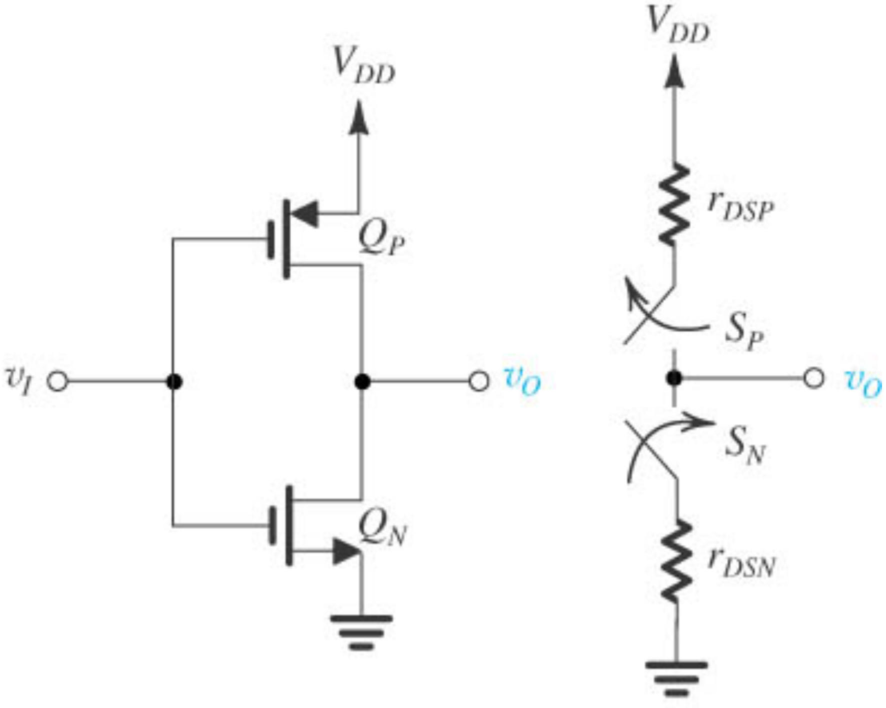
Power Consumption of Digital Circuits
4190.309
2008 Fall Semester

Naehyuck Chang
Dept. of EECS/CSE
Seoul National University
naehyuck@snu.ac.kr

Seoul National University

CMOS Inverter

- The CMOS inverter and (b) its representation as a pair of switches operated in a complementary fashion



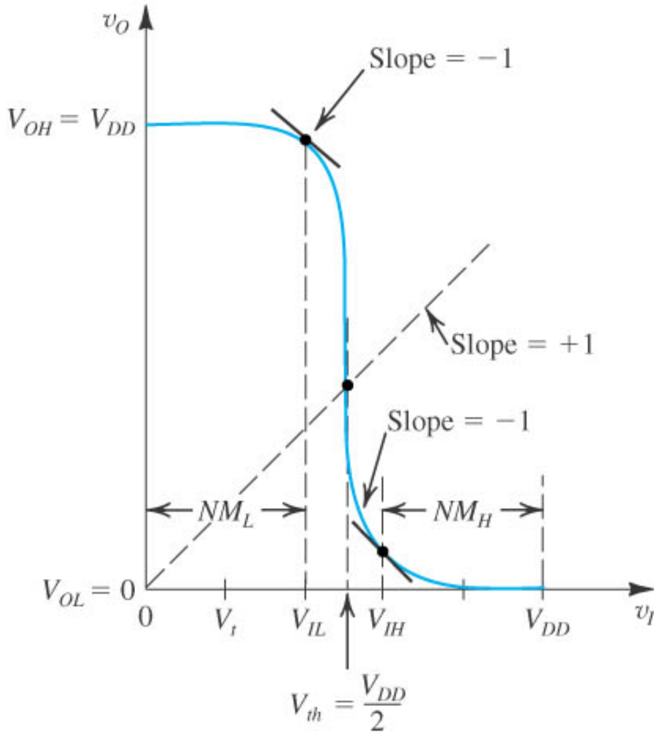
(a) (b)

$$r_{DSN} = \frac{1}{k_n'\left(\frac{W}{L}\right)_n (V_{DD} - V_t)}$$

$$r_{DSP} = \frac{1}{k_p'\left(\frac{W}{L}\right)_n (V_{DD} - V_t)}$$

CMOS Inverter

- Static operation
  - The voltage transfer characteristic (VTC) of the CMOS inverter when Q_N and Q_P are matched



$$V_{th} = \frac{V_{DD} - |V_{tp}| + \sqrt{\frac{K_n}{K_p}} V_{tn}}{1 + \sqrt{\frac{k_N}{k_p}}}$$

$$k_n = k_n' \left(\frac{W}{L}\right)_n$$

$$k_p = k_p' \left(\frac{W}{L}\right)_p$$

CMOS Inverter

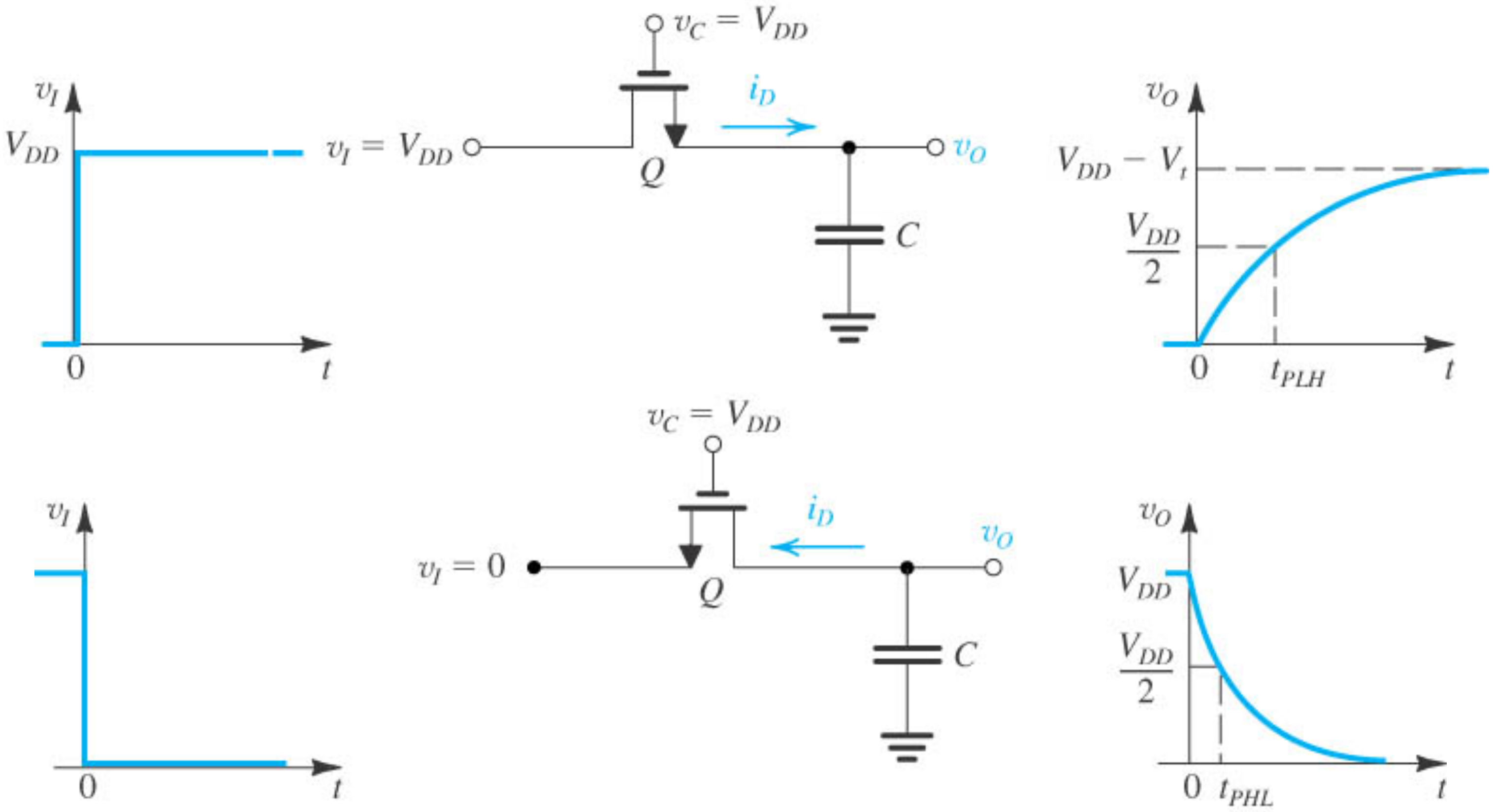
- Static operation
  - Matching for symmetrical transfer characteristic

$$\left(\frac{W}{L}\right)_p = \frac{\mu_n}{\mu_p}\left(\frac{W}{L}\right)_n$$

  - μ_n is 2 to 4 times larger than μ_p
  - Generally devices have the same channel length for a given technology
  - Device size: $(n+p)L^2$ where n=1.5 and p=4.5 for example

CMOS Inverter

- Dynamic operation

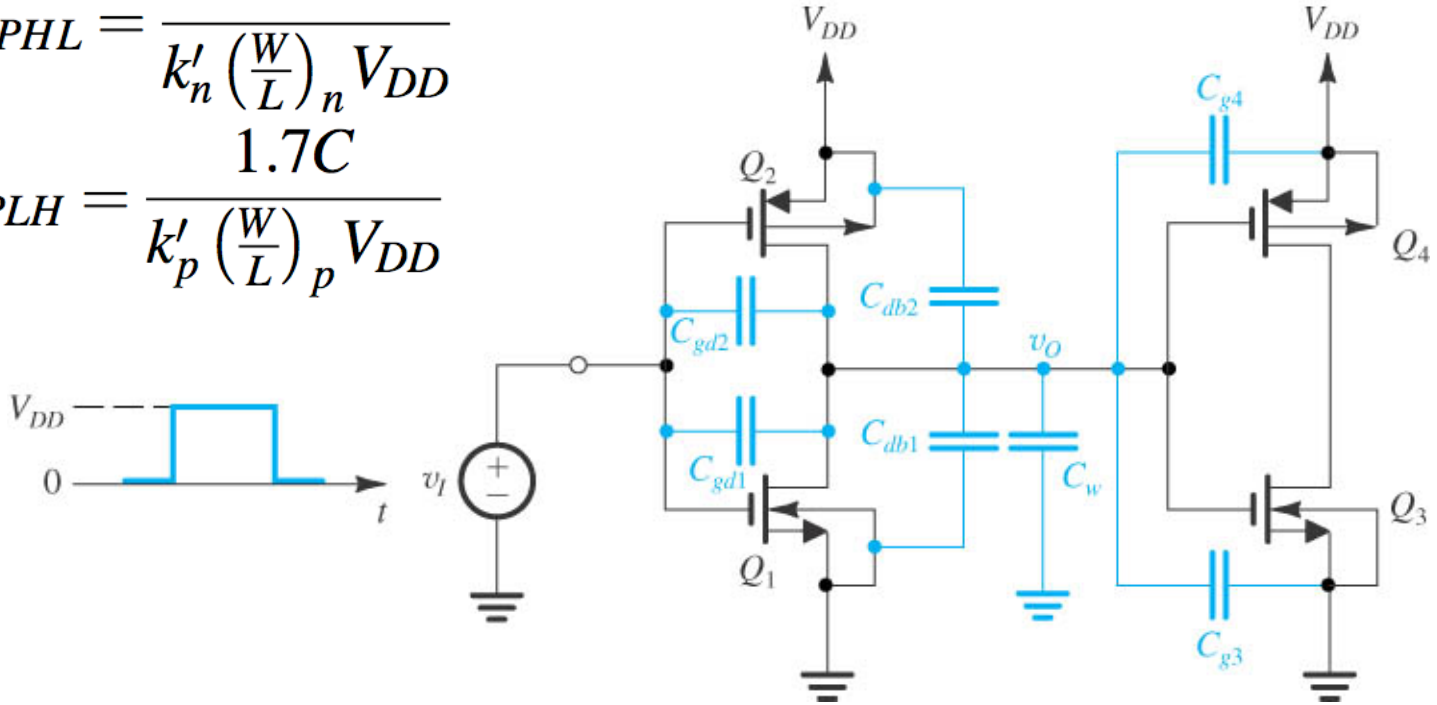


CMOS Inverter

- Dynamic operation

$$t_{PHL} = \frac{1.6C}{k'_n \left(\frac{W}{L}\right)_n V_{DD}}$$

$$t_{PLH} = \frac{1.7C}{k'_p \left(\frac{W}{L}\right)_p V_{DD}}$$



Power and Energy

- Power consumption of digital circuits is defined by the supply voltage times the current flow from V_{DD} to GND
  - Generally, V_{DD} is constant and I_{DD} is variable
- Instantaneous power: $P(t) = I_{DD}(t)V_{DD}$
- Energy: $E(T) = \int_0^T P(t)dt$
- Average power: $\overline{P(T)} = \frac{E(T)}{T}$

Source of Power Consumption

- Dynamic power
  - Current flow from VDD to GND when logic transition occurs
    - Switching power
    - Short-circuit power
    - Glitch power
- Static power
  - Current flow from VDD to GND regardless of logic transition
    - DC current
    - Leakage power

Source of Power Consumption

- Traditional CMOS circuits
  - Slow operation
    - Negligible dynamic power consumption
    - Electric watches, calculators, etc.
  - High V_{DD} and high V_T
    - Negligible leakage power consumption
    - Small short-circuit current

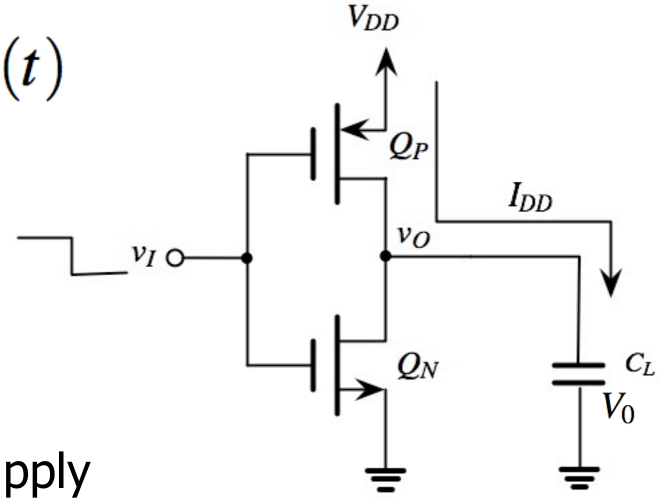
Source of Power Consumption

- Modern high-speed CMOS
  - Fast operation
    - High dynamic power
  - Low V_{DD} and low V_T
    - Less dynamic power but more leakage power per unit transistor
  - Power is the most important design constraints
    - Large-scale integration and thus power per unit area increase dramatically

Source of Power Consumption

- Switching power

$$P(t) = \frac{dE}{dt} = V_{DD} \times I_{DD}(t)$$



  - A step voltage is applied at t=0

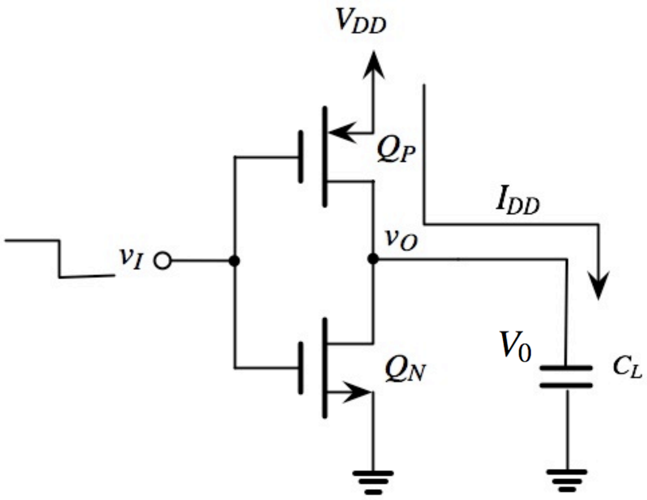
$$i_{DD}(t) = C_L \frac{dV_0}{dt}$$

  - Energy transferred from the power supply

$$E_{01} = \int_0^{t_d} P(t)dt = V_{DD}C_L \int_0^V dV_0 = C_L V_{DD} V$$

Source of Power Consumption

- Switching power
  - When $V = V_{DD},\ E_{0\to 1} = C_L V_{DD}^2$
  - $C_L V_{DD}^2/2$ is dissipated by heat
  - $C_L V_{DD}^2/2$ is stored in the capacitor
  - The remaining $C_L V_{DD}^2/2$ is dissipated by heat again when high-to-low transition occurs
  - High-to-low transition does not draw additional current from the power supply

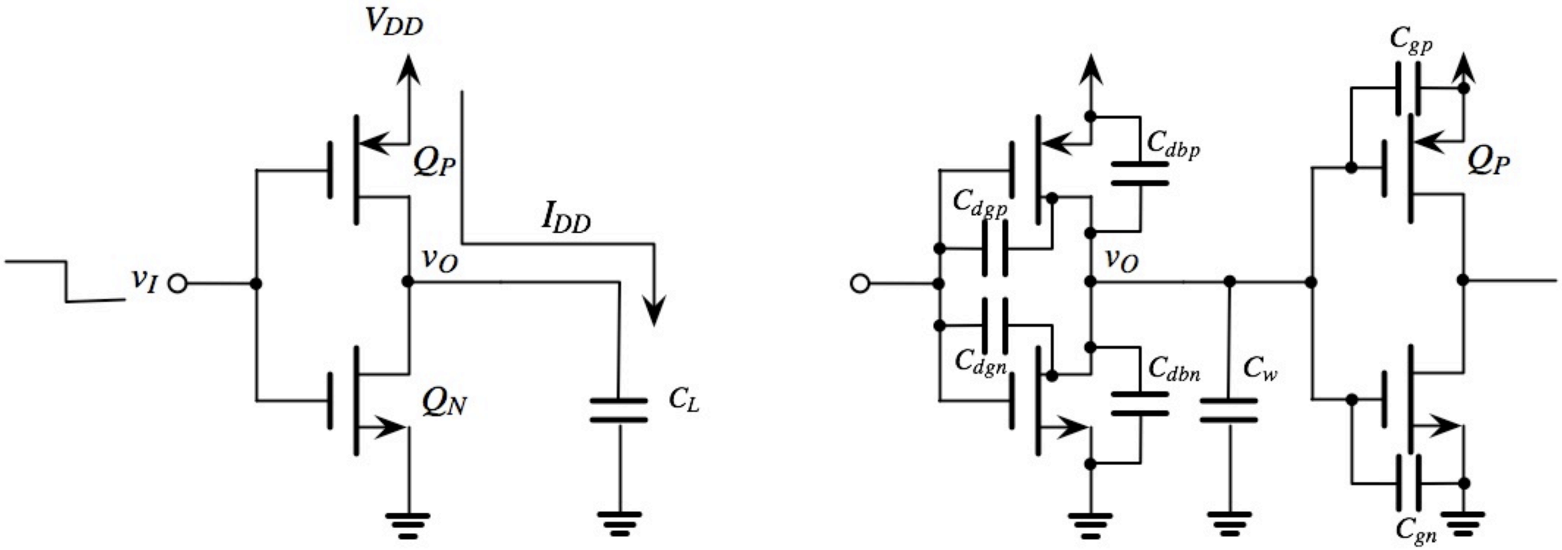


Source of Power Consumption

- Switching power:

$$P_{sw} = f_{sw} V_{DD}^2 C_L$$

$$E_{tot} = V_{DD} Q = V_{DD} C_L \Delta V = \frac{C_L C_{int}}{C_L + C_{int}} V_{DD}^2 = (C_L || C_{int}) V_{DD}^2$$



Source of Power Consumption

- Gate capacitance

$$C_g = C_{sg} + C_{dg} + C_{bg}$$

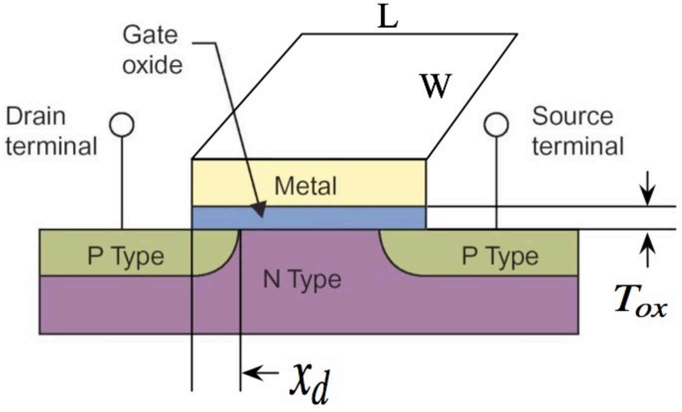
  - C_{gb} : sum of the gate-to-bulk capacitances

- Overlap capacitance

$$C_{ov} = C_{dg1} + C_{dg2} + C_{dg3} + C_{dg4} + C_{sg3} + C_{sg4}$$

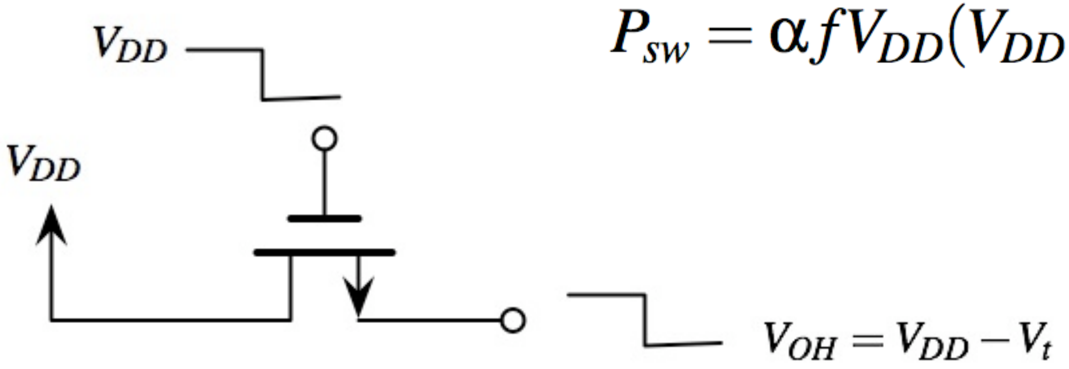
  - Due to Miller effect: $C_{dg1} = C_{dg2} = 2C_{ox}x + dW$
  - $C_{dg3} = C_{dg4} = C_{sg3} = C_{sg4} = C_{ox}x + dW$

- Diffusion capacitance
- Interconnect capacitance



Source of Power Consumption

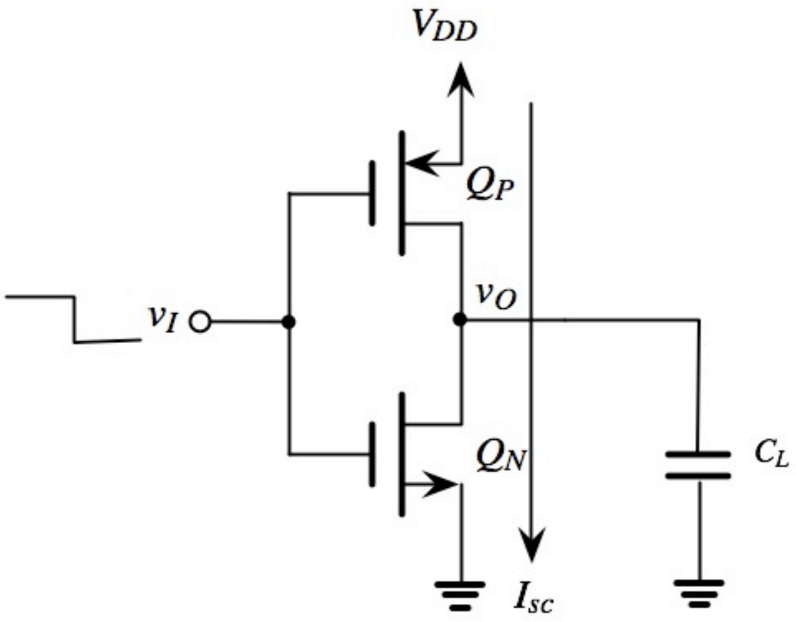
- Reduced swing switching power
  - Rail-to-rail swing: V_{DD} to GND
  - When $V_{OH} < V_{DD}$, swing is V_{OH} to GND
  - Reduced bit-line in memory

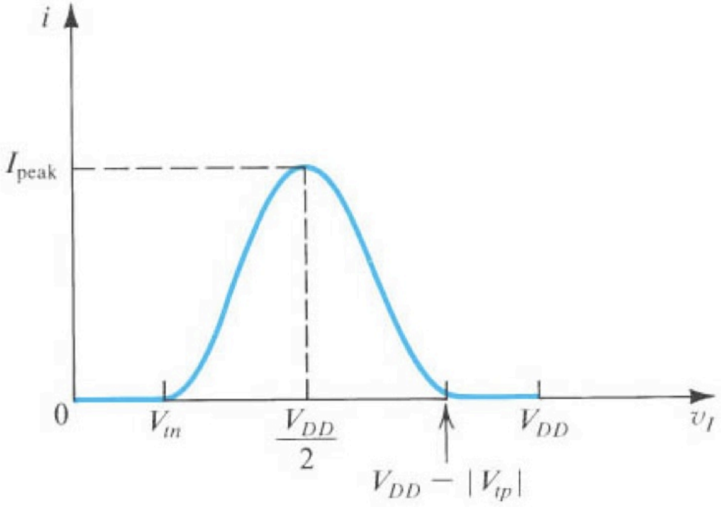


$$P_{sw} = \alpha f V_{DD}(V_{DD} - V_{th})C_L$$

Source of Power Consumption

- Short-circuit power
  - Transient current from VDD to GND when logic transition occurs





Source of Power Consumption

- Short-circuit power

when assuming $V_{thn} = V_{thp} = V_{th}$

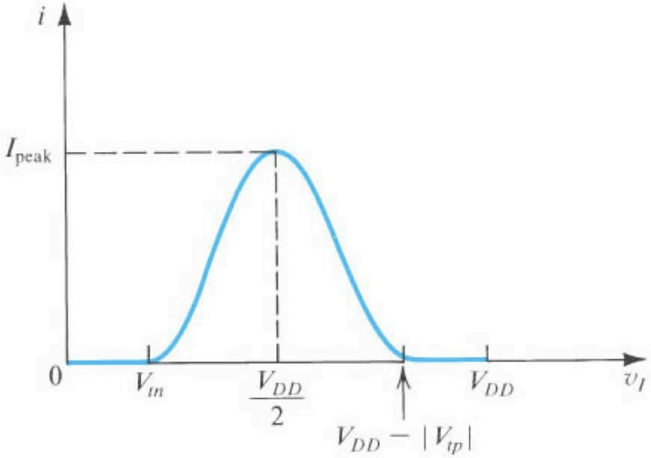
device parameter $\beta_n = \beta_p = \mu C_{ox}\frac{W}{L}$

μ : carrier mobility

C_{ox} : Oxide capacitance

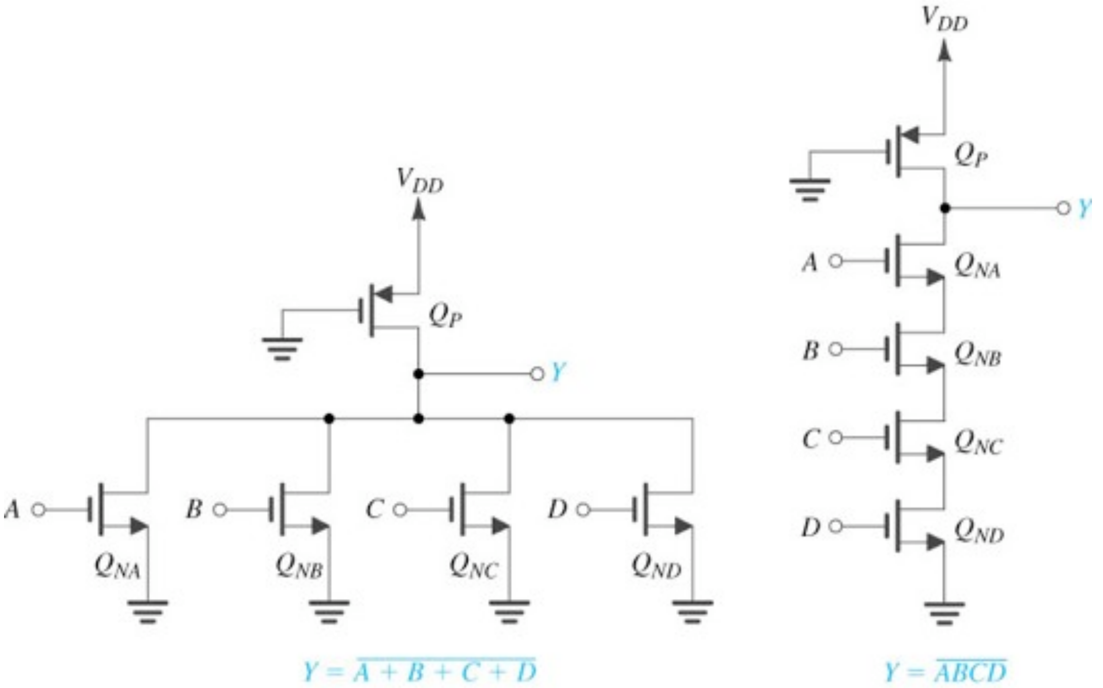
τ : rise and fall time

$$P_{sc} = \frac{\mu C_{ox}}{12}\frac{W}{L}(V_{DD} - 2V_{th})^3 \tau f$$



Static Power

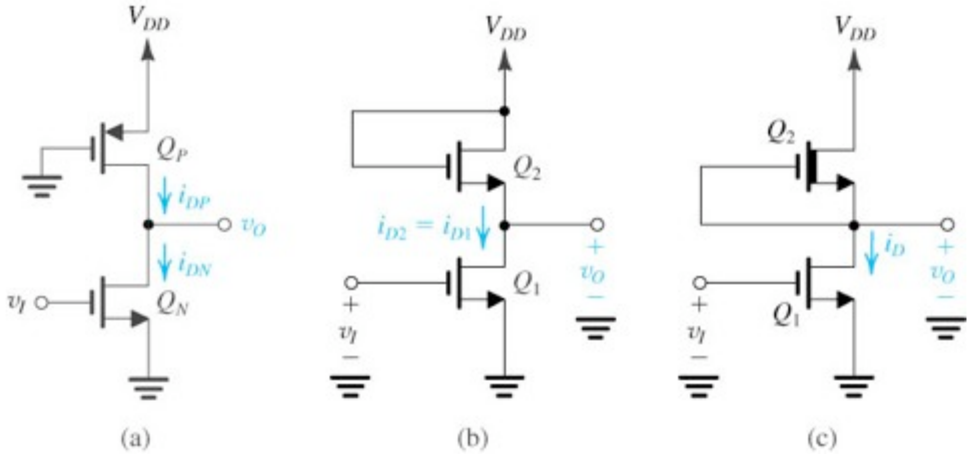
- DC current
  - Pseudo NMOS logic



(a) (b)

Static Power

- DC current
  - Steady current flow from VDD to GND
  - Either logic value is 0 or 1 depending on the logic structure
    - Mostly when the output is 0

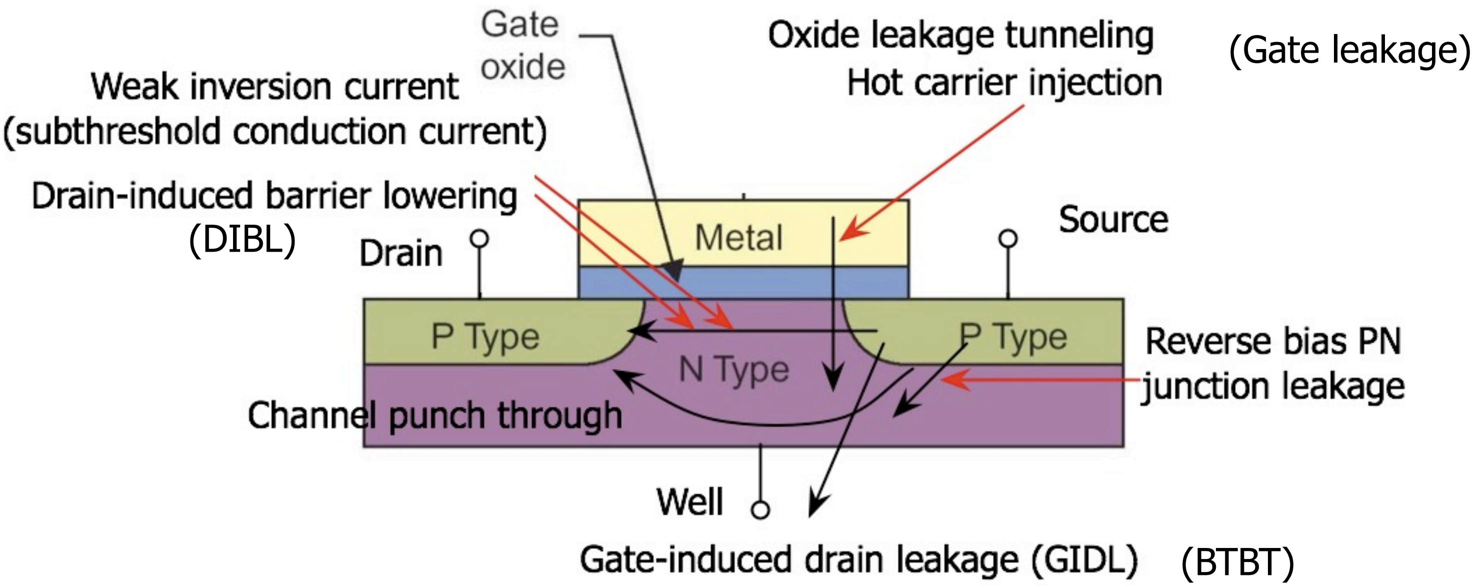


Static Power

- Leakage current
  - A transistor switch is a resistive-capacitive network between the power supply and GND
  - Non-ideal off-state characteristics (a finite resistance) makes current draw even when the transistor is in the cut-off state

Static Power

- Long channel (L>1um): negligible leakage
- Short channel (L>180nm, T_{ox}>30Å): subthreshold leakage
- Very short channel (L>90nm, Tox>20Å): subthreshold+gate leakage
- Nano scaled (L<90nm, T_{ox}<20Å): subthreshold+gate+BTBT leakage



Transistor Scaling

- Two key transistor scaling schemes
  - CE (Constant electric field) scaling
    - All the horizontal and vertical dimensions are scaled with the power supply to maintain constant electric fields throughout the device
    - Standard scaling methodology in industry in a 30% reduction (1/S=0.7) of all dimensions per generation
    - Supply and threshold voltages are scaled down by the factor of 1/S
    - Current, gate capacitances, and delay also scaled by 1/S
    - Results in 50% improvement in frequency
    - Improvement gradually degrades due to interconnect dominant delay
  - CV (Constant voltage) scaling
    - Maintains a constant power supply
    - Gradually scales the gate oxide thickness to slow down the growth of fields in the oxide

Transistor Scaling

- CE scaling
  - Switching energy scaled down by $1/S^3$
  - Dynamic power scaled down by $1/S^2$
  - Operating frequency scaled up by S
  - Dynamic power for a constant die size is the same
  - Number of switching elements scaled up by S^2
  - Leakage power increases exponentially
  - Total effective width of a device scaled up by S
- Example
  - Leakage power is 0.1% in 25um technology
  - Leakage power is 25% in 0.1um technology

Transistor Scaling

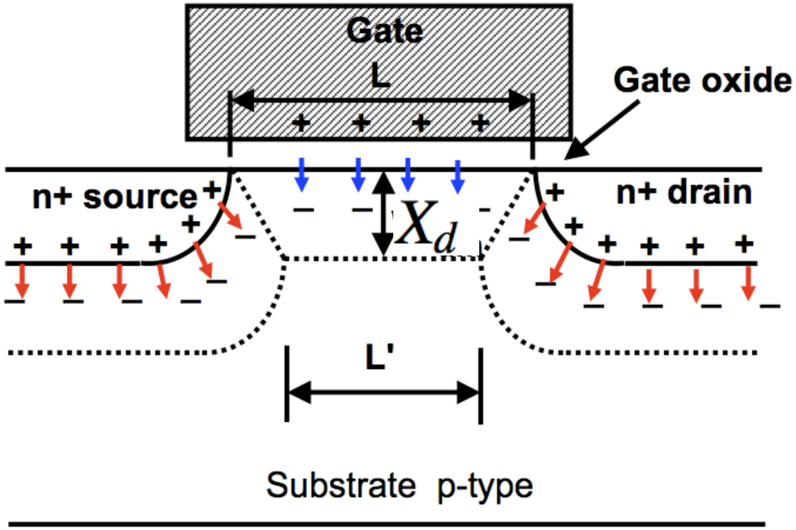
- Short-channel effect
  - V_T roll-off: charge partitioning model

$$V_T = V_{FB} - 2\phi_F - \frac{Q_B}{C_{ox}}$$

  - Long channel

$$Q_B \propto X_d \times L$$

  - Short channel



$$Q_B' \propto X_d \times \frac{L+L'}{2} < Q_B \rightarrow V_T \text{ decreases}$$

Transistor Scaling

- Alpha-Power model
  - Simple hand calculation model that empirically fits the real data

Measured data

$$I_{DS} = K_S \frac{W}{L} (V_{GS} - V_T)^{\alpha}$$

Measured data

  - α is close to 1 than 2, which is approximately 1.25, and continue to approach to 1 as technology scales

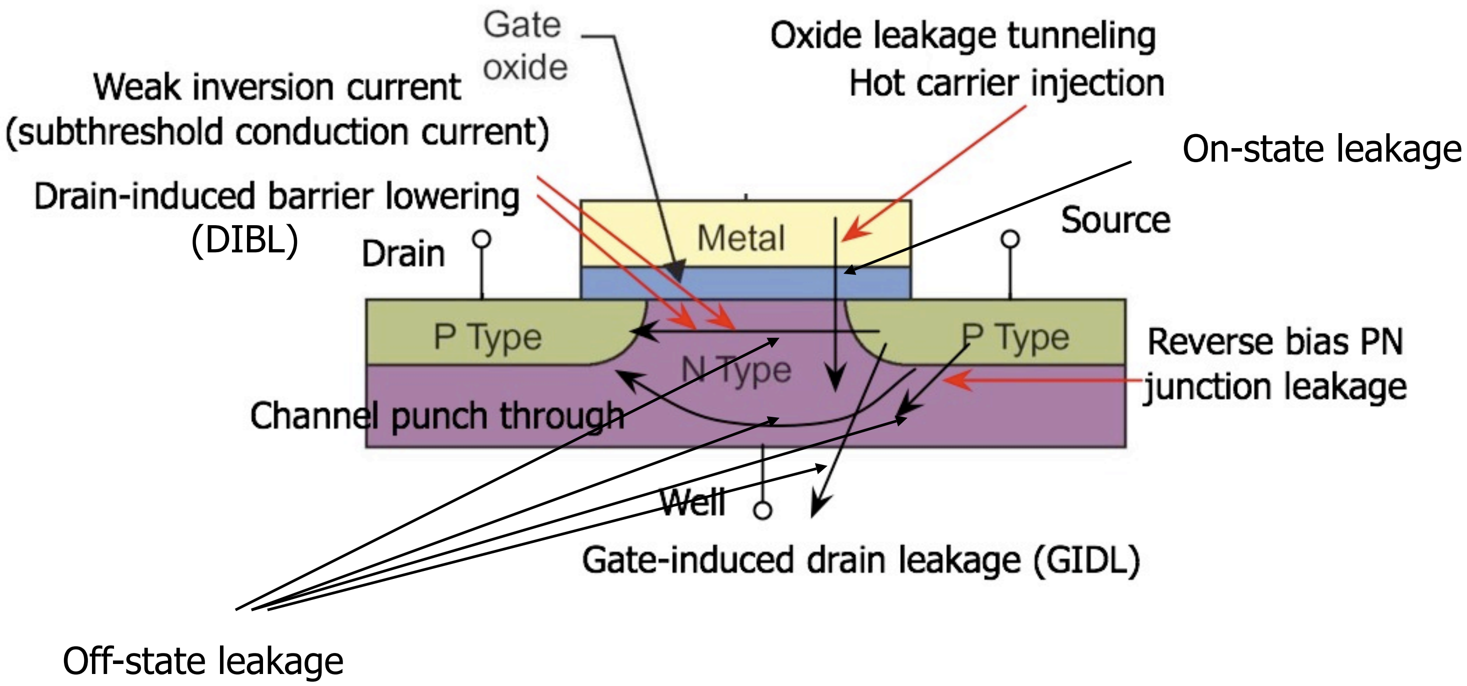
$$I_{ON} = I_0 (S\alpha)^{-\alpha} (V_{GS} - V_T)^{\alpha}$$

$$I_{sub} = I_0 e^{-} \quad e^{\frac{V_{GS}-V_T}{S}}$$

$$\text{Delay} \propto \frac{V_{DD}}{(V_{DD} - V_T)^{\alpha}}$$

Static Power

- Summary of leakage power components



Dynamic Power Reduction

Naehyuck Chang
Dept. of EECS/CSE
Seoul National University
naehyuck@snu.ac.kr

Seoul National University

Total Power Management

- Power minimization in both active and standby modes
  - Dynamic power in active mode
  - Subthreshold leakage power in standby mode

Switching Activity

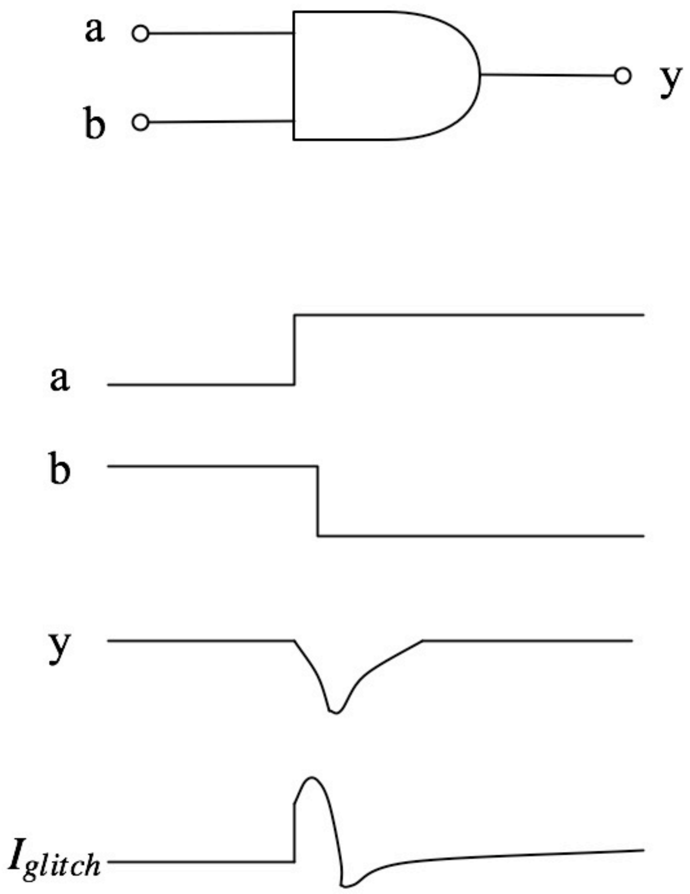
- Activity Factor: α
  - System clock frequency = f
  - Let $f_{sw} = \alpha f$, where α = activity factor
    - If the signal is a clock, $\alpha = 1$
    - If the signal switches once per cycle, $\alpha = ½$
    - Dynamic gates: switch either 0 or 2 times per cycle, $\alpha = ½$
    - Static gates: depending on design, but typically $\alpha = 0.1$

- Switching power:

$$P_{sw} = \alpha f V_{DD}^2 C_L$$

Switching Activity

- Abnormal switching activity
  - Glitch power
    - Power dissipated in intermediate transitions during the evaluation of the logic function
    - Unbalanced delay paths are principle cause
    - Usually 8% -25% of dynamic power



Path Balancing

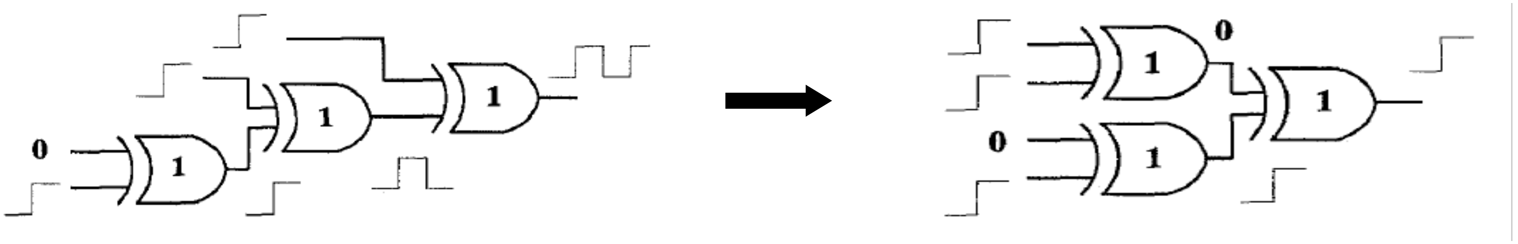
- Equalize the delay of input paths of each gate to reduce the possibility of spurious transitions
  - Spurious transitions are reported to amount to 10~40% of all switching activities
  - Balancing the paths

    ➔ Increase the possibility of simultaneous transition at the input

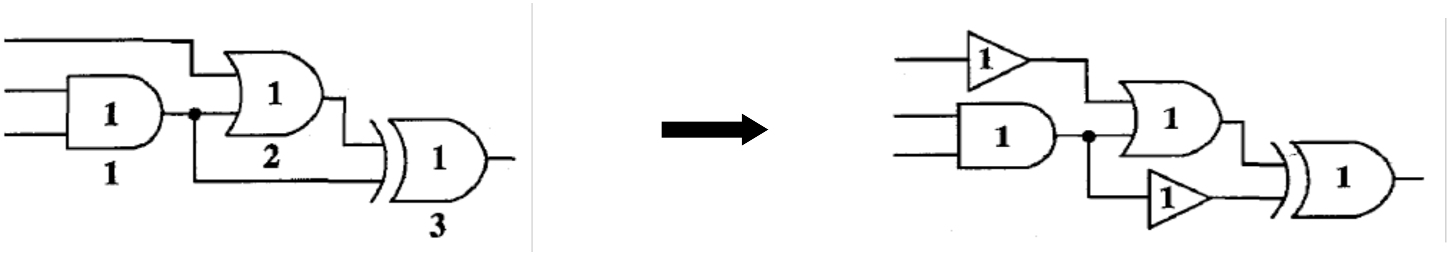
    ➔ Decrease the possibility of hazards at the output

Path Balancing

- Balance the paths by restructuring the logic circuit



- Balance the paths by inserting unit-delay buffers

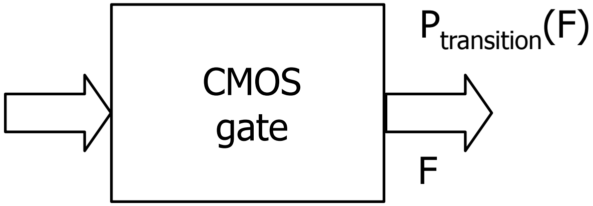


"Don't-care" Optimization

- Traditionally have been used for area minimization
  - Include appropriate "don't-care" sets in either the ON set or the OFF set
- Exploit the "don't-care" set so as to decrease the output transition probability
  - Include the "don't-care" set in the ON set if $P_{one}(F) > 0.5$
  - Include the "don't-care" set in the OFF set if $P_{one}(F) < 0.5$

* Transition probability of CMOS: $P_{transition}(F) = 2\ P_{one}(F)\ (1 - P_{one}(F))$

➔ Maximized when $P_{one}(F) = 0.5$ ($P_{one}(F)$: the probability of F being 1)



Logic Factorization

- Have been commonly used for area optimization
  - Reduce literal count to minimize the number of transistors being used to represent the target logic

$$a \times c + a \times d + b \times c + b \times d \longrightarrow (a+b) \times (c+d)$$

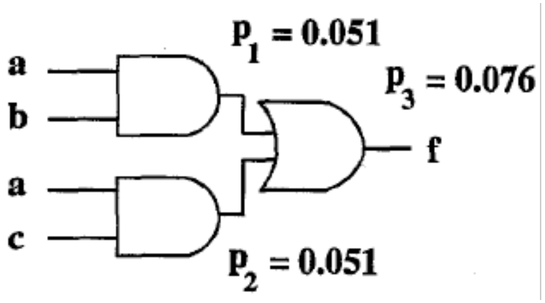
- Perform the factorization to reduce the switched capacitance
  - The smaller literal count does not guarantee the smaller switched capacitance unlike the case of area optimization
  - Should consider both the transition probability at the input and the load capacitance

Logic Factorization

- Should select the circuit (a) for area optimization
- Should select the circuit (b) for power optimization

Transition probability of input signals: $p_a = 0.1, p_b = 0.5, p_c = 0.5$

(a)

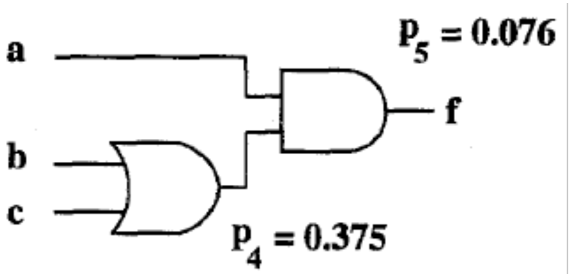


Switched capacitance

$$= (2p_a + p_b + p_c + p_1 + p_2 + p_3)C$$

$$= 1.378C$$

(b)



Switched capacitance

$$= (p_a + p_b + p_c + p_4 + p_5)C$$

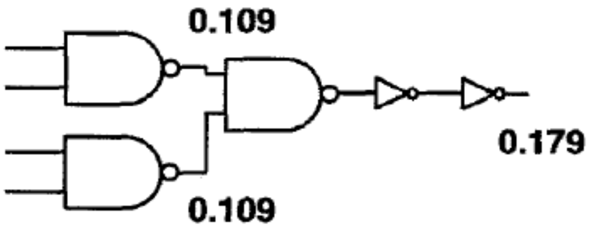
$$= 1.551C$$

Technology Mapping

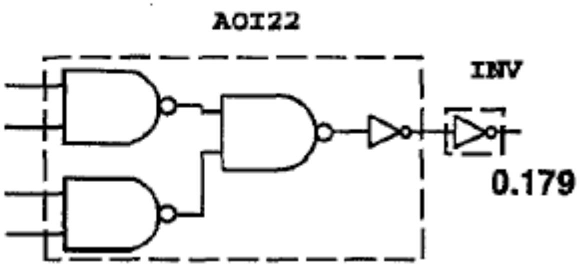
- The process of binding a set of logic equations to the gates in target cell library
  - Have been originally developed to optimize area and delay
- Hide nodes with high switching activity inside the gates
  - Generally, internal capacitances in gates are much smaller than external load capacitances
- Select the library with same function but different capacitances while meeting the delay constraints
  - Most technology libraries include the same logic element with different sizes

Technology Mapping

- Should select the circuit (a) for area optimization
- Should select the circuit (b) for power optimization



(a)

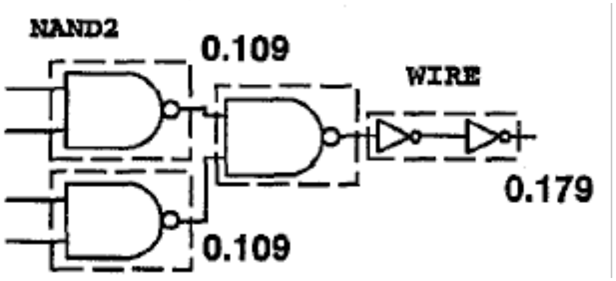


$$area = 2320 + 928 = 3248$$

$$sw.cap. = 0.179 \times (0.3410 + 0.0514) + 0.179 \times 0.1029$$
$$= 0.0887$$

Gate	Area	Intrinsic cap.	Input load cap.
INV	928	0.1029	0.0514
NAND2	1392	0.1421	0.0747
AOI22	2320	0.3410	0.1033

(b)



$$area = 1392 \times 3 = 4176$$

$$sw.cap. = 2 \times 0.109 \times (0.1421 + 0.0747) + 0.179 \times 0.1421$$
$$= 0.0725$$

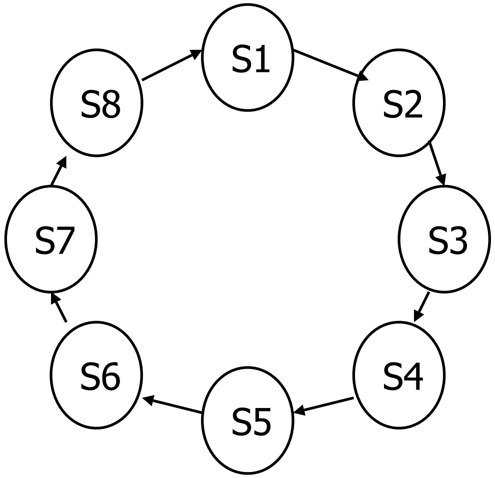
State Encoding

- The process of assigning a unique binary code to each state in a FSM (Finite State Machine)
  - Have been studied well for area minimization
- Assign codes with smaller Hamming distance to states with larger state transition probability when focusing on low power
  - Minimize the following cost function

$$f = \sum_{ij} w_{ij} \times H(c_i, c_j) \quad \left(\begin{array}{l} P_{ab}\text{: the transition probability from state } a \text{ to state } b, \\ c_a\text{: the codeword of state } a, \\ H(c_a, c_b)\text{: the Hamming distance between } c_a \text{ and } c_b, \\ w_{ij} = P_{ij} + P_{ji} \end{array} \right)$$

State Encoding

- Gray coding example

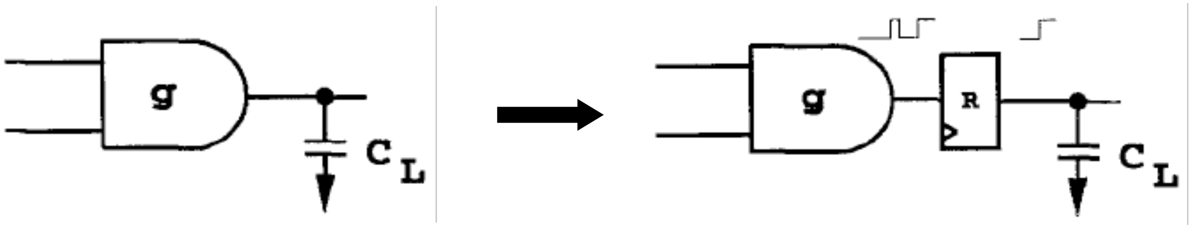


State	Gray	Binary
S1	0	0
S2	1	1
S3	11	10
S4	10	11
S5	110	100
S6	111	101
S7	101	110
S8	100	111
Total # of transitions	8	14
Max. transitions / cycle	1	3

- Need to consider not only the switching activity in the state registers but also in the combinational logic affected by assigned codewords for further optimization

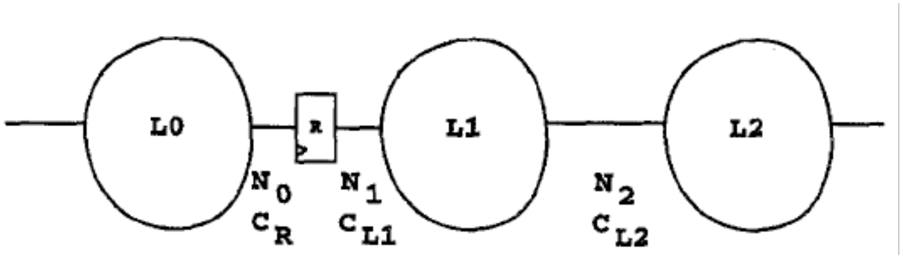
Retiming

- The process of repositioning registers (FFs) in a pipelined circuit (while maintaining I/O functionality)
  - First proposed to minimize the number of registers or the delay of the critical path (the longest pipeline stage)
- Pipeline the circuit by adding a register
  - Block the glitch propagation to the large load cap. (C_L)
    - Generally, input load cap. of registers are much smaller than C_L

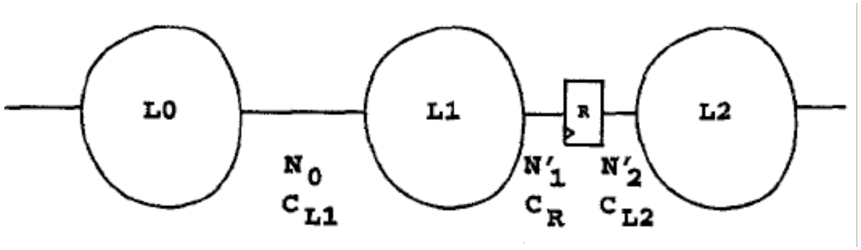


Retiming

- Move registers to nodes with higher switching activity
  - Maintain I/O functionality and (sequential) timing
  - May change the switching activity at one or more nodes
  - Choose the circuit with less switched capacitance



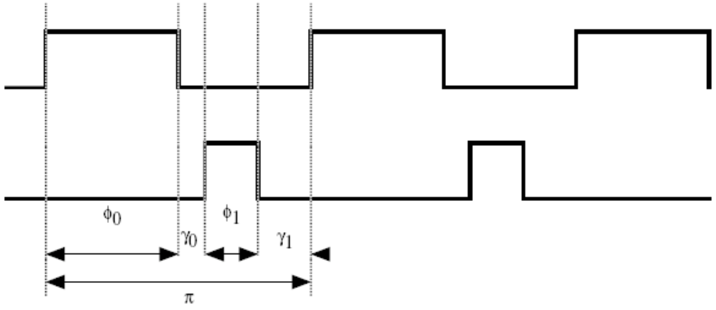
$$sw.cap. = N_0C_R + N_1C_{L1} + N_2C_{L2}$$

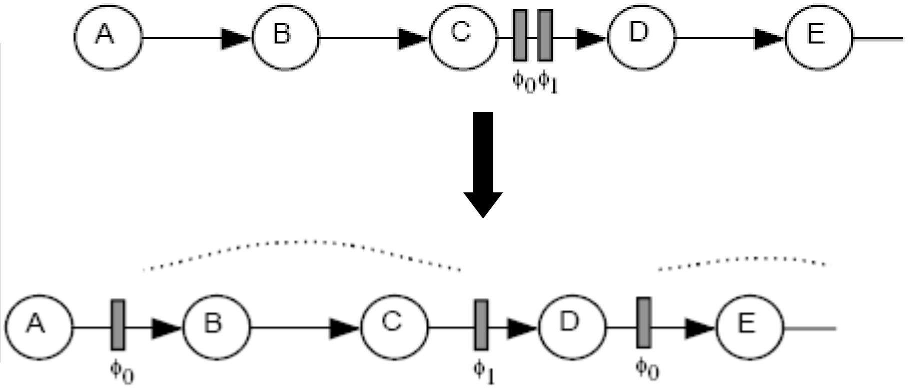


$$sw.cap. = N_0C_{L1} + N_1'C_R + N_2'C_{L2}$$

Retiming

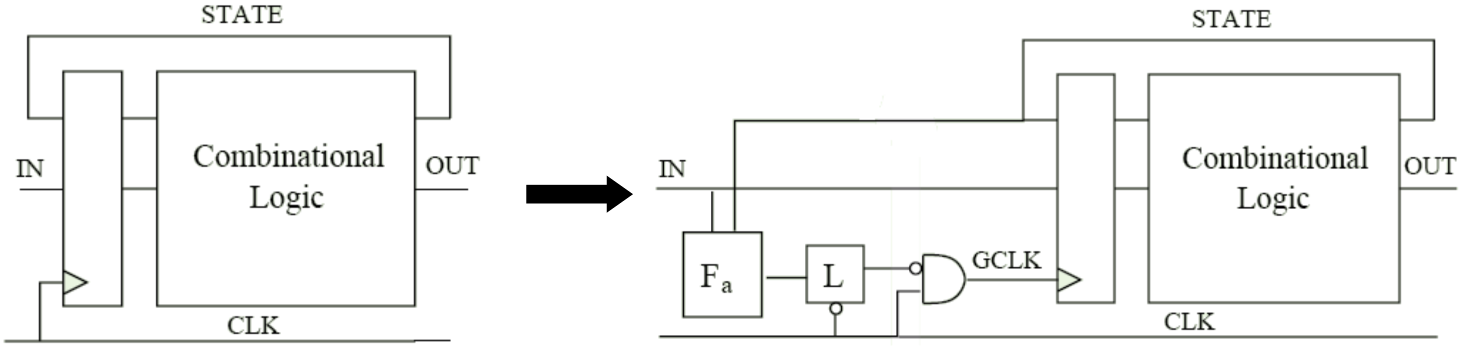
- Add a register with different clock phase
  - Maintain I/O functionality and (sequential) timing while placing more registers in the pipeline
  - Replace an existing register with multiple non-overlapping level-clocked latches or registers synchronous to different phase clocks and reposition them over the circuit





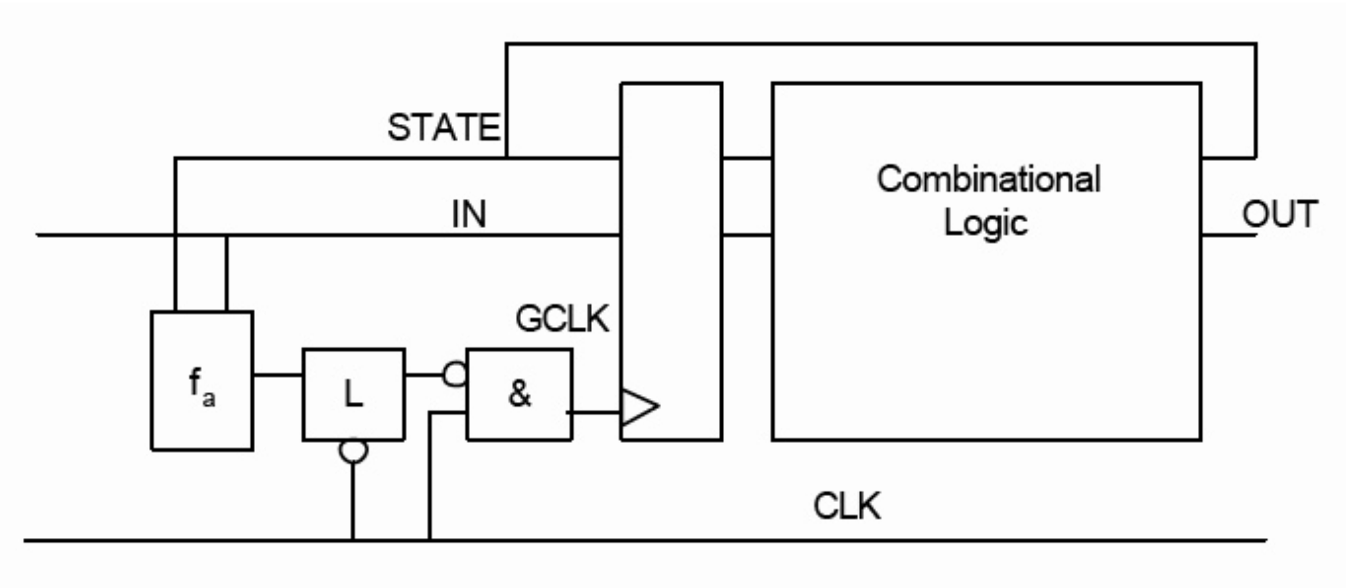
Clock Gating

- Provide a way to selectively stop the clock
  - Force the circuit to make no switching whenever the computation at the next cycle is unnecessary
  - Should be implemented as follows
    - Construct an idleness-detecting circuit which is small (i.e., consume little power) and accurate (i.e., able to stop the clock whenever idle)
    - Design gated-clock distribution circuit with minimum routing overhead
    - Keeps clock skew under tight control



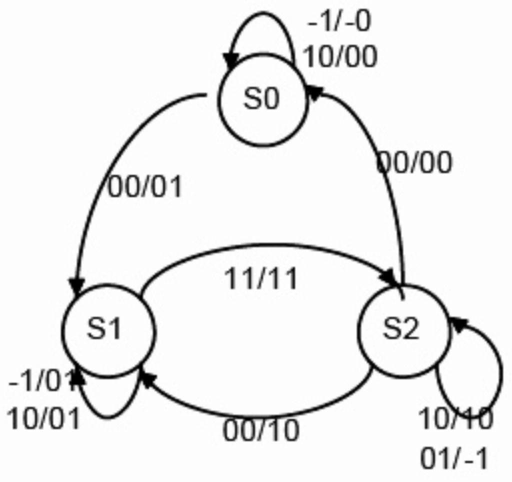
Clock Gating

- One activation function and a latch
  - f_a: activation function

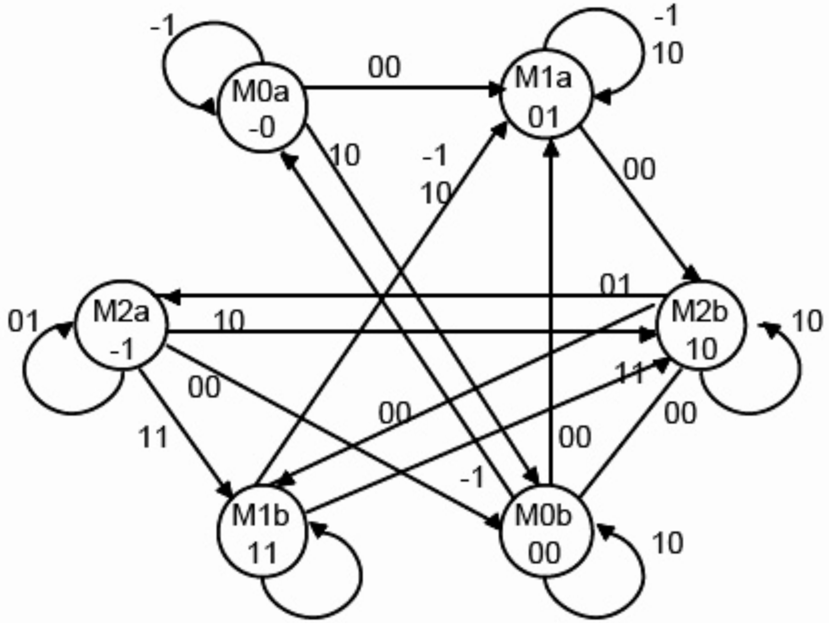


Clock Gating

- FSM conversion
  - Transform Mealy FSM to Moore FSM



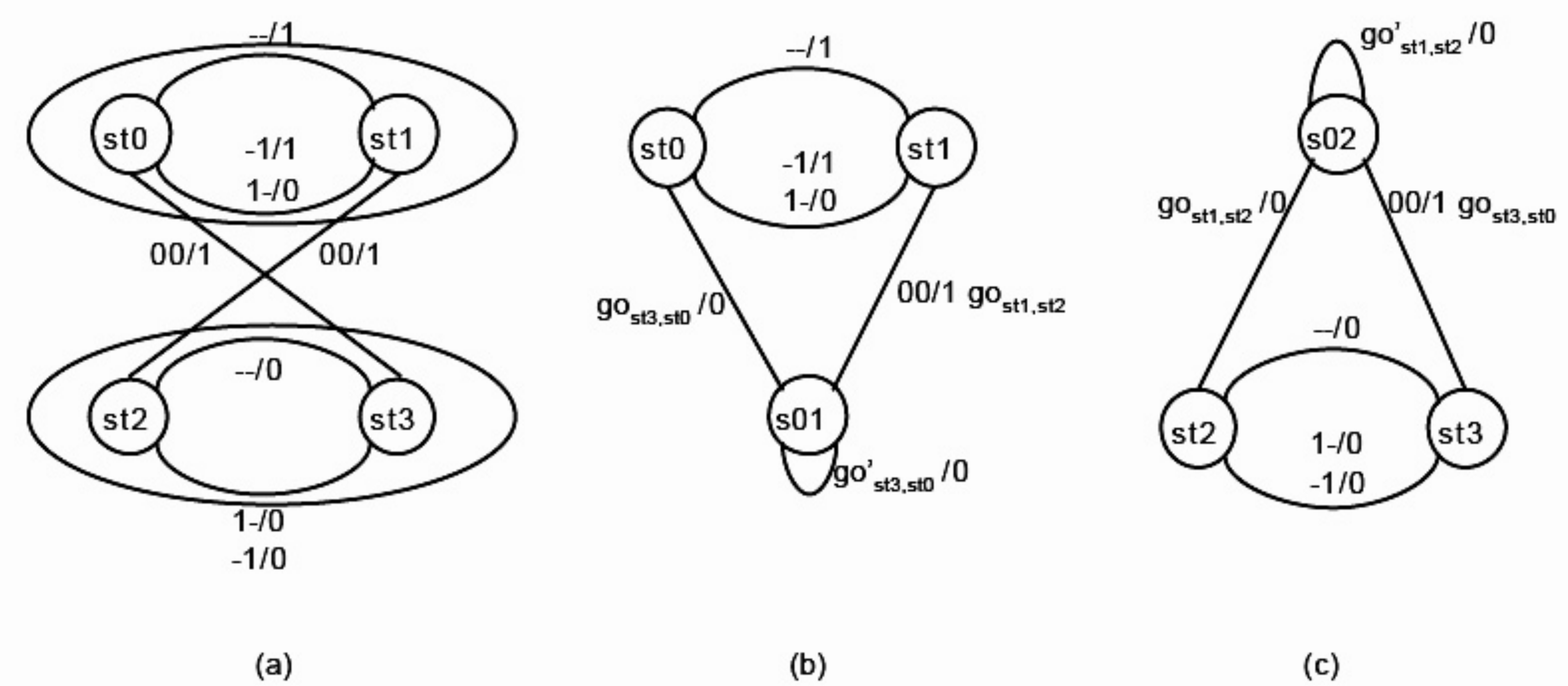
Mealy FSM



Moore FSM

Clock Gating

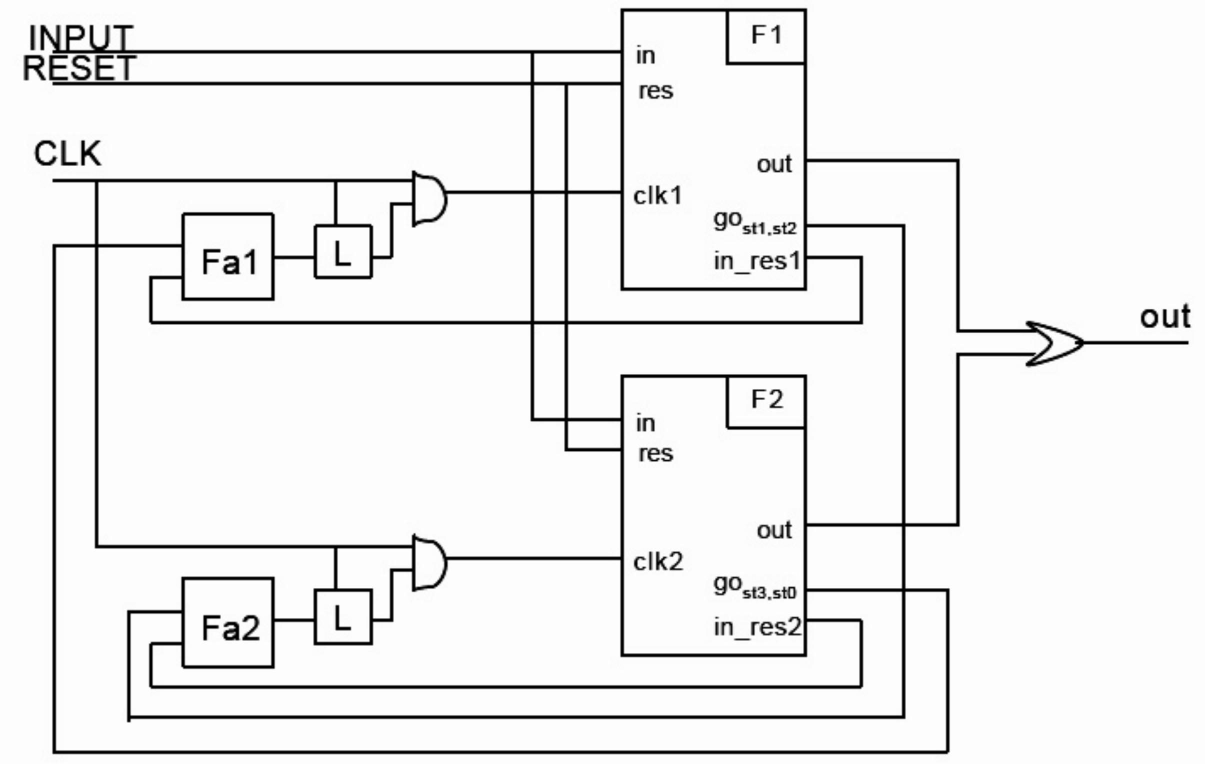
- Partitioned state diagram



(a) (b) (c)

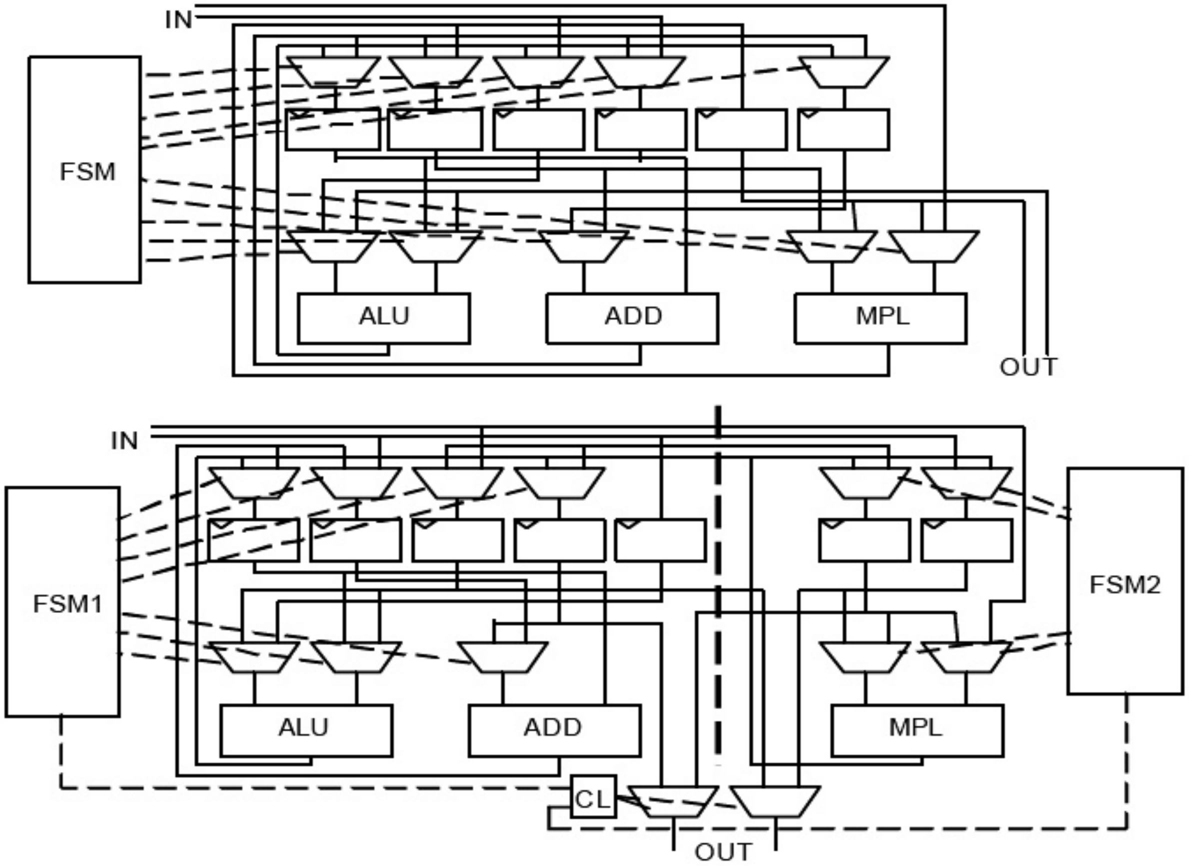
Clock Gating

- Partitioned control unit



Clock Gating

- Example



Leakage Power Reduction

Naehyuck Chang
Dept. of EECS/CSE
Seoul National University
naehyuck@snu.ac.kr

Seoul National University

Static Power

- Leakage power reduction
  - Lowering V_{DD} (voltage islands, dynamic voltage scaling)
  - Cooling and/or refrigeration
  - SOI technology
  - Dual V_T design
  - Body bias control (static and/or adaptive)
  - Input vector control during sleep mode
  - MTCMOS (sleep transistor)

Leakage Reduction Overview

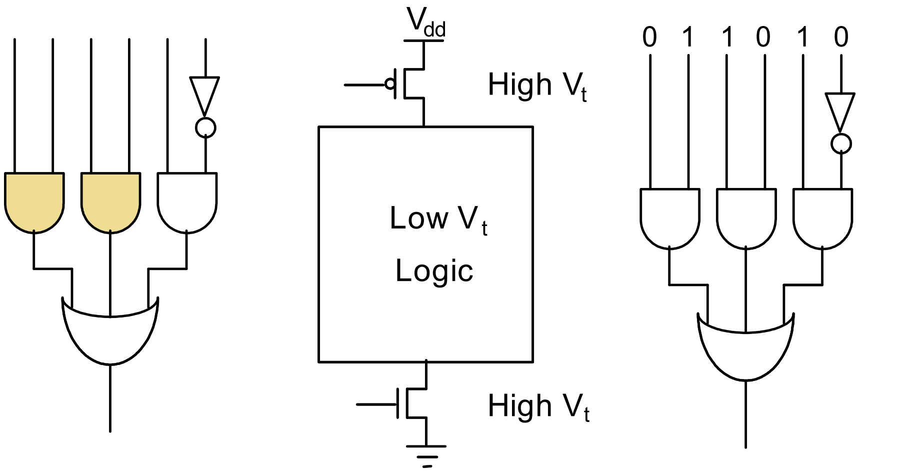
Active-mode leakage

Idle-mode leakage

Idle-mode leakage

High-V_T: slow but low leakage

Low-V_T: fast but high leakage



Dual Threshold

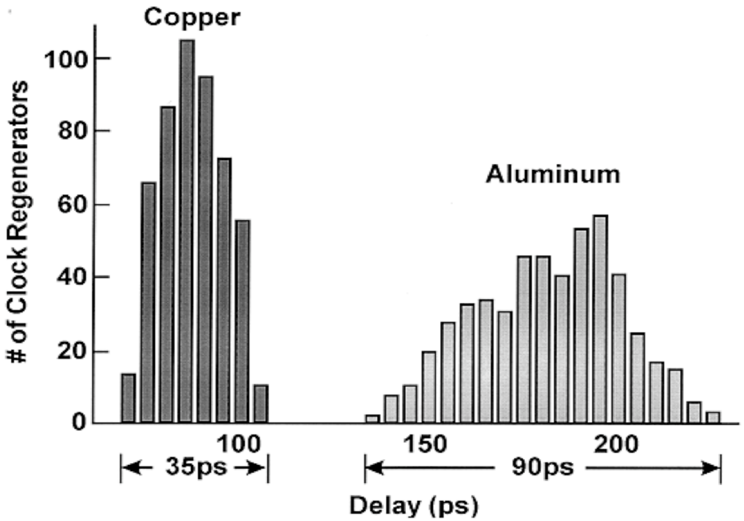
MTCMOS

State Assignment

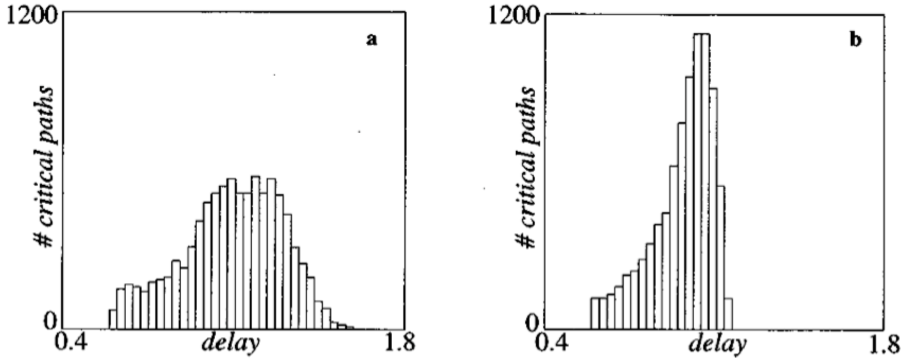
Source : [Johnson, et al. , DAC99]

Delay Estimation

- Locate the critical paths
- Example of delay distribution
  - 19% performance increase with only path delay improvement of 15% paths



Delay distributions

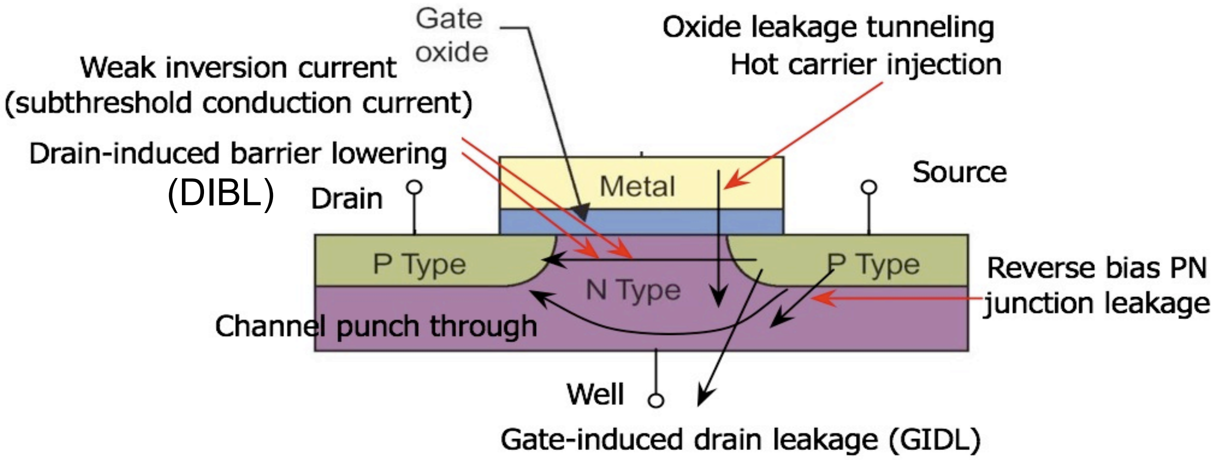


Delay optimization before and after

Leakage Estimation

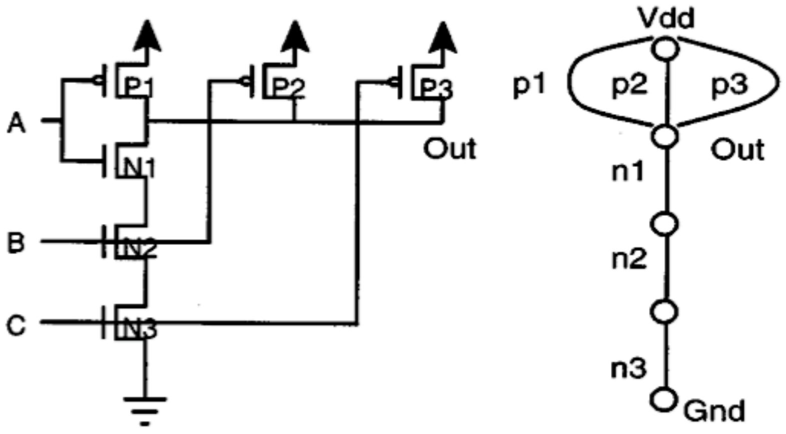
- Transistor leakage estimation
  - Leakage power components
    - Subthreshold leakage is the focus in leakage current modeling

$$I_{subthreshold} = \frac{\mu W C_{ox}}{L} V_T^2 e^{\frac{|V_{GS}|-|V_t|}{nV_T}} \left(1 - e^{\frac{-|V_{DS}|}{V_T}}\right)$$

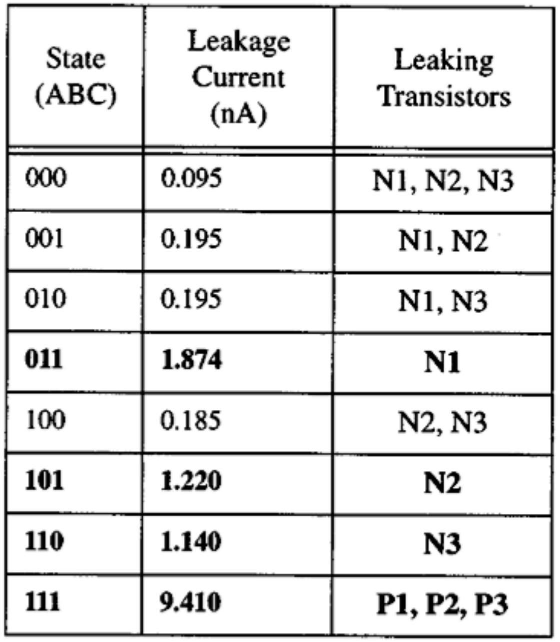


Leakage Estimation

- State probability
  - Three-input NAND SPICE leakage simulation



(a) (b)



State (ABC)	Leakage Current (nA)	Leaking Transistors
000	0.095	N1, N2, N3
001	0.195	N1, N2
010	0.195	N1, N3
011	**1.874**	**N1**
100	0.185	N2, N3
101	**1.220**	**N2**
110	**1.140**	**N3**
111	**9.410**	**P1, P2, P3**

Generalized Multiple V_T Problem

- Power minimization problem
  - Given:
    - A random logic network of N static CMOS gates
    - The critical path delay is less than equal to T_{max}
    - The device technology used
    - Activity profiles at each input node
  - Determine:
    - Supply voltage V_{DD}
    - Threshold voltage V_T
    - Channel width (size) W
  - Such that:
    - Static leakage and dynamic power are minimized
    - The area is within the bound
    - Generally subthreshold leakage is minimized
  - Solution:

Dual V_T Circuit Optimization

- Transistor is assigned either a high or low V_T
  - Low-V_T transistor
    - Reduced delay
    - Increased leakage
  - Speed critical path: low-V_T
  - Rest: high-V_T

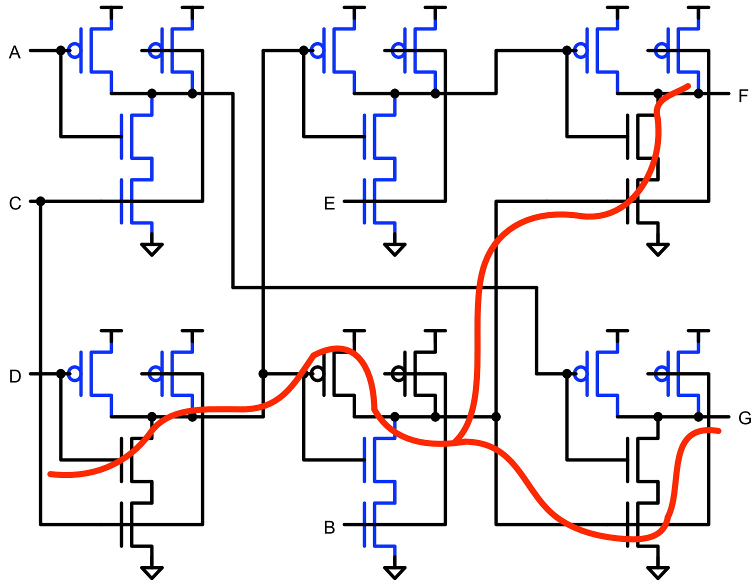
	Low-V_T: 0.8 V	High-V_T: 0.8 V	Low-V_T: 1.2 V	High-V_T: 1.2 V
Normalized leakage	1	0.05	1	0.049
Normalized delay	1	1.36	1	1.30

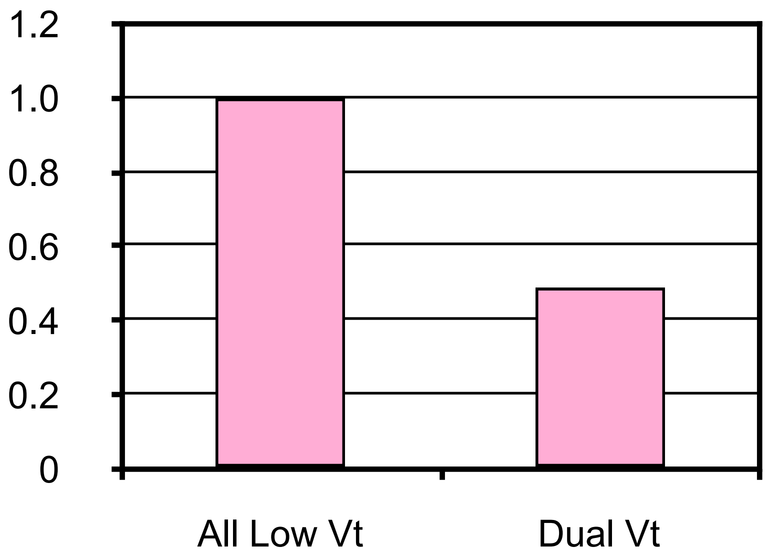
Dual V_T Circuit Optimization

- Objective
  - Find an implementation between the two extremes of all low VT, all high V_T, trading off leakage power for delay
  - Delay constraint must be met

Dual V_T Circuit Optimization

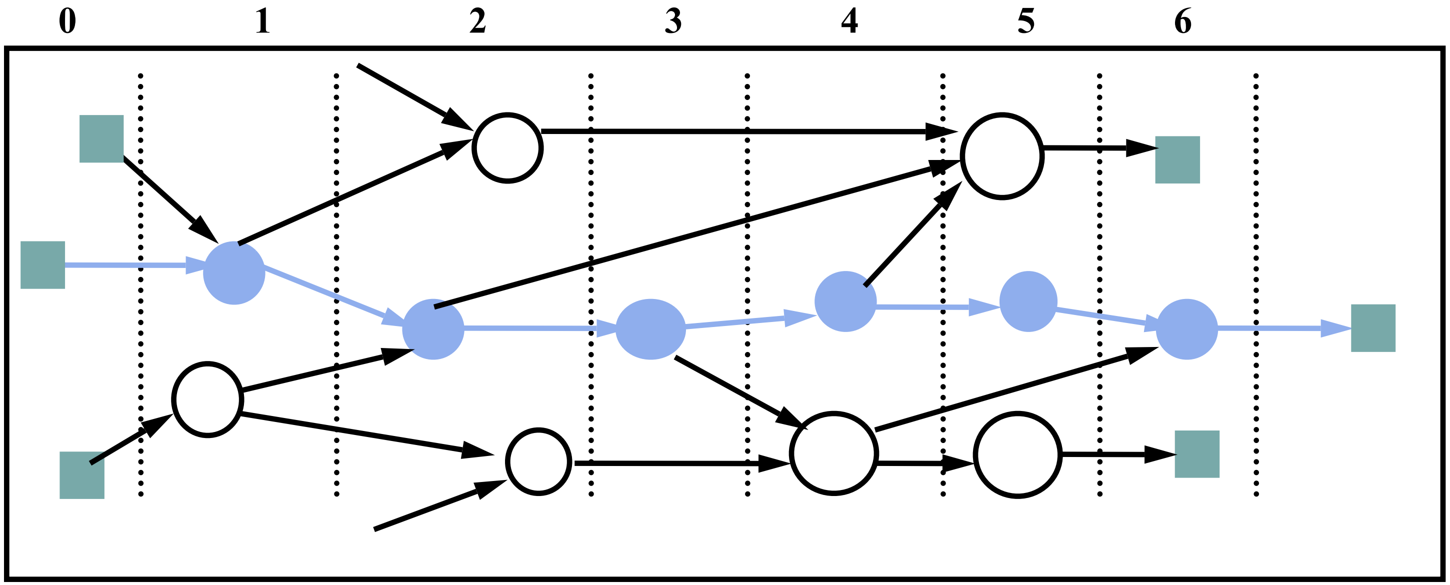
- Example
  - Dual V_T assignment approach
  - Transistor on critical path: low V_T
  - Non-critical transistor: high V_T





Dual V_T Circuit Optimization

- V_T assignment
  - Greedy approach: backward traversal of circuit
    - Select high V_T gate in critical path
    - Set gate to low V_T
    - Re-compute critical paths

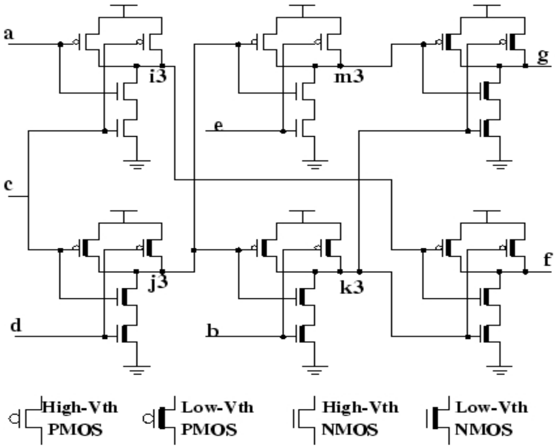


Dual V_T Circuit Optimization

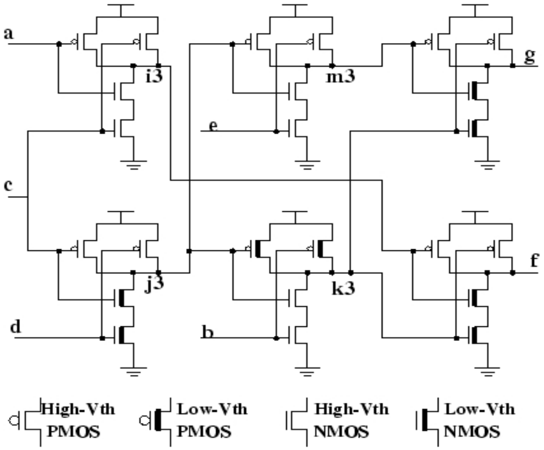
- V_T assignment granularity
  - Gate based assignment
  - Pull up network / Pull down network based assignment
    - Single V_T in P pull up or N pull down trees
  - Stack based assignment
    - Single V_T in series connected transistors
  - Individually assignment within transistor stacks
    - Possible area penalty

Dual V_T Circuit Optimization

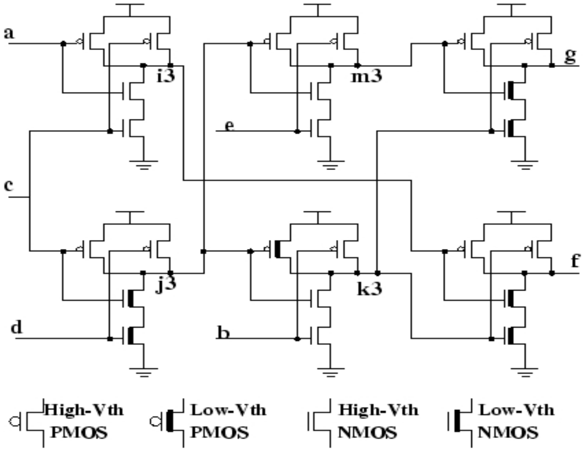
- Examples



Gate based



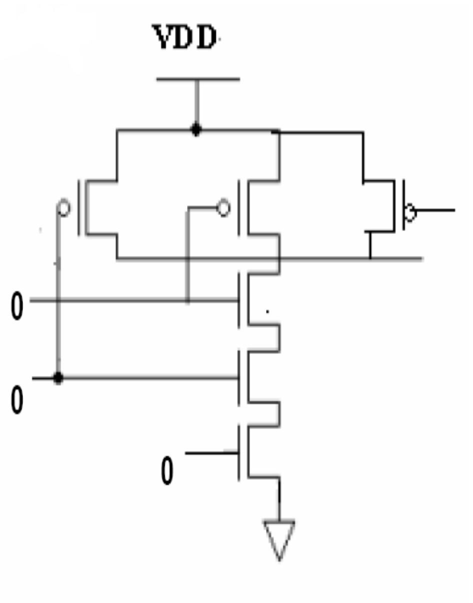
PU/PD based



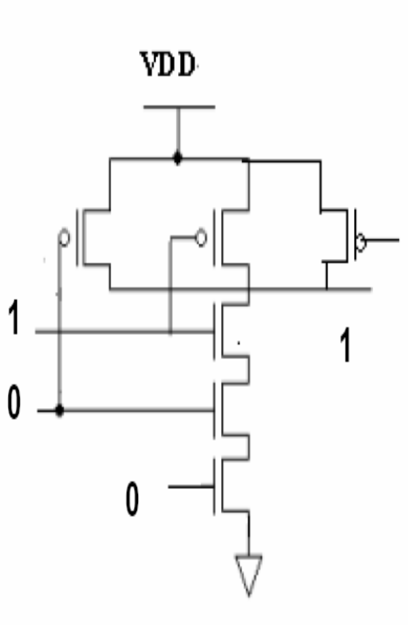
Stack based

IVC (Input Vector Control)

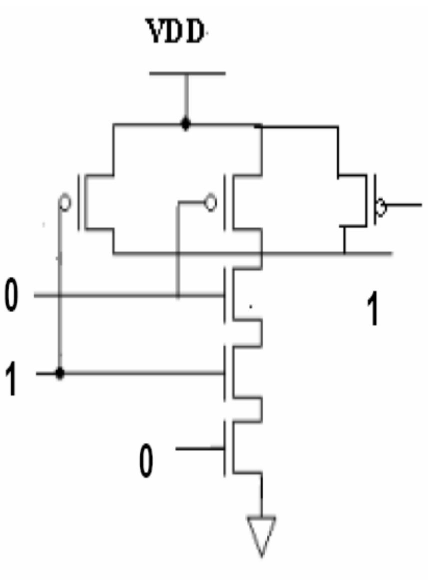
- The idea is based on the transistor stack effect



Least subthreshold leakage



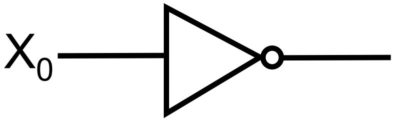
Least gate leakage

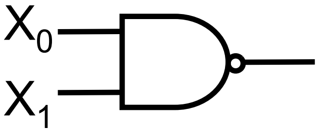


Largest gate leakage

IVC (Input Vector Control)

- Subthreshold leakage and gate leakage currents are dependent on the input vectors





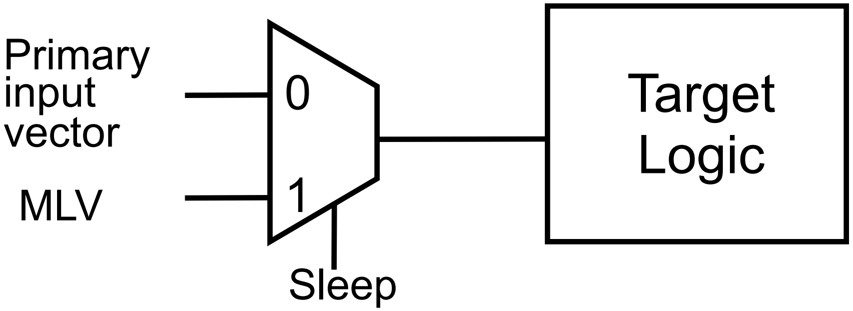
Input vector (X_0)	Leakage (nA)
0	100.3
1	227.2

Input (X_0X_1)	Leakage (nA)
0	37.84
1	100.30
10	95.17
11	454.50

Cadence spectra simulation, 0.18um technology

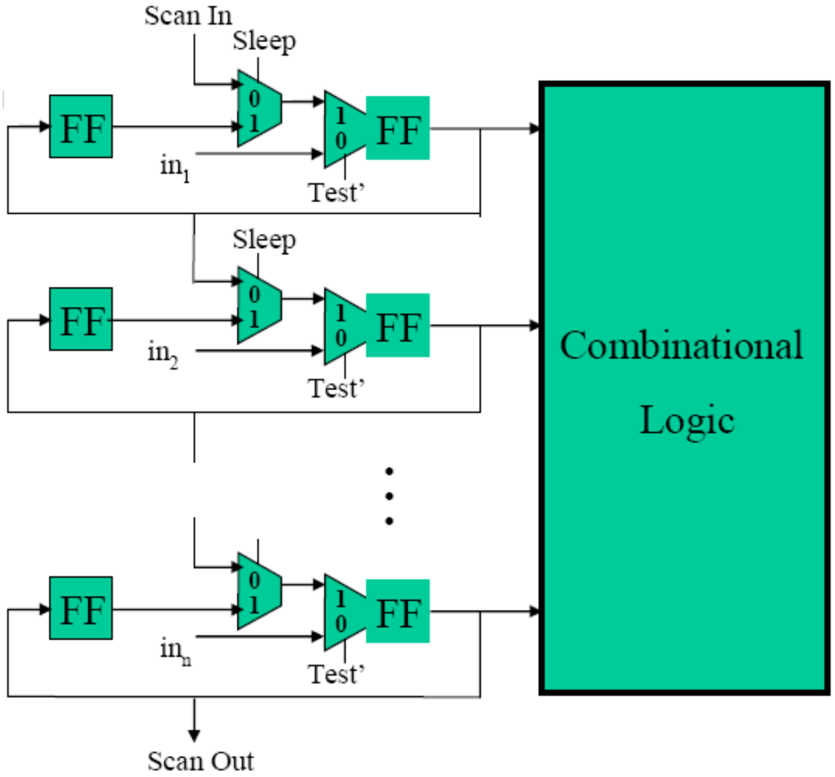
IVC (Input Vector Control)

- Implementation of IVC
  - Concept
    - IVC During the Sleep Mode
      - Providing the minimum leakage vector (MLV) to the target logics during the sleep (or standby) mode



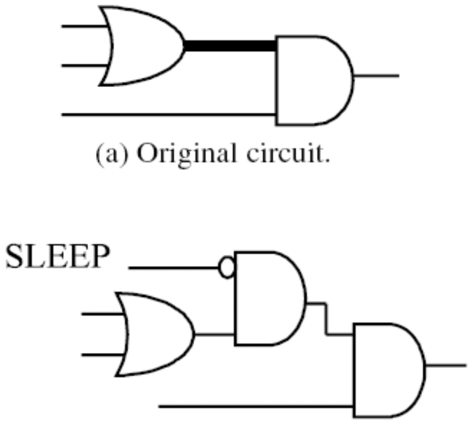
IVC (Input Vector Control)

- Implementation of IVC
  - Modification of a scan-chain registers
- Original MLV is stored in left FFs
  - Sleep mode
    - Sleep = 1
    - Test = 1
    - MLB is applied (right FF's)
  - Operational mode
    - Sleep= 0
    - Test = 0
    - Inputs are directly applied to the target logic



IVC (Input Vector Control)

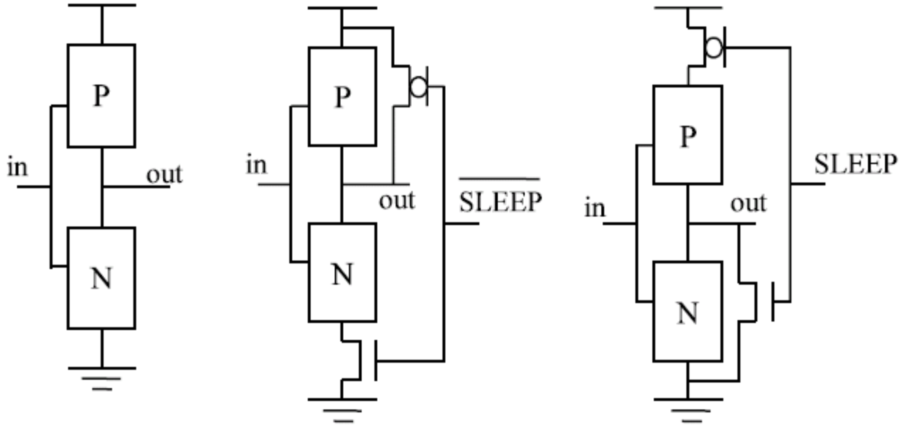
- Modifying the internal logic gates for further leakage reduction
  - Due to logic dependencies of the internal signals, driving a circuit with its MLB does not guarantee
  - Increase controllability in the standby mode



(a) Original circuit.

(b) Leakage-guarded circuit.

Replacing an internal signal line with a two-input AND gate



(A) out = g(in).

(B) out = NAND ($\overline{\text{SLEEP}}$, g(in)).

(C) out = NOR (SLEEP, g(in)).

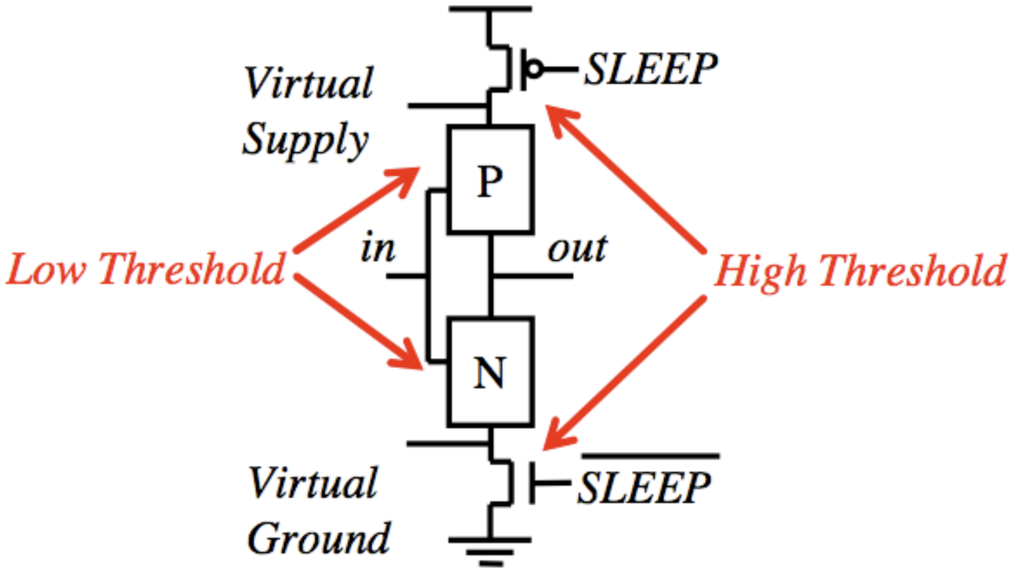
Modifying CMOS gate

MTCMOS: sleep transistor insertion

- Basic concept
  - Multi-threshold CMOS: sleep transistor insertion
    - To use both high-V_T and low-V_T cells in a logic block
    - Based on the observation that a circuit's overall performance is often determined by a few critical paths
      - Transistors and gates along the critical paths are set to a low-V_T
      - Transistor size is fixed
      - Overall circuit performance can be enhanced significantly
      - Leakage is kept within bounds
    - Operating frequency of a logic block is limited by the maximum path delay

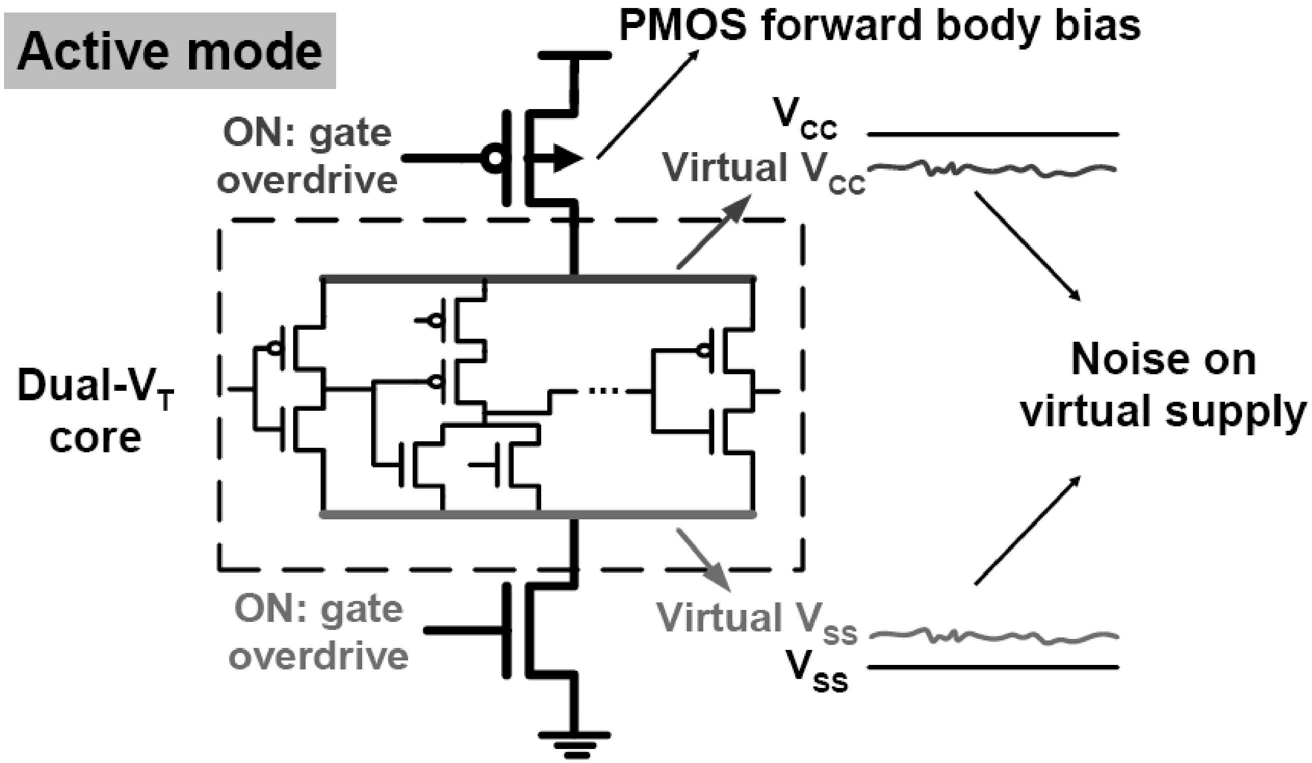
MTCMOS: Sleep Transistor Insertion

- Sleep transistor
  - Also called guarding, power gating, ground gating, using sleep transistor, etc.
  - Sleep transistor is inserted between the VDD and logic, and logic and GND.



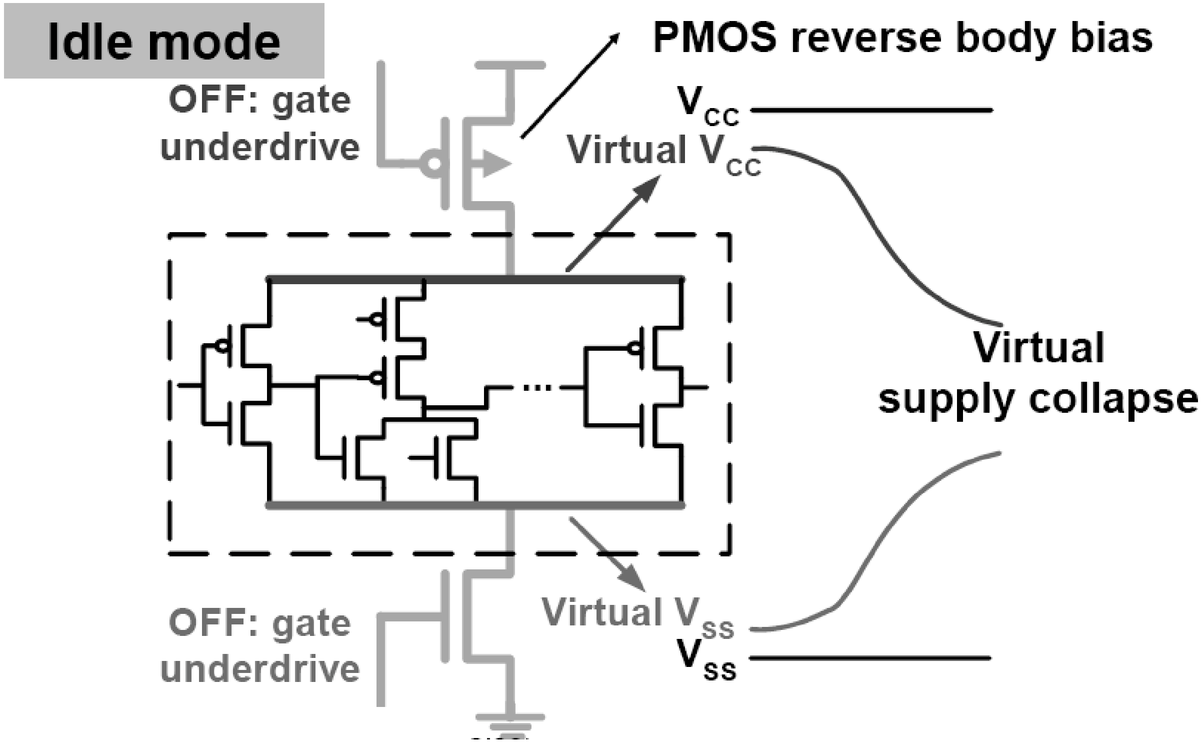
MTCMOS: Sleep Transistor Insertion

- Active-mode operation



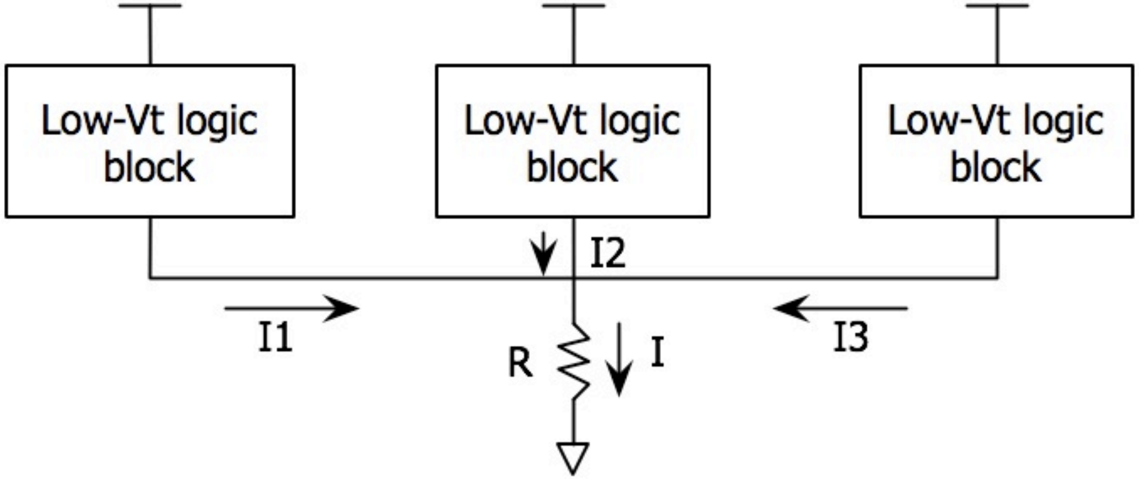
MTCMOS: Sleep Transistor Insertion

- Idle-mode operation



MTCMOS: Sleep Transistor Insertion

- The worst case
  - Low-V_T blocks switch at the same time
  - $I=I_1+I_2+I_3$
- The best case
  - Low-V_T blocks switch exclusively (no time overlap)
  - $I=\max(I_1, I_2, I_3)$
- In general
  - Partially overlap

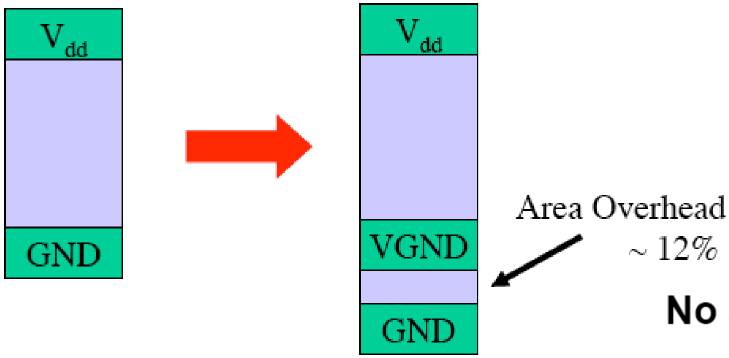


MTCMOS: Sleep Transistor Insertion

- Average current method
  - Estimation of the optimum size of the sleep transistor
  - If average current flow thought the sleep transistor and maximum speed penalty of the MTCMOS block are known, the minimum size of the sleep transistor can be estimated
  - The current consumed in the MTCMOS block is constant, the voltage drop across the sleep transistor is constant

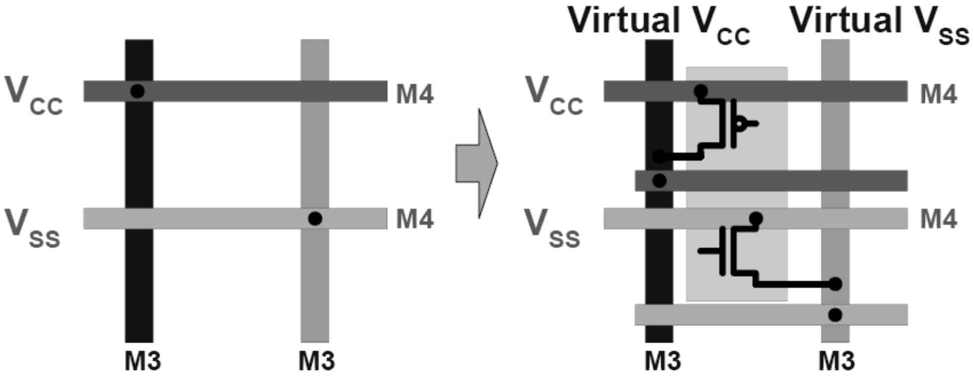
MTCMOS: Sleep Transistor Insertion

- Samsung's sleep transistor insertion
  - Use of conventional P&R



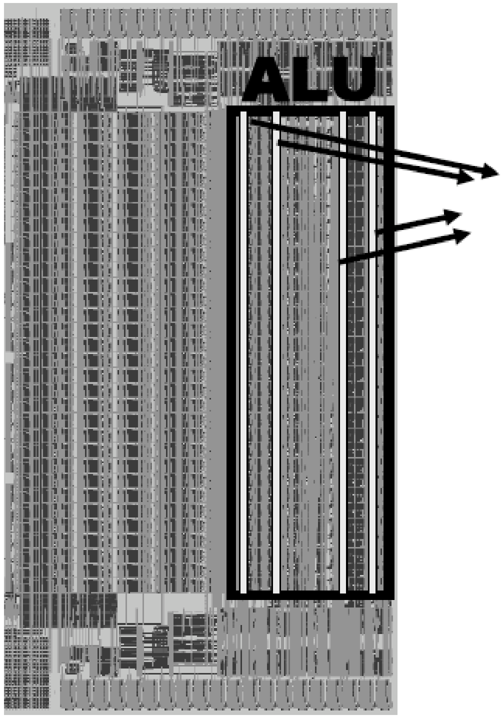
No sleep transistor

PMOS & NMOS sleep transistors



MTCMOS: Sleep Transistor Insertion

- Example of the sleep transistor placement



Sleep transistor cells

Area overhead	
PMOS	6%
NMOS	3%

Tschanz, ISSCC'03