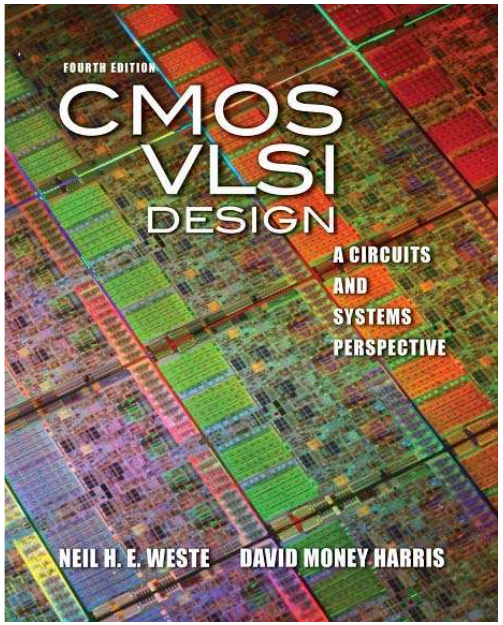
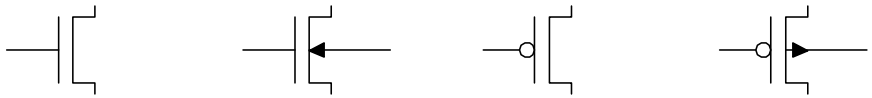


Chapter 2: MOS Transistor Theory

1. MOS transistor operation
2. I-V characteristics
3. C-V characteristics
4. DC transfer characteristics

1. MOS Transistor Operation

- So far, we have treated transistors as ideal switches
- An ON transistor passes a finite amount of current
  - Depends on terminal voltages
  - Derive current-voltage (I-V) relationships
- Transistor gate, source, drain all have capacitance
  - $I = C (\Delta V/\Delta t) \rightarrow \Delta t = (C/I)\, \Delta V$
  - Capacitance and current determine speed



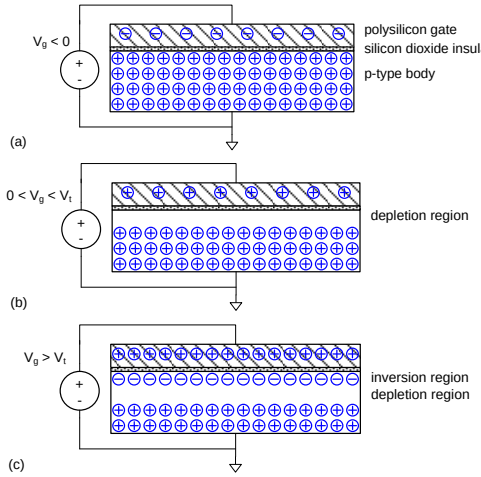
Electrical Property of MOS Devices

§ **Necessary to understand the basic electrical properties of the MOS transistor (geometry => electrical), e.g., delay/power**
- **Ensure that the circuits are robust**
- **Create working layouts**
- **Predict delays and power consumption**

§ **As technology advances and circuit dimensions scale down, Electrical effects become more important**
- **Secondary/non--ideal effects (next lecture)**

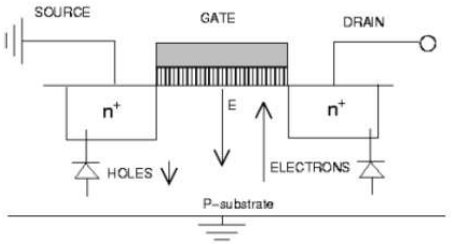
MOS Capacitor

- Gate and body form MOS capacitor
- Operating modes
  - Accumulation
  - Depletion
  - Inversion



MOS Capacitor

The nMOS transistor

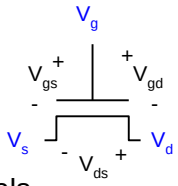


Moderately doped p- type substrate (or well) in which two heavily doped n+ regions, the Source and Drain are diffused

- Gate is insulated from substrate by thin oxide
  - Resistance of oxide is > 10^{12} Ω, so current ~ 0
- Two types of nMOS transistor
  - Enhancement mode: non conducting when gate voltage $V_{gs} = V_{sb}$ (source voltage) (normally used)
  - Depletion mode: conducting when $V_{gs} = V_{sb}$

Terminal Voltages

- Mode of operation depends on V_g, V_d, V_s
  - $V_{gs} = V_g - V_s$
  - $V_{gd} = V_g - V_d$
  - $V_{ds} = V_d - V_s = V_{gs} - V_{gd}$

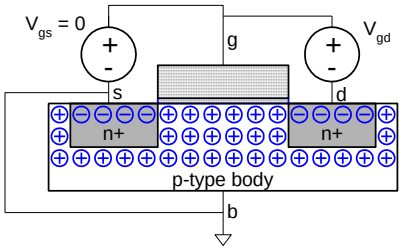


- Source and drain are symmetric diffusion terminals
  - By convention, source is terminal at lower voltage
  - Hence $V_{ds} \geq 0$
- nMOS body is grounded. First assume source is 0 too.
- Three regions of operation
  - *Cutoff*
  - *Linear*
  - *Saturation*

nMOS

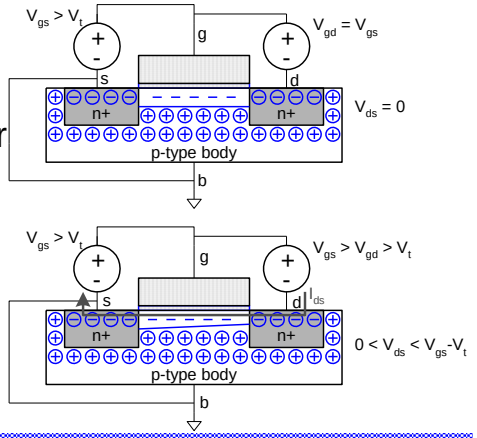
nMOS Cutoff

- No channel
- $I_{ds} \approx 0$



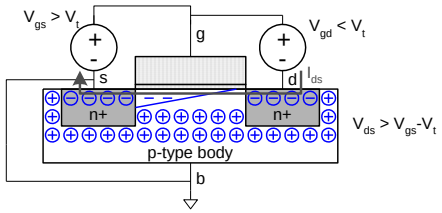
nMOS Linear

- Channel forms
- Current flows from d to s
  - e^- from s to d
- I_{ds} increases with V_{ds}
- Similar to linear resistor

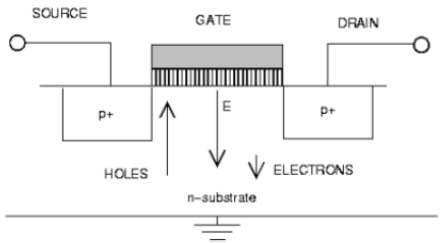


nMOS Saturation

- Channel pinches off
- I_{ds} independent of V_{ds}
- We say current *saturates*
- Similar to current source



The pMOS transistor

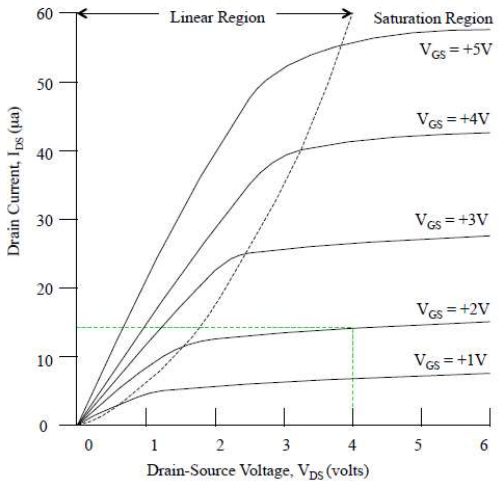


Moderately doped n- type substrate (or well) in which two heavily doped p+ regions, the Source and Drain are diffused

- Application of a negative gate voltage (w.r.t. source) draws holes into the region below the gate; channel changes from n to p-type (source-drain conduction path)
- Conduction due to holes; negative V_d sweeps holes from source (through channel) to drain

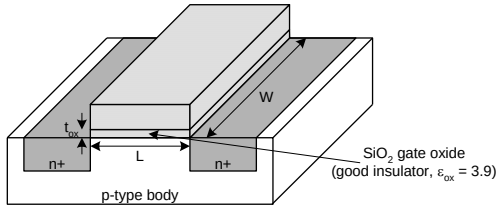
2. I-V Characteristics

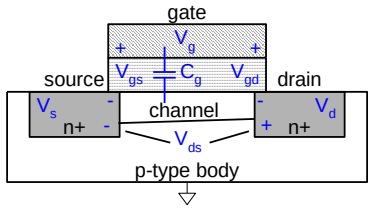
- In Linear region, I_{ds} depends on
  - How much charge is in the channel?
  - How fast is the charge moving?



Channel Charge

- MOS structure looks like parallel plate capacitor while operating in inversions
  - Gate – oxide – channel
- $Q_{channel}$ =
-
-

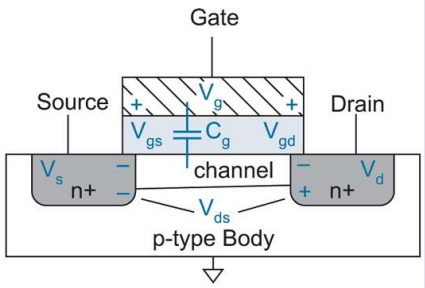




Channel Charge

- MOS structure looks like parallel plate capacitor while operating in inversion:
  - Gate – oxide – channel
- $Q_{channel} = CV$
- $C = C_g = \varepsilon_{ox}WL/t_{ox} = c_{ox}WL$
- $V = V_{gc} - V_t = (V_{gs} - V_{ds}/2) - V_t$

$$c_{ox} = \varepsilon_{ox} / t_{ox}$$



Average gate to channel potential:

$V_{gc} = (V_{gs} + V_{gd})/2 = V_{gs} - V_{ds}/2$

FIG 2.5 Average gate to channel voltage

Carrier velocity

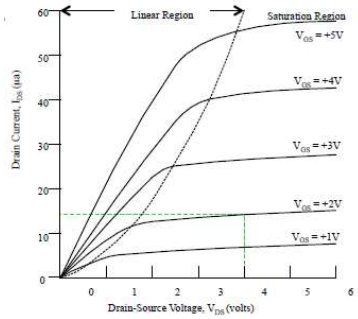
- Charge is carried by e-
- Electrons are propelled by the lateral electric field between source and drain
  - E =
- Carrier velocity *v* proportional to lateral E-field
  - *v* =
- Time for carrier to cross channel:
  - *t* =

nMOS Linear I-V

Now we know:

How much charge $Q_{channel}$ is in the channel

How much time *t* each carrier takes to cross



$$I_{ds} =$$

$$=$$

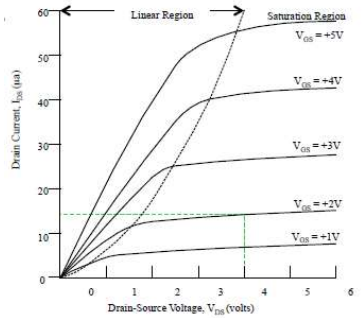
$$=$$

nMOS Saturation I-V

- If $V_{gd} < V_t$, channel pinches off near drain
  - When $V_{ds} > V_{dsat}$ =
- Now drain voltage no longer increases current

$$I_{ds} =$$

$$=$$

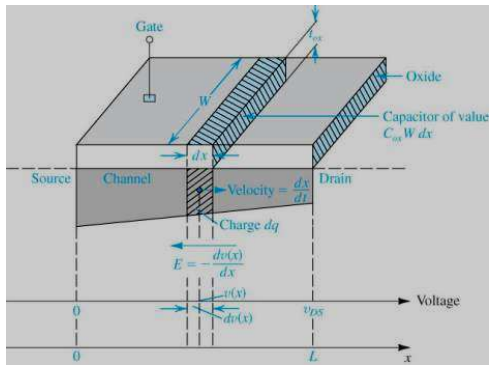


nMOS I-V Summary

- *Shockley* 1st order transistor models

$$I_{ds} = \begin{cases} 0 & V_{gs} < V_t \quad \text{cutoff} \\ \beta\left(V_{gs} - V_t - \frac{V_{ds}}{2}\right)V_{ds} & V_{ds} < V_{dsat} \quad \text{linear} \\ \frac{\beta}{2}\left(V_{gs} - V_t\right)^2 & V_{ds} > V_{dsat} \quad \text{saturation} \end{cases}$$

Tóm tắt cách tìm quan hệ dòng và áp trong N-EMOS



$$i = \frac{dq}{dt} = \frac{dq}{dx}\frac{dx}{dt} \quad (1)$$

$$Q = CV \qquad C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} \quad (2)$$

$$\varepsilon_{ox} = 3.9\varepsilon_0 = 3.9 \times 8.854 \times 10^{-12} = 3.45 \times 10^{-11} \text{ F/m}$$

$$dq = -C_{ox}(Wdx)\left[\upsilon_{GS} - \upsilon(x) - V_t\right] \quad (3)$$

$$E(x) = -\frac{d\upsilon(x)}{dx} \quad (\because E = -\nabla V)$$

$$\frac{dx}{dt} = -\mu_n E(x) = \mu_n \frac{d\upsilon(x)}{dx} \quad (4)$$

$$i = -\mu_n C_{ox} W\left[\upsilon_{GS} - \upsilon(x) - V_t\right]\frac{d\upsilon(x)}{dx}$$

$$i_D = -i = \mu_n C_{ox} W\left[\upsilon_{GS} - \upsilon(x) - V_t\right]\frac{d\upsilon(x)}{dx}$$

$$i_D dx = \mu_n C_{ox} W\left[\upsilon_{GS} - V_t - \upsilon(x)\right]d\upsilon(x)$$

$$\int_0^L i_D dx = \int_0^{\upsilon_{DS}} \mu_n C_{ox} W\left[\upsilon_{GS} - V_t - \upsilon(x)\right]d\upsilon(x)$$

$$i_D = (\mu_n C_{ox})\left(\frac{W}{L}\right)\left[(\upsilon_{GS} - V_t)\upsilon_{DS} - \frac{1}{2}\upsilon_{DS}^2\right] \quad (5)$$

Khi vào miền bão hòa, $\upsilon_{DS} = \upsilon_{GS} - V_t$

$$i_D = \frac{1}{2}(\mu_n C_{ox})\left(\frac{W}{L}\right)(\upsilon_{GS} - V_t)^2 \quad (6)$$

$$k_n' = \mu_n C_{ox} \quad (7) \qquad C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} \quad (2)$$

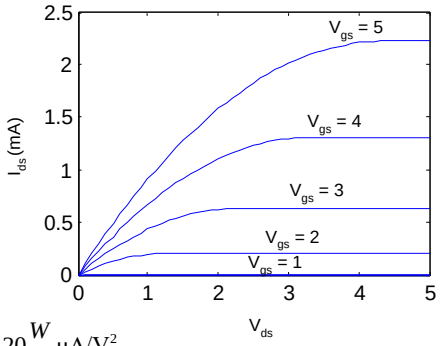
$$i_D = k_n'\left(\frac{W}{L}\right)\left[(\upsilon_{GS} - V_t)\upsilon_{DS} - \frac{1}{2}\upsilon_{DS}^2\right] \quad \textbf{(Triode region)} \quad (5a)$$

$$i_D = \frac{1}{2}k_n'\left(\frac{W}{L}\right)(\upsilon_{GS} - V_t)^2 \quad \textbf{(saturation region)} \quad (6a)$$

$\left(\frac{W}{L}\right)$: **Aspect ratio of the MOSFET**
(Tỉ số hình dạng của MOSFET)

Example

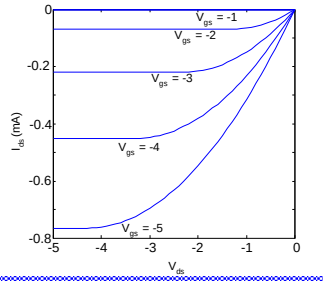
- We will be using a 0.6 μm process for your project
  - From AMI Semiconductor
  - t_{ox} = 100 Å
  - μ = 350 $cm^2/V*s$
  - V_t = 0.7 V
- Plot I_{ds} vs. V_{ds}
  - V_{gs} = 0, 1, 2, 3, 4, 5
  - Use W/L = 4/2 λ



$$\beta = \mu C_{ox}\frac{W}{L} = (350)\left(\frac{3.9\times 8.85\cdot 10^{-14}}{100\cdot 10^{-8}}\right)\left(\frac{W}{L}\right) = 120\frac{W}{L}\mu A/V^2$$

pMOS I-V

- All dopings and voltages are inverted for pMOS
  - Source is the more positive terminal
- Mobility μ_p is determined by holes
  - Typically 2-3x lower than that of electrons μ_n
  - 120 $cm^2/V•s$ in AMI 0.6 μm process
- Thus pMOS must be wider to provide same current
  - In this class, assume $\mu_n / \mu_p = 2$



Assignment

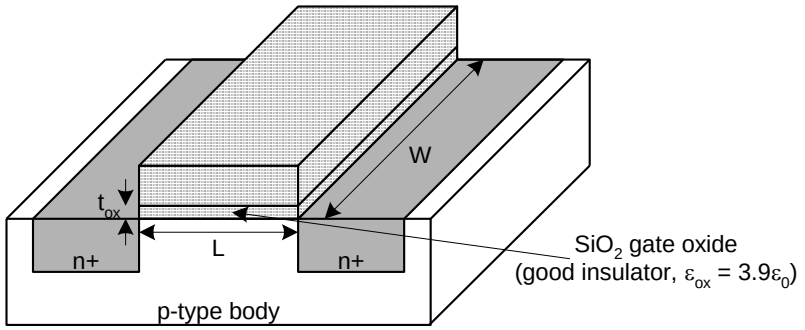
- Consider an nMOS transistor in a 0.6 μm process with *W*/*L* = 4/2 λ (i.e., 1.2/0.6 μm). In this process, the gate oxide thickness is 100 Aº and the mobility of electrons is 350 $cm^2/V\cdot s$. The threshold voltage is 0.7 V. Plot I_{ds} vs. V_{ds} for V_{gs} = 0, 1, 2, 3, 4, and 5 V.

3. C-V Characteristics

- Any two conductors separated by an insulator have capacitance
- Gate to channel capacitor is very important
  - Creates channel charge necessary for operation
- Source and drain have capacitance to body
  - Across reverse-biased diodes
  - Called diffusion capacitance because it is associated with source/drain diffusion

Gate Capacitance

- Approximate channel as connected to source
- $C_g = C_{ox}WL = C_{permicron}W$
- $C_{permicron} = C_{ox}L = \varepsilon_{ox}L/t_{ox}$
- $C_{permicron}$ is typically about 2 fF/μm



Gate Capacitance

- When the transistor is off, the channel is not inverted
  $C_g = C_{gb} = \varepsilon_{ox}WL/t_{ox} = C_{ox}WL$
- Let's call $C_{ox}WL = C_0$
- When the transistor is on, the channel extends from the source to the drain (if the transistor is unsaturated, or to the pinchoff point otherwise)
  $C_g = C_{gb} + C_{gs} + C_{gd}$

Gate Capacitance

Table 2.1 Approximation of intrinsic MOS gate capacitance

Parameter	Cutoff	Linear	Saturation
C_{gb}	C_0	0	0
C_{gs}	0	$C_0/2$	$2/3\ C_0$
C_{gd}	0	$C_0/2$	0
$C_g = C_{gs} + C_{gd} + C_{gb}$	C_0	C_0	$2/3\ C_0$

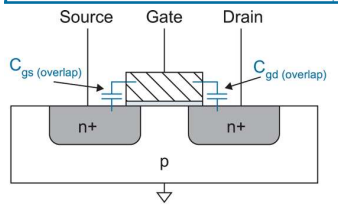


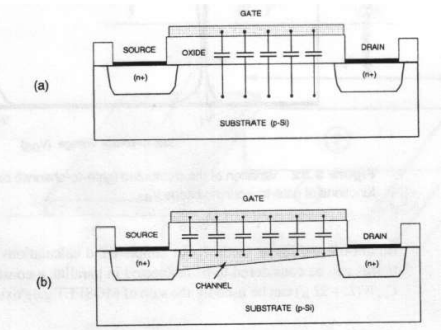
FIG 2.10 Overlap capacitance

In reality the gate overlaps source and drain. Thus, the gate capacitance should include not only the intrinsic capacitance but also parasitic overlap capacitances:

$C_{gs}(overlap) = C_{ox}\ W\ L_D$
$C_{gs}(overlap) = C_{ox}\ W\ L_D$

Detailed Gate Capacitance

Capacitance	Cutoff	Linear	Saturation
C_{gb} (total)	C_0	0	0
C_{gd} (total)	$C_{ox}WL_D$	$C_0/2 + C_{ox}WL_D$	$C_{ox}WL_D$
C_{gs} (total)	$C_{ox}WL_D$	$C_0/2 + C_{ox}WL_D$	$2/3\ C_0 + C_{ox}WL_D$



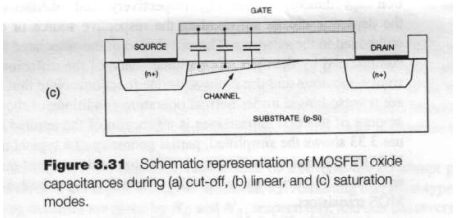
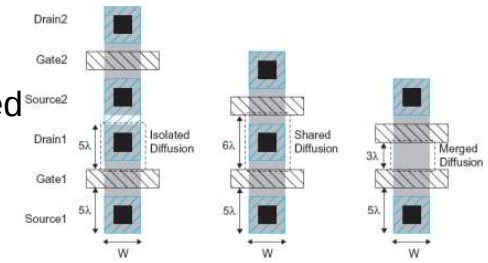


Figure 3.31 Schematic representation of MOSFET oxide capacitances during (a) cut-off, (b) linear, and (c) saturation modes.

Diffusion Capacitance

- C_{sb}, C_{db}
- Undesirable, called *parasitic* capacitance
- Capacitance depends on area and perimeter
  - Use small diffusion nodes
  - Comparable to C_g for contacted diff
  - ½ C_g for uncontacted
  - Varies with process



Lumped representation of the MOSFET capacitances

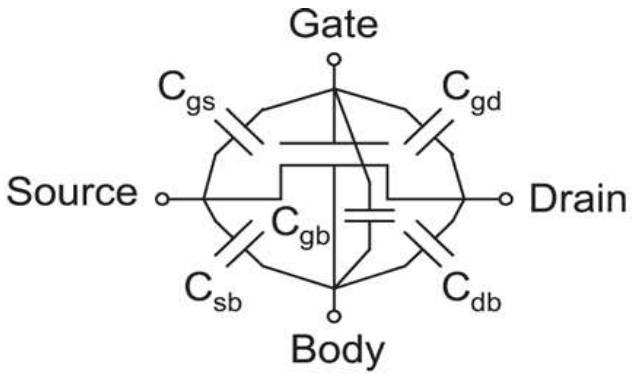


FIG 2.14 Capacitances of an MOS transistor

Review

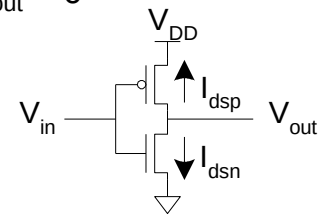
1. Which factors affect to switching speed of MOS transistors?
2. What are three operation modes of MOS transistors?
3. When is a MOS transistor OFF?
4. When does a MOS transistor operate in saturation region?
5. What does the drain-source current I_{ds} depend on?
6. What does the gate capacitance C_g depend on?

4. DC Transfer Characteristics

- Assuming the input changes **slowly enough** that capacitances have plenty of time to charge or discharge.
- Specific ranges of input and output voltages are defined as valid 0 and 1 logic levels

DC Response

- DC Response: V_{out} vs. V_{in} for a gate
- Ex: Inverter
  - When $V_{in} = 0$ -> $V_{out} = V_{DD}$
  - When $V_{in} = V_{DD}$ -> $V_{out} = 0$
  - In between, V_{out} depends on transistor size and current
  - By KCL, must settle such that $I_{dsn} = |I_{dsp}|$
  - We could solve equations
  - But graphical solution gives more insight

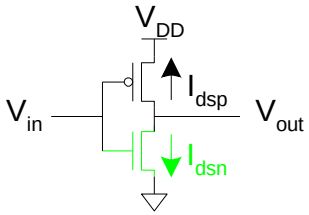


Transistor Operation

- Current depends on region of transistor behavior
- For what V_{in} and V_{out} are nMOS and pMOS in
  - Cutoff?
  - Linear?
  - Saturation?

nMOS Operation

Cutoff	Linear	Saturated
$V_{gsn} <$	$V_{gsn} >$	$V_{gsn} >$
	$V_{dsn} <$	$V_{dsn} >$



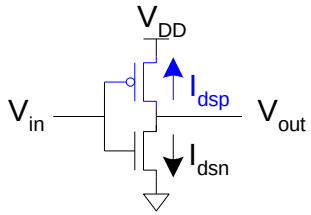
pMOS Operation

Cutoff	Linear	Saturated
$V_{gsp} > V_{tp}$	$V_{gsp} < V_{tp}$	$V_{gsp} < V_{tp}$
$V_{in} > V_{DD} + V_{tp}$	$V_{in} < V_{DD} + V_{tp}$	$V_{in} < V_{DD} + V_{tp}$
	$V_{dsp} > V_{gsp} - V_{tp}$	$V_{dsp} < V_{gsp} - V_{tp}$
	$V_{out} > V_{in} - V_{tp}$	$V_{out} < V_{in} - V_{tp}$

$V_{gsp} = V_{in} - V_{DD}$

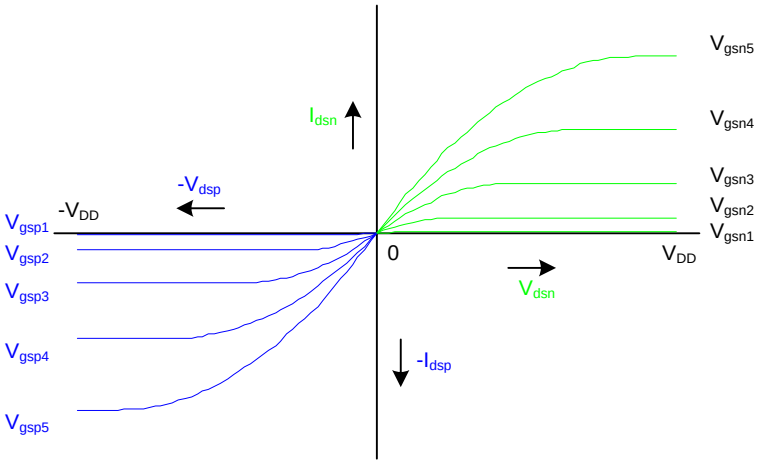
$V_{dsp} = V_{out} - V_{DD}$

$V_{tp} < 0$

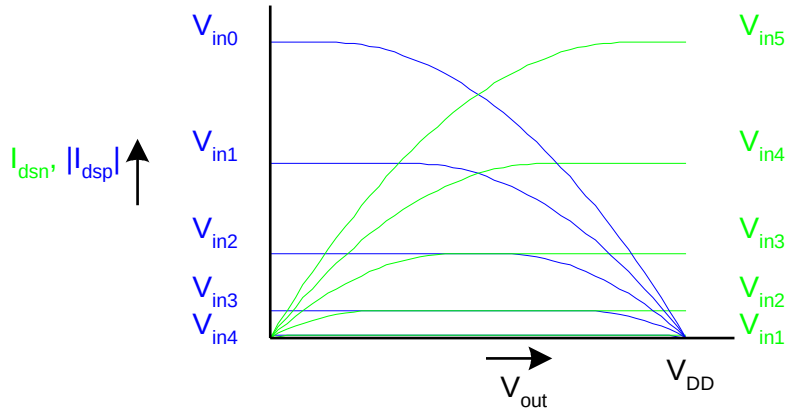


I-V Characteristics

- Make pMOS is wider than nMOS such that $\beta_n = \beta_p$

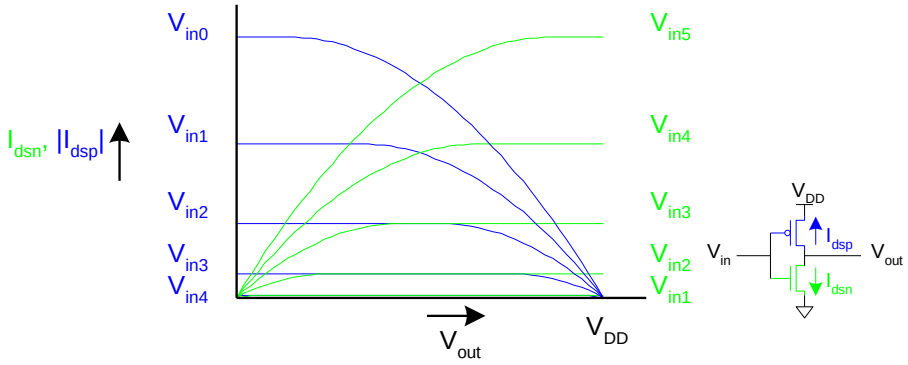


Current vs. V_{out}, V_{in}

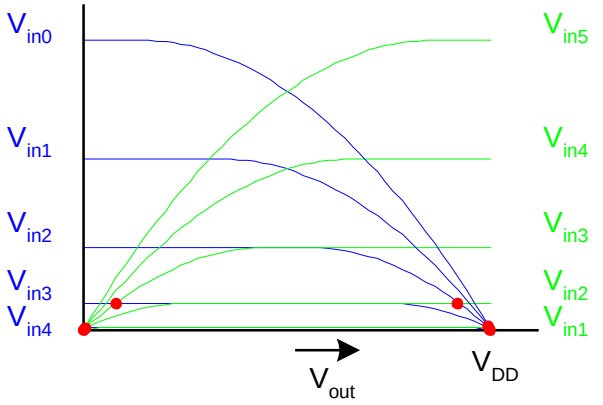


Load Line Analysis

- For a given V_{in}:
  - Plot I_{dsn}, I_{dsp} vs. V_{out}
  - V_{out} must be where |currents| are equal in

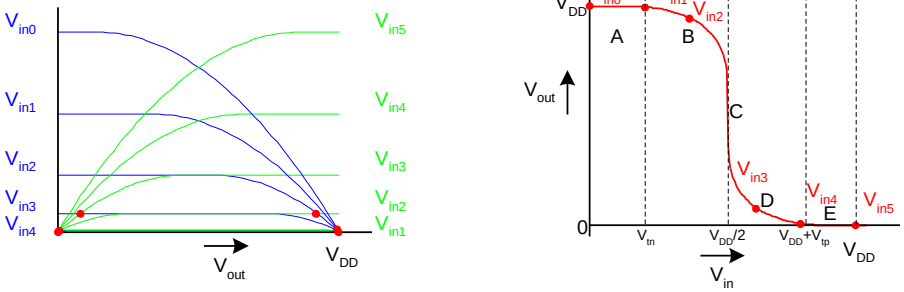


Load Line Analysis



DC Transfer Curve

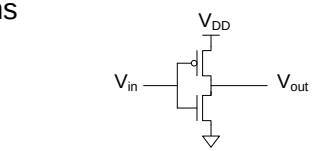
- Transcribe points onto V_{in} vs. V_{out} plot

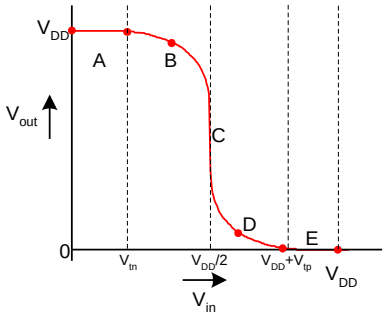


Operating Regions

- Revisit transistor operating regions

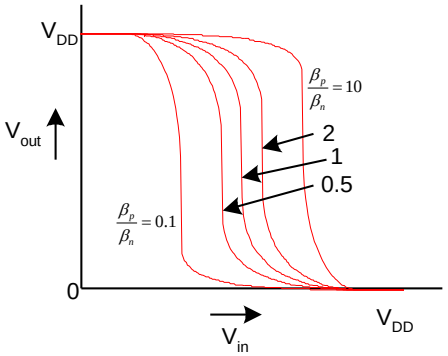
Region	nMOS	pMOS
A		
B		
C		
D		
E		





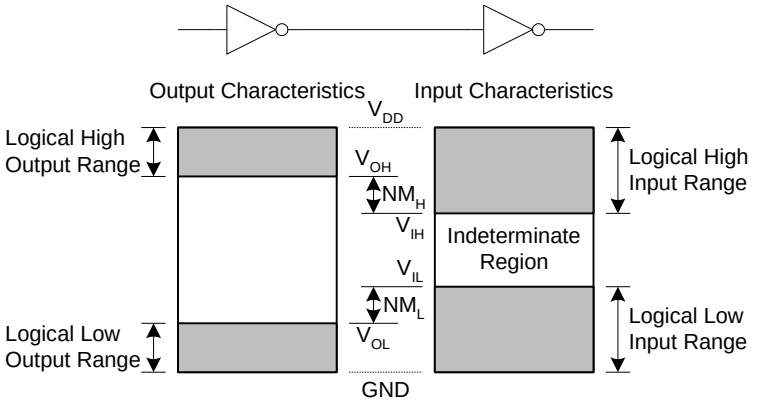
Beta Ratio

- If $\beta_p / \beta_n \neq 1$, switching point will move from $V_{DD}/2$
- Called *skewed* gate
- Other gates: collapse into equivalent inverter



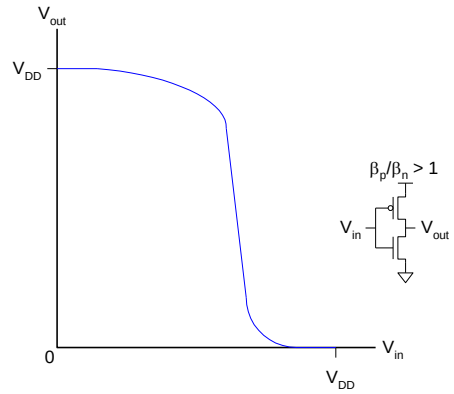
Noise Margins

- How much noise can a gate input see before it does not recognize the input?



Logic Levels

- To maximize noise margins, select logic levels at
  - unity gain point of DC transfer characteristic



Transient Response

- *DC analysis* tells us V_{out} if V_{in} is constant
- *Transient analysis* tells us $V_{out}(t)$ if $V_{in}(t)$ changes
  - Requires solving differential equations
- Input is usually considered to be a step or ramp
  - From 0 to V_{DD} or vice versa

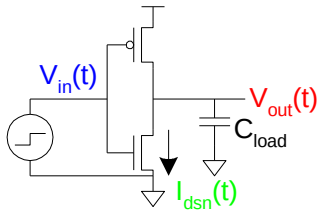
Inverter Step Response

- Ex: find step response of inverter driving load cap

$$V_{in}(t) =$$

$$V_{out}(t < t_0) =$$

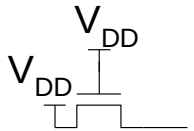
$$\frac{dV_{out}(t)}{dt} =$$



$$I_{dsn}(t) = \begin{cases} & t \le t_0 \\ & V_{out} > V_{DD} - V_t \\ & V_{out} < V_{DD} - V_t \end{cases}$$

Pass Transistors

- We have assumed source is grounded
- What if source > 0?
  - e.g. pass transistor passing V_{DD}
- $V_g = V_{DD}$
  - If $V_s > V_{DD}-V_t$, $V_{gs} < V_t$
  - Hence transistor would turn itself off
- nMOS pass transistors pull no higher than $V_{DD}-V_{tn}$
  - Called a degraded "1"
  - Approach degraded value slowly (low I_{ds})
- pMOS pass transistors pull no lower than V_{tp}
- **Transmission gates** are needed to pass both 0 and 1



Pass Transistor Ckts

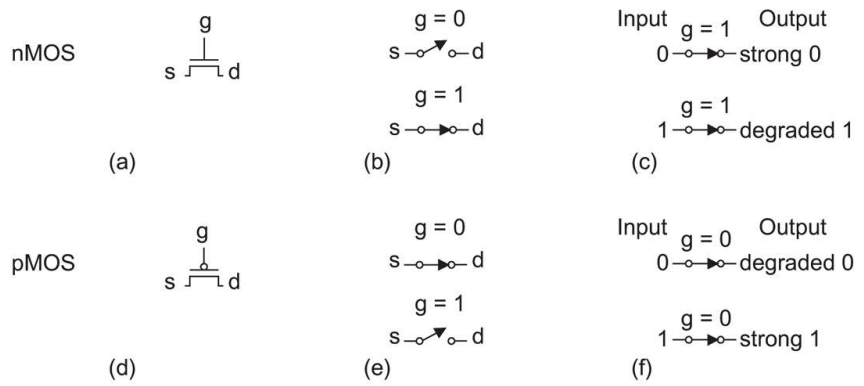
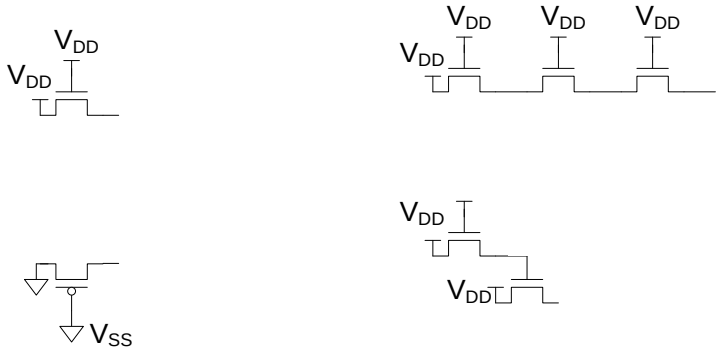
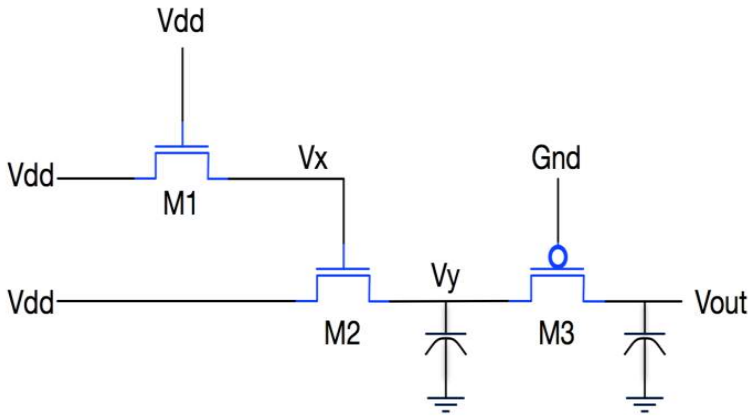


FIG 1.19 Pass transistor strong and degraded outputs

Pass Transistor Ckts



Pass Transistor Ckts



Transmission Gates

- Pass transistors produce degraded outputs
- Transmission gates pass both 0 and 1 well

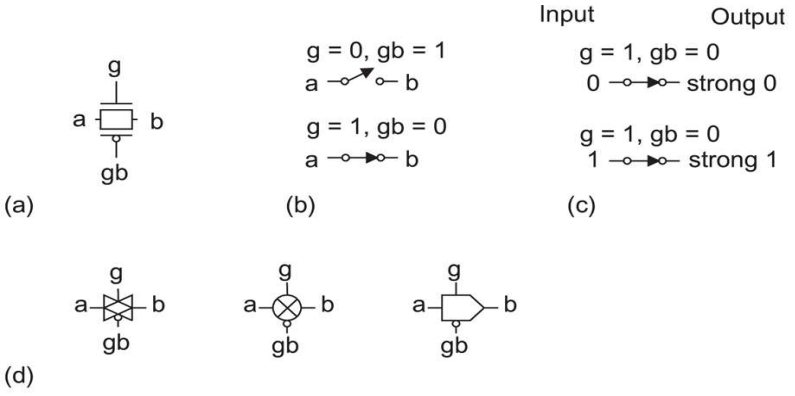


FIG 1.20 Transmission gate

Transmission gate ON resistance

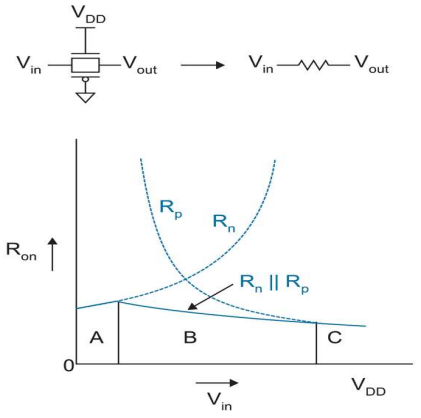
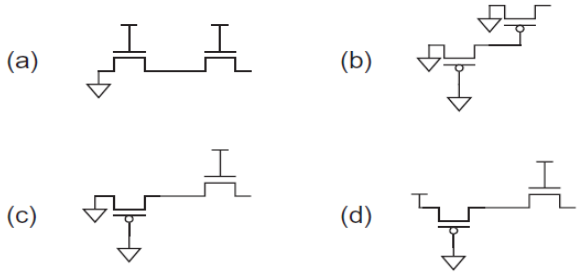
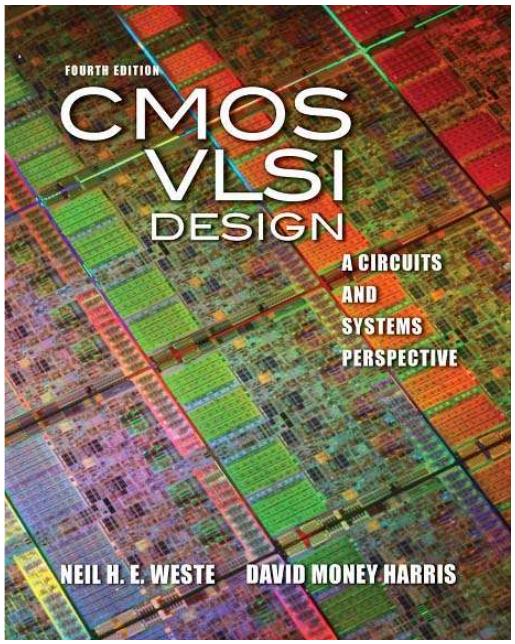


FIG 2.32 Resistance of a transmission gate as a function of input voltage

Review

1. What are conditions for V_{gs} when the nMOS is in cutoff, linear, and saturated modes?
2. What is noise margin?
3. What is transmission gate and its applications?
4. What are V_{tp} and V_{tn} ?
5. Give expressions for the output voltage.





Lecture 4: Nonideal Transistor Theory

Outline

- Nonideal Transistor Behavior
  - High Field Effects
    - Mobility Degradation
    - Velocity Saturation
  - Channel Length Modulation
  - Threshold Voltage Effects
    - Body Effect
    - Drain-Induced Barrier Lowering
    - Short Channel Effect
  - Leakage
    - Subthreshold Leakage
    - Gate Leakage
    - Junction Leakage
- Process and Environmental Variations

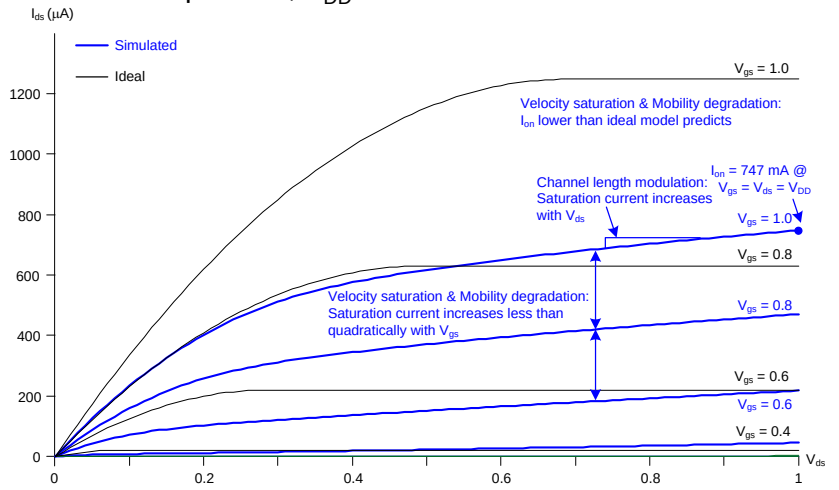
Ideal Transistor I-V

- *Shockley* long-channel transistor models

$$I_{ds} = \begin{cases} 0 & V_{gs} < V_t \quad \text{cutoff} \\ \beta\left(V_{gs} - V_t - \frac{V_{ds}}{2}\right)V_{ds} & V_{ds} < V_{dsat} \quad \text{linear} \\ \frac{\beta}{2}\left(V_{gs} - V_t\right)^2 & V_{ds} > V_{dsat} \quad \text{saturation} \end{cases}$$

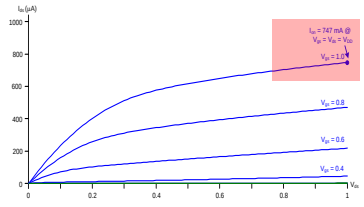
Ideal vs. Simulated nMOS I-V Plot

- 65 nm IBM process, V_{DD} = 1.0 V

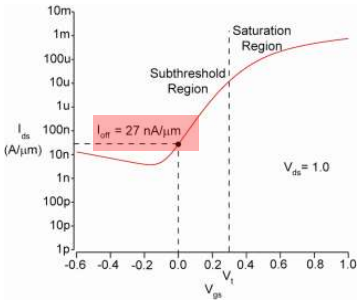


ON and OFF Current

- $I_{on} = I_{ds}$ @ $V_{gs} = V_{ds} = V_{DD}$
  - Saturation



- $I_{off} = I_{ds}$ @ $V_{gs} = 0$, $V_{ds} = V_{DD}$
  - Cutoff



Electric Fields Effects

- Vertical electric field: E_{vert} = _____
  - Attracts carriers into channel
  - Long channel: $Q_{channel} \propto E_{vert}$
- Lateral electric field: E_{lat} = _____
  - Accelerates carriers from drain to source
  - Long channel: $v = \mu E_{lat}$

Coffee Cart Analogy

- Tired student runs from VLSI lab to coffee cart
- Freshmen are pouring out of the physics lecture hall
- V_{ds} is how long you have been up
  - Your velocity = fatigue × mobility
- V_{gs} is a wind blowing you against the glass (SiO_2) wall
- At high V_{gs}, you are buffeted against the wall
  - *Mobility degradation*
- At high V_{ds}, you scatter off freshmen, fall down, get up
  - *Velocity saturation*
    - Don't confuse this with the saturation region

Mobility Degradation

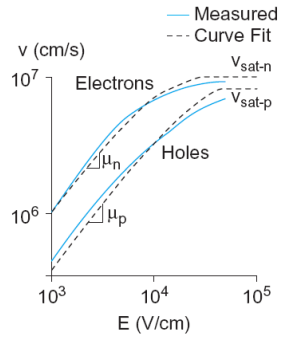
- High E_{vert} effectively reduces mobility
  - Collisions with oxide interface

$$\mu_{\text{eff}-n} = \frac{540\,\frac{\text{cm}^2}{\text{V}\cdot\text{s}}}{1+\left(\frac{V_{gs}+V_t}{0.54\,\frac{\text{V}}{\text{nm}}\,t_{\text{ox}}}\right)^{1.85}} \qquad \mu_{\text{eff}-p} = \frac{185\,\frac{\text{cm}^2}{\text{V}\cdot\text{s}}}{1+\frac{\left|V_{gs}+1.5V_t\right|}{0.338\,\frac{\text{V}}{\text{nm}}\,t_{\text{ox}}}}$$

Velocity Saturation

- At high E_{lat}, carrier velocity rolls off
  - Carriers scatter off atoms in silicon lattice
  - Velocity reaches v_{sat}
    - Electrons: 10^7 cm/s
    - Holes: 8 x 10^6 cm/s
  - Better model

$$v = \begin{cases} \dfrac{\mu_{\text{eff}} E}{1+\dfrac{E}{E_c}} & E < E_c \\ v_{\text{sat}} & E \geq E_c \end{cases} \qquad E_c = \frac{2v_{\text{sat}}}{\mu_{\text{eff}}}$$



Vel Sat I-V Effects

- Ideal transistor ON current increases with V_{DD}^2

$$I_{ds} = \mu C_{\text{ox}} \frac{W}{L} \frac{\left(V_{gs}-V_t\right)^2}{2} = \frac{\beta}{2}\left(V_{gs}-V_t\right)^2$$

- Velocity-saturated ON current increases with V_{DD}

$$I_{ds} = C_{\text{ox}} W \left(V_{gs}-V_t\right) v_{\max}$$

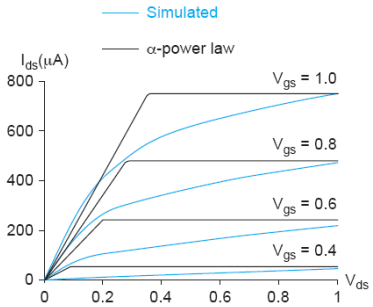
- Real transistors are partially velocity saturated
  - Approximate with α-power law model
  - $I_{ds} \propto V_{DD}^{\alpha}$
  - $1 < \alpha < 2$ determined empirically (≈ 1.3 for 65 nm)

α-Power Model

$$I_{ds} = \begin{cases} 0 & V_{gs} < V_t \quad \text{cutoff} \\ I_{dsat}\dfrac{V_{ds}}{V_{dsat}} & V_{ds} < V_{dsat} \quad \text{linear} \\ I_{dsat} & V_{ds} > V_{dsat} \quad \text{saturation} \end{cases}$$

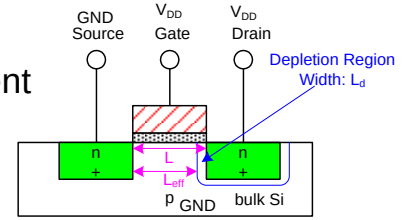
$$I_{dsat} = P_c \frac{\beta}{2}\left(V_{gs} - V_t\right)^{\alpha}$$

$$V_{dsat} = P_v \left(V_{gs} - V_t\right)^{\alpha/2}$$



Channel Length Modulation

- Reverse-biased p-n junctions form a *depletion region*
  - Region between n and p with no carriers
  - Width of depletion L_d region grows with reverse bias
  - $L_{eff} = L - L_d$
- Shorter L_{eff} gives _____ current
  - I_{ds} _____ with V_{ds}
  - Even in saturation



Channel Length Mod I-V

$$I_{ds} = \frac{\beta}{2}\left(V_{gs} - V_t\right)^2\left(1 + \lambda V_{ds}\right)$$

- λ = *channel length modulation coefficient*
  - not feature size
  - Empirically fit to I-V characteristics

Threshold Voltage Effects

- V_t is V_{gs} for which the channel starts to invert
- Ideal models assumed V_t is constant
- Really depends (weakly) on almost everything else:
  - Body voltage: *Body Effect*
  - Drain voltage: *Drain-Induced Barrier Lowering*
  - Channel length: *Short Channel Effect*

Body Effect

- Body is a fourth transistor terminal
- V_{sb} affects the charge required to invert the channel
  - Increasing V_s or decreasing V_b increases V_t

$$V_t = V_{t0} + \gamma\left(\sqrt{\phi_s + V_{sb}} - \sqrt{\phi_s}\right)$$

- ϕ_s = *surface potential* at threshold

$$\phi_s = 2v_T \ln\frac{N_A}{n_i}$$

  - Depends on doping level N_A
  - And intrinsic carrier concentration n_i
- γ = *body effect coefficient*

$$\gamma = \frac{t_{ox}}{\varepsilon_{ox}}\sqrt{2q\varepsilon_{si}N_A} = \frac{\sqrt{2q\varepsilon_{si}N_A}}{C_{ox}}$$

Body Effect Cont.

- For small source-to-body voltage, treat as linear

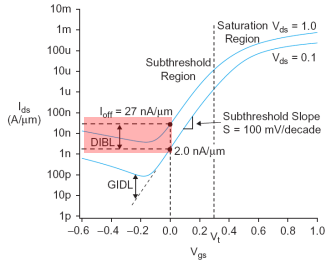
$$V_t = V_{t0} + k_\gamma V_{sb}$$

$$k_\gamma = \frac{\gamma}{2\sqrt{\phi_s}} = \frac{\sqrt{\frac{q\varepsilon_{si}N_A}{v_T \ln\frac{N_A}{n_i}}}}{2C_{ox}}$$

DIBL

- Electric field from drain affects channel
- More pronounced in small transistors where the drain is closer to the channel
- Drain-Induced Barrier Lowering
  - Drain voltage also affect V_t

$$V_t' = V_t - \eta V_{ds}$$



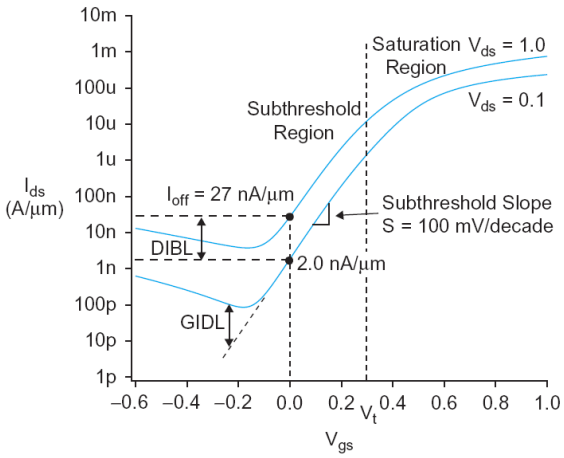
- High drain voltage causes current to _____

Short Channel Effect

- In small transistors, source/drain depletion regions extend into the channel
  - Impacts the amount of charge required to invert the channel
  - And thus makes V_t a function of channel length
- Short channel effect: V_t increases with L
  - Some processes exhibit a reverse short channel effect in which V_t decreases with L

Leakage

- What about current in cutoff?
- Simulated results
- What differs?
  –



Leakage Sources

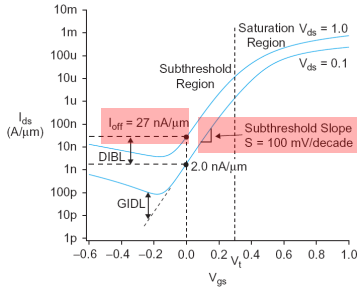
- Subthreshold conduction
  – Transistors can't abruptly turn ON or OFF
  – Dominant source in contemporary transistors
- Gate leakage
  – Tunneling through ultrathin gate dielectric
- Junction leakage
  – Reverse-biased PN junction diode current

Subthreshold Leakage

- Subthreshold leakage exponential with V_{gs}

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

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



$$I_{ds} = I_{off} 10^{\frac{V_{gs} + \eta(V_{ds} - V_{dd}) - k\gamma V_{sb}}{S}} \left(1 - e^{\frac{-V_{ds}}{v_t}}\right) \qquad S = \left[\frac{d(\log_{10} I_{ds})}{dV_{gs}}\right]^{-1} = n v_T \ln 10$$

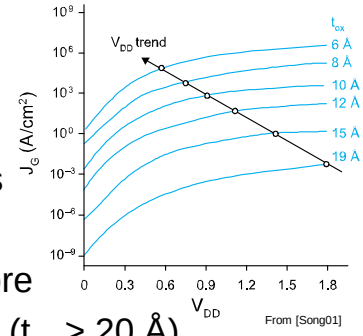
- S ≈ 100 mV/decade @ room temperature

Gate Leakage

- Carriers tunnel thorough very thin gate oxides
- Exponentially sensitive to t_{ox} and V_{DD}

$$I_{gate} = WA\left(\frac{V_{DD}}{t_{ox}}\right)^2 e^{-B\frac{t_{ox}}{V_{DD}}}$$

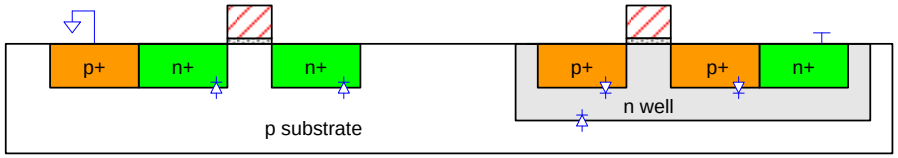
  – A and B are tech constants
  – Greater for electrons
    • So nMOS gates leak more



- Negligible for older processes (t_{ox} > 20 Å)
- Critically important at 65 nm and below (t_{ox} ≈ 10.5 Å)

Junction Leakage

- Reverse-biased p-n junctions have some leakage
  - Ordinary diode leakage
  - Band-to-band tunneling (BTBT)
  - Gate-induced drain leakage (GIDL)



Diode Leakage

- Reverse-biased p-n junctions have some leakage

$$I_D = I_S\left(e^{\frac{V_D}{v_T}} - 1\right)$$

- At any significant negative diode voltage, $I_D = -I_s$
- I_s depends on doping levels
  - And area and perimeter of diffusion regions
  - Typically < 1 fA/μm^2 (negligible)

Band-to-Band Tunneling

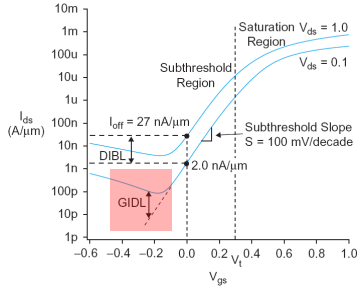
- Tunneling across heavily doped p-n junctions
  - Especially sidewall between drain & channel when *halo doping* is used to increase V_t
- Increases junction leakage to significant levels

$$I_{BTBT} = WX_j A\frac{E_j}{E_g^{0.5}}V_{dd}\, e^{-B\frac{E_g^{1.5}}{E_j}} \qquad E_j = \sqrt{\frac{2qN_{halo}N_{sd}}{\varepsilon(N_{halo}+N_{sd})}\left(V_{DD} + v_T \ln\frac{N_{halo}N_{sd}}{n_i^2}\right)}$$

  - X_j: sidewall junction depth
  - E_g: bandgap voltage
  - A, B: tech constants

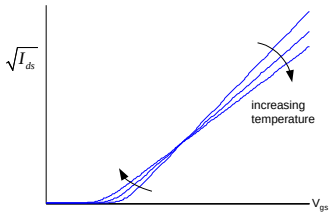
Gate-Induced Drain Leakage

- Occurs at overlap between gate and drain
  - Most pronounced when drain is at V_{DD}, gate is at a negative voltage
  - Thwarts efforts to reduce subthreshold leakage using a negative gate voltage



Temperature Sensitivity

- Increasing temperature
  - Reduces mobility
  - Reduces V_t
- I_{ON} _____ with temperature
- I_{OFF} _____ with temperature

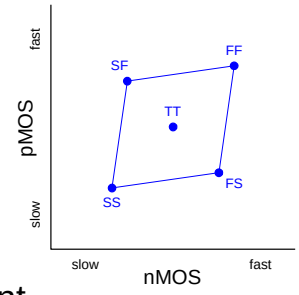


So What?

- So what if transistors are not ideal?
  - They still behave like switches.
- But these effects matter for…
  - Supply voltage choice
  - Logical effort
  - Quiescent power consumption
  - Pass transistors
  - Temperature of operation

Parameter Variation

- Transistors have uncertainty in parameters
  - Process: L_{eff}, V_t, t_{ox} of nMOS and pMOS
  - Vary around typical (T) values
- Fast (F)
  - L_{eff}: _____
  - V_t: _____
  - t_{ox}: _____
- Slow (S): opposite
- Not all parameters are independent for nMOS and pMOS



Environmental Variation

- V_{DD} and T also vary in time and space
- Fast:
  - V_{DD}: _____
  - T: _____

Corner	Voltage	Temperature
F		
T	1.8	70 C
S		

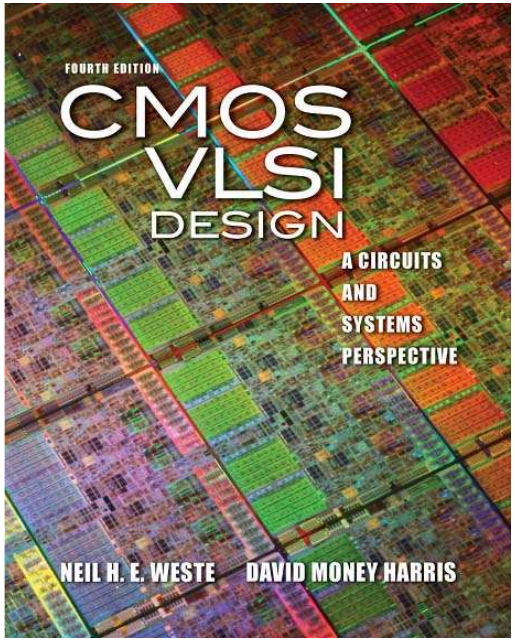
Process Corners

- Process corners describe worst case variations
  - If a design works in all corners, it will probably work for any variation.
- Describe corner with four letters (T, F, S)
  - nMOS speed
  - pMOS speed
  - Voltage
  - Temperature

Important Corners

- Some critical simulation corners include

Purpose	nMOS	pMOS	V_{DD}	Temp
Cycle time				
Power				
Subthreshold leakage				



CMOS Technology

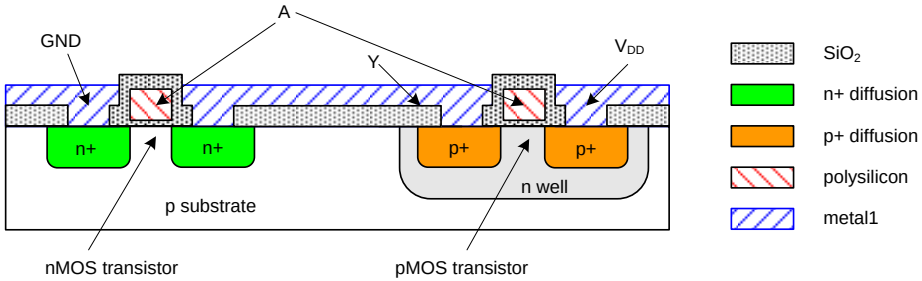
1. CMOS Fabrication
2. Layout Design Rules
3. CMOS gate design

1. CMOS Fabrication

- CMOS transistors are fabricated on silicon wafer
- Lithography process similar to printing press
- On each step, different materials are deposited or etched
- Easiest to understand by viewing both top and cross-section of wafer in a simplified manufacturing process

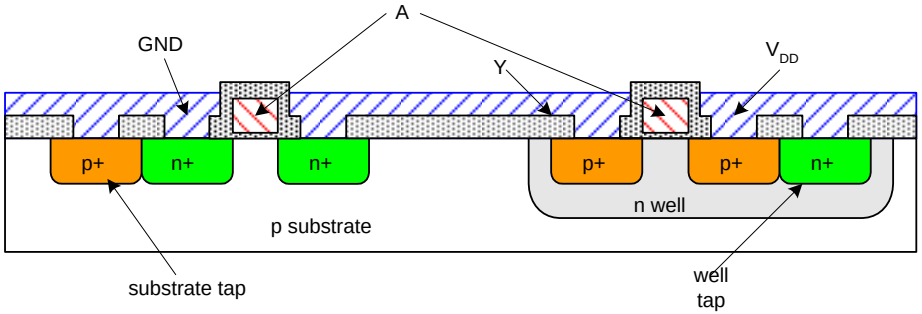
Inverter Cross-section

- Typically use p-type substrate for nMOS transistors
- Requires n-well for body of pMOS transistors



Well and Substrate Taps

- Substrate must be tied to GND and n-well to V_{DD}
- Metal to lightly-doped semiconductor forms poor connection called Schottky Diode
- Use heavily doped well and substrate contacts / taps



Well and Substrate Taps

What is Latchup: Latchup refers to short circuit formed between power and ground rails in an IC leading to high current and damage to the IC. Speaking about CMOS transistors, latch up is the phenomenon of low impedance path between power rail and ground rail due to interaction between parasitic pnp and npn transistors. The structure formed by these resembles a Silicon Controlled rectifier (SCR, usually known as a thyristor, a PNPN device used in power electronics). These form a +ve feedback loop, short circuit the power rail and ground rail, which eventually causes excessive current, and can even permanently damage the device.

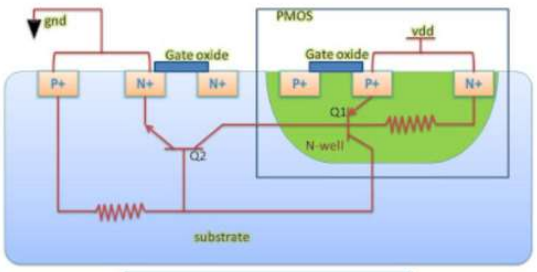
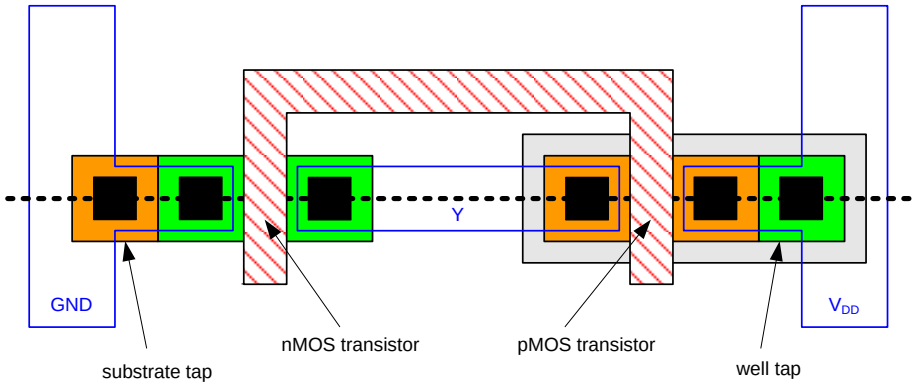


Figure 1 : Latchup formation in a CMOS device

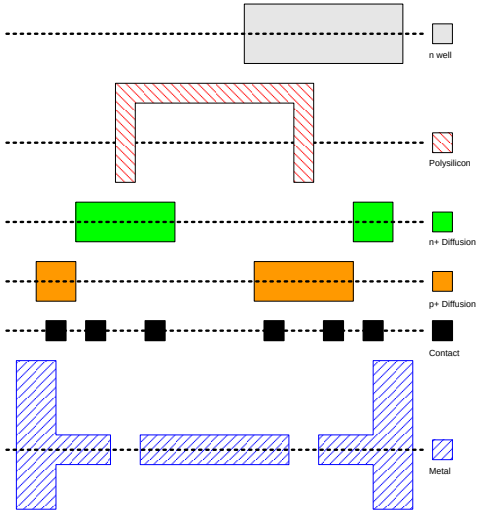
Inverter Mask Set

- Transistors and wires are defined by *masks*
- Cross-section taken along dashed line



Detailed Mask Views

- Six masks
  - n-well
  - Polysilicon
  - n+ diffusion
  - p+ diffusion
  - Contact
  - Metal



Fabrication

- Chips are built in huge factories called fabs
- Contain clean rooms as large as football fields

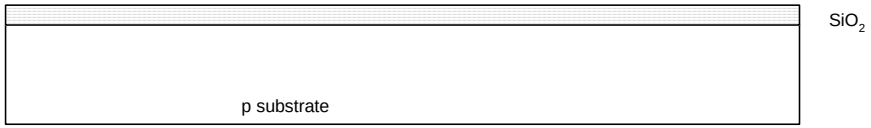
Courtesy of International Business Machines Corporation. Unauthorized use not permitted.

Fabrication Steps

- Start with blank wafer
- Build inverter from the bottom up
- First step will be to form the n-well
  - Cover wafer with protective layer of SiO_2 (oxide)
  - Remove layer where n-well should be built
  - Implant or diffuse n dopants into exposed wafer
  - Strip off SiO_2

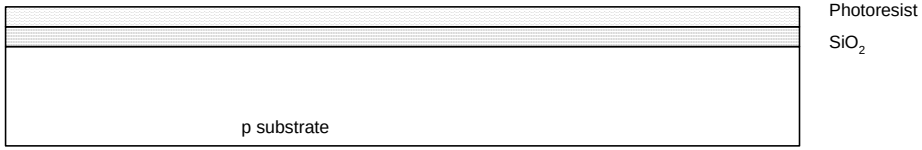
Oxidation

- Grow SiO_2 on top of Si wafer
  - 900 – 1200 C with H_2O or O_2 in oxidation furnace



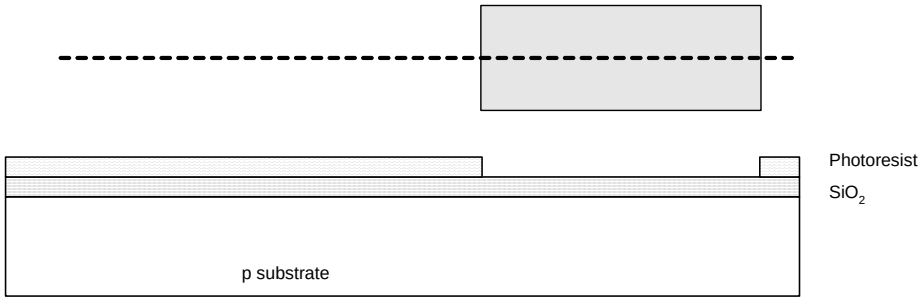
Photoresist

- Spin on photoresist
  - Photoresist is a light-sensitive organic polymer
  - Softens where exposed to light



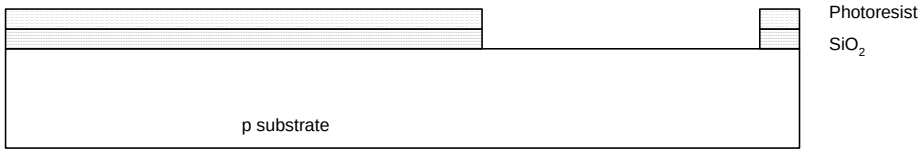
Lithography

- Expose photoresist through n-well mask
- Strip off exposed photoresist



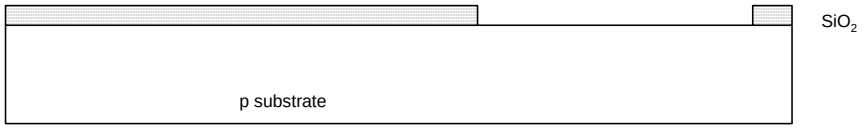
Etch

- Etch oxide with hydrofluoric acid (HF)
  - Seeps through skin and eats bone; nasty stuff!!!
- Only attacks oxide where resist has been exposed



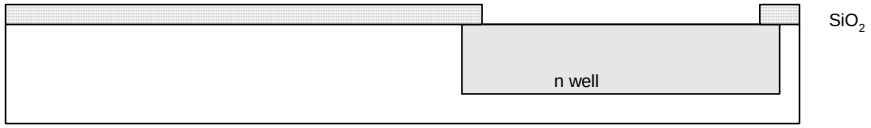
Strip Photoresist

- Strip off remaining photoresist
  - Use mixture of acids called piranah etch
- Necessary so resist doesn't melt in next step



n-well

- n-well is formed with diffusion or ion implantation
- Diffusion
  - Place wafer in furnace with arsenic gas
  - Heat until As atoms diffuse into exposed Si
- Ion Implanatation
  - Blast wafer with beam of As ions
  - Ions blocked by SiO_2, only enter exposed Si

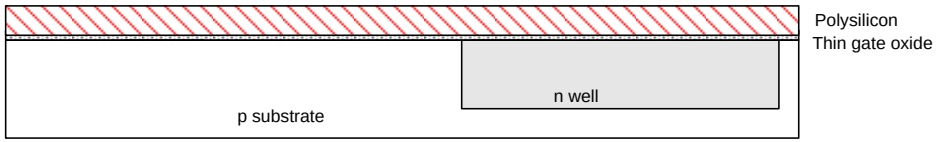


Strip Oxide

- Strip off the remaining oxide using HF
- Back to bare wafer with n-well
- Subsequent steps involve similar series of steps

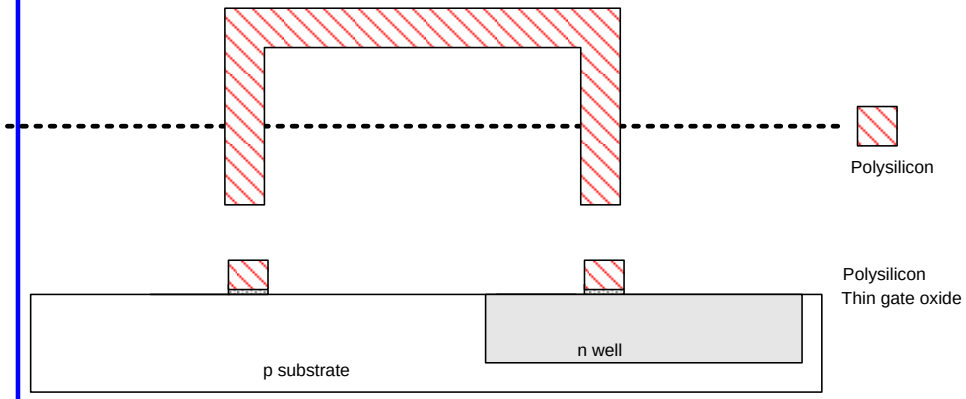
Polysilicon

- Deposit very thin layer of gate oxide
  - < 20 Å (6-7 atomic layers)
- Chemical Vapor Deposition (CVD) of silicon layer
  - Place wafer in furnace with Silane gas (SiH_4)
  - Forms many small crystals called polysilicon
  - Heavily doped to be good conductor



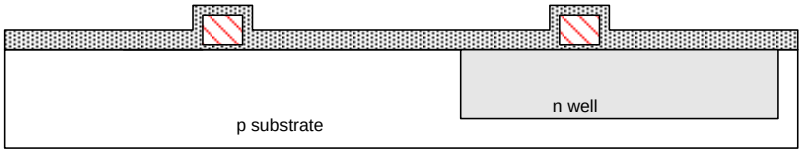
Polysilicon Patterning

- Use same lithography process to pattern polysilicon



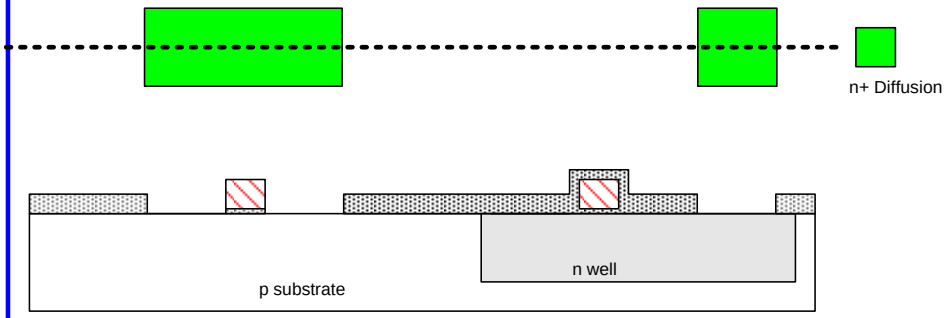
Self-Aligned Process

- Use oxide and masking to expose where n+ dopants should be diffused or implanted
- N-diffusion forms nMOS source, drain, and n-well contact



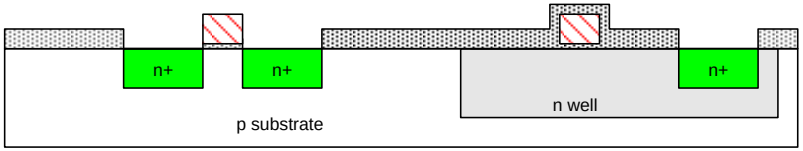
N-diffusion

- Pattern oxide and form n+ regions
- *Self-aligned process* where gate blocks diffusion
- Polysilicon is better than metal for self-aligned gates because it doesn't melt during later processing



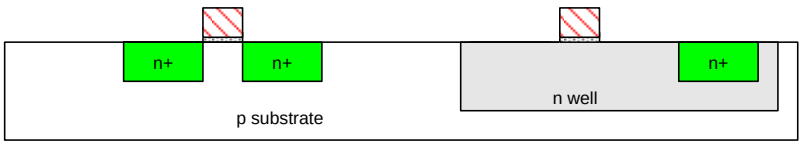
N-diffusion cont.

- Historically dopants were diffused
- Usually ion implantation today
- But regions are still called diffusion



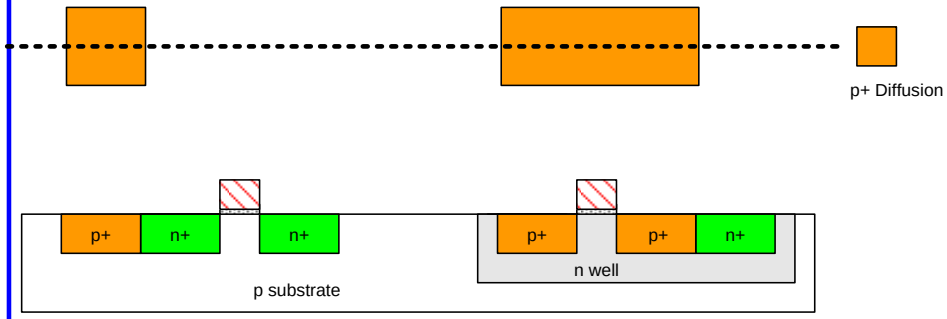
N-diffusion cont.

- Strip off oxide to complete patterning step



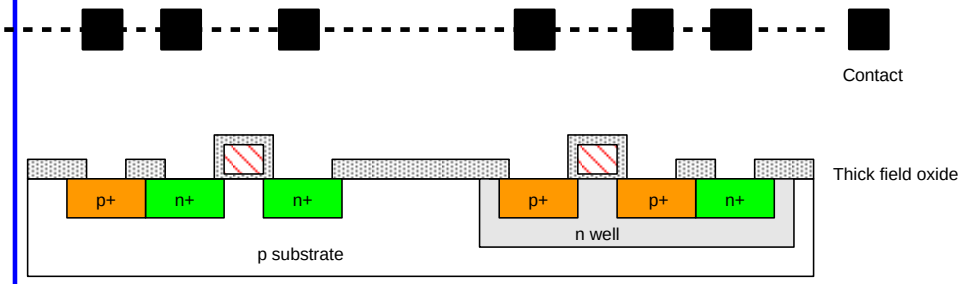
P-Diffusion

- Similar set of steps form p+ diffusion regions for pMOS source and drain and substrate contact



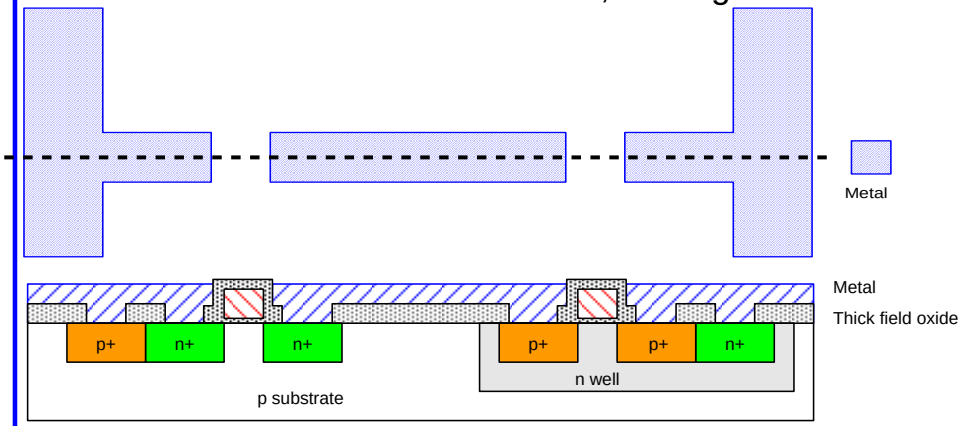
Contacts

- Now we need to wire together the devices
- Cover chip with thick field oxide
- Etch oxide where contact cuts are needed



Metalization

- Sputter on aluminum over whole wafer
- Pattern to remove excess metal, leaving wires



2. Layout Design Rules

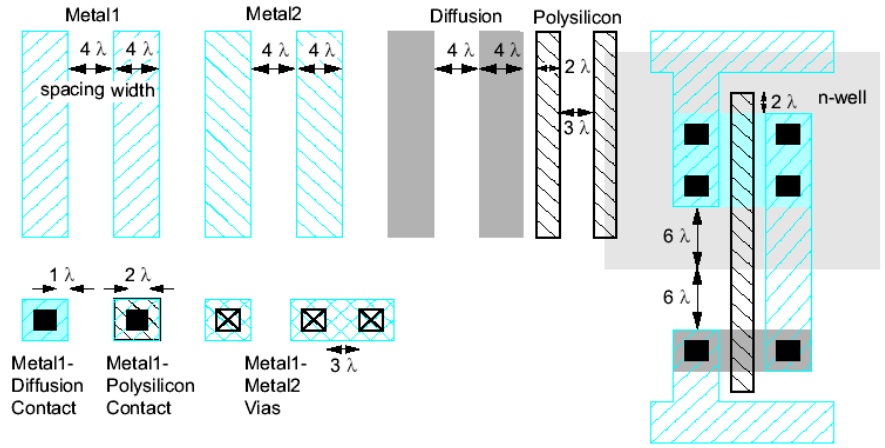
- **Describes actual layers and geometry on the silicon substrate to implement a function**
- **Need to define transistors, interconnection**
  - **Transistor widths (for performance)**
  - **Spacing, interconnect widths, to reduce defects, satisfy power requirements**
  - **Contacts (between poly or active and metal), and vias (between metal layers)**
  - **Wells and their contacts (to power or ground)**
- **Layout of lower--level cells constrained by higher-level requirements:**
  - **"floorplanning"**

2. Layout Design Rules

- Chips are specified with set of masks
- Minimum dimensions of masks determine transistor size (and hence speed, cost, and power)
- Feature size *f* = distance between source and drain
  - Set by minimum width of polysilicon
- Feature size improves 30% every 3 years or so
- Normalize for feature size when describing design rules
- Express rules in terms of $\lambda = f/2$
  - E.g. λ = 0.3 μm in 0.6 μm process

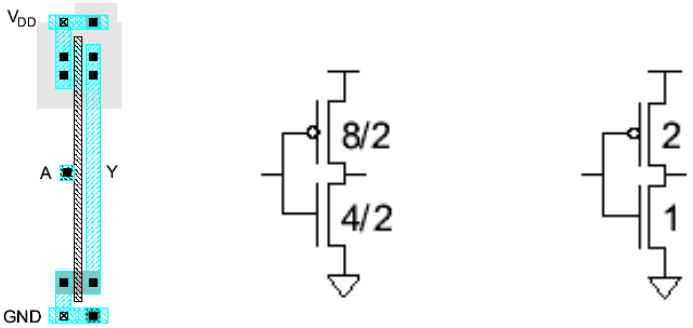
Simplified Design Rules

- Conservative rules to get you started



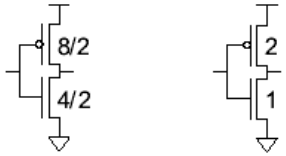
Inverter Layout

- Transistor dimensions specified as Width / Length
  - Minimum size is 4λ / 2λ, sometimes called 1 unit
  - In *f* = 0.6 μm process, this is 1.2 μm wide, 0.6 μm long



Review

1. Which material is for a blank wafer?
2. How many masks is for a basic CMOS inverter?
3. How to create SiO_2 layer on the wafer?
4. How to etch the SiO_2 layer?
5. How to form the n-well on the p substrate?
6. What is self-aligned process?
7. What is feature size? What is λ?
8. What are distances between two metals, two polysilicons, two metal vias?
9. Explain the following picture:

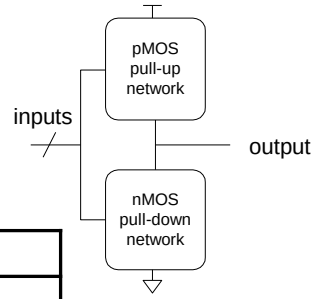


CMOS Gate Design

- Activity:
  - Sketch a 4-input CMOS NOR gate

Complementary CMOS

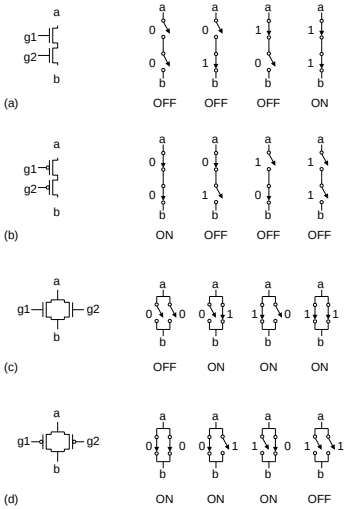
- Complementary CMOS logic gates
  - nMOS *pull-down network*
  - pMOS *pull-up network*
  - a.k.a. static CMOS



	Pull-up OFF	Pull-up ON
Pull-down OFF	Z (float)	1
Pull-down ON	0	X (crowbar)

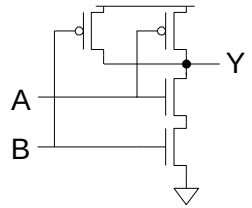
Series and Parallel

- nMOS: 1 = ON
- pMOS: 0 = ON
- *Series*: both must be ON
- *Parallel*: either can be ON



Conduction Complement

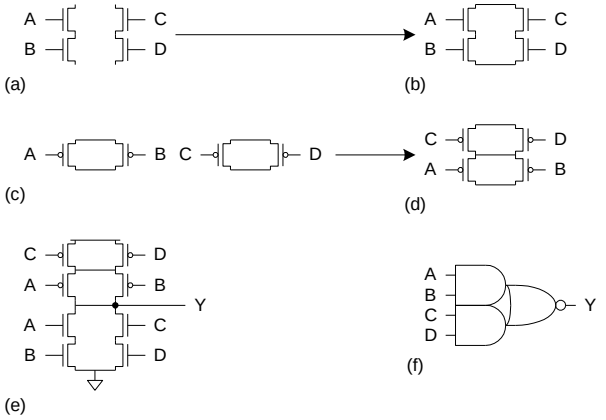
- Complementary CMOS gates always produce 0 or 1
- Ex: NAND gate
  - Series nMOS: Y=0 when both inputs are 1
  - Thus Y=1 when either input is 0
  - Requires parallel pMOS



- Rule of *Conduction Complements*
  - Pull-up network is complement of pull-down
  - Parallel -> series, series -> parallel

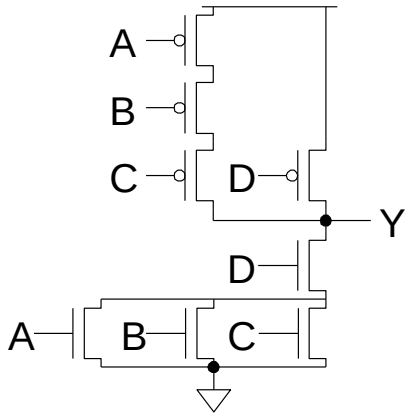
Compound Gates

- *Compound gates* can do any inverting function
- Ex: $Y = \overline{AB + CD}$ $(AND - AND - OR - INVERT, AOI22)$



Example: O3AI

- $Y = \overline{(A + B + C)D}$

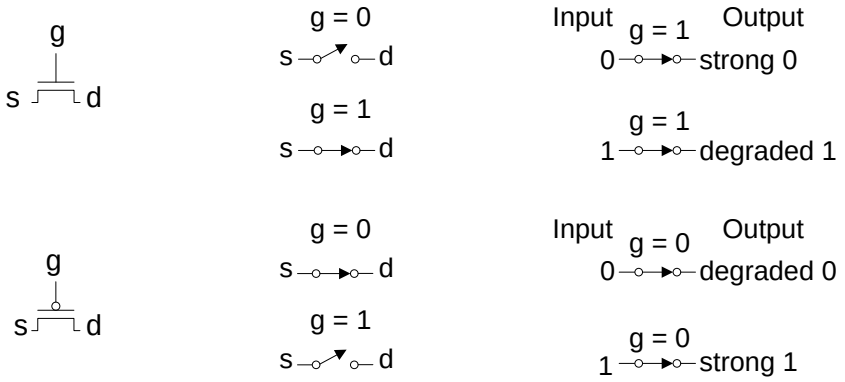


Signal Strength

- *Strength* of signal
  - How close it approximates ideal voltage source
- V_{DD} and GND rails are strongest 1 and 0
- nMOS pass strong 0
  - But degraded or weak 1
- pMOS pass strong 1
  - But degraded or weak 0
- Thus nMOS are best for pull-down network

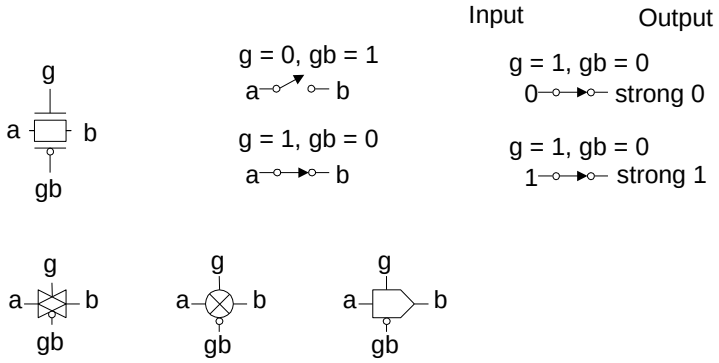
Pass Transistors

- Transistors can be used as switches



Transmission Gates

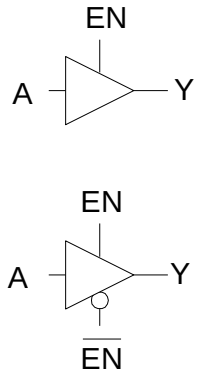
- Pass transistors produce degraded outputs
- *Transmission gates* pass both 0 and 1 well



Tristates

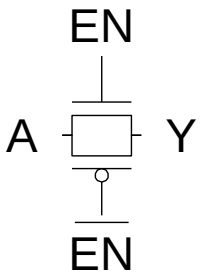
- *Tristate buffer* produces Z when not enabled

EN	A	Y
0	0	
0	1	
1	0	
1	1	



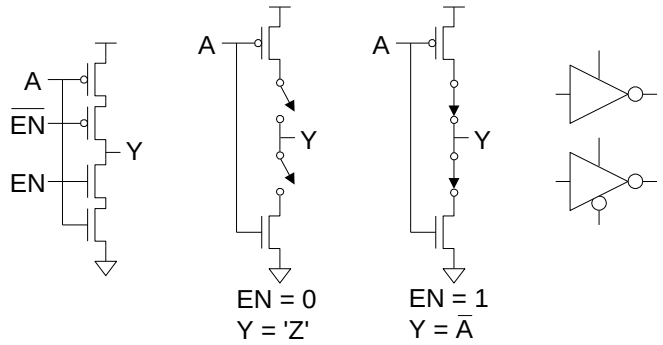
Nonrestoring Tristate

- Transmission gate acts as tristate buffer
  - Only two transistors
  - But *nonrestoring*
    - Noise on A is passed on to Y



Tristate Inverter

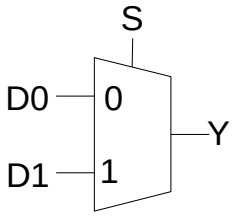
- Tristate inverter produces restored output
  - Violates conduction complement rule
  - Because we want a Z output



Multiplexers

- 2:1 multiplexer chooses between two inputs

S	D1	D0	Y
0	X	0	
0	X	1	
1	0	X	
1	1	X	

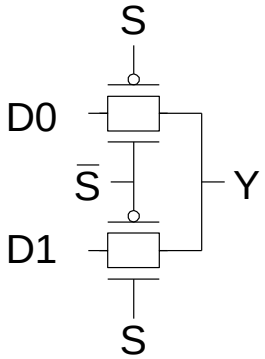


Gate-Level Mux Design

- $Y = SD_1 + \overline{S}D_0$ (too many transistors)
- How many transistors are needed?

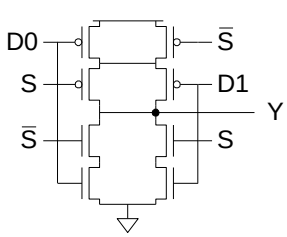
Transmission Gate Mux

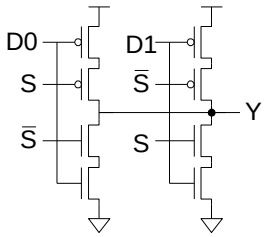
- Nonrestoring mux uses two transmission gates
  - Only 4 transistors

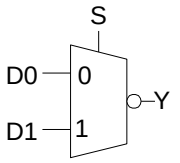


Inverting Mux

- Inverting multiplexer
  - Use compound AOI22
  - Or pair of tristate inverters
  - Essentially the same thing
- Noninverting multiplexer adds an inverter

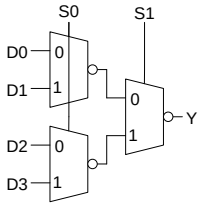


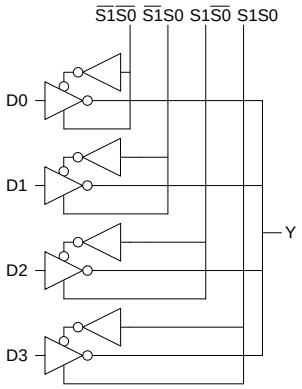




4:1 Multiplexer

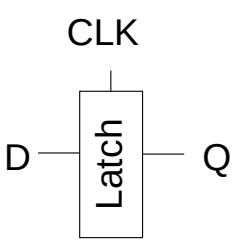
- 4:1 mux chooses one of 4 inputs using two selects
  - Two levels of 2:1 muxes
  - Or four tristates

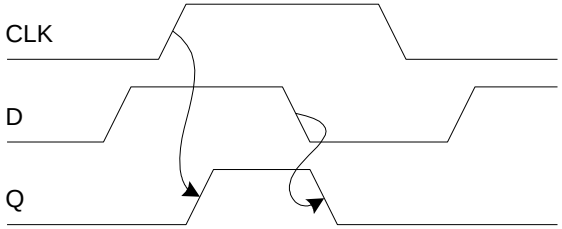




D Latch

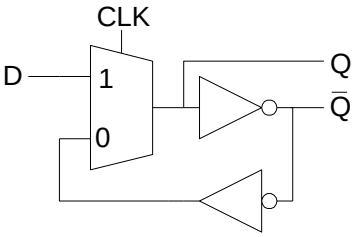
- When CLK = 1, latch is *transparent*
  - D flows through to Q like a buffer
- When CLK = 0, the latch is *opaque*
  - Q holds its old value independent of D
- a.k.a. *transparent latch* or *level-sensitive latch*

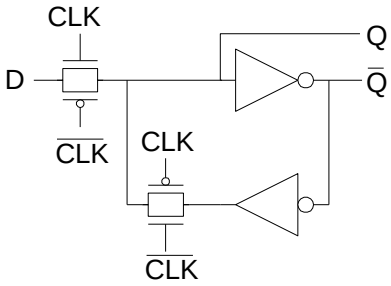




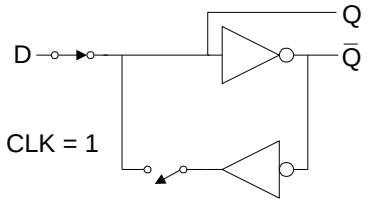
D Latch Design

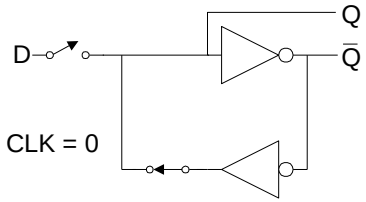
- Multiplexer chooses D or old Q

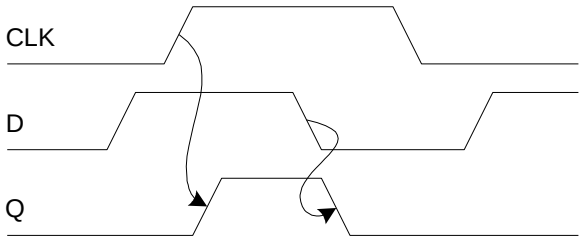




D Latch Operation

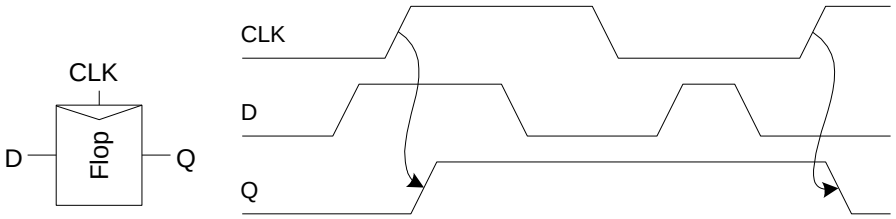






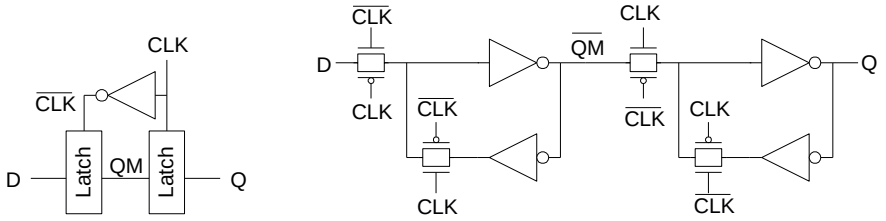
D Flip-flop

- When CLK rises, D is copied to Q
- At all other times, Q holds its value
- a.k.a. *positive edge-triggered flip-flop*, *master-slave flip-flop*

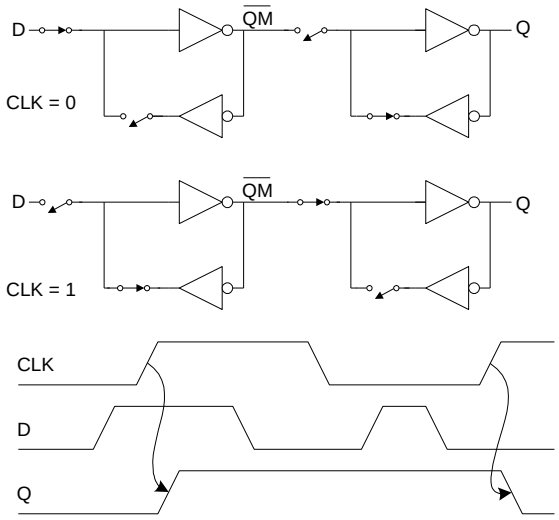


D Flip-flop Design

- Built from master and slave D latches

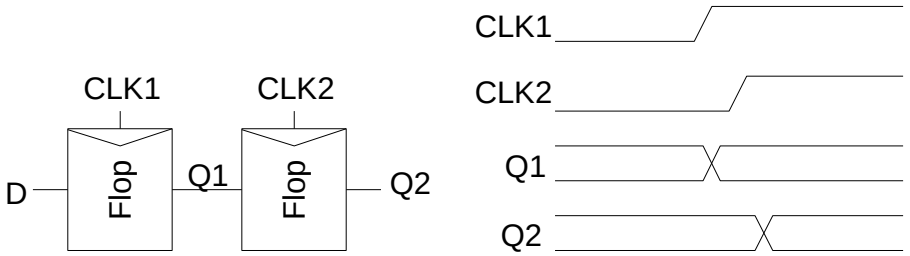


D Flip-flop Operation



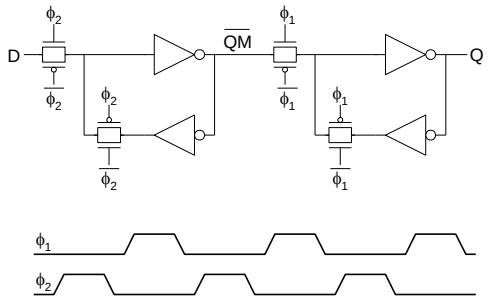
Race Condition

- Back-to-back flops can malfunction from clock skew
  - Second flip-flop fires late
  - Sees first flip-flop change and captures its result
  - Called *hold-time failure* or *race condition*



Nonoverlapping Clocks

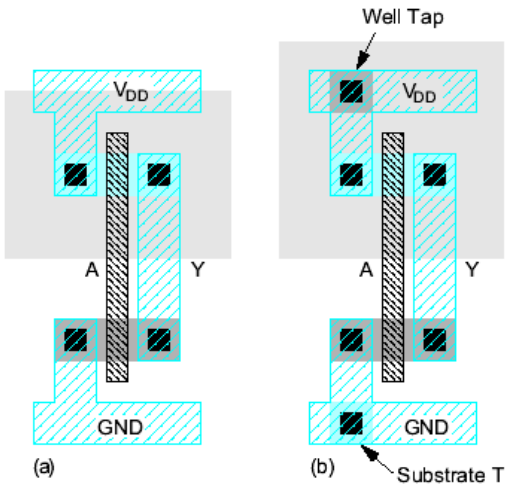
- Nonoverlapping clocks can prevent races
  - As long as nonoverlap exceeds clock skew
- We will use them in this class for safe design
  - Industry manages skew more carefully instead



Gate Layout

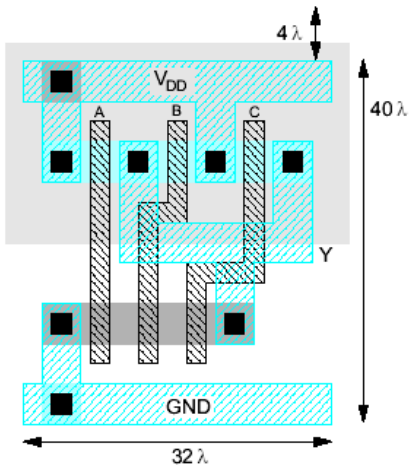
- Layout can be very time consuming
  - Design gates to fit together nicely
  - Build a library of standard cells
- Standard cell design methodology
  - V_{DD} and GND should abut (standard height)
  - Adjacent gates should satisfy design rules
  - nMOS at bottom and pMOS at top
  - All gates include well and substrate contacts

Example: Inverter



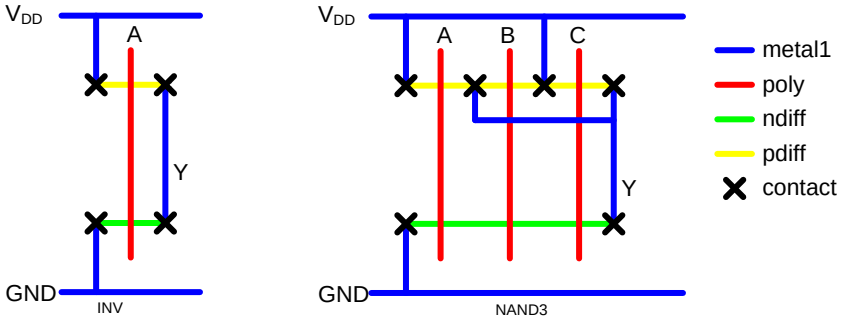
Example: NAND3

- Horizontal N-diffusion and p-diffusion strips
- Vertical polysilicon gates
- Metal1 V_{DD} rail at top
- Metal1 GND rail at bottom
- 32 λ by 40 λ



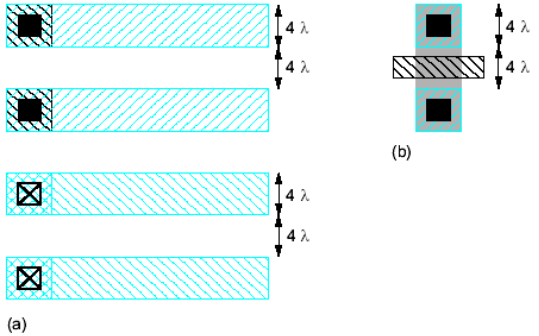
Stick Diagrams

- *Stick diagrams* help plan layout quickly
  - Need not be to scale
  - Draw with color pencils or dry-erase markers



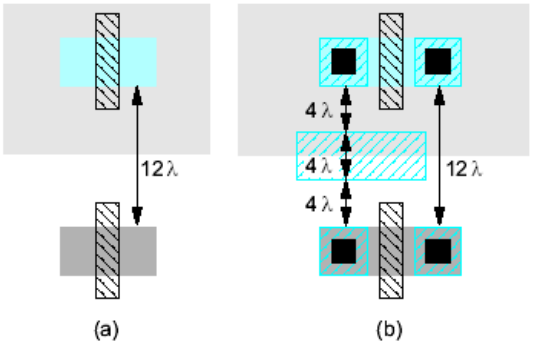
Wiring Tracks

- A *wiring track* is the space required for a wire
  - 4 λ width, 4 λ spacing from neighbor = 8 λ pitch
- Transistors also consume one wiring track



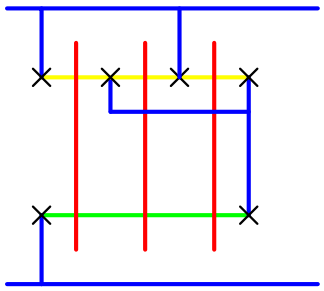
Well spacing

- Wells must surround transistors by 6 λ
  - Implies 12 λ between opposite transistor flavors
  - Leaves room for one wire track



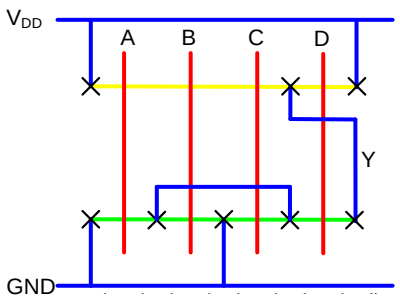
Area Estimation

- Estimate area by counting wiring tracks
  - Multiply by 8 to express in λ



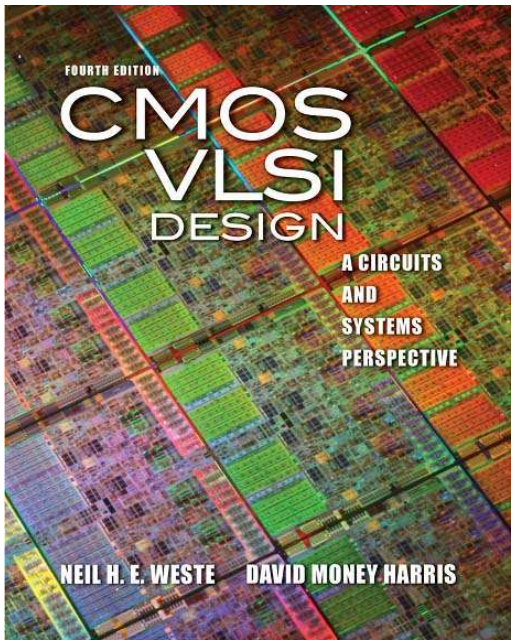
Example: O3AI

- Sketch a stick diagram for O3AI and estimate area
  - $Y = \overline{(A+B+C)D}$



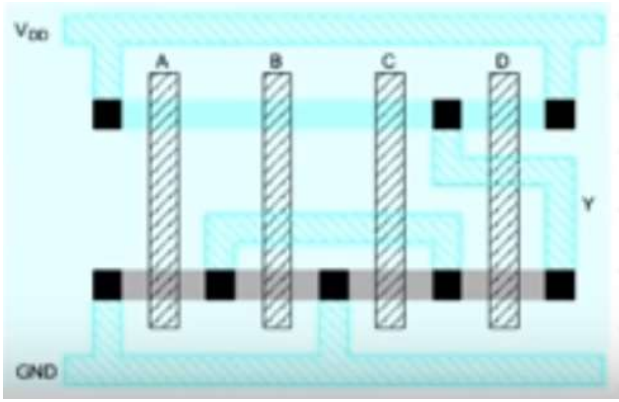
Review

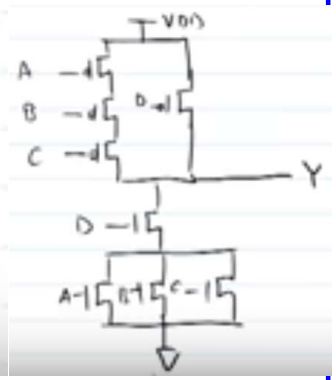
1. Draw the transistor-level schematics of the following compound gates:
   1. Y = $\overline{A + B.C.D}$
   2. Y = $\overline{(A+B)C + DE}$
2. Draw the transistor-level schematics of the multiplexer 4:1
3. Design the transistor-level schematics for SR flip-flop
4. Draw stick diagrams and calculate areas for the following gates:
   1. OAI22
   2. AOI31



Lecture 4: Delay

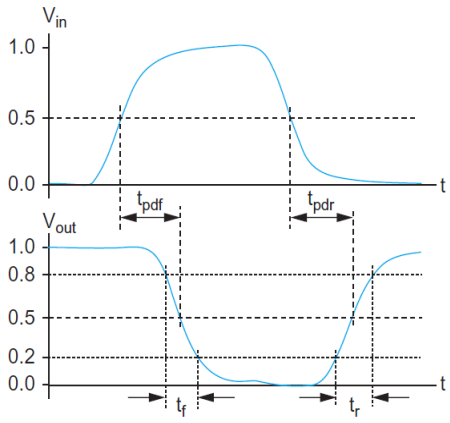
1. Delay definition
2. Transient response
3. RC delay models
4. Linear delay models





1. Delay Definitions

- t_{pdr}: *rising propagation delay*
  - From input to rising output crossing $V_{DD}/2$
- t_{pdf}: *falling propagation delay*
  - From input to falling output crossing $V_{DD}/2$
- t_{pd}: *average propagation delay*
  - $t_{pd} = (t_{pdr} + t_{pdf})/2$
- t_r: *rise time*
  - From output crossing 0.2 V_{DD} to 0.8 V_{DD}
- t_f: *fall time*
  - From output crossing 0.8 V_{DD} to 0.2 V_{DD}

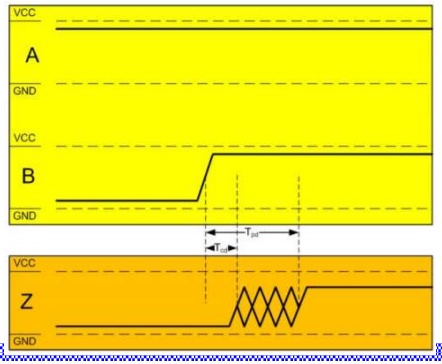


1. Delay Definitions

- t_{cdr}: *rising contamination delay*
  - From input to rising output crossing $V_{DD}/2$
- t_{cdf}: *falling contamination delay*
  - From input to falling output crossing $V_{DD}/2$
- t_{cd}: *average contamination delay*
  - $t_{cd} = (t_{cdr} + t_{cdf})/2$

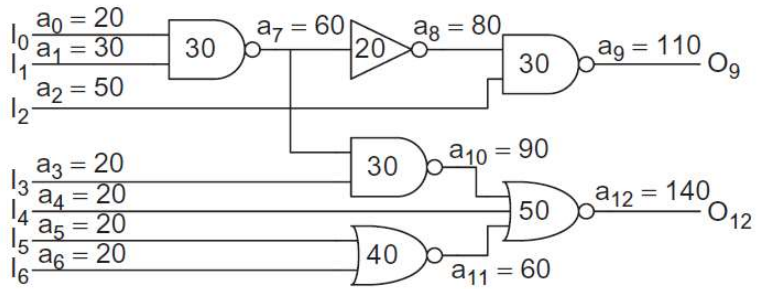
1. Delay Definitions

The contamination delay, t_{cd}, is the fastest that the logic gate will change output based on a changed input.
The propagation delay, t_{pd}, is the slowest that the logic gate will change output based on a changed input.



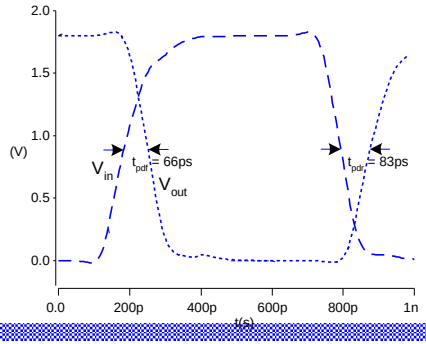
Arrival time

- Arrival time is the latest time at which each node in a block of logic will switch
- The *slack* is the difference between the required and arrival times.
- *Positive slack* means that the circuit meets timing.
- *Negative slack* means that the circuit is not fast enough.



Simulated Inverter Delay

- Solving differential equations by hand is too hard
- SPICE simulator solves the equations numerically
  - Uses more accurate I-V models too!
- But simulations take time to write, may hide insight



Delay Estimation

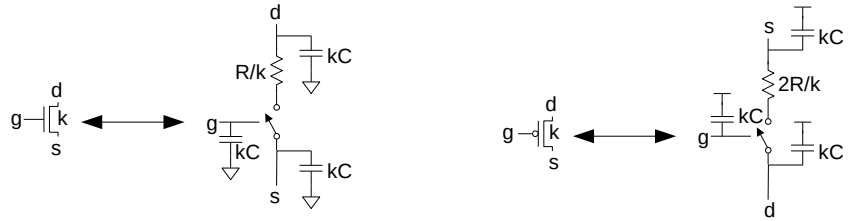
- We would like to be able to easily estimate delay
  - Not as accurate as simulation
  - But easier to ask "What if?"
- The step response usually looks like a 1st order RC response with a decaying exponential.
- Use RC delay models to estimate delay
  - C = total capacitance on output node
  - Use *effective resistance* R
  - So that t_{pd} = RC
- Characterize transistors by finding their effective R
  - Depends on average current as gate switches

Effective Resistance

- Shockley models have limited value
  - Not accurate enough for modern transistors
  - Too complicated for much hand analysis
- Simplification: treat transistor as resistor
  - Replace $I_{ds}(V_{ds}, V_{gs})$ with effective resistance R
    - $I_{ds} = V_{ds}/R$
  - R averaged across switching of digital gate
- Too inaccurate to predict current at any given time
  - But good enough to predict RC delay

3. RC Delay Model

- Use equivalent circuits for MOS transistors
  - Ideal switch + capacitance and ON resistance
  - Unit nMOS has resistance R, capacitance C
  - Unit pMOS has resistance 2R, capacitance C
- Capacitance proportional to width
- Resistance inversely proportional to width

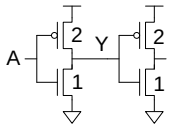


RC Values

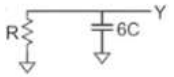
- Capacitance
  - $C = C_g = C_s = C_d = 2$ fF/μm of gate width in 0.6 μm
  - Gradually decline to 1 fF/μm in nanometer techs.
- Resistance
  - $R \approx 6$ KΩ*μm in 0.6 μm process
  - Improves with shorter channel lengths
- Unit transistors
  - May refer to minimum contacted device (4/2 λ)
  - Or maybe 1 μm wide device
  - Doesn't matter as long as you are consistent

Inverter Delay Estimate

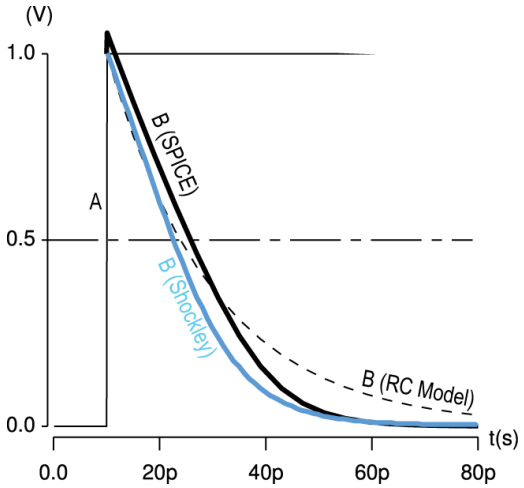
- Estimate the delay of a fanout-of-1 inverter



d = 6RC

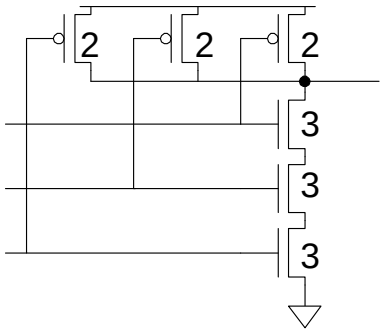


Delay Model Comparison



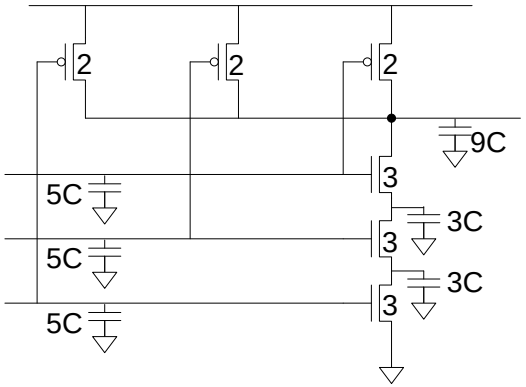
Example: 3-input NAND

- Sketch a 3-input NAND with transistor widths chosen to achieve effective rise and fall resistances equal to a unit inverter (R).



3-input NAND Caps

- Annotate the 3-input NAND gate with gate and diffusion capacitance.

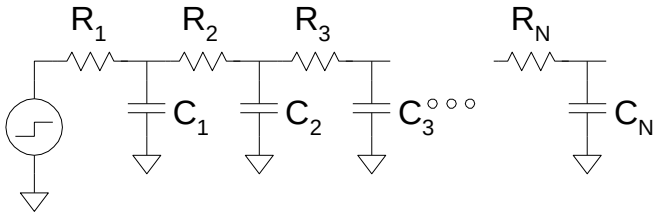


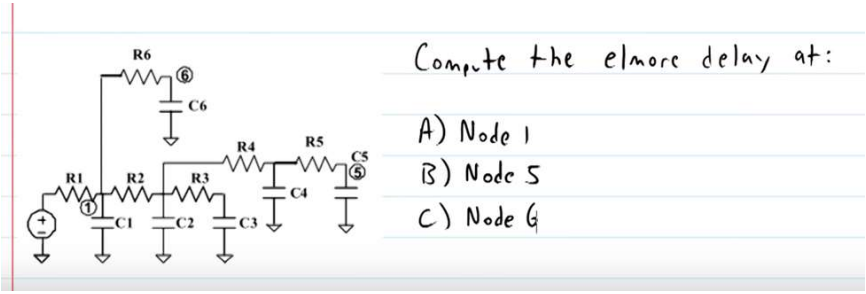
Elmore Delay

- ON transistors look like resistors
- Pullup or pulldown network modeled as *RC ladder*
- Elmore delay of RC ladder

$$t_{pd} \approx \sum_{\text{nodes } i} R_{i-to-source} C_i$$

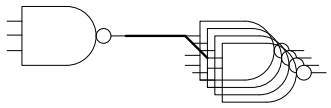
$$= R_1 C_1 + (R_1 + R_2) C_2 + \ldots + (R_1 + R_2 + \ldots + R_N) C_N$$

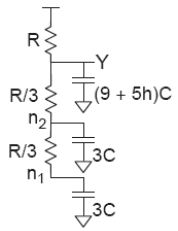




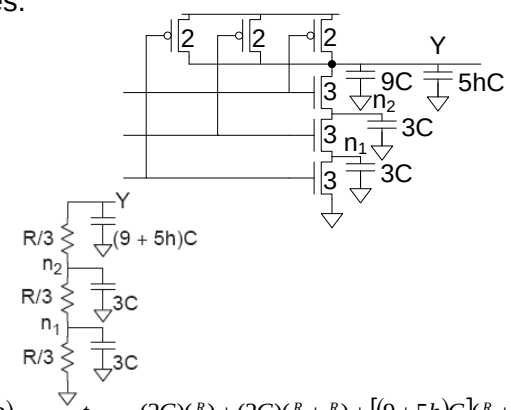
Example: 3-input NAND

- Estimate worst-case rising and falling delay of 3-input NAND driving *h* identical gates.





$$t_{pdr} = [(9+5h)C](R) + (3C)(R) + (3C)(R)$$
$$= (15+5h)RC$$



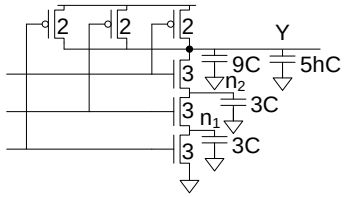
$$t_{pdf} = (3C)(\tfrac{R}{3}) + (3C)(\tfrac{R}{3} + \tfrac{R}{3}) + [(9+5h)C](\tfrac{R}{3} + \tfrac{R}{3} + \tfrac{R}{3})$$
$$= (12+5h)RC$$

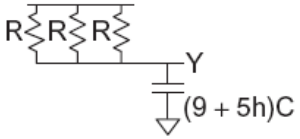
Delay Components

- Delay has two parts
  - *Parasitic delay*
    - 15 or 12 RC
    - Independent of load
  - *Effort delay*
    - 5h RC
    - Proportional to load capacitance

Contamination Delay

- Best-case (contamination) delay can be substantially less than propagation delay.
- Ex: If all three inputs fall simultaneously

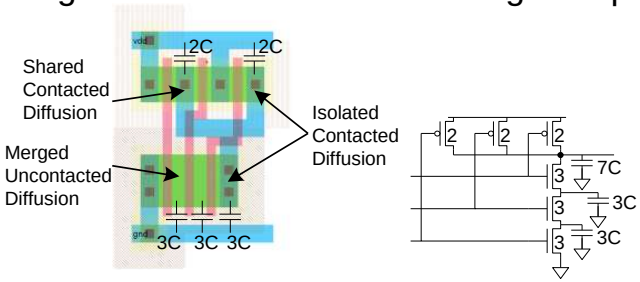




$$t_{cdr} = \left[(9+5h)C\right]\left(\frac{R}{3}\right) = \left(3+\frac{5}{3}h\right)RC$$

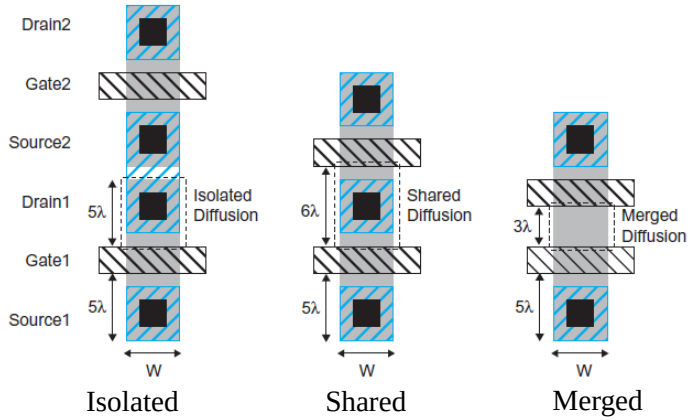
Diffusion Capacitance

- We assumed contacted diffusion on every s / d.
- Good layout minimizes diffusion area
- Ex: NAND3 layout shares one diffusion contact
  - Reduces output capacitance by 2C
  - Merged uncontacted diffusion might help too



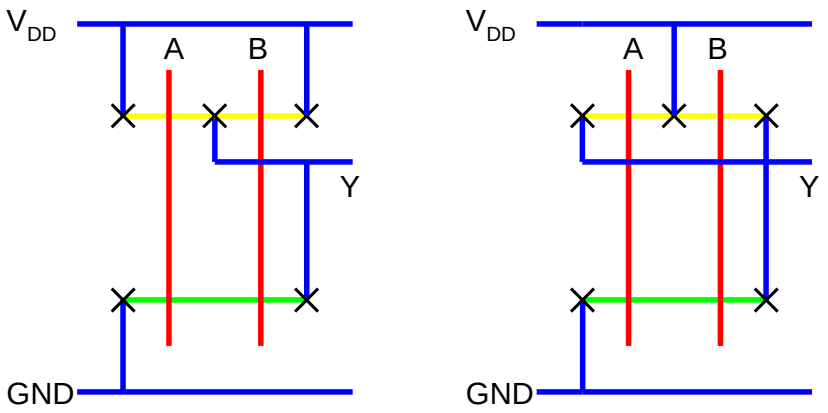
Isolated/Shared/Merged Diffusion

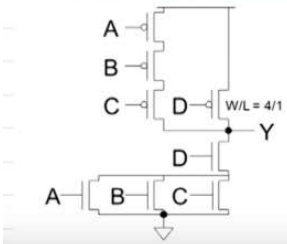
- Shared contacted diffusion can reduce the diffusion capacitance
- Un-contacted diffusion nodes can reduce more capacitance



Layout Comparison

- Which layout is better?



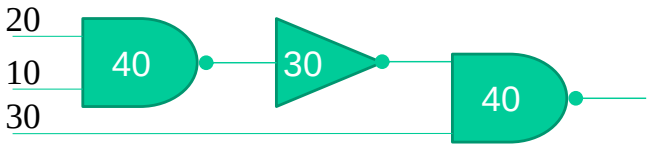


A) Size the transistors so the pull-up and pull-down resistances are matched

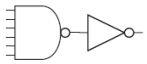
B) Find the input patterns that produce the longest pull-up and pull-down propagation delays

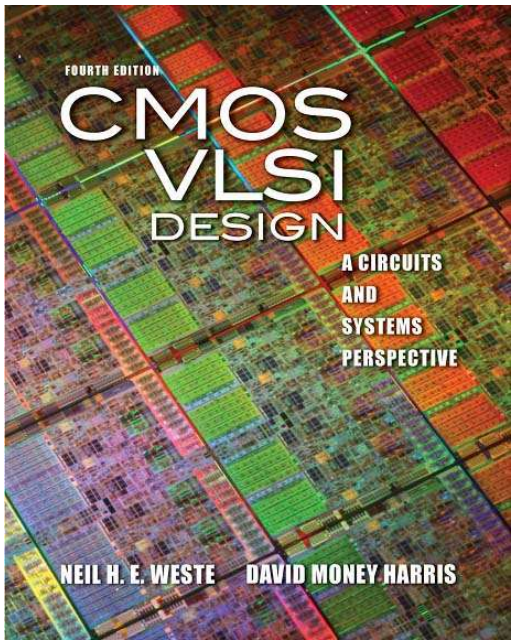
Review

1. What are t_{pdr}, t_{pdf}, t_f, t_r, t_{cdr}, t_{cdf}?
2. Calculate arrive time of the following circuit:



3. Explain the delay estimation of a fanout-of-1 inverter (slide 10)
4. Explain the t_{pdr} and t_{pdf} delay estimation of 3-input NAND driving h identical gates (slide 15).
5. Estimate delay for the gates: AOI21, OAI31
6. What is logical effort?
7. What is parasitic delay?
8. Estimate the delay of the following gate:





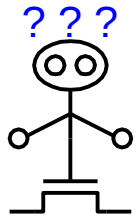
Lecture 6: Logical Effort

Outline

- Logical Effort
- Delay in a Logic Gate
- Multistage Logic Networks
- Choosing the Best Number of Stages
- Example
- Summary

Introduction

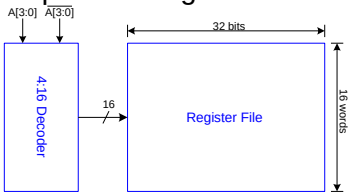
- Chip designers face a bewildering array of choices
  - What is the best circuit topology for a function?
  - How many stages of logic give least delay?
  - How wide should the transistors be?



- Logical effort is a method to make these decisions
  - Uses a simple model of delay
  - Allows back-of-the-envelope calculations
  - Helps make rapid comparisons between alternatives
  - Emphasizes remarkable symmetries

Example

- Ben Bitdiddle is the memory designer for the Motoroil 68W86, an embedded automotive processor. Help Ben design the decoder for a register file.



- Decoder specifications:
  - 16 word register file
  - Each word is 32 bits wide
  - Each bit presents load of 3 unit-sized transistors
  - True and complementary address inputs A[3:0]
  - Each input may drive 10 unit-sized transistors
- Ben needs to decide:
  - How many stages to use?
  - How large should each gate be?
  - How fast can decoder operate?

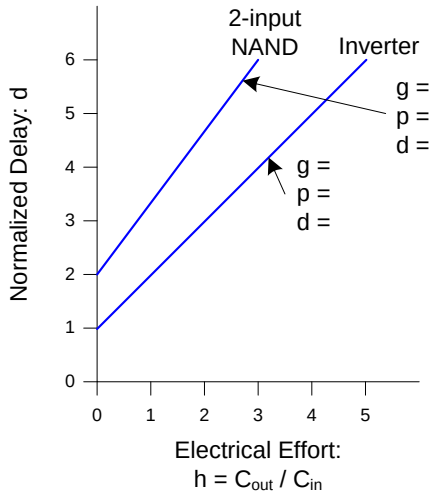
Delay in a Logic Gate

- Express delays in process-independent unit $d = \frac{d_{abs}}{\tau}$
- Delay has two components: d = f + p
- τ = 3RC
  - ≈ 3 ps in 65 nm process
  - 60 ps in 0.6 μm process
- *f*: *effort delay* = *gh* (a.k.a. stage effort)
  - Again has two components
- *g*: *logical effort*
  - Measures relative ability of gate to deliver current
  - $g \equiv 1$ for inverter
- *h*: *electrical effort* = C_{out} / C_{in}
  - Ratio of output to input capacitance
  - Sometimes called fanout
- *p*: parasitic delay
  - Represents delay of gate driving no load
  - Set by internal parasitic capacitance

Delay Plots

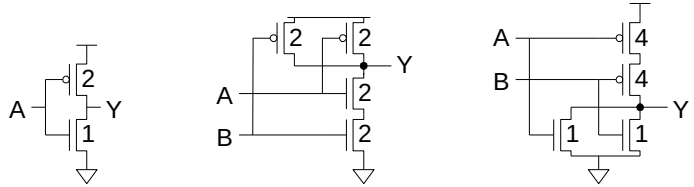
$d = f + p$

$= gh + p$



Computing Logical Effort

- DEF: *Logical effort is the ratio of the input capacitance of a gate to the input capacitance of an inverter delivering the same output current*.
- Measure from delay vs. fanout plots
- Or estimate by counting transistor widths



$C_{in} = 3$
$g = 3/3$

$C_{in} = 4$
$g = 4/3$

$C_{in} = 5$
$g = 5/3$

Catalog of Gates

- Logical effort of common gates

Gate type	Number of inputs				
	1	2	3	4	n
Inverter	1				
NAND		4/3	5/3	6/3	(n+2)/3
NOR		5/3	7/3	9/3	(2n+1)/3
Tristate / mux	2	2	2	2	2
XOR, XNOR		4, 4	6, 12, 6	8, 16, 16, 8	

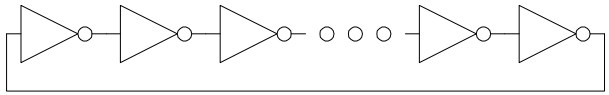
Catalog of Gates

- Parasitic delay of common gates
  - In multiples of p_{inv} (≈ 1)

Gate type	Number of inputs				
	1	2	3	4	n
Inverter	1				
NAND		2	3	4	n
NOR		2	3	4	n
Tristate / mux	2	4	6	8	2n
XOR, XNOR		4	6	8	

Example: Ring Oscillator

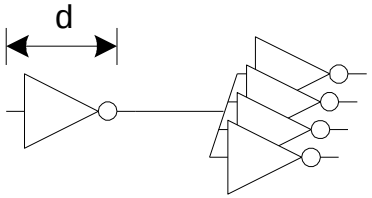
- Estimate the frequency of an N-stage ring oscillator



Logical Effort: g =
Electrical Effort: h =
Parasitic Delay: p =
Stage Delay: d =
Frequency: f_{osc} =

Example: FO4 Inverter

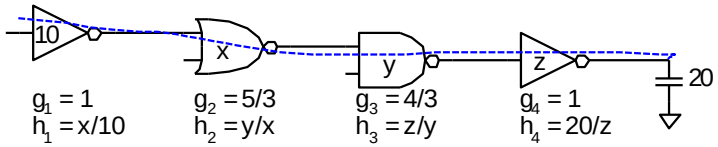
- Estimate the delay of a fanout-of-4 (FO4) inverter



Logical Effort: g =
Electrical Effort: h =
Parasitic Delay: p =
Stage Delay: d =

Multistage Logic Networks

- Logical effort generalizes to multistage networks
- *Path Logical Effort* $G = \prod g_i$
- *Path Electrical Effort* $H = \frac{C_{\text{out-path}}}{C_{\text{in-path}}}$
- *Path Effort* $F = \prod f_i = \prod g_i h_i$



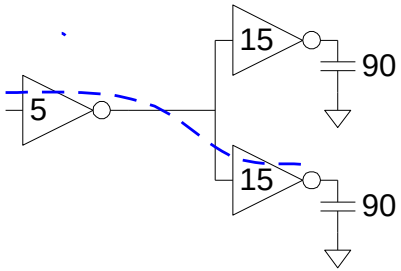
Multistage Logic Networks

- Logical effort generalizes to multistage networks
- *Path Logical Effort* $G = \prod g_i$
- *Path Electrical Effort* $H = \frac{C_{out-path}}{C_{in-path}}$
- *Path Effort* $F = \prod f_i = \prod g_i h_i$
- Can we write F = GH?

Paths that Branch

- No! Consider paths that branch:

G =
H =
GH =
h_1 =
h_2 =
F =



Branching Effort

- Introduce *branching effort*
  - Accounts for branching between stages in path

$$b = \frac{C_{\text{on path}} + C_{\text{off path}}}{C_{\text{on path}}}$$

$$B = \prod b_i$$

Note:
$$\prod h_i = BH$$

- Now we compute the path effort
  - F = GBH

Multistage Delays

- Path Effort Delay $D_F = \sum f_i$
- Path Parasitic Delay $P = \sum p_i$
- Path Delay $D = \sum d_i = D_F + P$

Designing Fast Circuits

$$D = \sum d_i = D_F + P$$

- Delay is smallest when each stage bears same effort

$$\hat{f} = g_i h_i = F^{\frac{1}{N}}$$

- Thus minimum delay of N stage path is

- This is a key result of logical effort
  - Find fastest possible delay
  - Doesn't require calculating gate sizes

Gate Sizes

- How wide should the gates be for least delay?

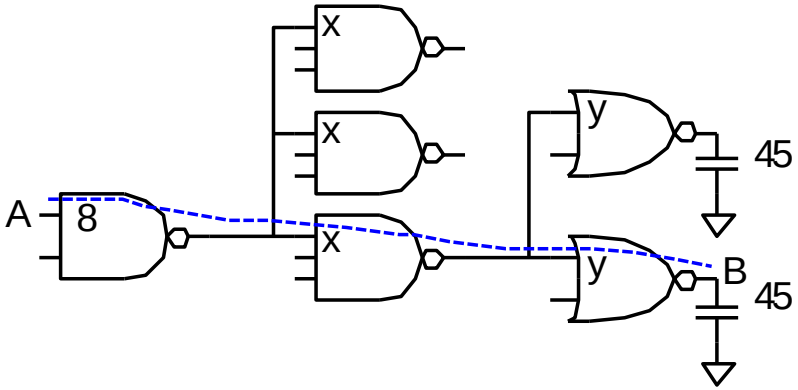
$$\hat{f} = gh = g\frac{C_{out}}{C_{in}}$$

$$\Rightarrow C_{in_i} = \frac{g_i C_{out_i}}{\hat{f}}$$

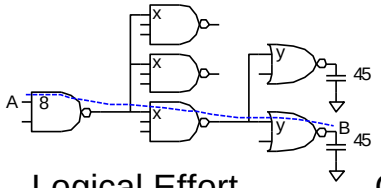
- Working backward, apply capacitance transformation to find input capacitance of each gate given load it drives.
- Check work by verifying input cap spec is met.

Example: 3-stage path

- Select gate sizes x and y for least delay from A to B



Example: 3-stage path



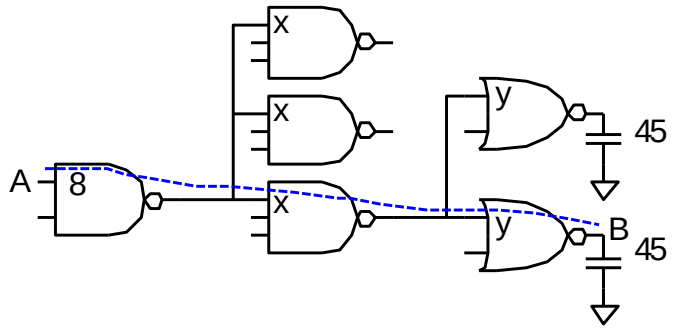
Logical Effort	G =
Electrical Effort	H =
Branching Effort	B =
Path Effort	F =
Best Stage Effort	$\hat{f}$ =
Parasitic Delay	P =
Delay	D =

Example: 3-stage path

- Work backward for sizes

y =

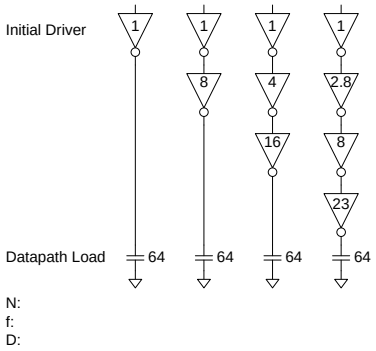
x =



Best Number of Stages

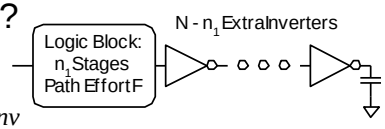
- How many stages should a path use?
  - Minimizing number of stages is not always fastest
- Example: drive 64-bit datapath with unit inverter

D =



Derivation

- Consider adding inverters to end of path
  - How many give least delay?



$$D = NF^{\frac{1}{N}} + \sum_{i=1}^{n_1} p_i + \left(N - n_1\right) p_{inv}$$

$$\frac{\partial D}{\partial N} = -F^{\frac{1}{N}} \ln F^{\frac{1}{N}} + F^{\frac{1}{N}} + p_{inv} = 0$$

- Define best stage effort $\rho = F^{\frac{1}{N}}$

$$p_{inv} + \rho\left(1 - \ln \rho\right) = 0$$

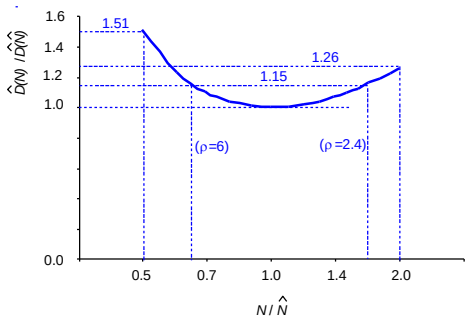
Best Stage Effort

- $p_{inv} + \rho\left(1 - \ln \rho\right) = 0$ has no closed-form solution

- Neglecting parasitics (p_{inv} = 0), we find ρ = 2.718 (e)
- For p_{inv} = 1, solve numerically for ρ = 3.59

Sensitivity Analysis

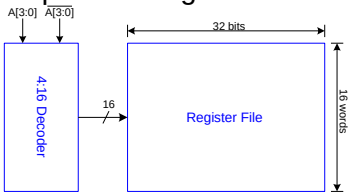
- How sensitive is delay to using exactly the best number of stages?



- 2.4 < ρ < 6 gives delay within 15% of optimal
  - We can be sloppy!
  - I like ρ = 4

Example, Revisited

- Ben Bitdiddle is the memory designer for the Motoroil 68W86, an embedded automotive processor. Help Ben design the decoder for a register file.



- Decoder specifications:
  - 16 word register file
  - Each word is 32 bits wide
  - Each bit presents load of 3 unit-sized transistors
  - True and complementary address inputs A[3:0]
  - Each input may drive 10 unit-sized transistors
- Ben needs to decide:
  - How many stages to use?
  - How large should each gate be?
  - How fast can decoder operate?

Number of Stages

- Decoder effort is mainly electrical and branching
  - Electrical Effort: H =
  - Branching Effort: B =
- If we neglect logical effort (assume G = 1)
  - Path Effort: F =
  - Number of Stages: N =
- Try a -stage design

Gate Sizes & Delay

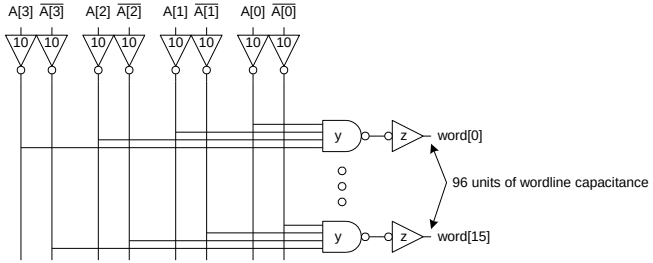
Logical Effort: G =

Path Effort: F =

Stage Effort: $\hat{f} =$

Path Delay: $D =$

Gate sizes: z = y =



Comparison

- Compare many alternatives with a spreadsheet
- $D = N(76.8\ G)^{1/N} + P$

Design	N	G	P	D
NOR4	1	3	4	234
NAND4-INV	2	2	5	29.8
NAND2-NOR2	2	20/9	4	30.1
INV-NAND4-INV	3	2	6	22.1
NAND4-INV-INV-INV	4	2	7	21.1
NAND2-NOR2-INV-INV	4	20/9	6	20.5
NAND2-INV-NAND2-INV	4	16/9	6	19.7
INV-NAND2-INV-NAND2-INV	5	16/9	7	20.4
NAND2-INV-NAND2-INV-INV-INV	6	16/9	8	21.6

Review of Definitions

Term	Stage	Path
number of stages	1	N
logical effort	g	$G = \prod g_i$
electrical effort	$h = \frac{C_{out}}{C_{in}}$	$H = \frac{C_{out\text{-}path}}{C_{in\text{-}path}}$
branching effort	$b = \frac{C_{on\text{-}path} + C_{off\text{-}path}}{C_{on\text{-}path}}$	$B = \prod b_i$
effort	$f = gh$	$F = GBH$
effort delay	f	$D_F = \sum f_i$
parasitic delay	p	$P = \sum p_i$
delay	$d = f + p$	$D = \sum d_i = D_F + P$

Method of Logical Effort

1) Compute path effort $F = GBH$
2) Estimate best number of stages $N = \log_4 F$
3) Sketch path with N stages
4) Estimate least delay $D = NF^{\frac{1}{N}} + P$
5) Determine best stage effort $\hat{f} = F^{\frac{1}{N}}$
6) Find gate sizes $C_{in_i} = \frac{g_i C_{out_i}}{\hat{f}}$

Limits of Logical Effort

- Chicken and egg problem
  - Need path to compute G
  - But don't know number of stages without G
- Simplistic delay model
  - Neglects input rise time effects
- Interconnect
  - Iteration required in designs with wire
- Maximum speed only
  - Not minimum area/power for constrained delay

Summary

- Logical effort is useful for thinking of delay in circuits
  - Numeric logical effort characterizes gates
  - NANDs are faster than NORs in CMOS
  - Paths are fastest when effort delays are ~4
  - Path delay is weakly sensitive to stages, sizes
  - But using fewer stages doesn't mean faster paths
  - Delay of path is about $\log_4$F FO4 inverter delays
  - Inverters and NAND2 best for driving large caps
- Provides language for discussing fast circuits
  - But requires practice to master

Practice 1

Consider the two designs of a 2-input AND gate shown in Figure 4.39. Give an intuitive argument about which will be faster. Back up your argument with a calculation of the path effort, delay, and input capacitances x and y to achieve this delay.

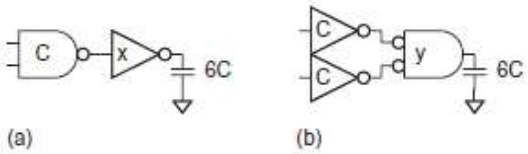


FIGURE 4.39 2-input AND gate

(a) should be faster than (b) because the NAND has the same parasitic delay but lower logical effort than the NOR. In each design, $H = 6$, $B = 1$, $P = 1 + 2 = 3$. For (a), $G = (4/3) * 1 = (4/3)$. $F = GBH = 8$. $f = 8^{1/2} = 2.8$. $D = 2f + P = 8.6\ \tau$. $x = 6C * 1 / f = 2.14C$. For (b), $G = 1 * (5/3)$. $F = GBH = 10$. $f = 10^{1/2} = 3.2$. $D = 2f + P = 9.3\ \tau$. $x = 6C * (5/3) / f = 3.16C$.

Practice 2

Consider four designs of a 6-input AND gate shown in Figure 4.40. Develop an expression for the delay of each path if the path electrical effort is H. What design is fastest for $H = 1$? For $H = 5$? For $H = 20$? Explain your conclusions intuitively.

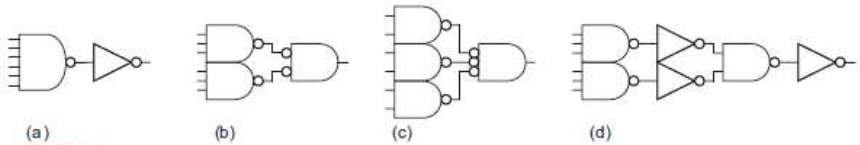


FIGURE 4.40 6-input AND gate

Practice 2

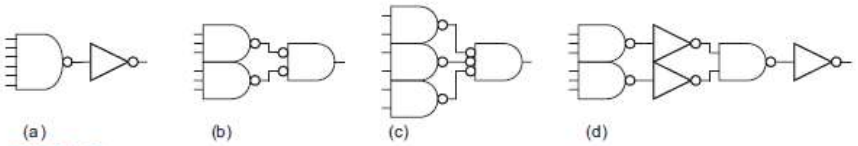


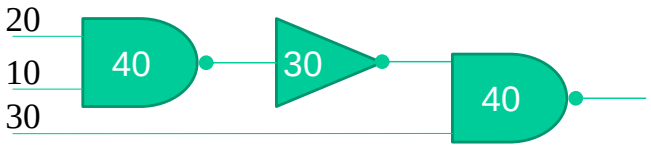
FIGURE 4.40 6-input AND gate

$D = N(GH)^{1/N} + P$. Compare in a spreadsheet. Design (b) is fastest for $H = 1$ or 5. Design (d) is fastest for $H = 20$ because it has a lower logical effort and more stages to drive the large path effort. (c) is always worse than (b) because it has greater logical effort, all else being equal.

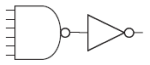
Design	G	P	N	D (H=1)	D (H=5)	D (H=20)
(a)	8/3 * 1	6 + 1	2	10.3	14.3	21.6
(b)	5/3 * 5/3	3 + 2	2	8.3	12.5	19.9
(c)	4/3 * 7/3	2 + 3	2	8.5	12.9	20.8
(d)	5/3 * 1 * 4/3 * 1	3 + 1 + 2 + 1	4	11.8	14.3	17.3

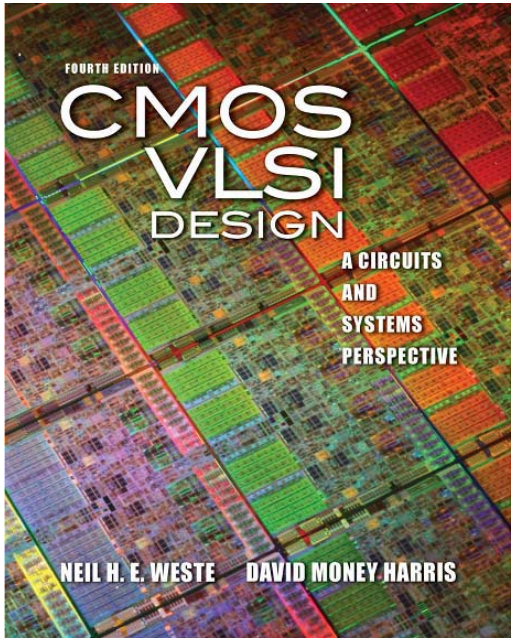
Review

1. What are t_{pdr}, t_{pdf}, t_f, t_r, t_{cdr}, t_{cdf}?
2. Calculate arrive time of the following circuit:



3. Explain the delay estimation of a fanout-of-1 inverter (slide 10)
4. Explain the t_{pdr} and t_{pdf} delay estimation of 3-input NAND driving h identical gates (slide 15).
5. Estimate delay for the gates: AOI21, OAI31
6. What is logical effort?
7. What is parasitic delay?
8. Estimate the delay of the following gate:





Lecture 5: Power

1. Power and Energy
2. Dynamic Power
3. Static Power
4. Low Power Techniques

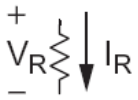
1. Power and Energy

- Power is drawn from a voltage source attached to the V_{DD} pin(s) of a chip.
- Instantaneous Power: $P(t) =$
- Energy: $E =$
- Average Power: $P_{avg} =$

Power in Circuit Elements

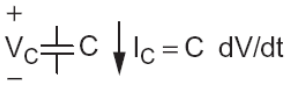
$$P_{VDD}(t) = I_{DD}(t) V_{DD}$$

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



$$E_C = \int_0^\infty I(t) V(t)\, dt = \int_0^\infty C \frac{dV}{dt} V(t)\, dt$$

$$= C \int_0^{V_C} V(t)\, dV = \tfrac{1}{2} C V_C^2$$



2. Dynamic Power

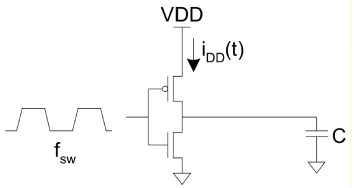
Charging a Capacitor

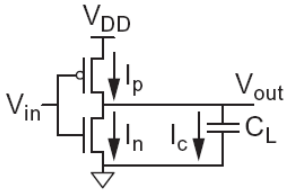
- When the gate output rises
  - Energy stored in capacitor is $E_C = \frac{1}{2} C_L V_{DD}^2$
  - But energy drawn from the supply is

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

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

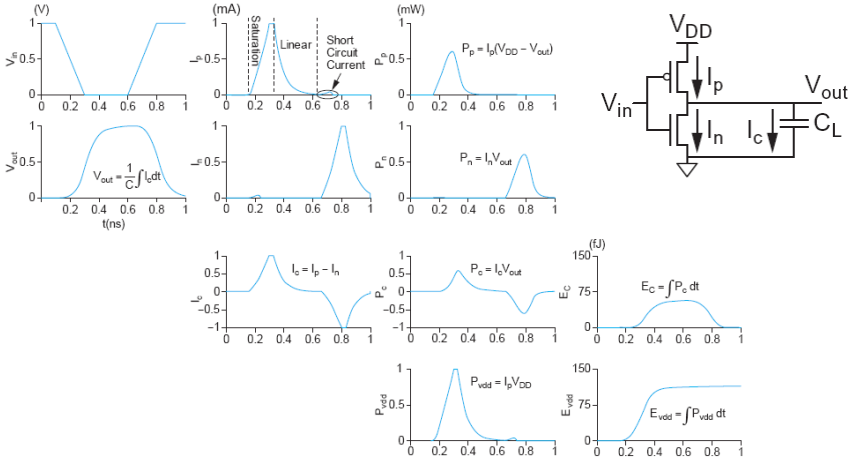
  - Half the energy from V_{DD} is dissipated in the pMOS transistor as heat, other half stored in capacitor
- When the gate output falls
  - Energy in capacitor is dumped to GND
  - Dissipated as heat in the nMOS transistor





Switching Waveforms

- Example: V_{DD} = 1.0 V, C_L = 150 fF, f = 1 GHz



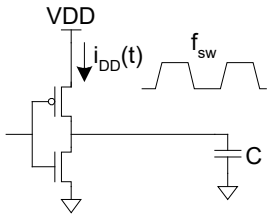
Switching Power

$$P_{\text{switching}} = \frac{1}{T}\int_0^T i_{DD}(t)V_{DD}\,dt$$

$$= \frac{V_{DD}}{T}\int_0^T i_{DD}(t)\,dt$$

$$= \frac{V_{DD}}{T}\left[Tf_{sw}CV_{DD}\right]$$

$$= CV_{DD}^{\ 2}f_{sw}$$



Activity Factor

- Suppose the system clock frequency = f
- Let $f_{sw} = \alpha f$, where α = activity factor
  - If the signal is a clock, α = 1
  - If the signal switches once per cycle, α = ½

- Dynamic power:

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

Lowering Dynamic Power

Capacitance:
Function of fan-out, wire length, transistor sizes

Supply voltage:
Has been dropping with successive generations

$$P_{dyn} = C_L\, V_{DD}^{\ 2}\, P_{0\to1}\, f$$

Activity factor:
How often, on average, do wires switch?

Clock frequency:
Increasing…

Short Circuit Current

- When transistors switch, both nMOS and pMOS networks may be momentarily ON at once
- Leads to a blip of "short circuit" current.
- < 10% of dynamic power if rise/fall times are comparable for input and output
- We will generally ignore this component

Power Dissipation Sources

- $P_{total} = P_{dynamic} + P_{static}$
- Dynamic power: $P_{dynamic} = P_{switching} + P_{shortcircuit}$
  - Switching load capacitances
  - Short-circuit current
- Static power: $P_{static} = (I_{sub} + I_{gate} + I_{junct} + I_{contention})V_{DD}$
  - Subthreshold leakage
  - Gate leakage
  - Junction leakage
  - Contention current

CMOS Energy & Power Equations

$$E = C_L V_{DD}^2 P_{0\rightarrow1} + t_{sc} V_{DD} I_{peak} P_{0/1\rightarrow1/0} + V_{DD} I_{leak}$$

$$f = P * f_{clock}$$

$$P = C_L V_{DD}^2 f + t_{sc} V_{DD} I_{peak} f + V_{DD} I_{leak}$$

Dynamic power (~90% today and decreasing relatively)

Short-circuit power (~8% today and decreasing absolutely)

Leakage power (~2% today and increasing)

Power and Energy Design Space

	Constant Throughput/Latency		Variable Throughput/Latency
Energy	Design Time	Non-active Modules	Run Time
Active (Dynamic)	Logic design Reduced V_{dd} TSizing Multi-V_{dd}	Clock Gating	DFS, DVS (Dynamic Freq, Voltage Scaling)
Leakage (Standby)	Multi-V_T Stack effect Pin ordering	Sleep Transistors Multi-V_{dd} Variable V_T Input control	Variable V_T

Estimate dynamic power

- 1 billion transistor chip
  - 50M logic transistors
    - Average width: 12 λ
    - Activity factor = 0.1
  - 950M memory transistors
    - Average width: 4 λ
    - Activity factor = 0.02
  - 1.0 V 65 nm process
  - C = 1 fF/μm (gate) + 0.8 fF/μm (diffusion)
- Estimate dynamic power consumption @ 1 GHz. Neglect wire capacitance and short-circuit current.

Solution

$$C_{\text{logic}} = \left(50\times10^{6}\right)\left(12\lambda\right)\left(0.025\mu m/\lambda\right)\left(1.8fF/\mu m\right) = 27 \text{ nF}$$

$$C_{\text{mem}} = \left(950\times10^{6}\right)\left(4\lambda\right)\left(0.025\mu m/\lambda\right)\left(1.8fF/\mu m\right) = 171 \text{ nF}$$

$$P_{\text{dynamic}} = \left[0.1C_{\text{logic}} + 0.02C_{\text{mem}}\right](1.0)^{2}\left(1.0 \text{ GHz}\right) = 6.1 \text{ W}$$

Dynamic Power Reduction

- $P_{\text{switching}} = \alpha C V_{DD}^{2} f$
- Try to minimize:
  - Activity factor
  - Capacitance
  - Supply voltage
  - Frequency

Activity Factor Estimation

- Let P_i = Prob(node i = 1)
  - $\bar{P}_i = 1-P_i$
- $\alpha_i = \bar{P}_i * P_i$
- Completely random data has P = 0.5 and α = 0.25
- Data is often not completely random
  - e.g. upper bits of 64-bit words representing bank account balances are usually 0
- Data propagating through ANDs and ORs has lower activity factor
  - Depends on design, but typically $\alpha \approx 0.1$

Dynamic Power Consumption is Data Dependent

- Switching activity, $P_{0\rightarrow1}$, has two components
  - A static component – function of the logic topology
  - A dynamic component – function of the timing behavior (glitching)

2-input NOR Gate

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

Static transition probability

$$P_{0\rightarrow1} = P_{out=0} \times P_{out=1}$$
$$= P_0 \times (1-P_0)$$

With input signal probabilities

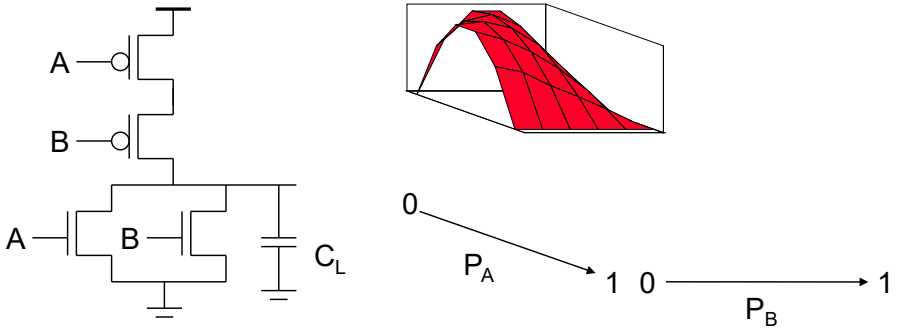
$$P_{A=1} = 1/2$$
$$P_{B=1} = 1/2$$

NOR static transition probability

$$= 3/4 \times 1/4 = 3/16$$

NOR Gate Transition Probabilities

- Switching activity is a strong function of the input signal statistics
  - P_A and P_B are the probabilities that inputs A and B are one



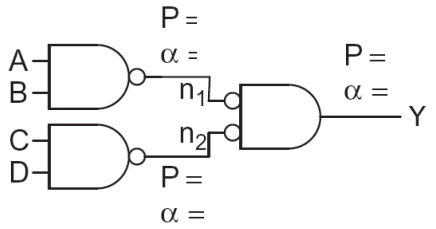
$$P_{0\rightarrow1} = P_0 \times P_1 = (1-(1-P_A)(1-P_B))\,(1-P_A)(1-P_B)$$

Switching Probability

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

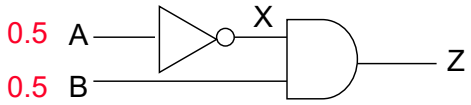
Example

- A 4-input AND is built out of two levels of gates
- Estimate the activity factor at each node if the inputs have P = 0.5



Transition Probabilities for Some Basic Gates

	$P_{0\to1} = P_{out=0} \times P_{out=1}$
NOR	$(1 - (1 - P_A)(1 - P_B)) \times (1 - P_A)(1 - P_B)$
OR	$(1 - P_A)(1 - P_B) \times (1 - (1 - P_A)(1 - P_B))$
NAND	$P_A P_B \times (1 - P_A P_B)$
AND	$(1 - P_A P_B) \times P_A P_B$
XOR	$(1 - (P_A + P_B - 2P_A P_B)) \times (P_A + P_B - 2P_A P_B)$



For X: $P_{0\to1} = P_0 \times P_1 = (1-P_A) P_A$

$= 0.5 \times 0.5 = 0.25$

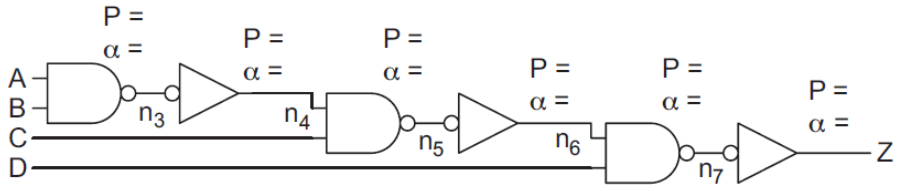
For Z: $P_{0\to1} = P_0 \times P_1 = (1-P_X P_B) P_X P_B$

$= (1 - (0.5 \times 0.5)) \times (0.5 \times 0.5) = 3/16$

Example

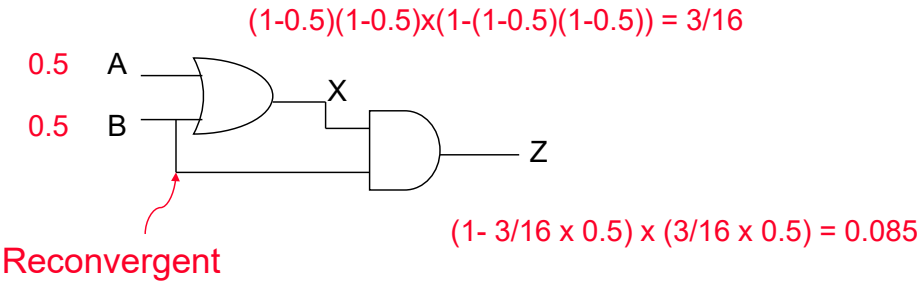
- Determine the activity factors at each node in the circuit assuming the input probabilities

  $P_A = P_B = P_C = P_D = 0.5$.



Inter-signal Correlations

- Determining switching activity is complicated by the fact that signals exhibit correlation in space and time
  - reconvergent fan-out



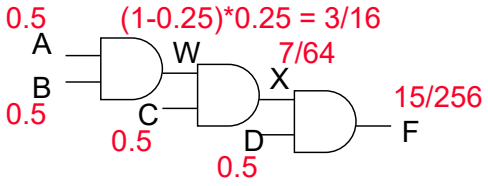
$P(Z=1) = P(B=1) \;\&\; P(A=1 \mid B=1)$

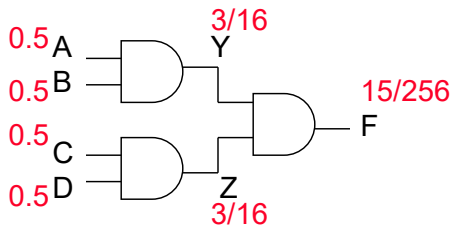
- Have to use conditional probabilities

Logic Restructuring

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

AND: $P_{0\to1} = P_0 \times P_1 = (1 - P_A P_B) \times P_A P_B$

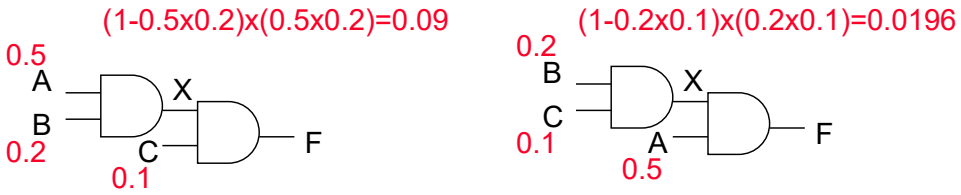




Chain implementation has a lower overall switching activity than the tree implementation for random inputs

Ignores glitching effects

Input Ordering

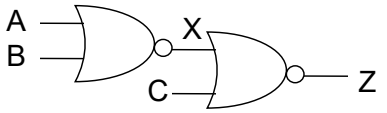


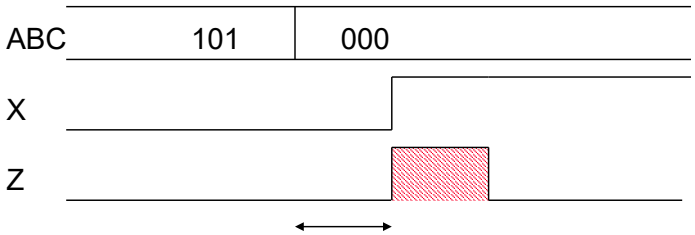
Which is better wrt transition probabilities?

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

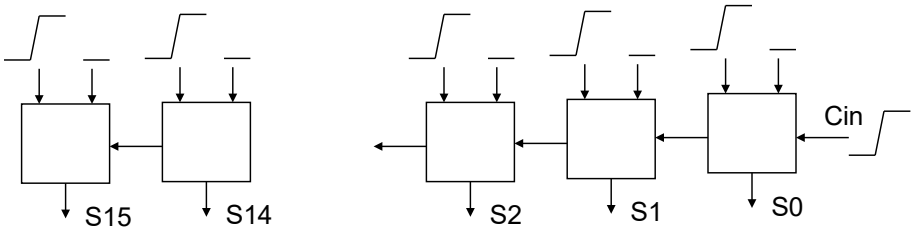
Glitching in Static CMOS Networks

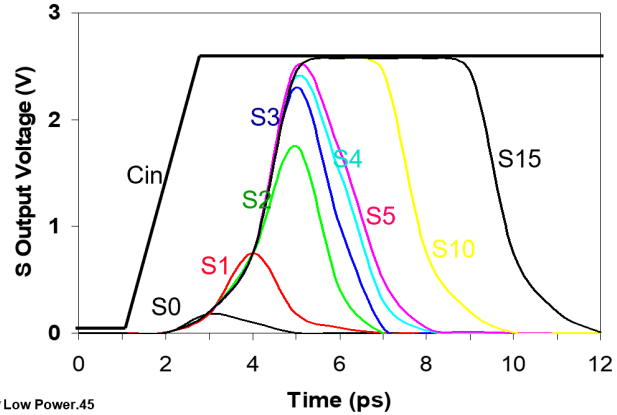
- Gates have a nonzero propagation delay resulting in spurious transitions or glitches (dynamic hazards)
  - glitch: node exhibits multiple transitions in a single cycle before settling to the correct logic value





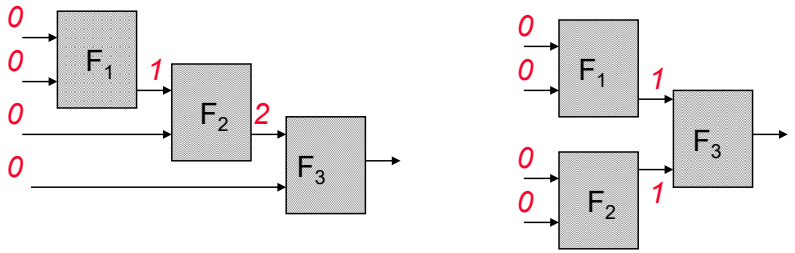
Glitching in an RCA





Balanced Delay Paths to Reduce Glitching

- Glitching is due to a mismatch in the path lengths in the logic network; if all input signals of a gate change simultaneously, no glitching occurs

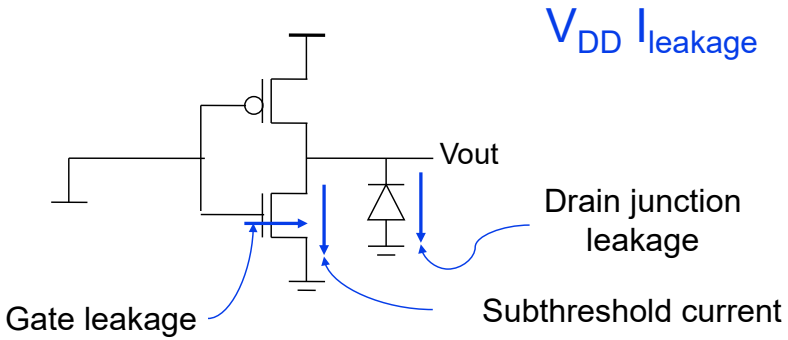


So equalize the lengths of timing paths through logic

3. Static Power

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

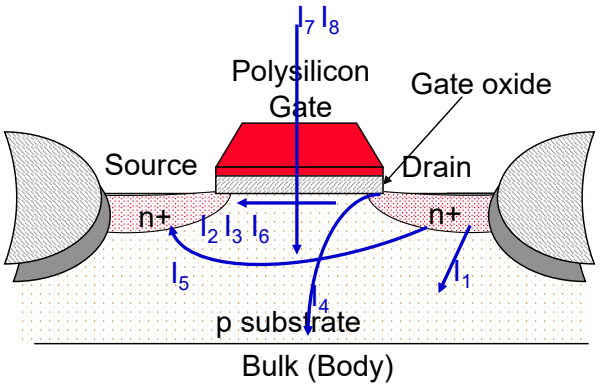
Leakage (Static) Power Consumption



Sub-threshold current is the dominant factor.

All increase exponentially with temperature!

Leakage Current Mechanisms



- I_1 p-n junction reverse bias current (drain junction)
- I_2 weak inversion (subthreshold current)
- I_3 DIBL
- I_4 GIDL
- I_5 punchthrough
- I_6 narrow width effect
- I_7 gate oxide tunneling (gate leakage)
- I_8 hot carrier injection

Static Power Example

- Revisit power estimation for 1 billion transistor chip
- Estimate static power consumption
  - Subthreshold leakage
    - Normal V_t: 100 nA/μm
    - High V_t: 10 nA/μm
    - High Vt used in all memories and in 95% of logic gates
  - Gate leakage 5 nA/μm
  - Junction leakage negligible

Solution

$$W_{\text{normal-V}_t} = (50\times10^6)(12\lambda)(0.025\mu\text{m}/\lambda)(0.05) = 0.75\times10^6\ \mu\text{m}$$

$$W_{\text{high-V}_t} = \left[(50\times10^6)(12\lambda)(0.95)+(950\times10^6)(4\lambda)\right](0.025\mu\text{m}/\lambda) = 109.25\times10^6\ \mu\text{m}$$

$$I_{sub} = \left[W_{\text{normal-V}_t}\times100\ \text{nA}/\mu\text{m}+W_{\text{high-V}_t}\times10\ \text{nA}/\mu\text{m}\right]/2 = 584\ \text{mA}$$

$$I_{gate} = \left[\left(W_{\text{normal-V}_t}+W_{\text{high-V}_t}\right)\times5\ \text{nA}/\mu\text{m}\right]/2 = 275\ \text{mA}$$

$$P_{static} = (584\ \text{mA}+275\ \text{mA})(1.0\ \text{V}) = 859\ \text{mW}$$

Subthreshold Leakage

- For $V_{ds} > 50$ mV

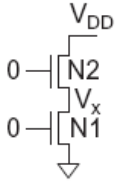
$$I_{sub} \approx I_{off}\,10^{\frac{V_{gs}+\eta(V_{ds}-V_{DD})-k_\gamma V_{sb}}{S}}$$

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

Typical values in 65 nm
I_{off} = 100 nA/μm @ V_t = 0.3 V
I_{off} = 10 nA/μm @ V_t = 0.4 V
I_{off} = 1 nA/μm @ V_t = 0.5 V
η = 0.1
k_γ = 0.1
S = 100 mV/decade

Stack Effect

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



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

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

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

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

Leakage Control

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

Gate Leakage

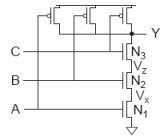
- Extremely strong function of t_{ox} and V_{gs}
  - Negligible for older processes
  - Approaches subthreshold leakage at 65 nm and below in some processes
- An order of magnitude less for pMOS than nMOS
- Control leakage in the process using t_{ox} > 10.5 Å
  - High-k gate dielectrics help
  - Some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

NAND3 Leakage Example

- 100 nm process

  I_{gn} = 6.3 nA I_{gp} = 0

  I_{offn} = 5.63 nA I_{offp} = 9.3 nA



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

Data from [Lee03]

Junction Leakage

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

Power and Energy Design Space

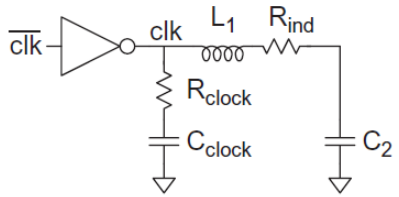
	Constant Throughput/Latency		Variable Throughput/Latency
Energy	Design Time	Non-active Modules	Run Time
Active (Dynamic)	Logic design Reduced V_{dd} TSizing Multi-V_{dd}	Clock Gating	DFS, DVS (Dynamic Freq, Voltage Scaling)
Leakage (Standby)	Multi-V_T Stack effect Pin ordering	Sleep Transistors Multi-V_{dd} Variable V_T Input control	Variable V_T

4. Low Power Methodology

- Resonant circuits: reduce switching power consumption
- Clock gating: turn off the clock to registers in unused blocks to reduce the activity
- Reduce capacitance: wire/gate capacitance
- Dynamic Voltage Scaling: Adjust VDD and f according to workload
- Power gating: Turn OFF power to blocks when they are idle to save leakage

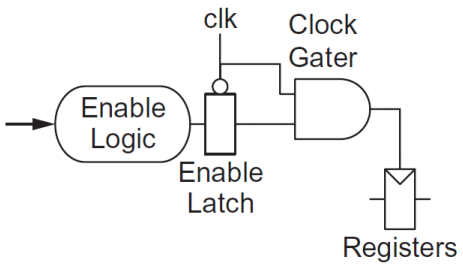
Resonant Circuits

- Letting energy slosh back and forth between storage elements such as capacitors and inductors rather than dumping the energy to ground.
- The technique is best suited to applications such as clocks that operate at a constant frequency.



Clock Gating

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

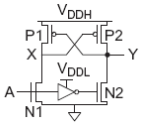


Capacitance

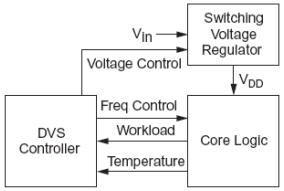
- Gate capacitance
  - Fewer stages of logic
  - Small gate sizes
- Wire capacitance
  - Good floorplanning to keep communicating blocks close to each other
  - Drive long wires with inverters or buffers rather than complex gates

Voltage / Frequency

- Run each block at the lowest possible voltage and frequency that meets performance requirements
- Voltage Domains
  - Provide separate supplies to different blocks
  - Level converters required when crossing from low to high V_{DD} domains

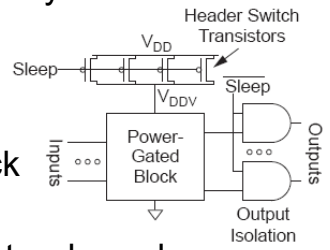


- Dynamic Voltage Scaling
  - Adjust V_{DD} and f according to workload



Power Gating

- Turn OFF power to blocks when they are idle to save leakage
  - Use virtual V_{DD} (V_{DDV})
  - Gate outputs to prevent invalid logic levels to next block

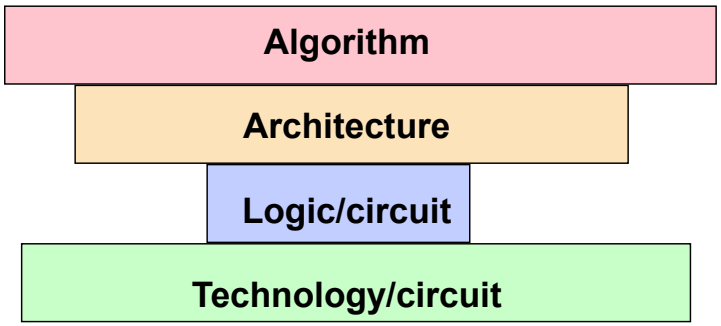


- Voltage drop across sleep transistor degrades performance during normal operation
  - Size the transistor wide enough to minimize impact
- Switching wide sleep transistor costs dynamic power
  - Only justified when circuit sleeps long enough

Low Power Design

- Reduce dynamic power
  - α: clock gating, sleep mode
  - C: small transistors (esp. on clock), short wires
  - V_{DD}: lowest suitable voltage
  - f: lowest suitable frequency
- Reduce static power
  - Selectively use ratioed circuits
  - Selectively use low V_t devices
  - Leakage reduction: stacked devices, body bias, low temperature, …

Low Power Design Techniques



- Need combination of techniques at all levels

Review

1. What is dynamic power?
2. What is static power?
3. Why does switching probability affect to dynamic power?
4. Describe some low power techniques
5. Describe resonant circuits
6. What are difference between clock gating and power gating
7. Calculate activity factors of the following circuits:

($P_A = P_B = P_C = P_D = 0.5$)