Source of Power Consumption

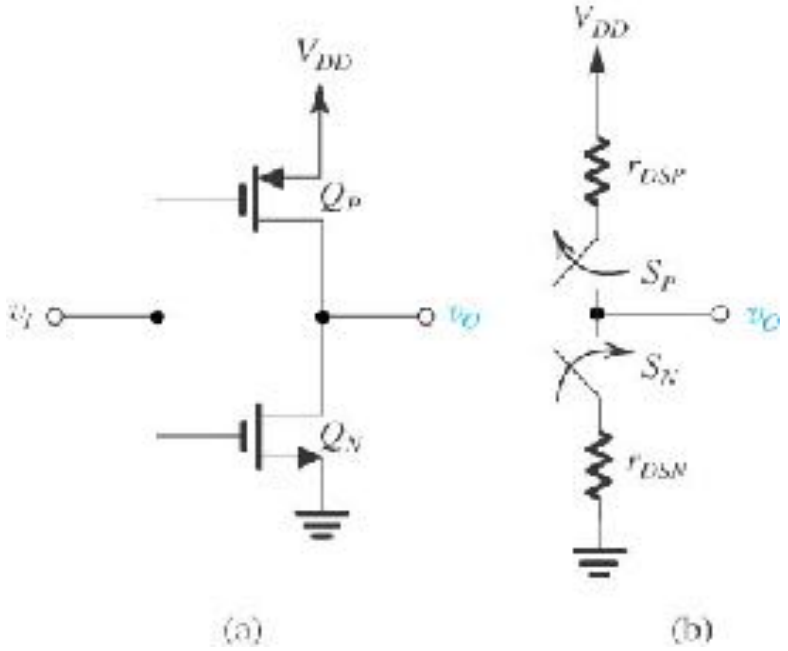
Sangyoung Park
Chair for Real-Time Computer Systems
Technical University of Munich
sangyoung.park@tum.de

Contents

- CMOS Inverter
- Power and Energy
- Source of Power Consumption
- Dynamic Power
- Static Power
- (VDD-VT) Design Space
- Total Power Management

CMOS Inverter

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

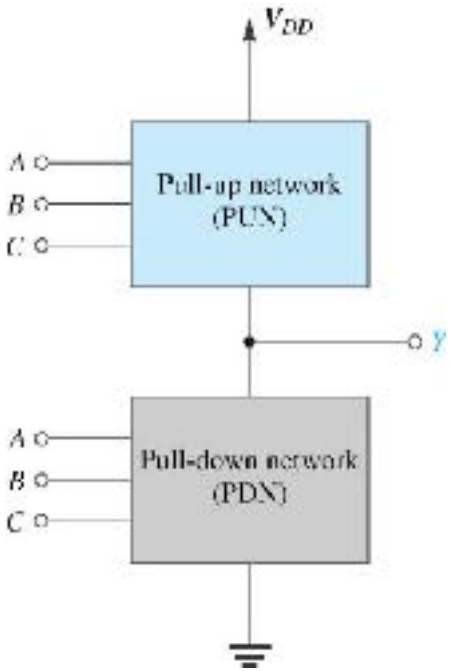


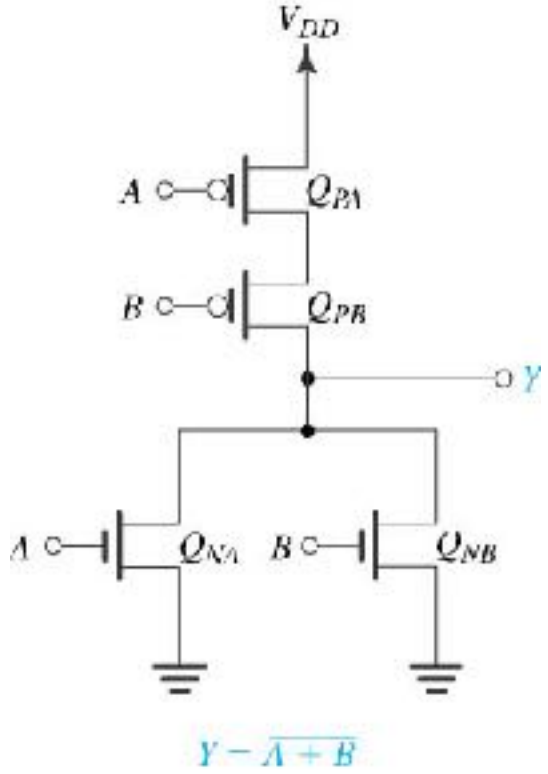
$$r_{DSN} = \frac{1}{k_n'(\frac{W}{L})_n(V_{DD} - V_t)}$$

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

CMOS Inverter

- Pull- up and down networks





$Y = \overline{A + B}$

CMOS Inverter

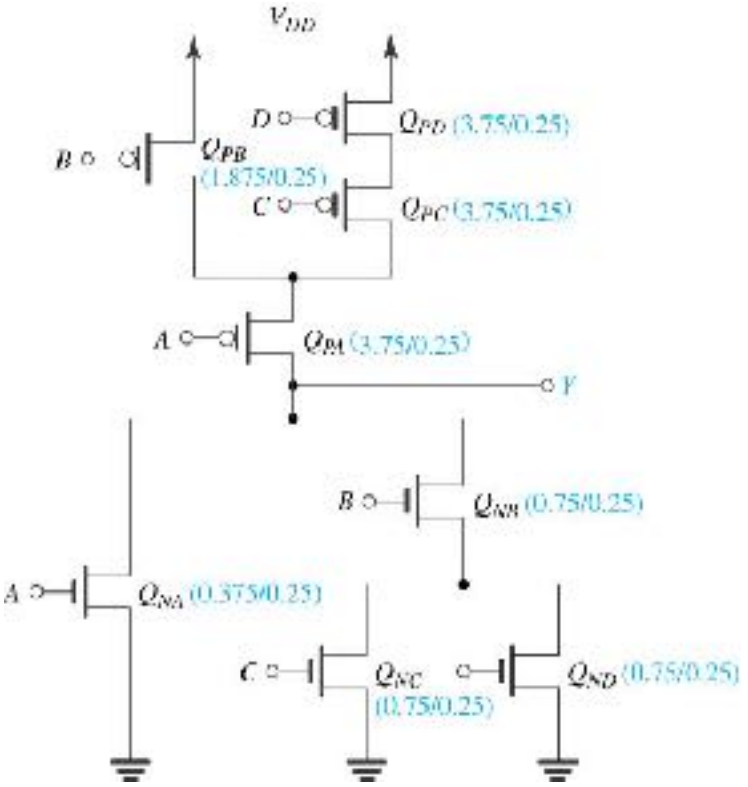
- Static operation
  - Matching for symmetrical transfer characteristic

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

  - 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

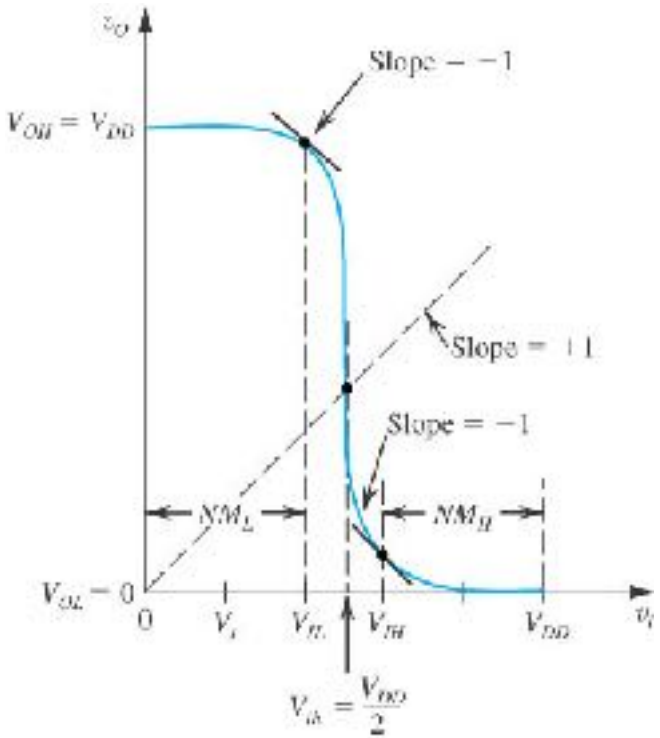
CMOS Inverter

- Transistor sizing
  - Determination of the W/L ratio
  - Provide the same current-driving capability in both directions equal to that of the basic inverter



CMOS Inverter

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



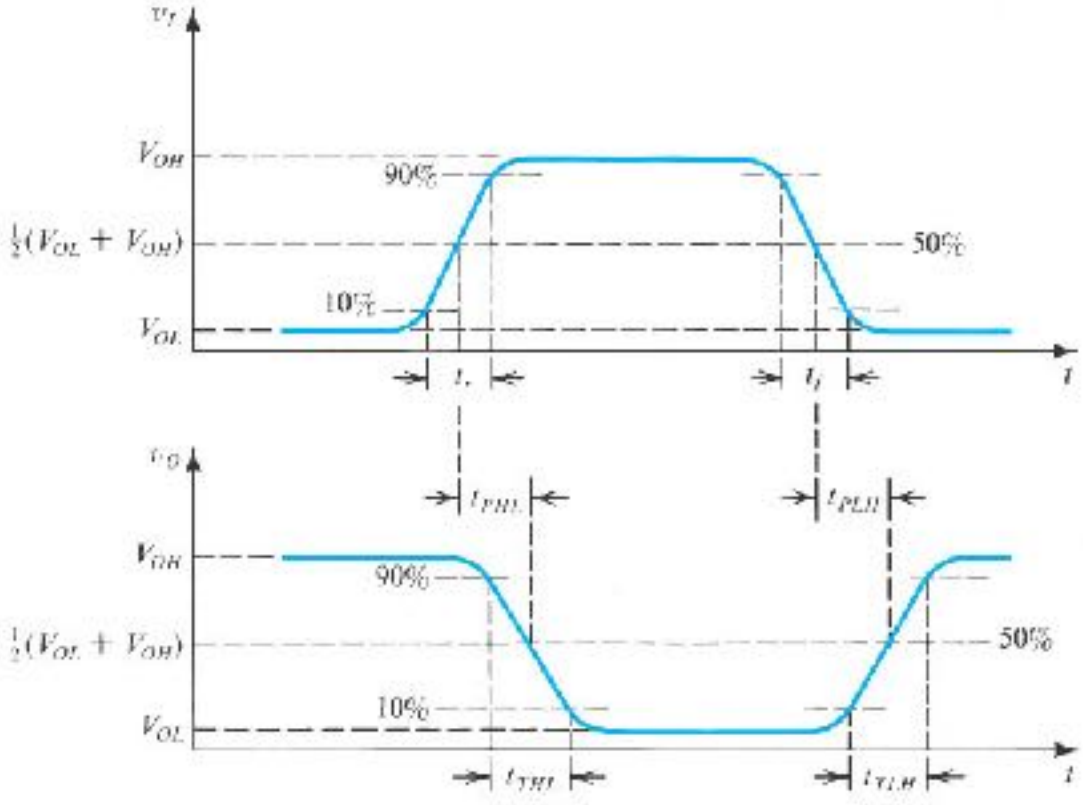
$$V_{th} = \frac{V_{DD} - |V_{tp}| + \sqrt{\frac{K_n}{K_p}} V_{in}}{1 + \sqrt{\frac{k_N}{k_p}}}$$

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

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

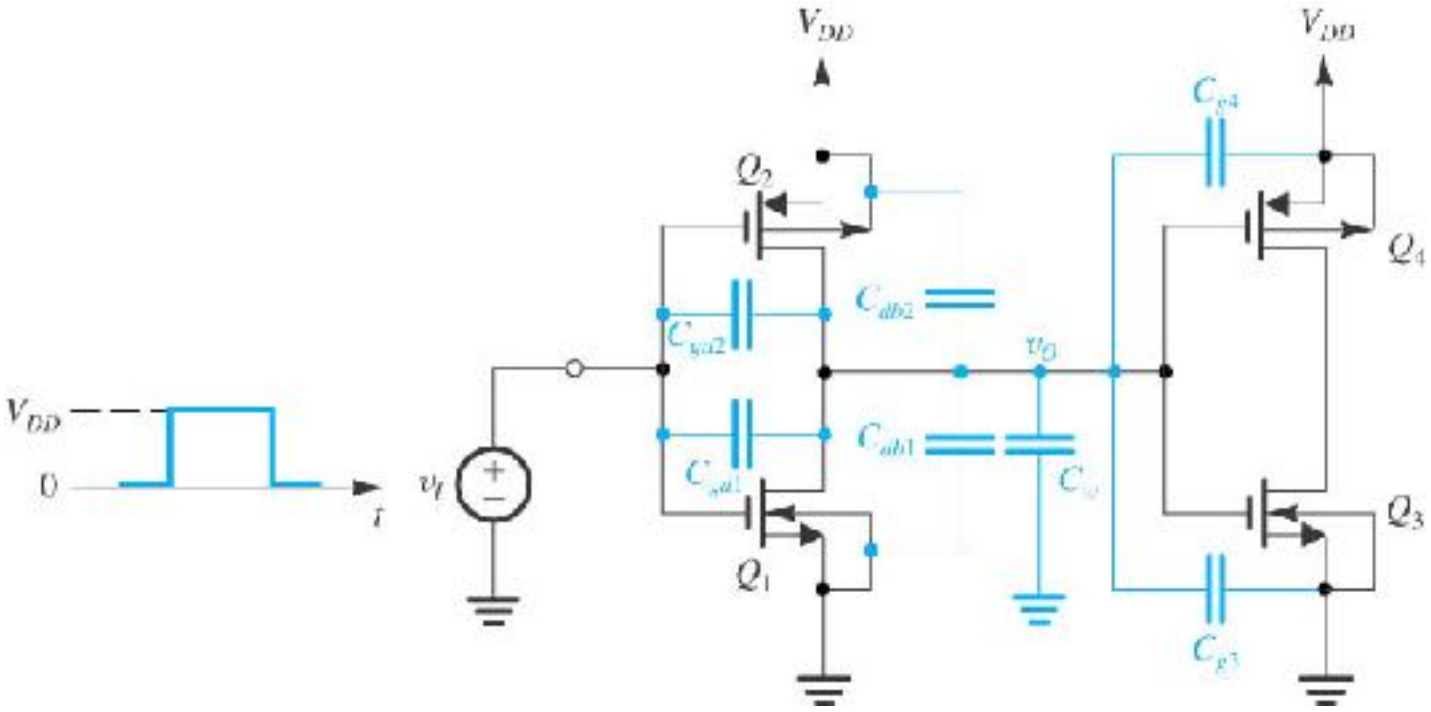
CMOS Inverter

- Definitions of propagation delays and switching times of the logic inverter



CMOS Inverter

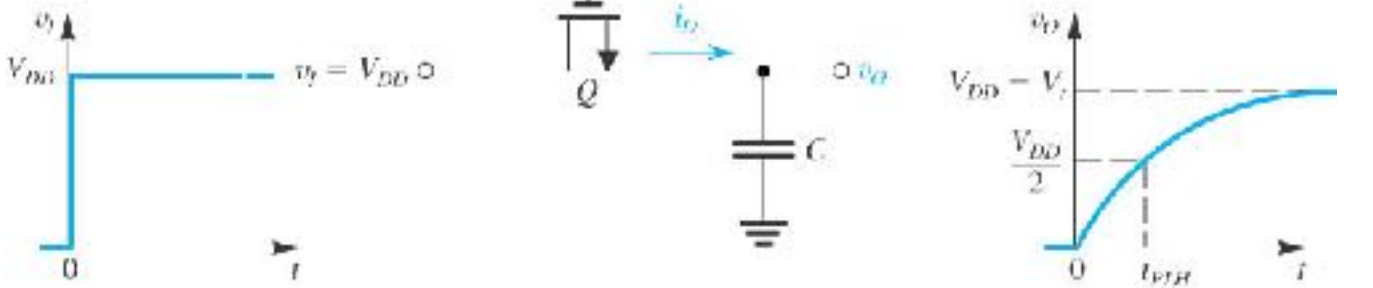
- Dynamic operation
  - Parasitic capacitance



CMOS Inverter

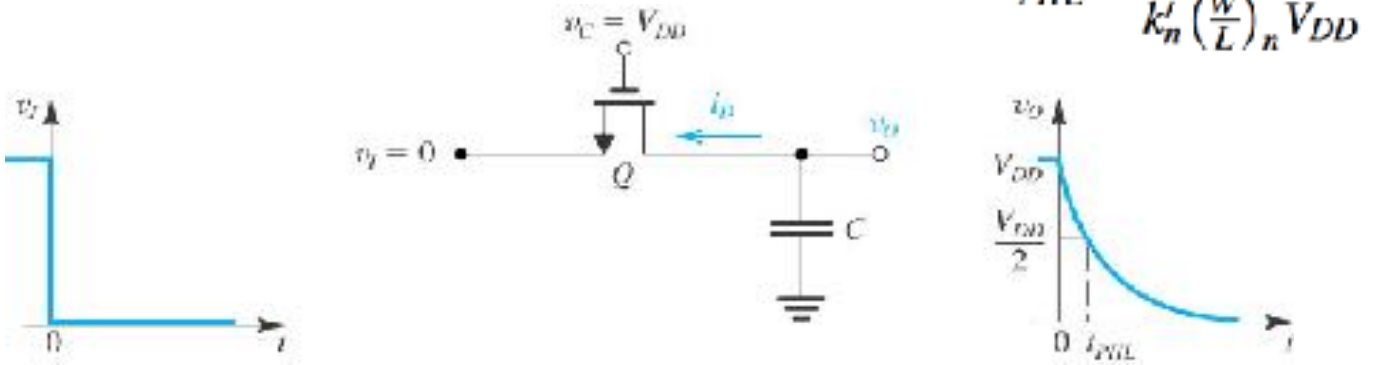
- Dynamic operation
  - Low to high transition

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



  - High to low transition

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



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 V_{DD} to GND when logic transition occurs
    - Switching power
    - Short-circuit power
    - Glitch power
- Static power
  - Current flow from V_{DD} 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
- 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

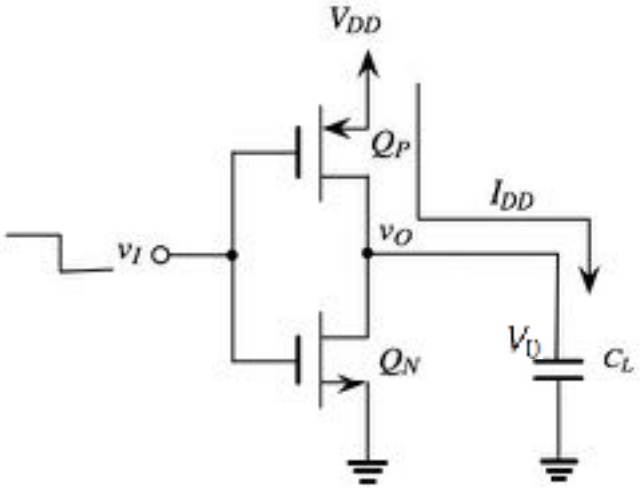
Dynamic Power

- Switching power

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

- A step voltage is applied at t=0

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

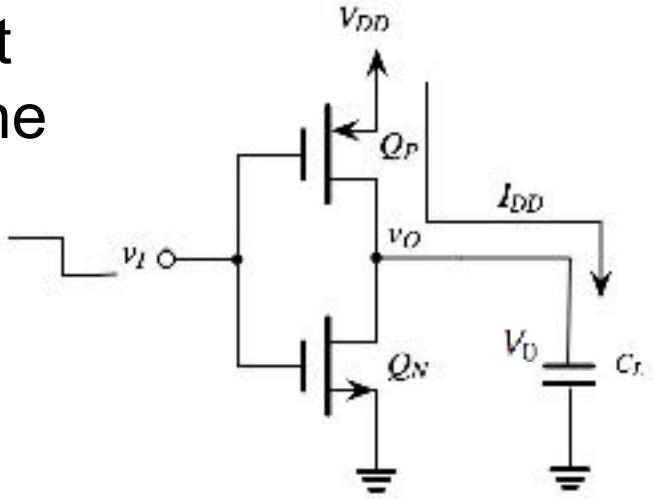


- Energy transferred from the power supply

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

Dynamic Power

- Switching power
  - When $V = V_{DD},\ E_{0\to1} = C_L V_{DD}^2$
  - $C_L V_{DD}^2/2$ 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

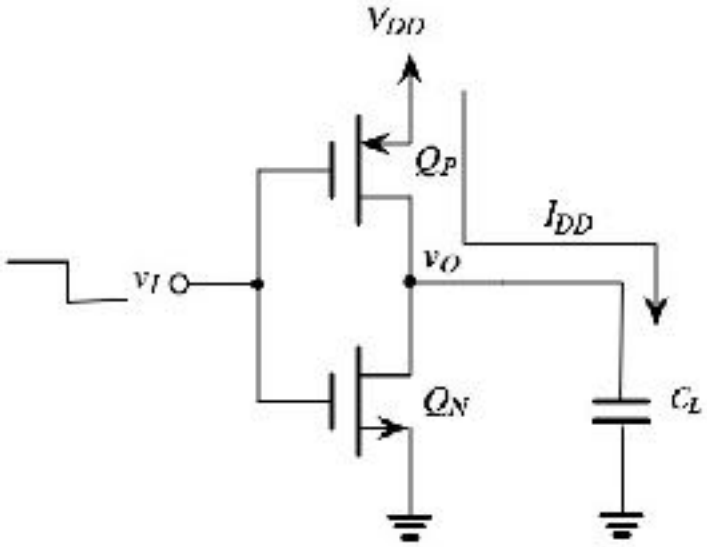


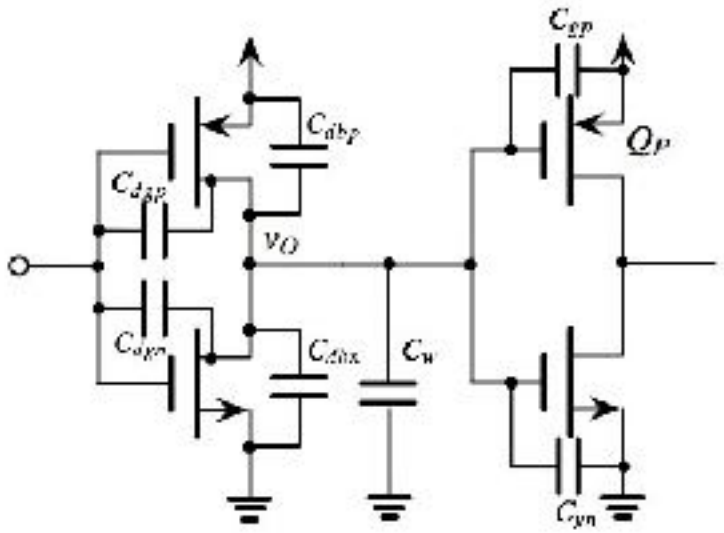
Dynamic Power

- Switching power:

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

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





Dynamic Power

- Gate capacitance

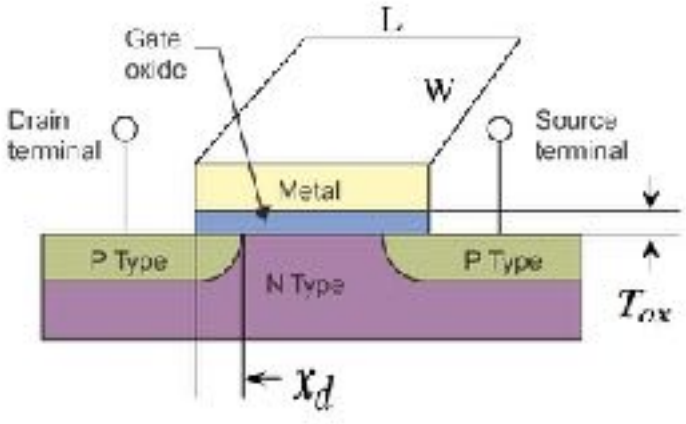
$$C_g = C_{sg} + C_{dg} + C_{bg}$$

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

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

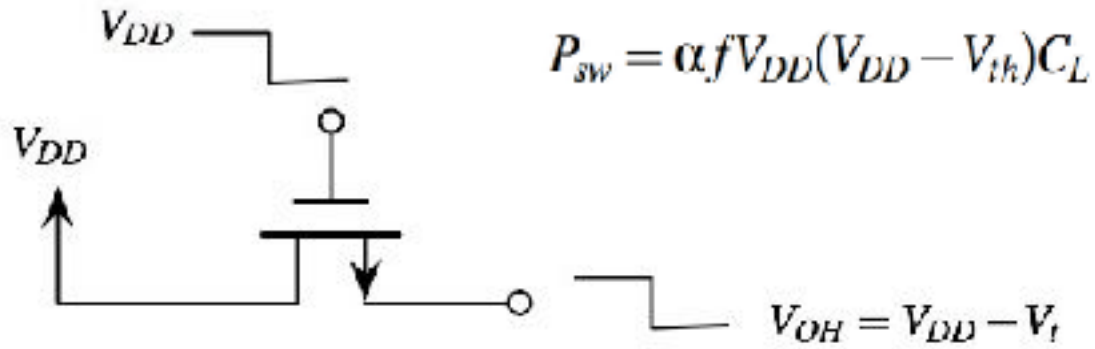
  - Due to Miller effect: $C_{dg1} = C_{dg2} = 2C_{ox}x + dW$
  - $C_{dg3} = C_{dg4} = C_{sg3} = C_{sg4} = C_{ox}x + dW$

- Diffusion capacitance
- Interconnect capacitance



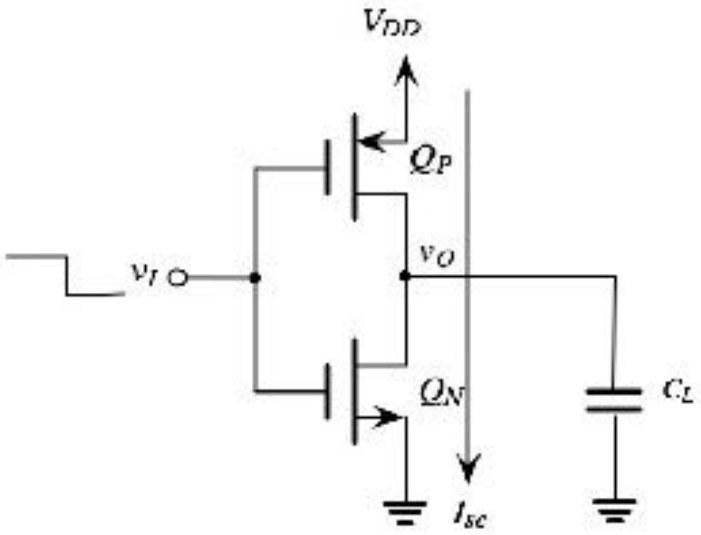
Dynamic Power

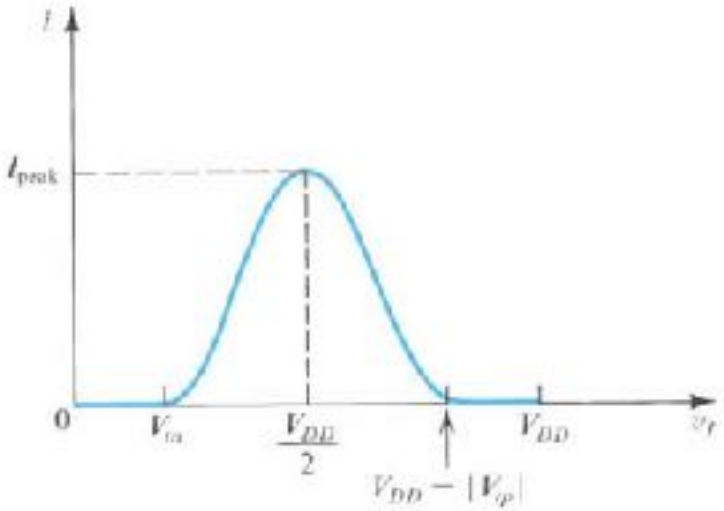
- Reduced swing switching power
  - Rail-to-rail swing: V_{DD} to GND
  - When $V_{OH} < V_{DD}$, swing is V_{OH} to GND
  - Reduced bit-line in memory



Dynamic Power

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





Dynamic Power

- Short-circuit power

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

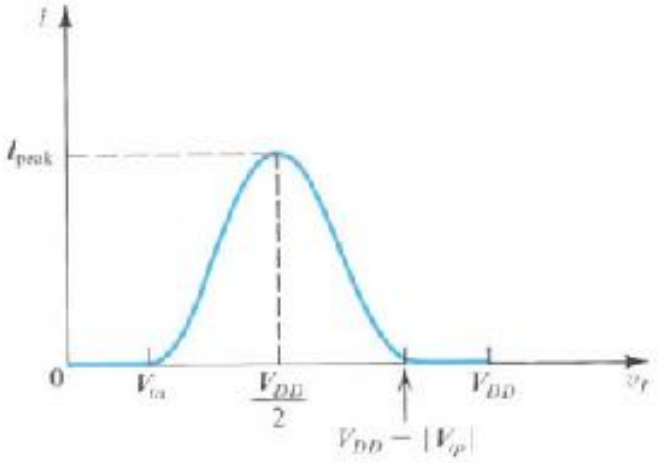
device parameter $\beta_n = \beta_p = \mu C_{ox} \frac{W}{L}$

μ : carrier mobility

C_{ox} : Oxide capacitance

τ : rise and fall time

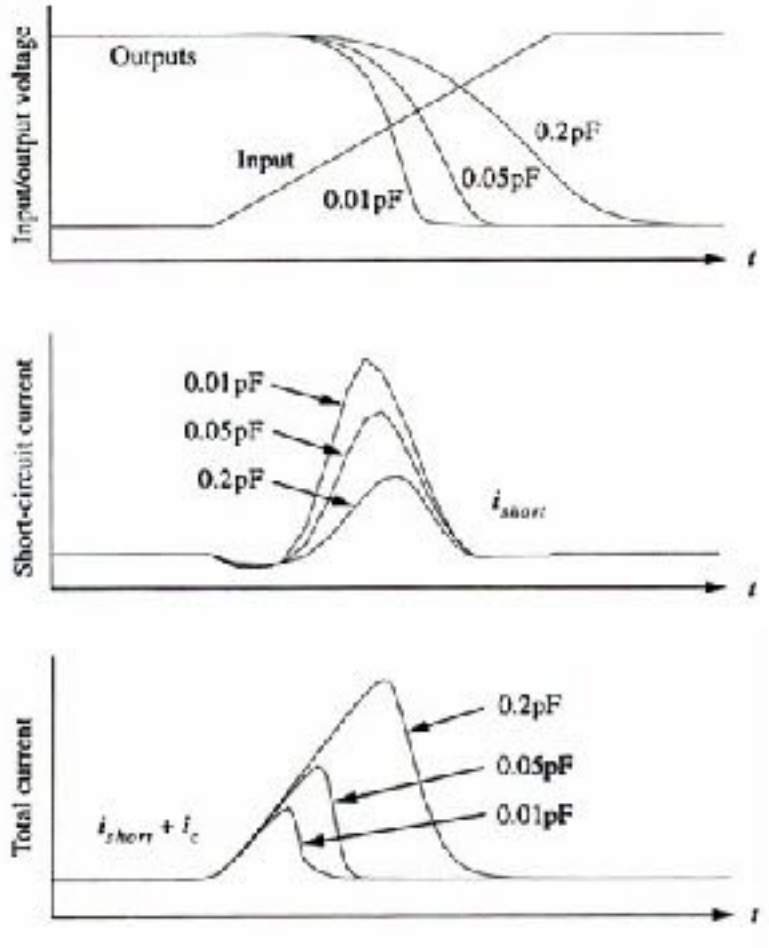
$$P_{sc} = \frac{\mu C_{ox} W}{12 \; L} (V_{DD} - 2V_{th})^3 \tau f$$



Dynamic Power

- Impact of load capacitance
  - As output loading increases:

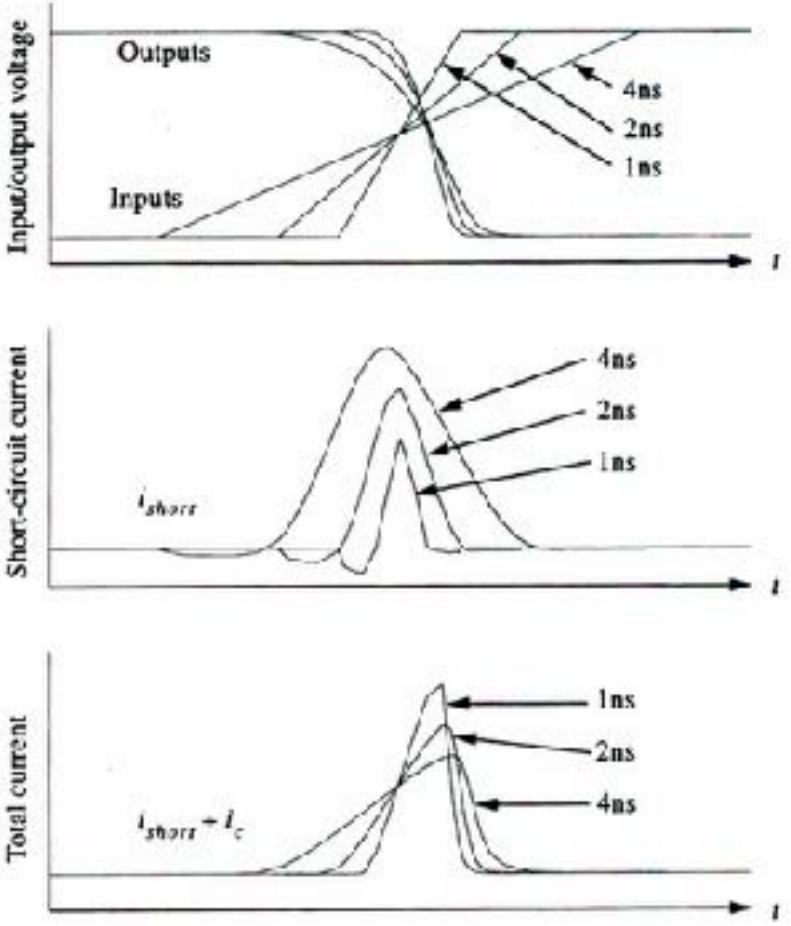
Current envelope	width	peak	integration
i_{short}	no change	decrease	decrease
i_c	increase	increase	increase
$i_{short}+i_c$	increase	increase	increase



Dynamic Power

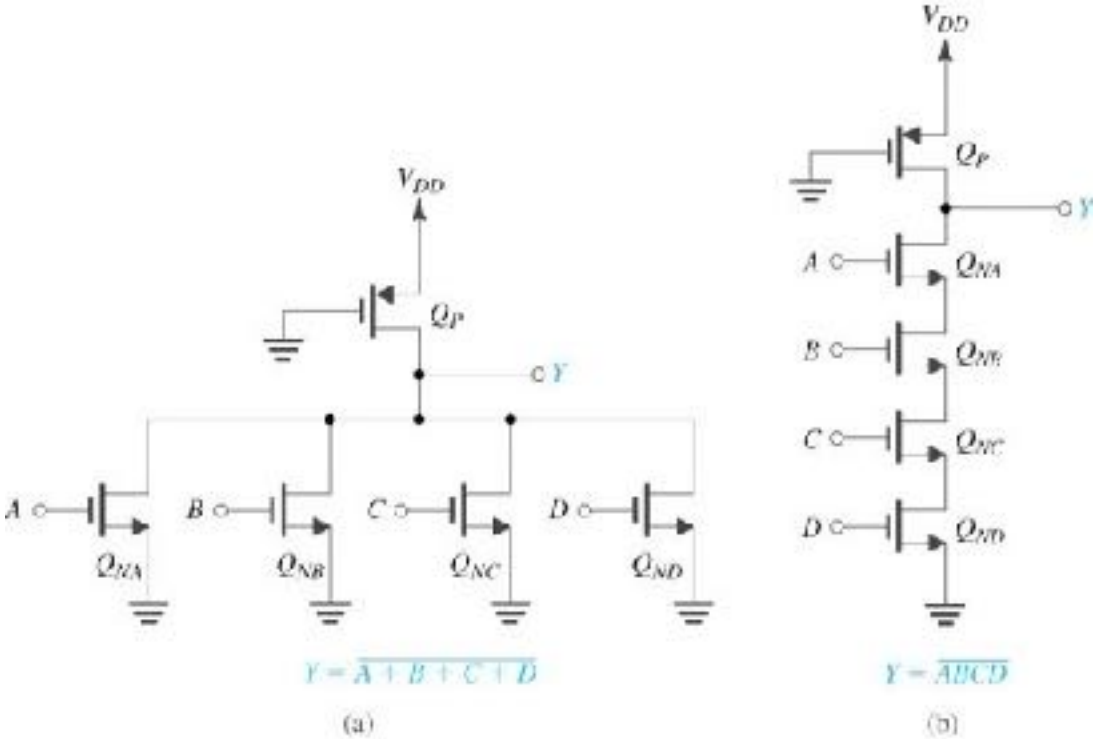
- Impact of input slope
  - As input signal slope deteriorates:

Current envelope	width	peak	integration
i_{short}	increase	increase	increase
i_c	increase	decrease	no change
$i_{short}+i_c$	increase	decrease	increase



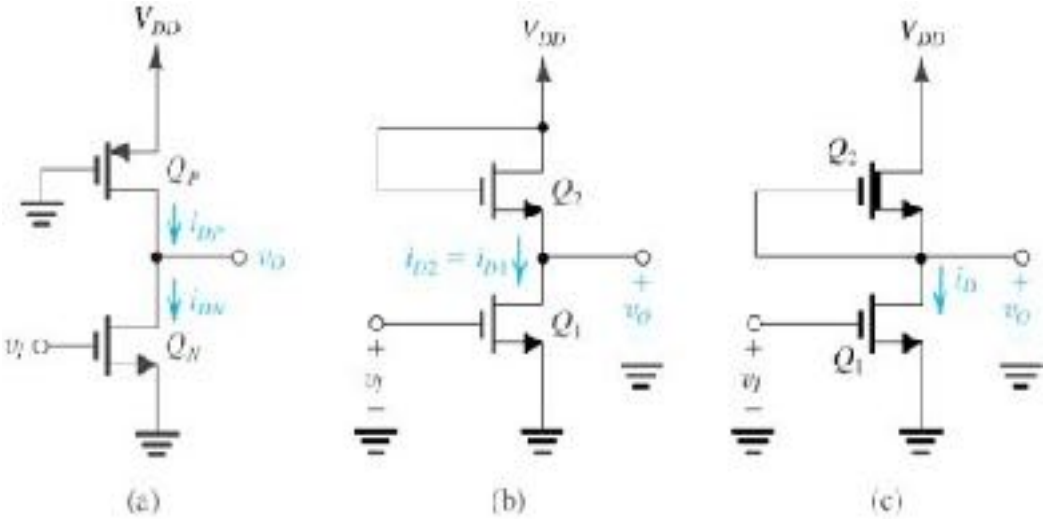
Static Power

- DC current
  - Pseudo NMOS logic



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



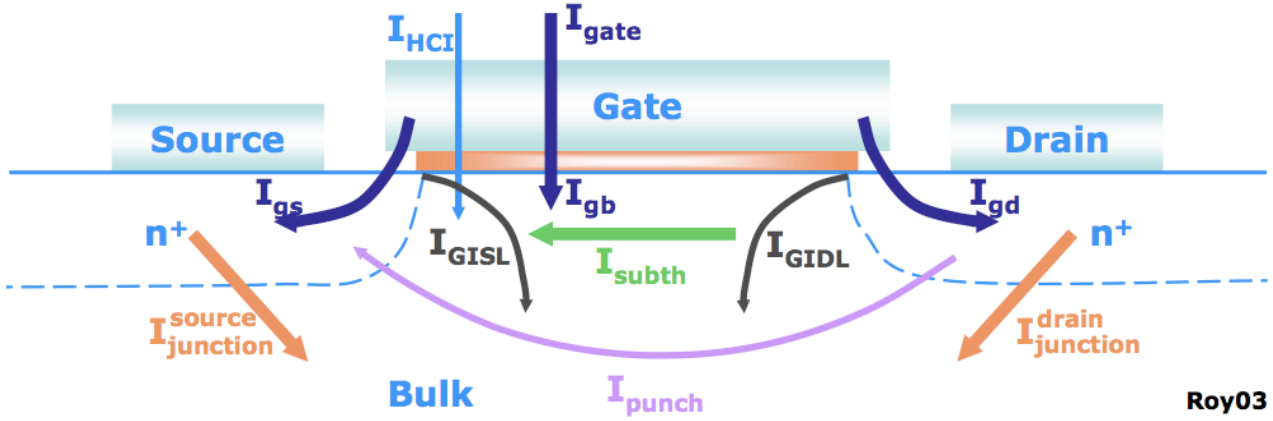
Static Power

- 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

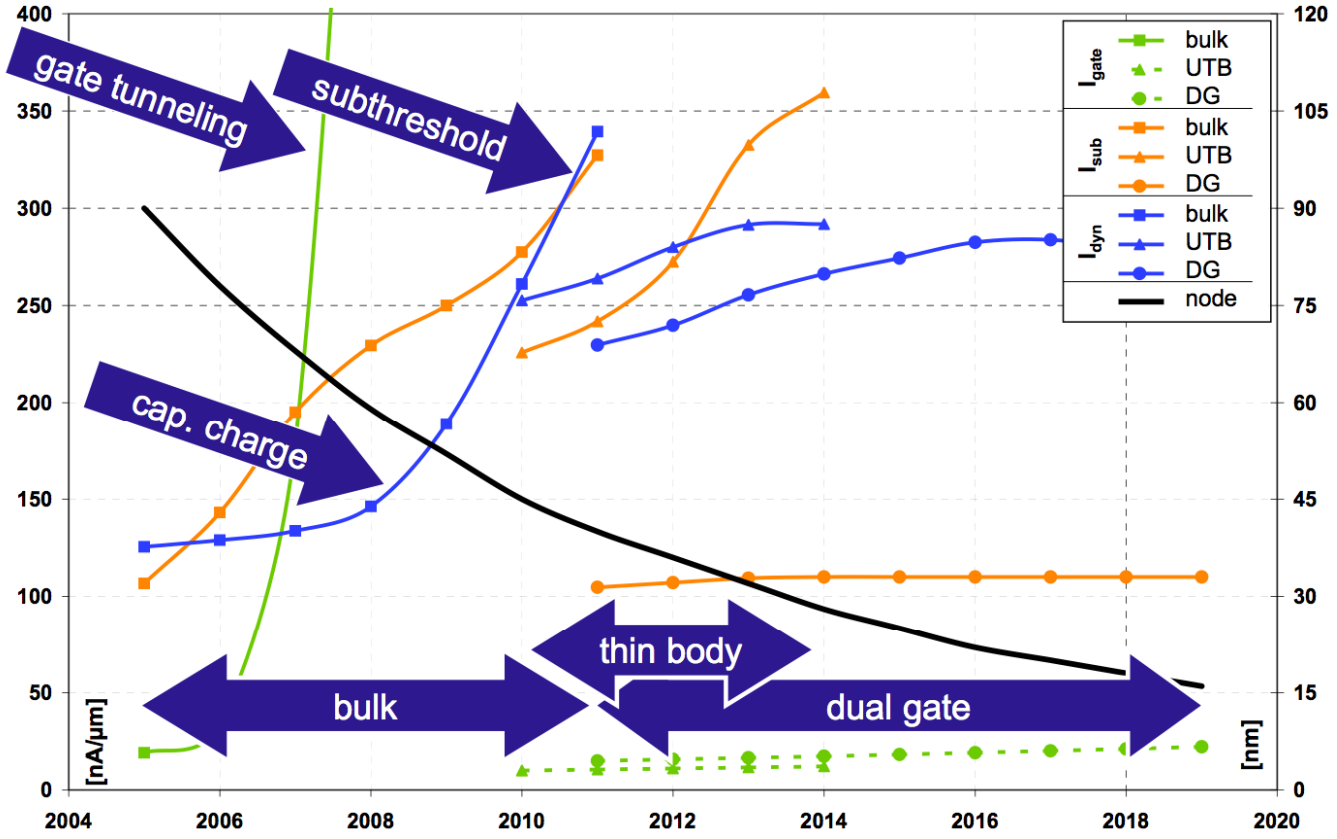
- Leakage current overview

- If channel is locking:
  - Subthreshold current (I_{subth}) ≤180nm
  - Gate tunneling to S/D (I_{gate}) ≤90nm
  - PN-junction leakage ($I_{junction}$)
  - Gate induced drain leakage (I_{GIDL}) ≤65nm
  - Depletion punch-through (I_{punch})
- If channel is conducting:
  - Gate tunneling (I_{gate})
  - PN-junction leakage ($I_{junction}$)
- If channel is switching:
  - Hot carrier injection (I_{HCI})



Static Power

- ITRS 2006 prognosis



Static Power

- Subthreshold current

$$I_{sub} = I_s e^{\frac{q(V_{GS} - V_T - V_{offset})}{nkT}} (1 - e^{\frac{-qV_{DS}}{kT}})$$

  - Weak inversion current
  - A MOSFET operates in the weak inversion (subthreshold) region when $V_{GS} < V_T$
  - Source to drain current conduction is primarily due to diffusion of the carriers
  - Subthreshold current is exponentially dependent on threshold voltage and the threshold voltage again depends on several parameters

Static Power

- Simplified equation of V_T from BSIM4 manual

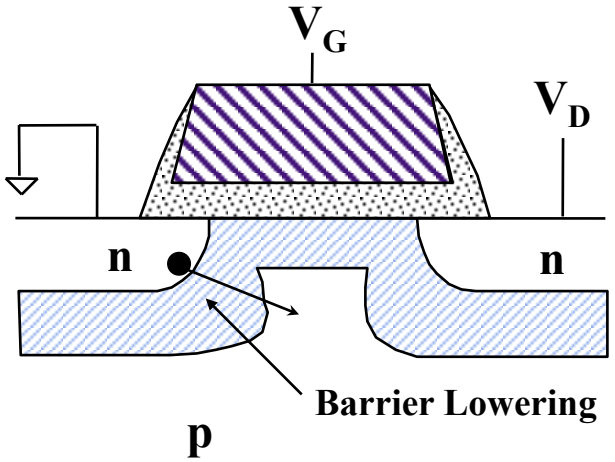
$$V_{th} = \underbrace{V_{FB}}_{\text{Flatband Voltage}} + \underbrace{\Phi_S(T)}_{\text{Surface Potential}} + \underbrace{\gamma\left(\sqrt{\Phi_S + V_{bs}} - \sqrt{\Phi_S}\right)}_{\text{Body Effect}} - \underbrace{\frac{(V_{bi}(T) - \Phi_S) + V_{ds}/2}{\cosh(L/l_c) - 1}}_{\text{Drain Induced Barrier Lowering}}$$

$$\underbrace{+\alpha(V_{bs})\frac{\Phi_S T_{ox}}{W + \Delta W}}_{\text{Narrow Width Effect}} \underbrace{- k_{retro} V_{bs}}_{\text{Non Uniform Lateral Doping}} \underbrace{+ k_{halo}(L)\sqrt{\Phi_S}}_{V_{th}\text{ Roll-up}} \underbrace{+ \Delta_{DITS}(V_{ds}, T)}_{\text{Drain Induced Threshold Shift}}$$

- First two: zero bias threshold
- Body effect
  - If bulk source voltage is 0, body effect is 0
  - A positive source voltage result in a positive term thus threshold is higher and leakage lower
- DIBL factor: the negative DIBL factor basically depends linearly on V_{DS} and exponentially on the channel length

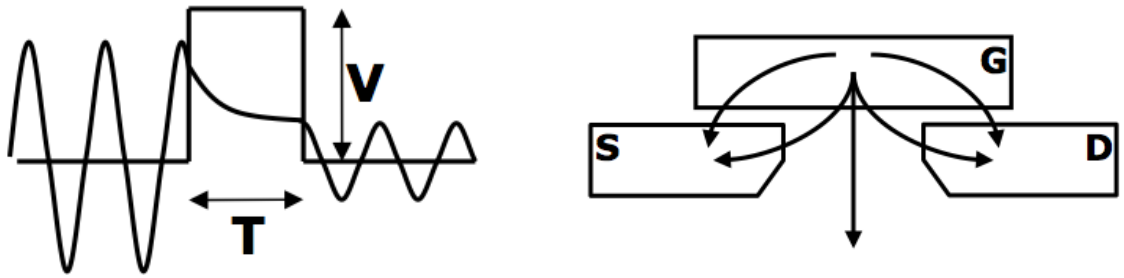
Static Power

- Drain induced barrier lowering (DIBL)
  - The depth of the junction depletion layer increases as the reverse bias voltage across the drain-to-body PN junction increases
  - Increased drain-to-body reverse bias voltage enhances the short-channel effects and lowers V_T
  - Drain current is influenced by drain voltage not just gate voltage
  - A significant portion of the subthreshold leakage current of a DSM MOSFET can be due to DIBL at high reverse bias voltage across the drain-to-body PN junction



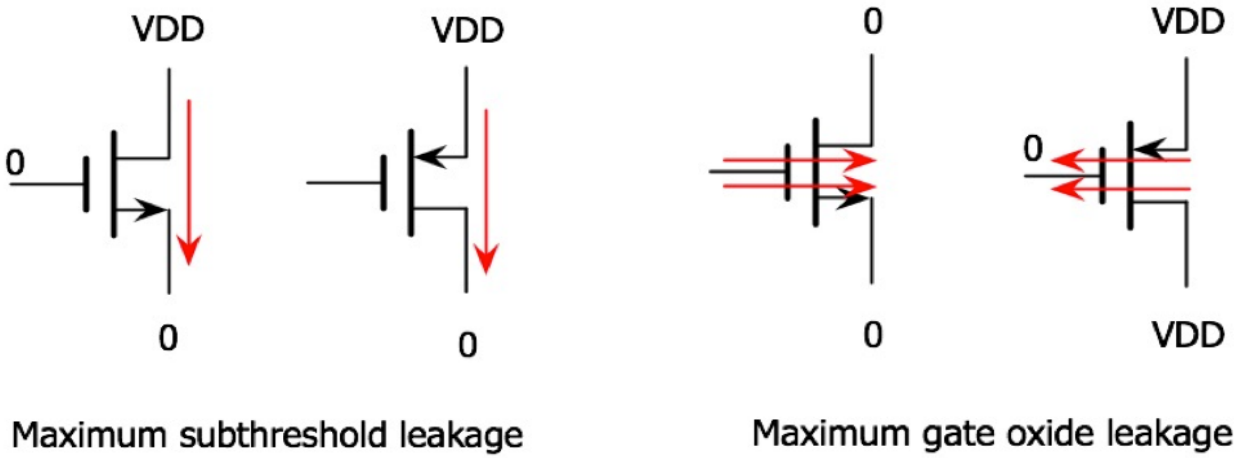
Static Power

- Gate leakage
  - Tunneling: an electron on the left can pass a barrier higher than its energy with a certain probability (classically impossible)
  - Tunneling current exponentially depends on barrier height V and width T and on carrier's mass m_{eff}
  - Leakage current can be carried by tunneling electrons or holes
    - Direct tunneling: from gate to channel
    - Fowler-Nordheim tunneling: from gate to oxide
  - In the overlap region, the tunneling can carry current from the gate to source and drain directly. The current tunneling to the cannel will go to source, drain or bulk



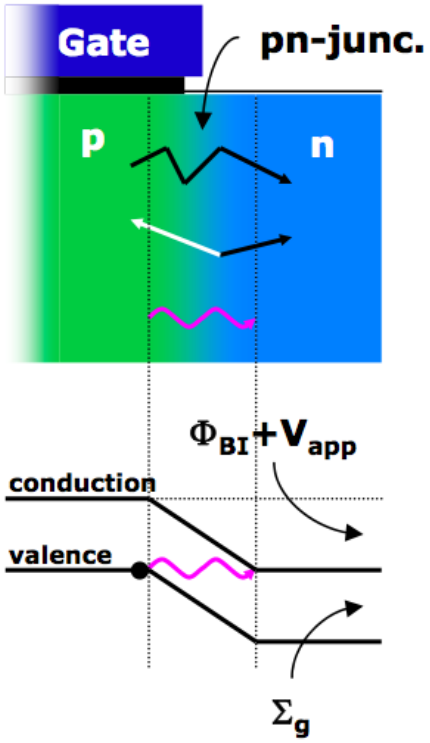
Static Power

- Maximum subthreshod leakage
  - Cut-off transistor
- Maximum gate oxide leakage
  - A transistor operates in the active region with the maximum voltage difference across the gate-to-source and the gate-to-drain terminals



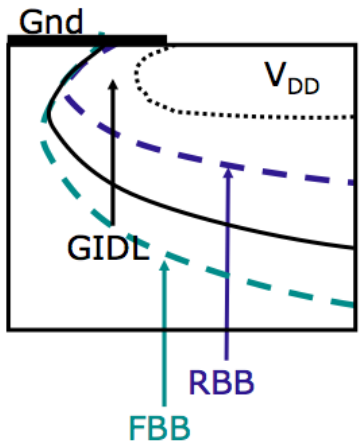
Static Power

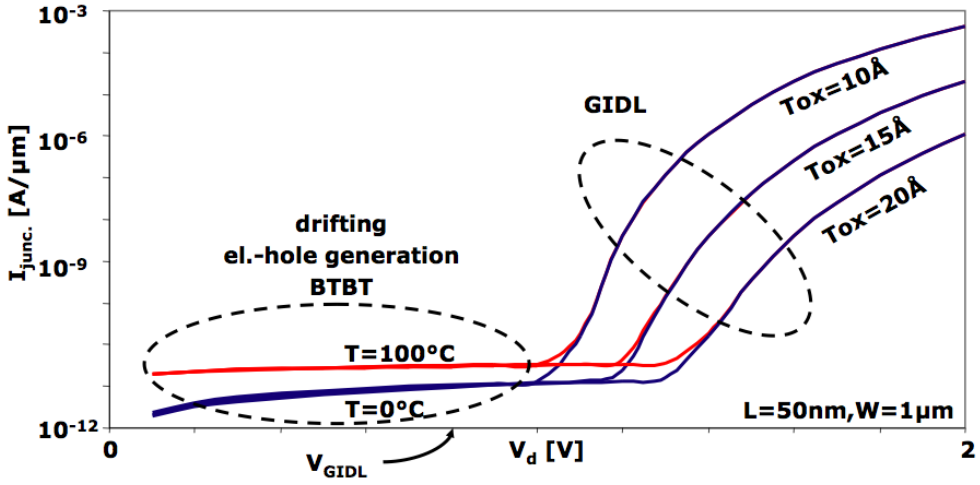
- Junction leakage
  - Schematic view of a PN-junction in reverse bias
  - As known from diodes: small currents are carried by
    - Drifting carriers
    - Electron-hole generation in junction
  - As soon as build in potential plus applied reverse bias are higher than the band gap, electrons can directly tunnel from the P-side's valence band to the N-side's conduction band → Band-To-Band-Tunneling (BTBT)



Static Power

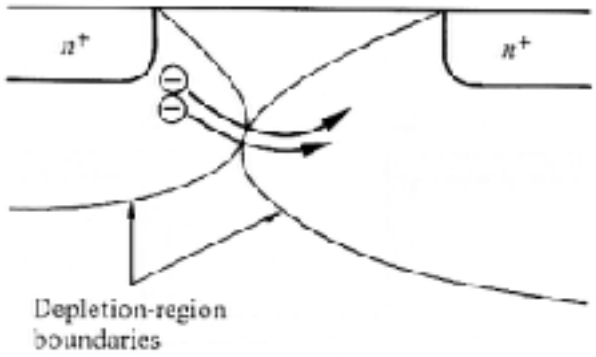
- Junction leakage
  - Smaller barrier means exponentially higher BTBT:
    - Technology scaling: steeper doping profiles
  - Gate Induced Drain Leakage (GIDL)
    - The potential difference between drain and gate makes the PN junction steeper
    - The tunneling distance smaller, and thus the current exponentially higher
  - Body biasing also influences the BTBT current





Static Power

- Depletion punch-through
  - In a sufficiently small device
  - When the drain and source depletion regions approach each other and electrically touch deep in the channel
    - Source and drain are actually merged together
  - A space-charge condition that allows the channel current to exist deep in the sub-gate region
  - Causing the gate to lose control of the sub-gate channel region



Depletion-region boundaries

Static Power

- Hot carrier injection (HCI)
  - Short-channel transistors are more susceptible to the injection of hot carriers (holes and electrons) into the oxide
  - Theses charges are a reliability risk and are measurable as gate and substrate currents
  - Can occur in the off-state, but more typically occurs during the transistor bias states in transition

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)

(V_{DD}-V_T) Design Space

- 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 S 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

(V_{DD}-V_T) Design Space

- 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
- Example
  - Leakage power is 0.1% in 25um technology
  - Leakage power is 25% in 0.1um technology

(V_{DD}-V_T) Design Space

Parameter	*Relation*	*Full Scaling*	*General Scaling*	*Fixed-V Scaling*
W, L, t_{ox}		***1/S***	***1/S***	***1/S***
V_{DD}, V_T		***1/S***	***1/U***	***1***
N_{SUB}	V/W_{depl}^2	S	S^2/U	S^2
Area/Device	WL	$1/S^2$	$1/S^2$	$1/S^2$
C_{ox}	$1/t_{ox}$	S	S	S
C_{gate}	$C_{ox}WL$	1/S	1/S	1/S
k_n, k_p	$C_{ox}W/L$	S	S	S
I_{sat}	$C_{ox}WV$	1/S	1/U	1
Current density	I_{sat}/Area	S	S^2/U	S^2
R_{on}	V/I_{sat}	1	1	1
Intrinsic Delay	$R_{on}C_{gate}$	**1/S**	**1/S**	**1/S**
P	$I_{sat}V$	**$1/S^2$**	**$1/U^2$**	**1**
Power density	P/Area	1	S^2/U^2	S^2

http://www.csee.umbc.edu/~cpatel2/links/640/lectures/lect11_scaling.pdf

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

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

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

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

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

(V_{DD}-V_T) Design Space

- (V_{DD}-V_T) design space
  - Delay of a gate

$$t_{pd} \propto \frac{C_L V_{DD}}{(V_{DD} - V_T)^{\alpha}}$$

Short-channel effect

  - Subthreshold leakage current increases exponentially as V_T decreases

$$I_{sub} - I_S^{\frac{-|V_T| \ln 10}{S}}$$

  - For a given process and V_{DD}/V_T ratio, the energy efficient V_T point is significantly below the typical threshold levels of today's technology
    - Excessive headroom for (V_{DD}-V_T) scaling
    - But lowering V_T results in bad noise margin, short-channel effect, and V_T variation

Total Power Management

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

Note

- Slides are modified from lecture notes of “Advanced Computer System Design” from Seoul National University (Lecturer: Prof. Naehyuck Chang)