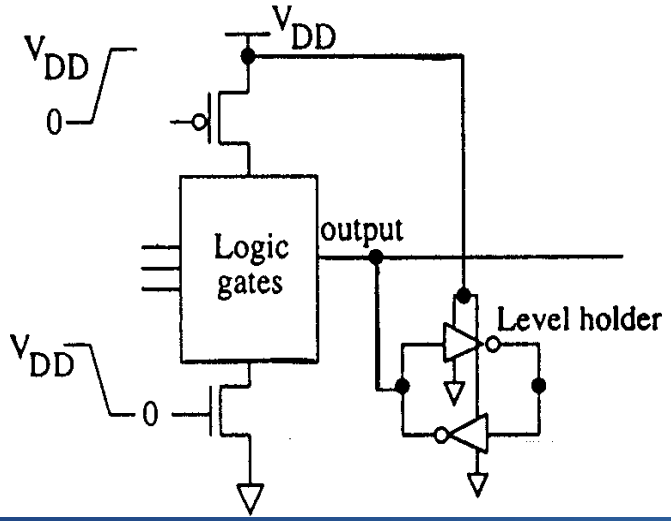
High-V_T pMOS: Prevents subthreshold leakage in stand-by mode

CMOS logic with low-V_T: high-speed operation with low power consumption

High-V_T nMOS: Prevents subthreshold leakage in stand-by mode

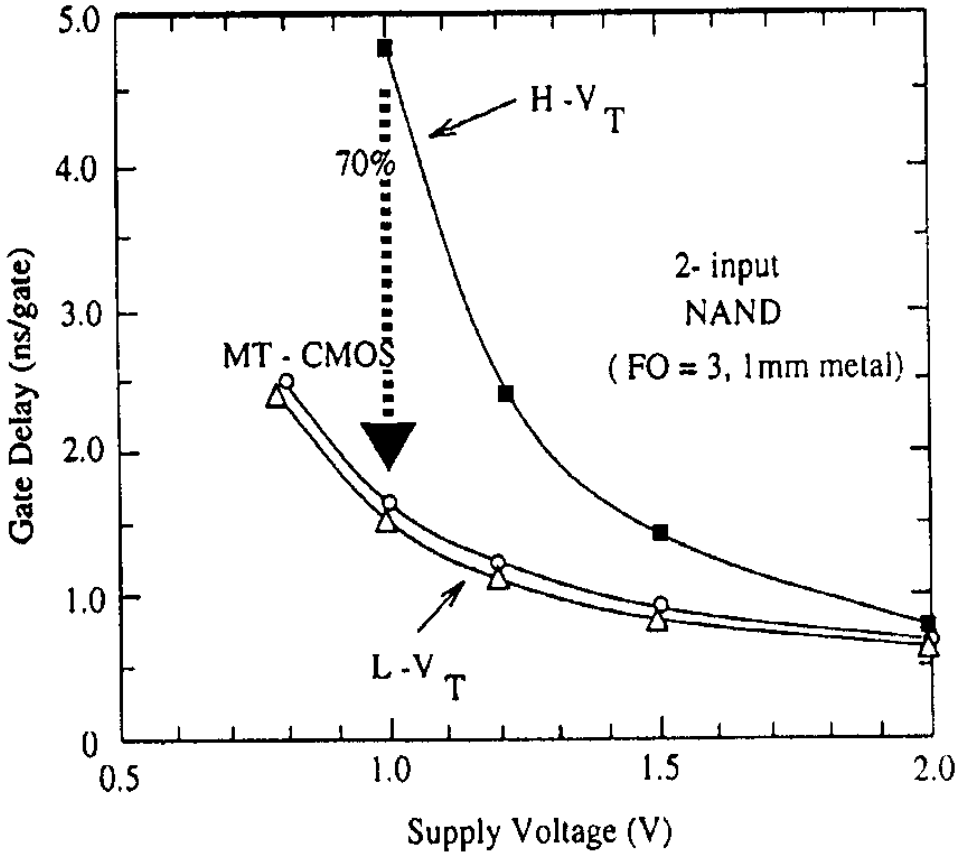
CMOS (MTCMOS) Gate with Level Holder

- A level holder is necessary to preserve the data (hold the output level) in the stand-by mode.
- It consists of a cross-couple inverter pair with high-V_T devices powered from the power supply V_{DD}.
- Easier to apply and to use. Does not require a twin-well or triple-well CMOS process. The only significant process-related overhead is the fabrication of MOS transistors with different V_T on the same chip.
- Presence of series-connected stand-by transistors increases overall circuit area and add extra parasitic capacitance and delay.



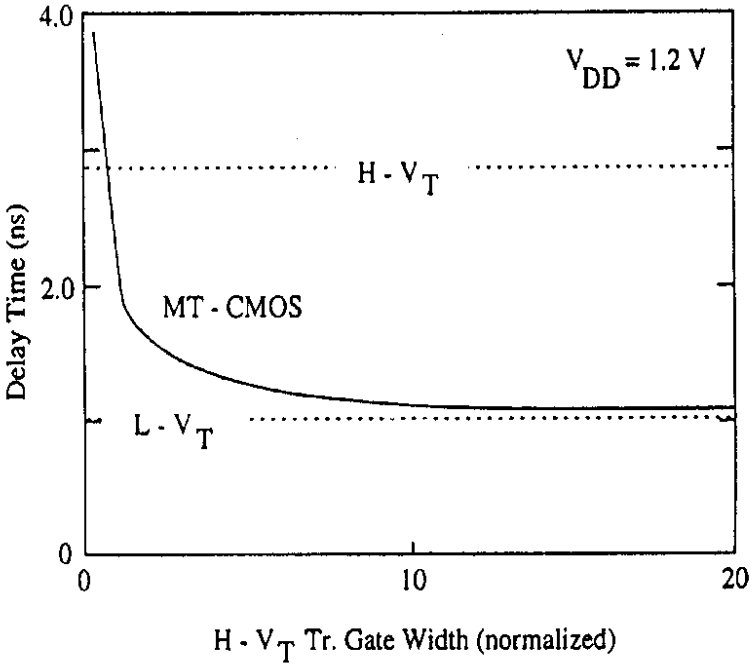
Effect of Gate Delays on MT-CMOS

- MT-CMOS logic has almost the same speed as the full low-V_T *logic*.
- Its logic delay time is reduced by 70% at 1 V compared with that of the high-V_T circuit.



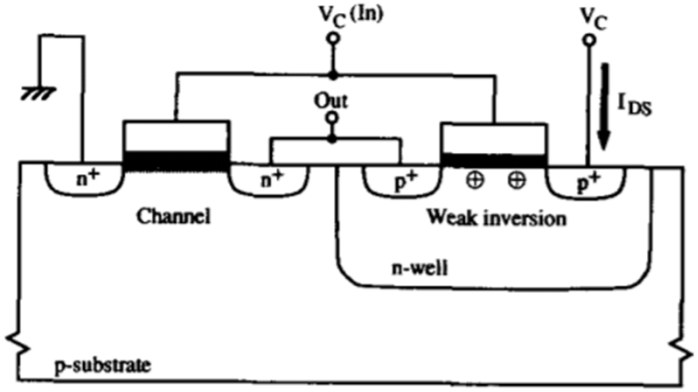
Effect of High-V_T MOS Width on MT-CMOS

- Width of P/N should be at least 10 times larger than that of logic cells.
- This condition depends greatly on the parasitic capacitances C_1 and C_2 of virtual supply lines.
- If C_1 and C_2 are large, then the width of P and N transistors can be reduced, because these capacitance tend to suppress the bouncing of V_{DDV} and V_{SSV} and hence improve the speed.



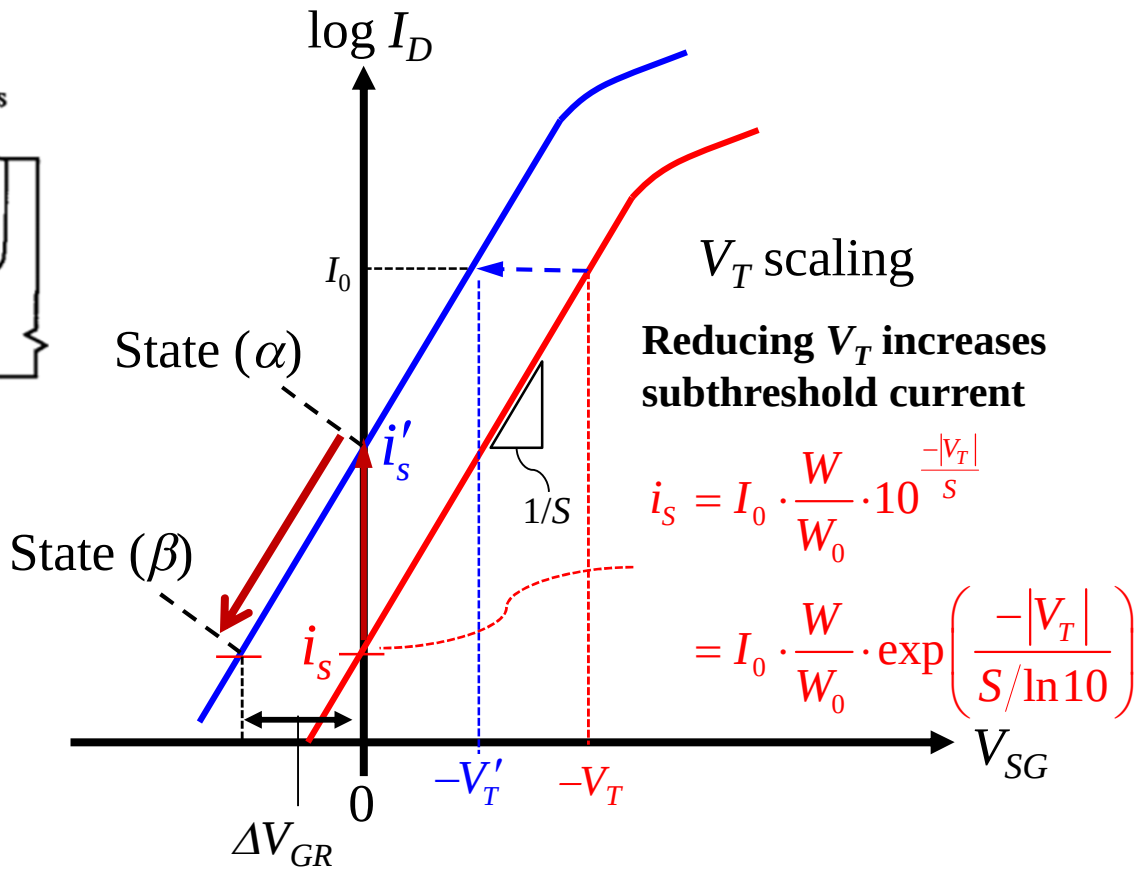
Reduce Subthreshold Current by Reverse Biasing

When $-V_{GS} = \Delta V_{GR}$, the stand-by state of the pMOS transistor moves from state α to state β.



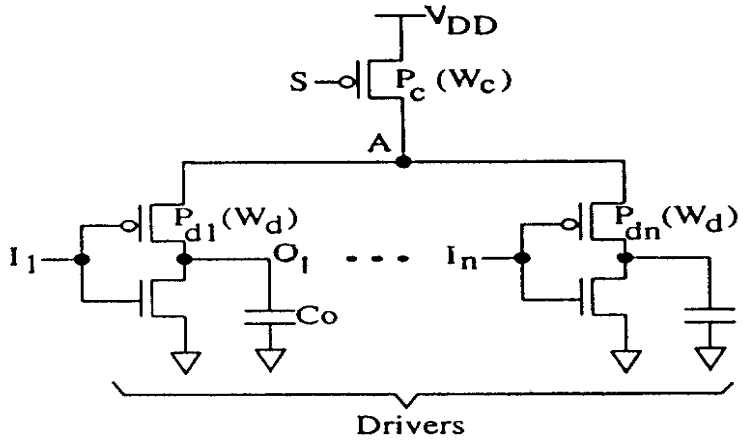
$$I_D = I_0 \cdot \frac{W}{W_0} \cdot 10^{\frac{-V_{GS}-|V_T|}{S}}$$

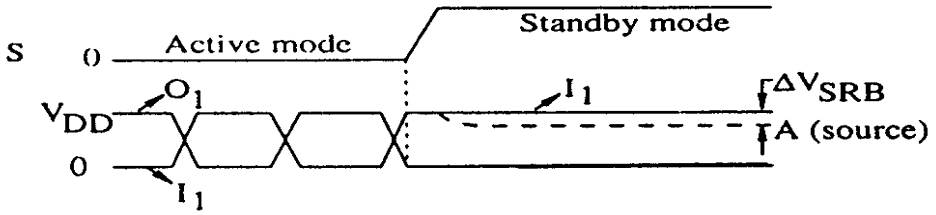
V_T : Threshold voltage
I_0 : Drain current at threshold voltage
W_0 : Gate width at threshold voltage
W: Gate width
S: Subthreshold swing



Self-Reverse Biasing

- Drivers, in memory, are replicated in large number, but only a few of them operate simultaneously.
- All PMOS (P_{d1}, P_{d2}, …, P_{dn}) of the drivers have the same size W_d and common source (node A). The number of drivers n can be between a few hundreds to a few thousands.
- A PMOS transistor P_c with size W_c is placed between V_{DD} and the common source node A.
- The MOS transistors in the drivers have low $|V_{Td}|$ (e.g., 0.1V). P_c has a threshold voltage $|V_{Tc}|$ slightly higher than $|V_{Td}|$ (e.g., 0.2 – 0.4 V).





$$I_o \frac{W_c}{W_o} \exp\left(\frac{-|V_{Tc}|}{S/\ln 10}\right) = nI_o \frac{W_d}{W_o} \exp\left(\frac{-\Delta V_{SRB} - |V_{Td}|}{S/\ln 10}\right)$$

$$\Delta V_{SRB} = -\left(|V_{Td}| - |V_{Tc}|\right) + \frac{S}{\ln 10} \cdot \ln \frac{nW_d}{W_c}$$

Self-Reverse Biasing

- In active mode, input V_S at node S is low and P_c does not affect the drive current.
- W_c should be larger than W_d, depending on the capacitance of the common source, which is huge for large n.
- In stand-by mode, S is high and P_c is OFF. The inputs of all drivers are set to high (V_{DD}).
- Without P_c, the total subthreshold current would be n times the current of each driver. The total subthreshold current is given by:

$$I_{sub1} = nI_o \frac{W_d}{W_o} \exp\left(\frac{-|V_{Td}|}{S/\ln 10}\right) = nI_o \frac{W_d}{W_o} 10^{\frac{-|V_{Td}|}{S}}$$

- With P_c, the voltage of common source node A is reduced by an amount ΔV_{SRB} (a few hundreds of mV), causing the PMOS transistors of all drivers to have self-reverse-biasing gate-source voltage.

Self-Reverse Biasing

- With P_c, the total subthreshold current is given by:

$$I_{sub2} = I_o \frac{W_c}{W_o} \exp\left(\frac{-|V_{Tc}|}{S/\ln 10}\right) = I_o \frac{W_c}{W_o} 10^{\frac{-|V_{Tc}|}{S}}$$

- Assuming the devices have the same I_o, W_o and S, the reduction factor of subthreshold current is:

$$\gamma = \frac{I_{sub1}}{I_{sub2}} = n \frac{W_d}{W_c} \exp\left(\frac{|V_{Tc}| - |V_{Td}|}{S/\ln 10}\right) = n \frac{W_d}{W_c} 10^{\frac{|V_{Tc}| - |V_{Td}|}{S}}$$

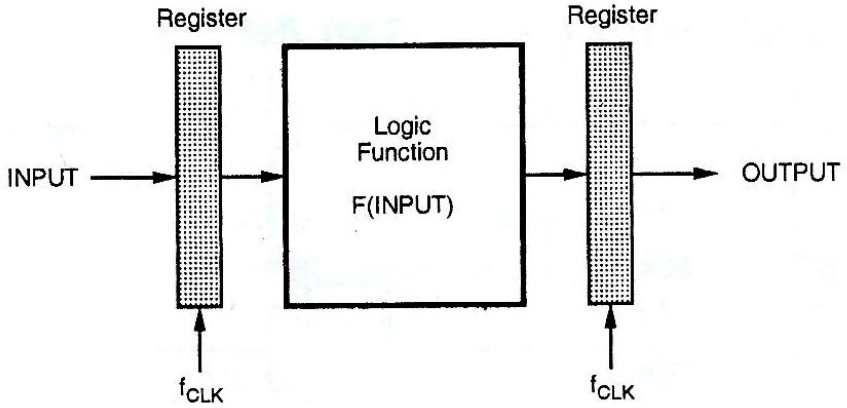
- Example: For n = 512, W_c = 10 W_d (speed is not affected with this ratio), V_{Tc} = 0.3 V, V_{Td} = 0.1 V and S = 90 mV/decade of drain current, $\gamma = 8.5 \times 10^3$. So the saving in subthreshold current is significant.

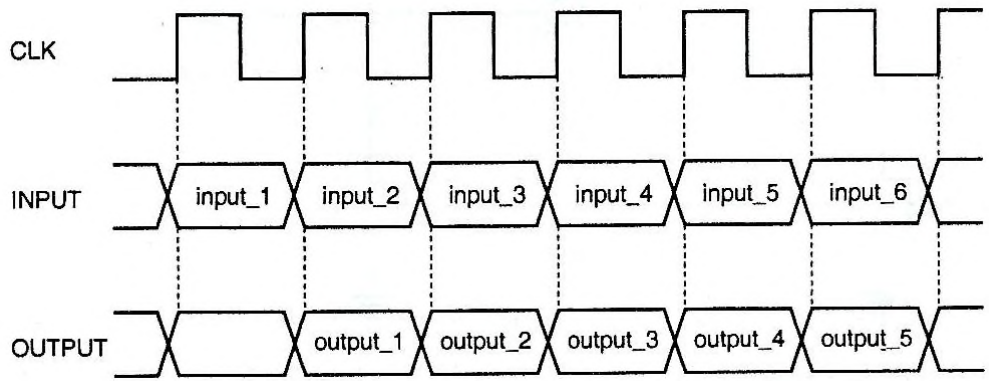
Power Reduction by Pipelining

- Let f_{CLK} be the maximum sampling frequency. The critical path delay of F(INPUT), $T_{ref} \leq T_{CLK} = 1/f_{CLK}$.
- A new input vector is latched into the input register array at each clock cycle and the output data become valid with a latency of one clock cycle.
- The dynamic power consumption is:

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

where the total capacitance switched C_{total} every cycle consists of capacitance switched in the input and output register arrays and capacitance switched to implement F(INPUT).

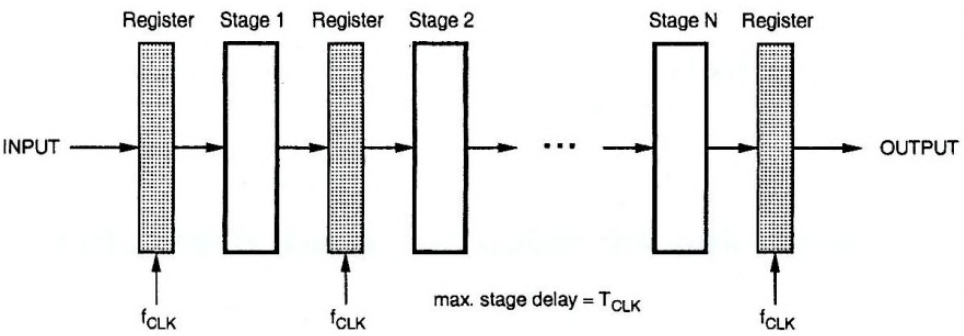


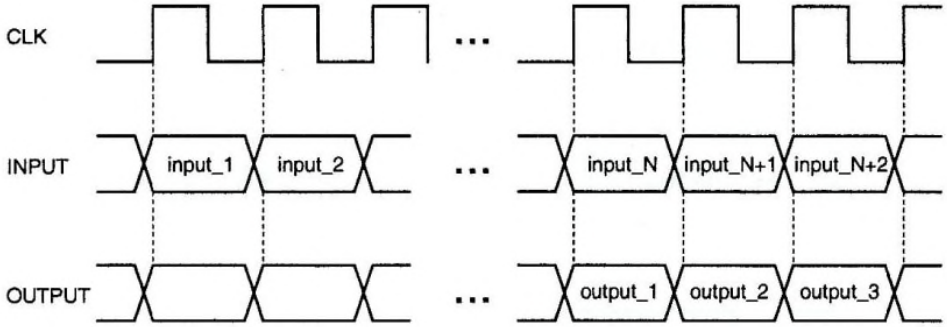


Power Reduction by Pipelining

- Partition F(INPUT) into *N* successive stages with (*N*–1) register arrays inserted. All registers are clocked at the original sample rate f_{CLK}.
- If all stages of partitioned function have approximately equal delay of
  $T_{pipe} = T_{ref} / N$
- The logic blocks between two successive registers can operate *N*-times slower while maintaining the same throughput. Thus, the supply voltage can be reduced to βV_{DD}, where $0 < \beta < 1$.
- The dynamic power consumption of the *N*-stage pipelined structure is:

$$P_{pipe} = \left[C_{total} + (N-1) C_{reg} \right] \cdot \beta^2 \cdot V_{DD}^2 \cdot f_{CLK}$$





Power Reduction by Pipelining

- The power reduction factor achieved by an *N*-stage pipeline is:

$$\frac{P_{pipe}}{P_{ref}} = \left[1 + \frac{C_{reg}}{C_{total}}(N-1)\right] \cdot \beta^2$$

- Example: Replacing a single-stage logic block at V_{DD} = 5 V, f_{CLK} = 20 MHz with a 4-stage pipelined structure running at the same clock frequency. Assuming $|V_T|$ = 0.8V, each pipeline stage can operate at 4 times slower speed at approximately 2V. With C_{reg}/C_{total} = 0.1, P_{pipe}/P_{ref} ≈ 0.2, which means a switching power saving of about 80%.
- There is an area overhead of (N – 1) register arrays. The latency increases from one to N clock cycles, but the throughput is preserved.
- In many applications, such as signal processing and data encoding, latency is not a major concern.

Estimation of Voltage Reduction Factor β

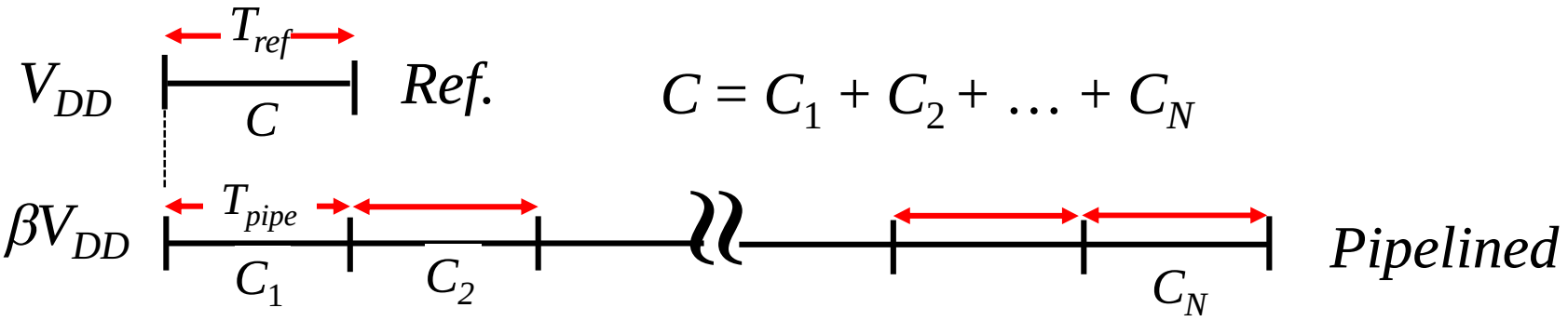
- The propagation delay of the original function and the pipelined function are:

$$T_{ref} = \frac{C_{charge} \cdot V_{DD}}{k\left(V_{DD} - V_t\right)^2}; \quad T_{pipe} = \frac{\left(C_{charge} / N\right) \cdot \beta V_{DD}}{k\left(\beta V_{DD} - V_t\right)^2};$$

C_{charge}: capacitance charged or discharged in a single clock cycle.

- To maintain the throughput, i.e., $T_{pipe} = T_{ref} \leq T_{CLK}$. Thus,

$$N\left(\beta V_{DD} - V_t\right)^2 = \beta\left(V_{DD} - V_t\right)^2$$



Example of β estimation

- V_{DD} = 5 V, V_t = 0.8 V, N = 4.

$$4\left(5\beta - 0.8\right)^2 = \beta\left(5 - 0.8\right)^2$$

$$4\left(25\beta^2 - 8\beta + 0.64\right) = 17.64\beta$$

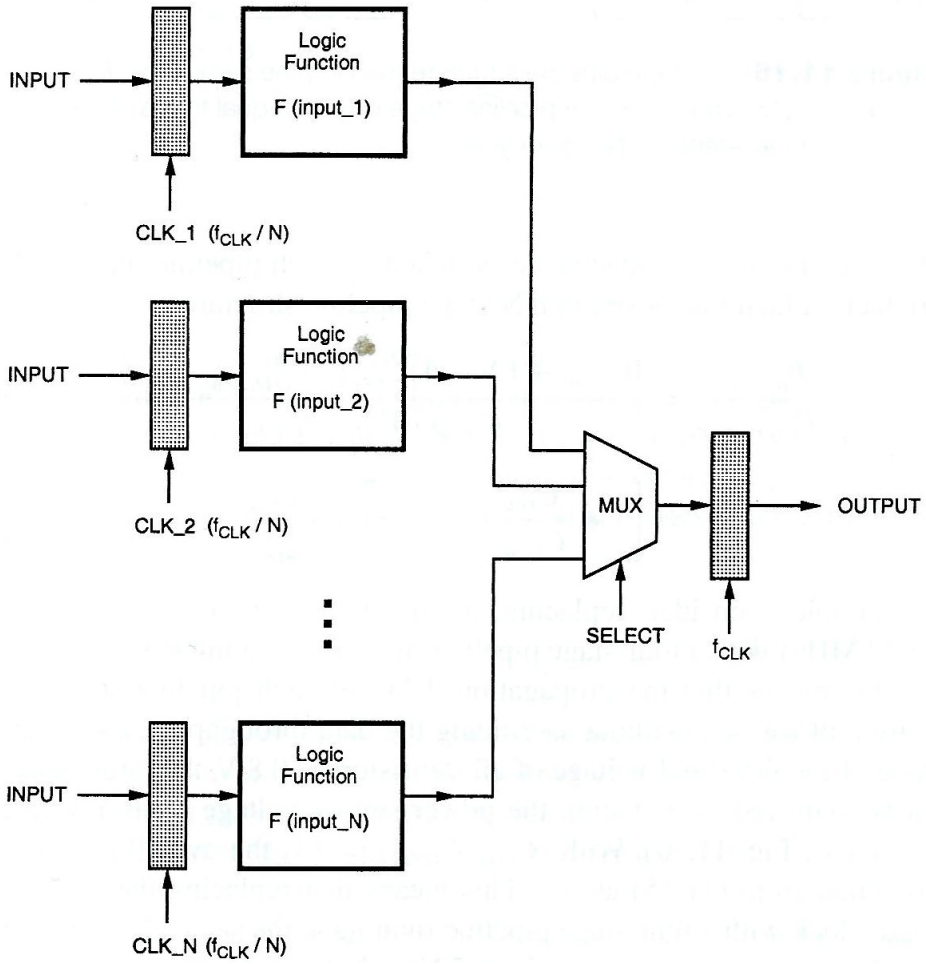
$$100\beta^2 - 49.64\beta + 2.56 = 0$$

$$\beta = \frac{49.64 \pm \sqrt{49.64^2 - 4 \times 100 \times 2.56}}{2 \times 100} = 0.44 \text{ or } 0.058$$

- If β = 0.44, reduced V_{DD} = 2.2 V (feasible)
- If β = 0.058, reduced V_{DD} = 0.29 V (infeasible).

Parallel Processing (Hardware Replication)

- *N* replicates of F(INPUT).
- Consecutive input vectors arrive at the same rate as in single-stage.
- Input vectors are routed to all the registers of the *N* parallel blocks.
- Gated clock signal, each with clock period of (NT_{CLK}), are used to load each register every *N* clock cycles such that each of the *N* consecutive input vectors is loaded into a different input register.
- The time to compute F(INPUT) for each input vector is increased to NT_{CLK} and V_{DD} can be reduced to βV_{DD}, where $0 < \beta < 1$.
- Output of the *N* processing blocks are multiplexed.
- The output register operates at f_{CLK}.

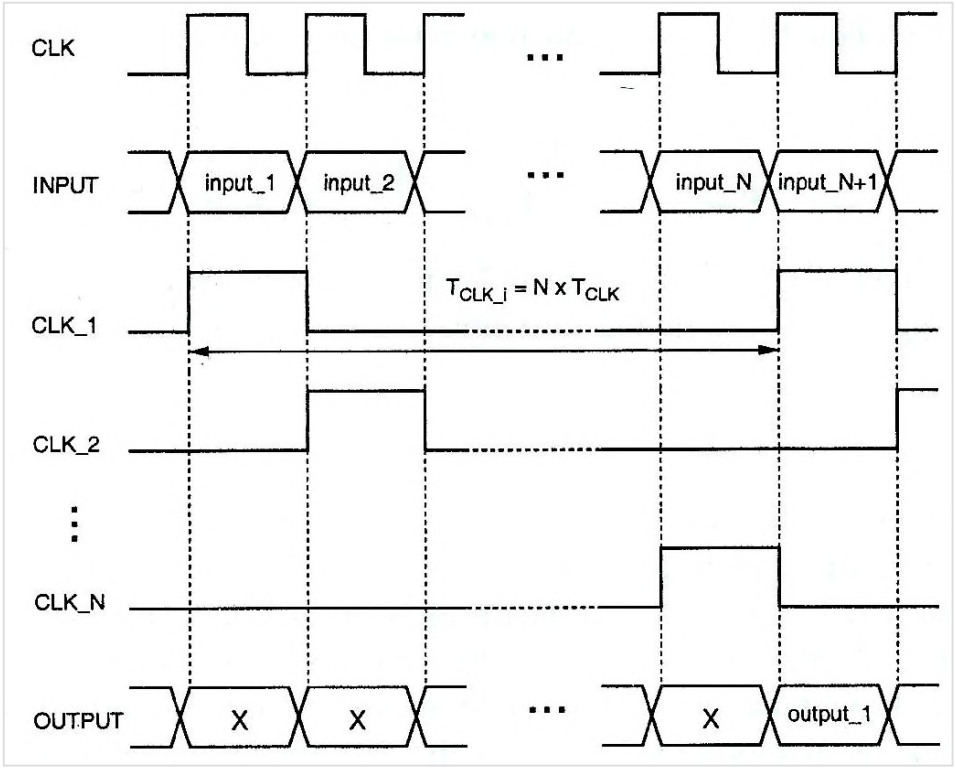


Power Reduction by Parallel Processing

- Neglecting the overhead of input and output routing capacitance and the output multiplexor structure, all of which are increasing function of *N*,

$$P_{para} \approx N \cdot C_{total} \cdot \beta^2 \cdot V_{DD}^2 \cdot \frac{f_{CLK}}{N} - NC_{reg} \cdot \beta^2 \cdot V_{DD}^2 \cdot \frac{f_{CLK}}{N}$$

$$\approx \left(1 - \frac{C_{reg}}{C_{total}}\right) \cdot C_{total} \cdot \beta^2 \cdot V_{DD}^2 \cdot f_{CLK}$$



- Power reduction achievable:

$$\frac{P_{para}}{P_{ref}} = \left[1 - \frac{C_{reg}}{C_{total}}\right] \beta^2$$

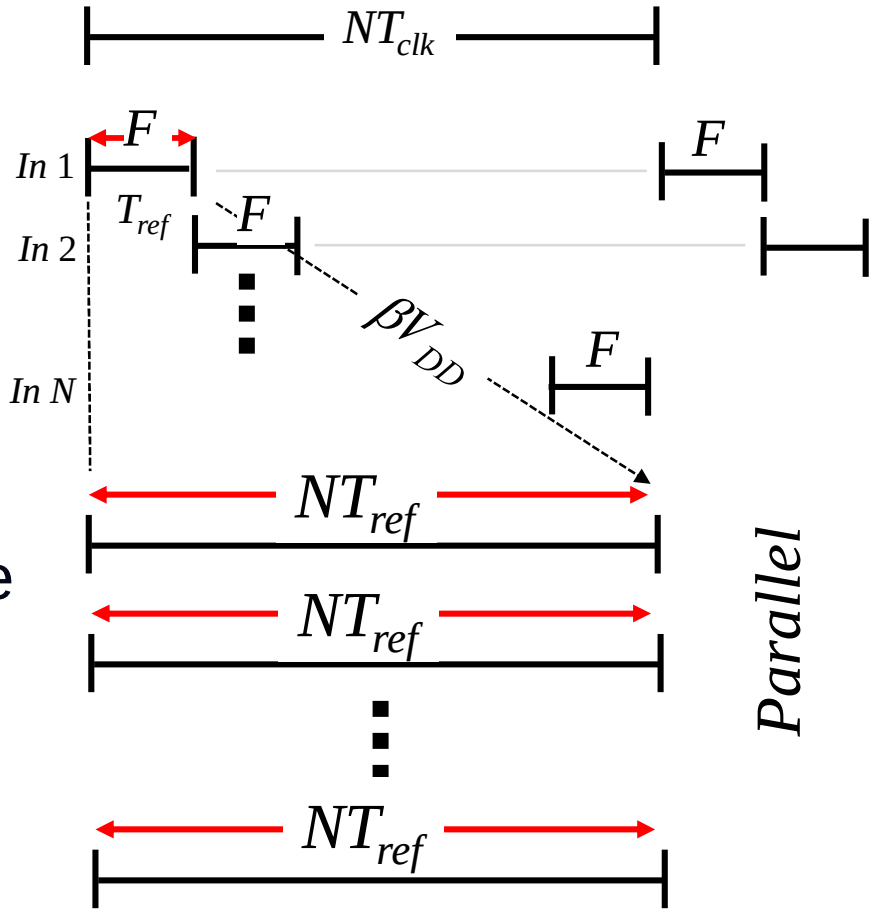
$$\geq \frac{1}{N^2}$$

Estimation of Voltage Reduction Factor β

- The propagation of the original function and the parallel function are:

$$T_{ref} = \frac{C_{charge} \cdot V_{DD}}{k\left(V_{DD} - V_t\right)^2};$$

$$T_{para} = \frac{C_{charge} \cdot \beta V_{DD}}{k\left(\beta V_{DD} - V_t\right)^2}$$

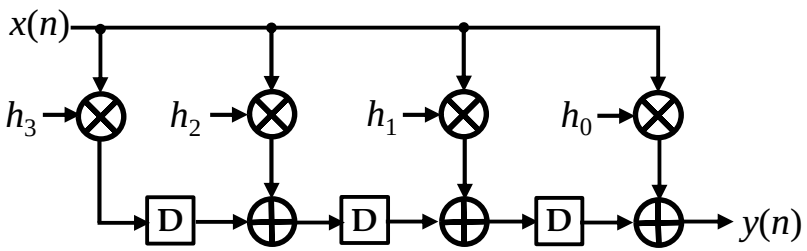


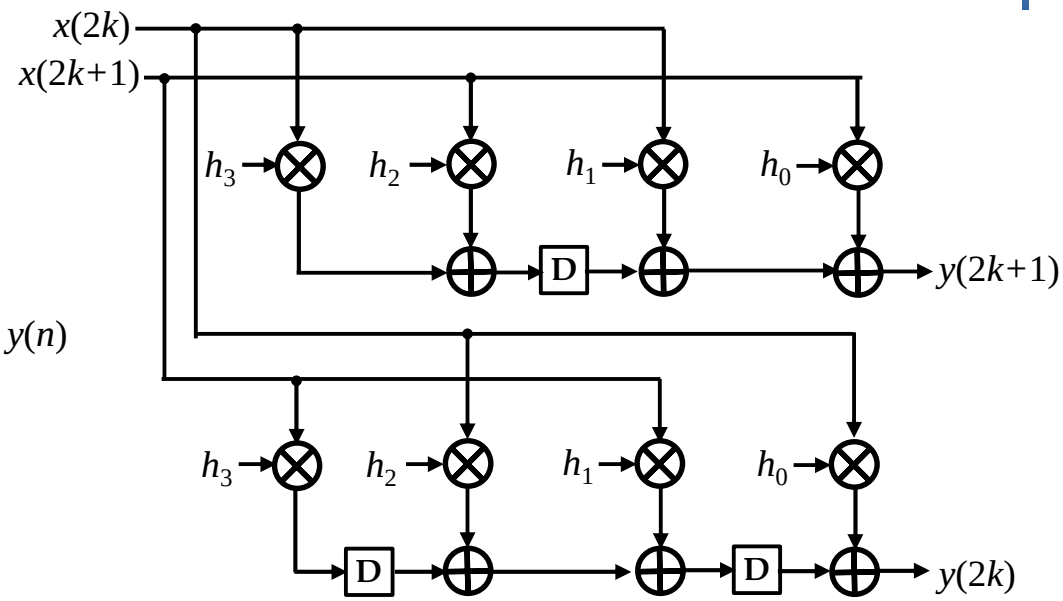
- To maintain the same throughput, $T_{para} = NT_{ref}$.

$$N\left(\beta V_{DD} - V_t\right)^2 = \beta\left(V_{DD} - V_t\right)^2$$

Example: Parallel FIR Filter

4-tap FIR filter





Given $T_M = 8$, $T_A = 1$, $V_t = 0.45$V, $V_{DD} = 3.3$V, $C_M = 8C_A$. If the two architectures are operating at the same throughput.

(a) What is the supply voltage of the 2-parallel filter?

(b) What is the power consumption of the 2-parallel filter as a percentage of the original filter?

Example: Parallel FIR Filter

Original filter: $T_{delay} = T_M + T_A = 9$ unit time

2-parallel filter: $T_{delay} = T_M + 2T_A = 10$ unit time

$T_{min} = 10$ => max. throughput that both designs can work properly

Original: $C_{charge} = C_M + C_A = 9C_A$

2-parallel: $C_{charge} = C_M + 2C_A = 10C_A$

$$\frac{10C_A \times \beta \times 3.3}{k(\beta \times 3.3 - 0.45)^2} = \frac{2 \times 9C_A \times 3.3}{k(3.3 - 0.45)^2}$$

$$9(3.3\beta - 0.45)^2 = 5\beta(3.3 - 0.45)^2$$

$\beta = 0.6589$ or 0.0282

$V_{para} = 0.66 \times 3.3 = 2.17$ V

%power reduction $\approx 1 - (0.66)^2 = 56.44\%$

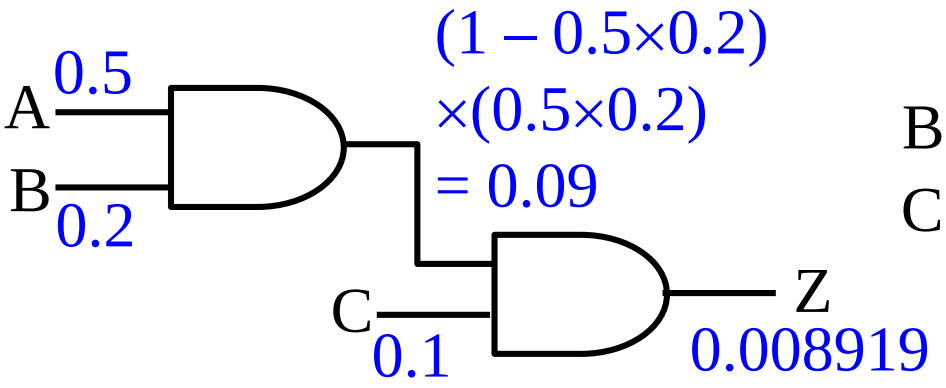
Ignoring registers,

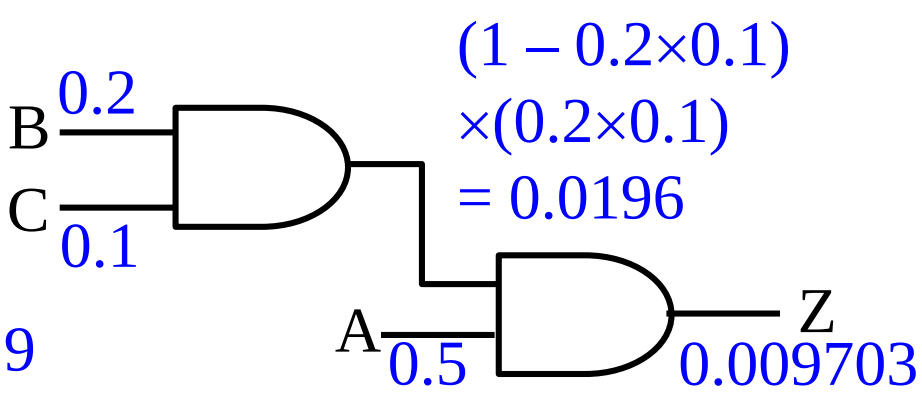
$$\frac{P_{para}}{P_{orig}} \approx \frac{70C_A \times 2.17^2 \times \frac{1}{20}}{35C_A \times 3.3^2 \times \frac{1}{10}} = 0.43$$

Reduction of Switching Activity

	$P_{0\to1}$
AND	$(1 - P_A P_B) P_A P_B$
OR	$(1 - P_A)(1 - P_B)\ \{1 - (1 - P_A)(1 - P_B)\}$
EXOR	$\{1 - (P_A + P_B - 2P_A P_B)\}\ (P_A + P_B - 2P_A P_B)$

Input ordering: The switching activity of a logic gate is a strong function of input signal statistics. It is better to postpone the introduction of signals with high transition rate. 0.09892 ⇨ 0.02930





Switching Power is Data-dependent

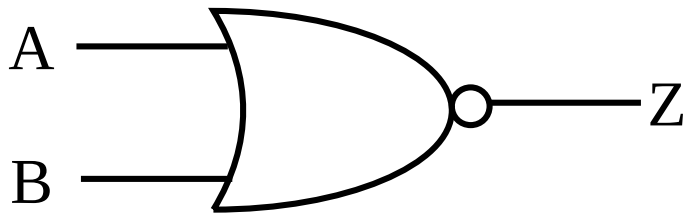
Let P_A and P_B be probability of logic '1' at inputs A and B, respectively of a NOR2 gate.

Then, $P_1 = (1 - P_A)(1 - P_B)$ and

$P_{0\to1} = (1 - P_1)P_1 = \{1 - (1 - P_A)(1 - P_B)\}(1 - P_A)(1 - P_B)$

If $P_A = P_B = ½$, then $P_{0\to1} = ¾$.

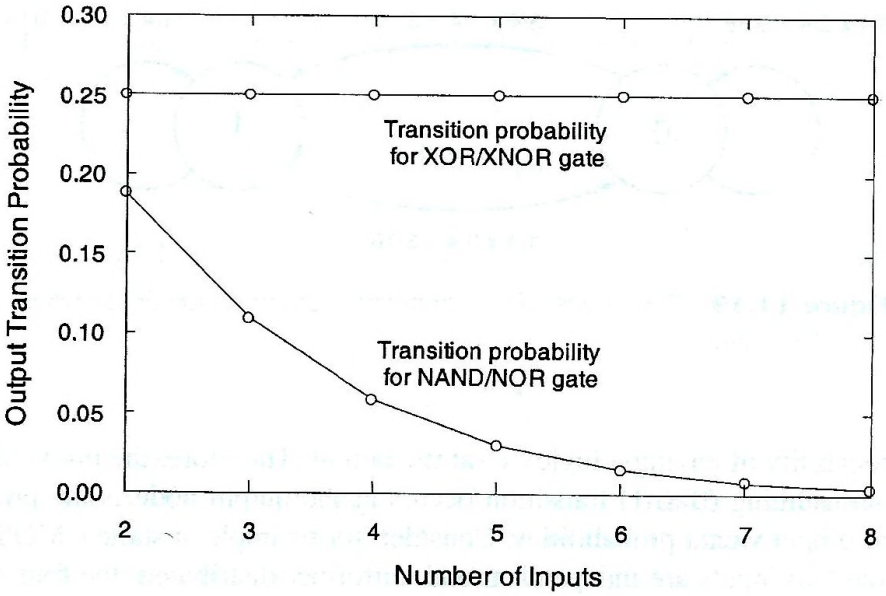
Suppose now that only patterns 00 and 11 can be applied (w/ equal probabilities) to a NOR gate. Then $P_{0\to1} = 1/4$

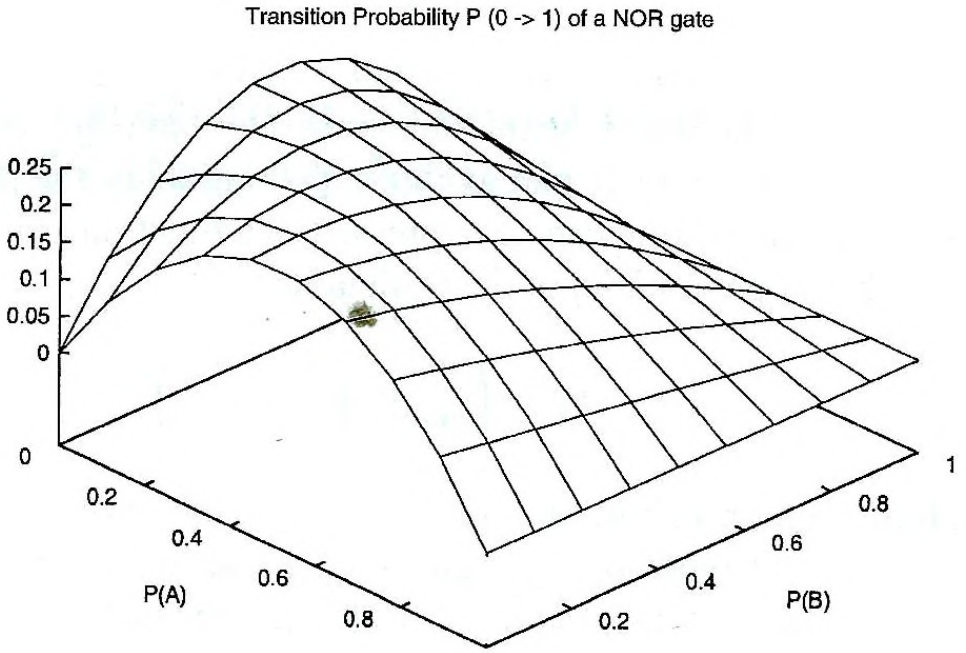


A	B	Z
0 → 0	0 → 0	1 → 1
0 → 1	0 → 1	1 → 0
1 → 0	1 → 0	0 → 1
1 → 1	1 → 1	0 → 0

Reduction of Switching Activity

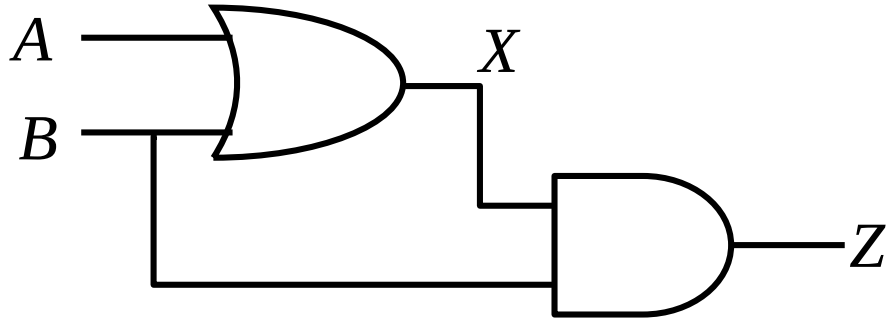
In multi-level logic circuits, distribution of input signal probabilities is typically non-uniform. The output transition probability becomes a function of the input probability distributions.





Re-convergent Fanout

Evaluation of switching activity becomes complicated in large circuits, with sequential elements, reconvergent nodes and feedback loops.



In this case, $Z = B$ as it can be easily seen.
The previous analysis simply fails because the signals are not independent!

$P(Z = 1) = P(B = 1) \cdot P(X = 1 \mid B = 1) = P(B = 1)$

Becomes complex and intractable real fast!

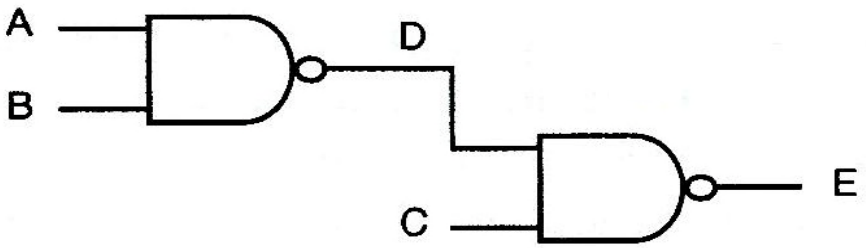
Logic Restructuring

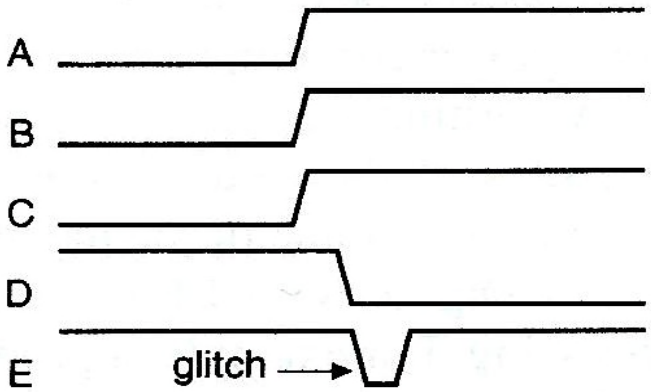
- Ignore glitching and assume uncorrelated random inputs, chain implementation has overall lower switching activity than the tree implementation. However, tree topology experiences lower glitching activity since the signal paths are more balanced.

- Glitch Reduction by balancing signal paths. By equalizing the arrival times of the input signals to a gate can dramatically reduce the number of spurious transitions due to critical race or dynamic hazards.

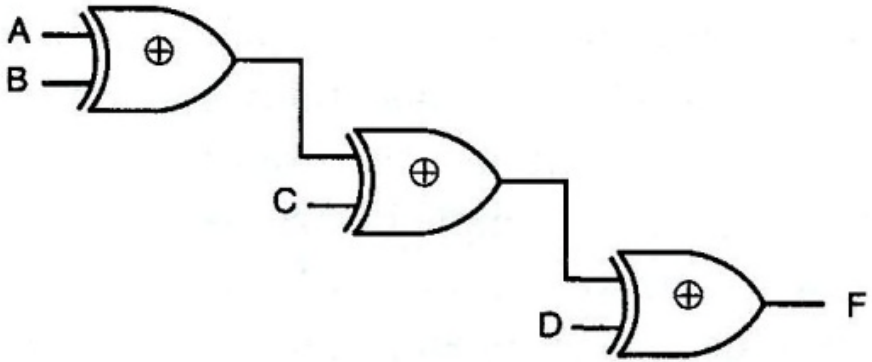
Examples: Reduction of Switching Activity

Dynamic hazard or glitch can occur if input changes at different times due to gate delay.

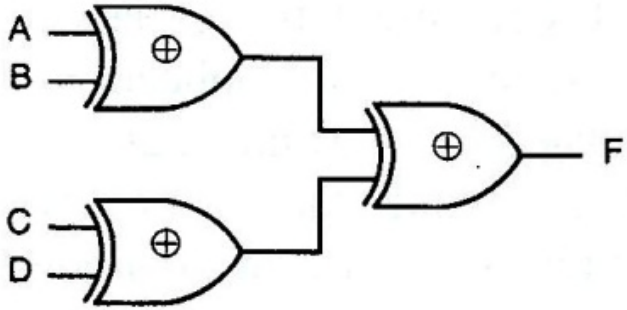




Implementation of four-input parity function



Chain structure



Tree structure reduces glitches

System-Level Switching Activity Reduction

- Algorithmic optimization: dynamic range, correlation, and statistics of data transmission. Examples: in vector quantization algorithm, number of memory accesses, number of multiplications and additions can be reduced by a factor of 30 if differential instead of full tree search algorithm is used.
- In applications where data bits change sequentially and are highly correlated (such as address bits to access instructions), Gray code (as opposed to binary code) reduces the number of transitions.
- Change of sign causes transitions of higher-order bits in 2's complement representation but only the sign bit will change in sign-magnitude representation. Switching activity can be reduced by using sign-magnitude in applications where data sign changes are frequent.

Dynamic Power Management

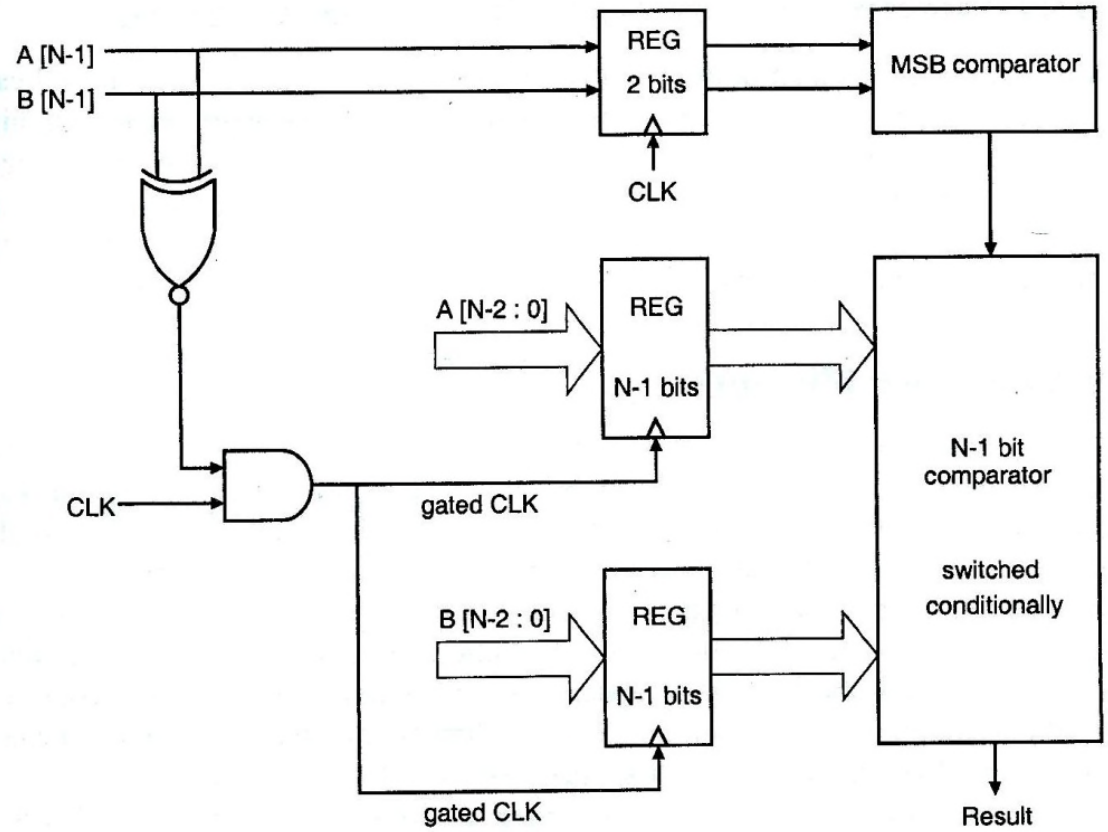
- Dynamically reconfigures an electronic system to provide the requested services and performance levels with a minimum number of active components or a minimum load on such components.
- Selectively turns off or reduce the performance of idle or partially unexploited components.
- DPM techniques:
  - Predictive techniques.
  - Static techniques.
  - Adaptive techniques.
  - Power shut down.
  - Clock gating.

Power Down Mode

- Power-down mode: put the modules that are not needed for processing in standby mode.
- It is more control oriented than architecture driven.
- It requires three major changes to the system architecures:
  - Conditioning the clock in the power-down section by a power down control signal.
  - Adding a state to the affected section's control which corresponds to the power-down mode.
  - Modify control logic to prevent the power-down and power-up from corrupting the states of the machine.

Simple Example of Gated Clock Scheme

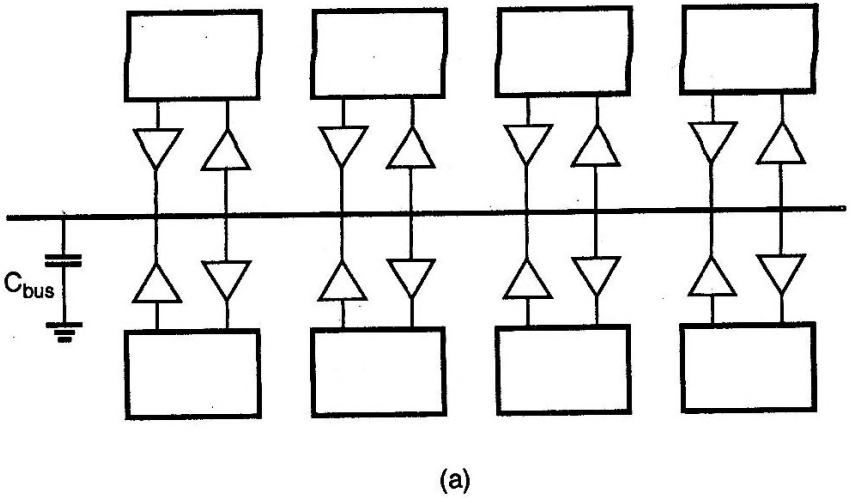
- Magnitude comparison of two unsigned *N*-bit binary numbers.
- Overall switching power reduced by ≈ 50%, since a large portion of the system is disabled for half of all input combinations.



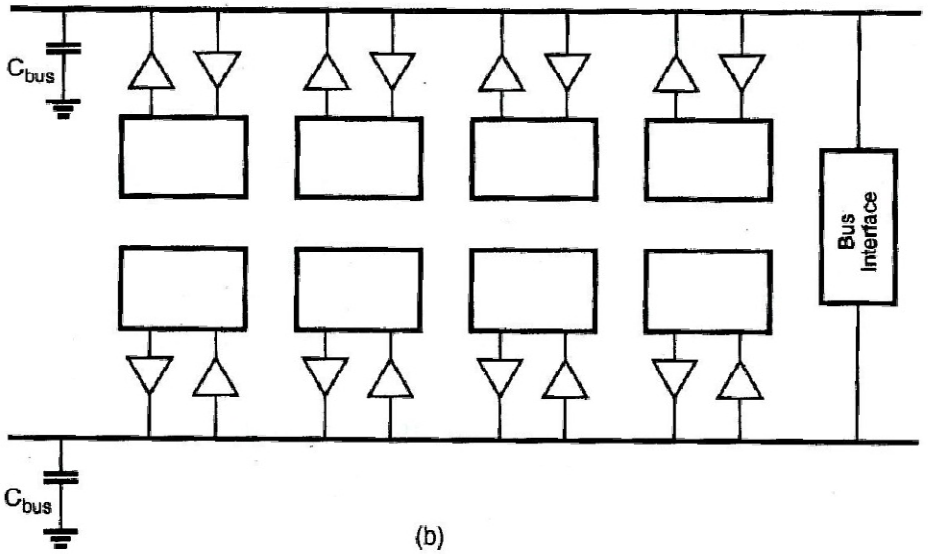
If the two MSBs are different, clock signal to lower-order registers are disable. Otherwise, the gated clock signal is applied and decision is made by the $(N-1)$-bit comparator circuit.

System-level Switched Capacitance Reduction

- At system level, limit the use of shared resources.
- Partition a global bus into a number of smaller dedicated local buses to reduce the switched capacitance during each bus access.



(a)



(b)

A single global bus structure:

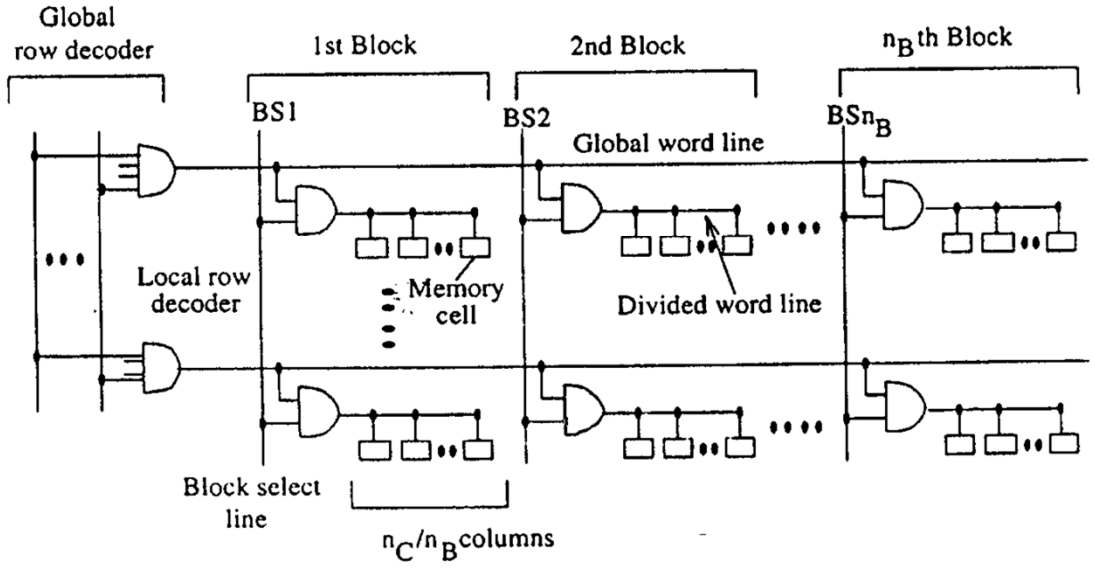
(i) large number of drivers and receivers sharing the transmission medium.

(ii) parasitic capacitance of long bus line.

Using smaller local buses reduces the amount of switched capacitance, at the expense of additional chip area.

Divided Word-Line (DWL)

- Cell array and word-line is divided into n_B blocks. If a SRAM has n_C columns, Each block has n_C/n_B columns.
- DWL of each block is activated by the main word-line and the corresponding block select signal. Only the memory cells connected to one divided word-line within a selected block are accessed in a cycle.
- The column current is reduced, since only the selected columns switch.
- With reduced total capacitance of connected transistors, word-line selection delay is also reduced.



$$P_{mem\_array} = m \times P_{actm}$$

$$P_{actm} = C_{BL} \Delta V_{BL} V_{DD} f$$

m: number of selected cells

Physical-level Switched Capacitance Reduction

- The effect of interconnect parasitic capacitance on the latency, power consumption and VLSI area to become increasingly non-trivial at deep submicron process.
- The parasitic capacitance can be reduced by circuit style selection, transistor sizing and incorporating power-conscious layout scheme where fewer and smaller devices as well as fewer and shorter interconnects are used at the physical/layout level.
- The physical place and route should be optimized such that signals that have high switching activity (such as clocks) should be assigned short wires and signals with lower switching activity are allowed to use relatively longer wires.

How about energy consumption?

- A battery-operated 65nm digital CMOS device is found to consume equal amounts (P) of dynamic power and leakage power while the short-circuit power is negligible. The energy consumed by a computing task, that takes T seconds, is $2PT$.
- Compare two power reduction strategies for extending the battery life:
  A. Clock frequency is reduced to half, keeping all other parameters constant.
  B. Supply voltage is reduced to half. This slows the gates down and forces the clock frequency to be lowered to half of its original (full voltage) value. Assume that leakage current is held unchanged by modifying the design of transistors.

Solution A: Reduce clock frequency

- Reducing the clock frequency will reduce dynamic power to $P/2$, keep the static power the same as P, and double the execution time of the task.
- Energy consumption for the task will be,

  Energy = $(P/2 + P)\ 2T = 3PT$

  which is greater than the original 2PT.

Solution B: Supply Voltage Reduction

- When the supply voltage and clock frequency are reduced to half their values, dynamic power is reduced to $P/8$ and static power to $P/2$. The time of task is doubled and the total energy consumption is,

  Energy = $(P/8 + P/2)\ 2T = 5PT/\ 4 = 1.25PT$

- *The voltage reduction strategy reduces energy consumption while a simple frequency reduction consumes more energy.*

All The Best!