4. MOS Transistor Theory

Jacob Abraham

Department of Electrical and Computer Engineering
The University of Texas at Austin

VLSI Design
Fall 2020

September 8, 2020

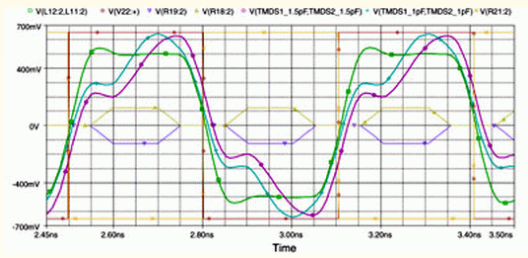
Electrical Properties

Necessary to understand basic electrical properties of the MOS transistor to design useful circuits

- Deal with non-ideal devices
- Ensure that the circuits are robust
- Create working layouts
- Predict delays and power consumption

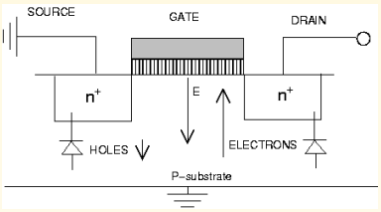
As circuit dimensions scale down, electrical effects become more important, even for digital circuits

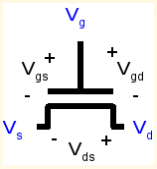
1.65 GHz square wave from an HDMI Interface (Source: Dunnihoo, *EE Times Asia*, 8/25/2005)



ECE Department, University of Texas at Austin Lecture 4. MOS Transistor Theory Jacob Abraham, September 8, 2020 1 / 31

The nMOS Transistor

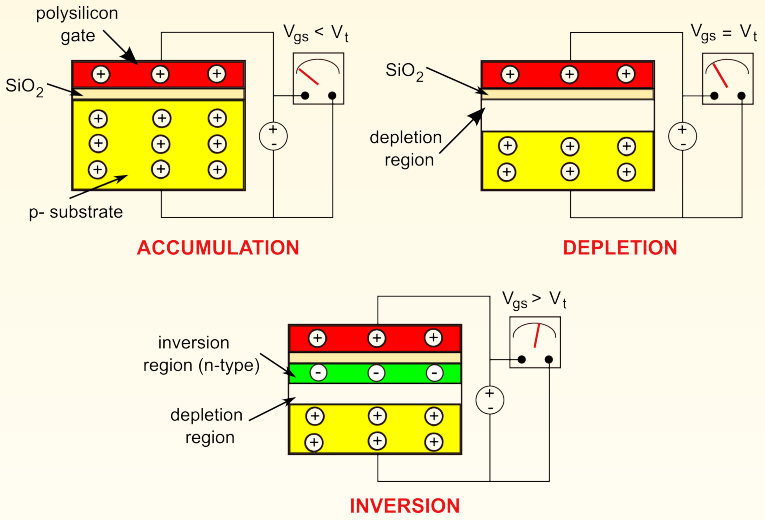




Terminal Voltages

- Modes of operation depend 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, so $V_{ds} \geq 0$
- nMOS body is grounded for simple designs; assume source is 0
- Three regions of operation: *Cutoff, Linear, Saturated*

Modes in nMOS Structure



nMOS Transistor Operation

Positive voltage on Gate produces electric field across substrate – attracts electrons to the Gate and repels holes

- With sufficient voltage, region under Gate changes from p- to n-Type – conducting path between the Source and Drain
- Inversion layer is field-induced junction, unlike a PN junction which is metallurgical
- Horizontal component of electric field associated with $V_{ds} > 0$ is responsible for sweeping electrons from channel to drain

Threshold Voltage

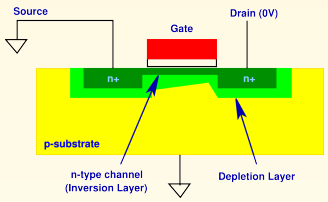
- The gate voltage at which conduction takes place is the **Threshold Voltage,** V_t
- Current flow occurs when the drain to source voltage $V_{ds} > 0$, and consists almost entirely of majority-carriers (electrons), that flow through **the channel**
- A depletion region insulates the channel from the substrate

Conducting nMOS Transistor

- Conduction when $V_{gs} > V_t$ and $V_{ds} > 0$
  - No significant current through the substrate because of reverse biased PN junction with the channel
  - As the voltage from drain to source is increased, the resistive drop along the channel begins to change the shape of the channel characteristic
  - At source end of the channel, the full gate voltage is effective in inverting the channel
  - At drain end of the channel, only the difference between the gate and the drain voltage is effective
- If $V_{ds} > V_{gs} - V_t$, then $V_{gd} < V_t$, and the channel is "pinched down" (the inversion layer no longer reaches the drain)
  - In this case, conduction is brought about by the drift mechanism of electrons under the influence of positive drain voltage; as the negative electrons leave the channel, they are accelerated towards the drain
- Voltage across the pinchdown channel tends to remain fixed at $(V_{gs} - V_t)$, and the channel current remains constant with increasing V_{ds}

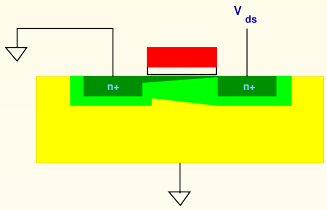
nMOS Device Behavior

$V_{gs} > V_t$, $V_{ds} = 0$



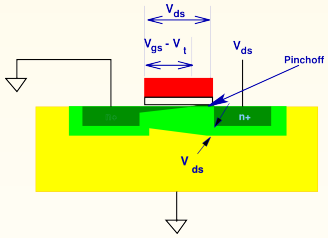
Nonsaturated Mode

$V_{ds} < V_{gs} - V_t$

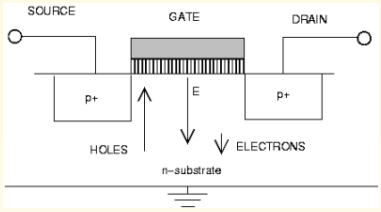


Saturated Mode

$(V_{ds} > V_{gs} - V_t)$



The pMOS Transistor



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

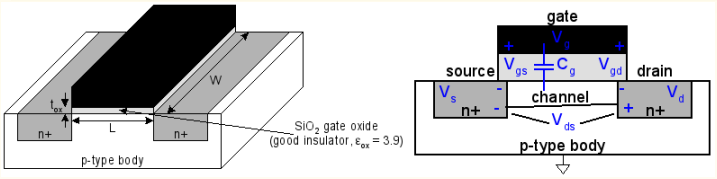
- Application of a negative gate voltage (w.r.t. source) draws holes into the region below the gate; channel changes from n to p-type (source-drain conduction path)
- Conduction due to holes; negative V_d sweeps holes from source (through channel) to drain

Current in the Channel

In the Linear region, I_{ds} depends on how much charge there is in the channel and how fast the charge is moving

Channel Charge

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



Carrier Velocity

- Charge is carried by electrons
- Carrier velocity ν proportional to lateral E- field between source and drain
- $\nu = \mu E$
  - μ is called **mobility**
- $E = V_{ds}/L$
- Time for carrier to cross channel: $t = L/\nu$

I-V Characteristics

nMOS Linear I-V

Current can be obtained from charge in channel and the time t each carrier takes to cross

$$\begin{aligned} I_{ds} &= \frac{Q_{channel}}{t} \\ &= \mu C_{ox} \frac{W}{L} \left(V_{gs} - V_t - V_{ds}/2\right) V_{ds} \\ &= \beta \left(V_{gs} - V_t - V_{ds}/2\right) V_{ds} \end{aligned}$$

nMOS Saturation I-V

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

$$\begin{aligned} I_{ds} &= \beta \left(V_{gs} - V_t - V_{dsat}/2\right) V_{dsat} \\ &= \frac{\beta}{2} (V_{gs} - V_t)^2 \end{aligned}$$

nMOS I-V Summary

Shockley First Order transistor models

$$I_{ds} = \begin{cases} 0 & V_{gs} < V_t & \text{Cutoff} \\ \beta \left(V_{gs} - V_t - V_{ds}/2\right) V_{ds} & V_{ds} < V_{dsat} & \text{Linear} \\ \frac{\beta}{2} (V_{gs} - V_t)^2 & V_{ds} > V_{dsat} & \text{Saturation} \end{cases}$$

pMOS I-V

All dopings and voltages are inverted for pMOS (compared with nMOS)

- Mobility μ_p is determined by holes
  - Typically 2x-3x lower than that of electrons μ_n
- Thus pMOS must be wider to provide the same current
  - Simple assumption, $\frac{\mu_n}{\mu_p} = 2$

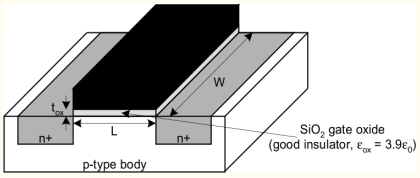
Capacitance

Capacitance in CMOS circuits

- 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
- Interconnection wires also have (distributed) capacitance

Gate Capacitance

- Approximate channel as connected to source
- $C_{gs} = \epsilon_{ox} WL/t_{ox} = C_{ox} WL = C_{permicron} W$
- Typical $C_{permicron} \approx 2\text{fF}/\mu\text{m}$

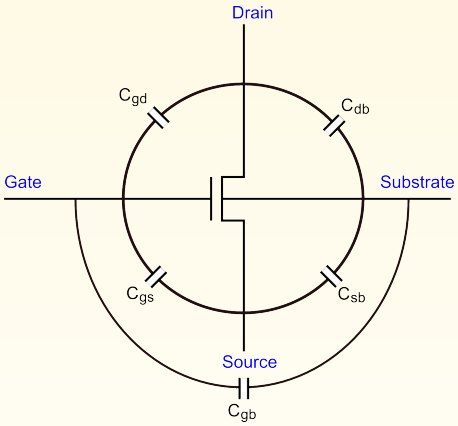


Device Capacitances

The **dynamic response** (switching speed) of a CMOS circuit is very dependent on parasitic capacitances associated with the circuit

Use a simple approximation for quick estimates of capacitances; use tools for extraction of more accurate values from actual layouts

Consider the capacitances seen during the different regions of operation



Device Capacitances, Cont'd

Off Region

- $V_{gs} \leq V_t$; when the MOS device is off, only C_{gb} (due to the series combination of gate oxide and depletion layer capacitance) is non-zero.
- $C_{gb} = C_{ox} = \epsilon A/t_{ox}$, where A is the gate area, and $\epsilon = \epsilon_0 \epsilon_{SiO_2}$
  - ϵ_0 is the permittivity of free space (8.854×10^4 F/m), and ϵ_{SiO_2} is the dielectric constant of SiO_2 (about 3.9)

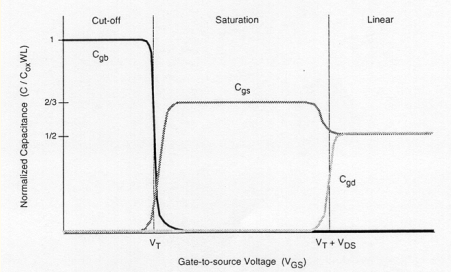
Linear Region

- Depletion region exists, forming dielectric of depletion capacitance, C_{dep} in series with C_{ox}
- As the device turns on, C_{gb} reduces to 0
- The gate capacitance is now a function of the gate voltage

Device Capacitances, Cont'd

Saturated Region

- Region under the gate is heavily inverted, and drain region of channel pinched off, with C_{gd} reducing to zero
- Gate capacitance is now less than C_{ox}



Approximation of Gate Capacitance

- For simplicity, we can assume the gate capacitance to be constant, $C_g = \epsilon A/t_{ox}$

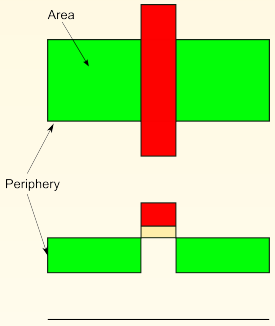
Diffusion (Source/Drain) Capacitance

Capacitance at the drain (C_{db}) or source (C_{sb}) of a device, or when diffusion is used as a wire

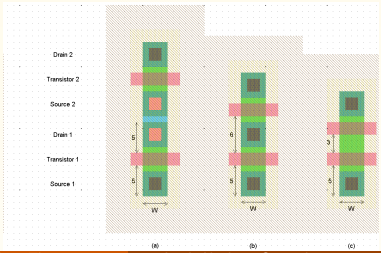
Two components:

1. An **Area** component
2. A **Peripheral** (sidewall) component

The peripheral component comes from the depth of the diffusion



- Assume diffusion capacitance is **approximately C_g for contacted diffusion**
- It is $1/2C_g$ for **uncontacted diffusion**

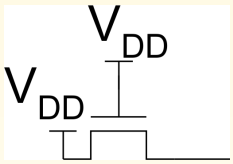


Pass Transistors

Have assumed that source is grounded

What happens if source > 0?

- Example, 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 $V_{DD} - V_t$

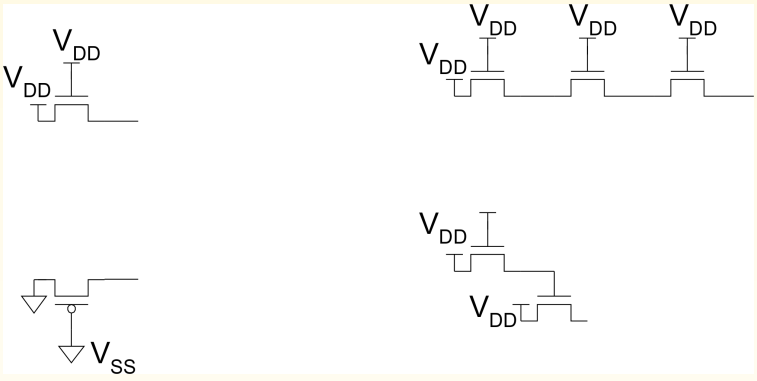
- Called a degraded "1"
- Degraded value reached slowly in a transition (low I_{ds})

pMOS pass transistors pull no lower than V_{tp}

- Degraded "0"

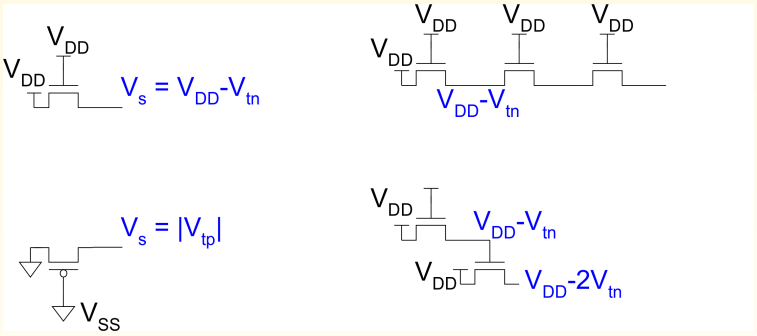
Pass Transistor Circuits

What would be the voltages on the different nodes?



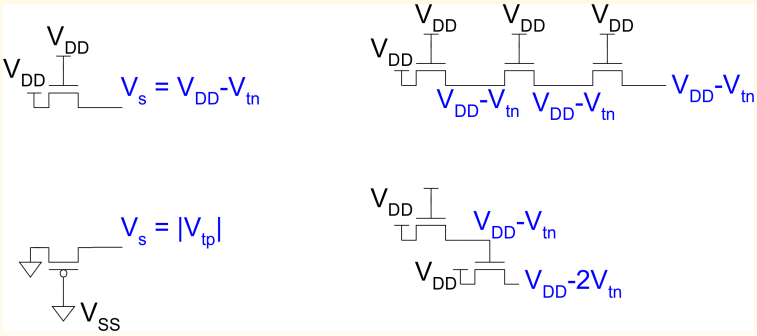
Pass Transistor Circuits

What would be the voltages on the different nodes?



Pass Transistor Circuits

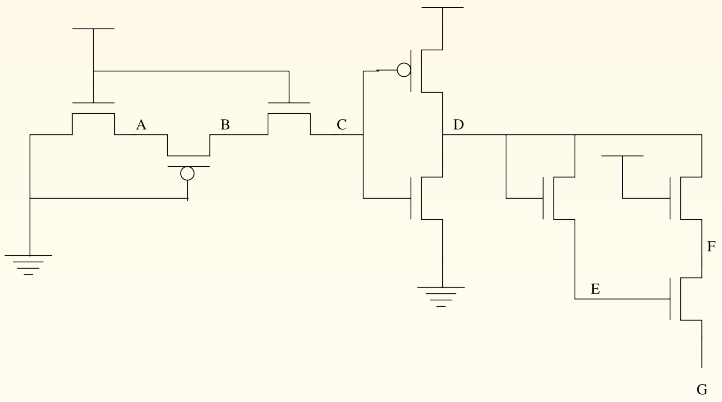
What would be the voltages on the different nodes?



Example 1

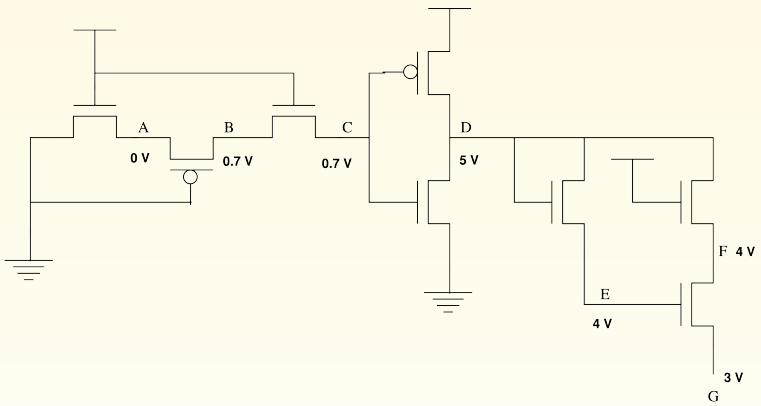
Assumption: initial voltage on each node is 2.5 volts

Relevant transistor parameters are, $V_{dd} = 5V$, $V_{tn} = 1V$ and $|V_{tp}| = 0.7V$



Example 1, Cont'd

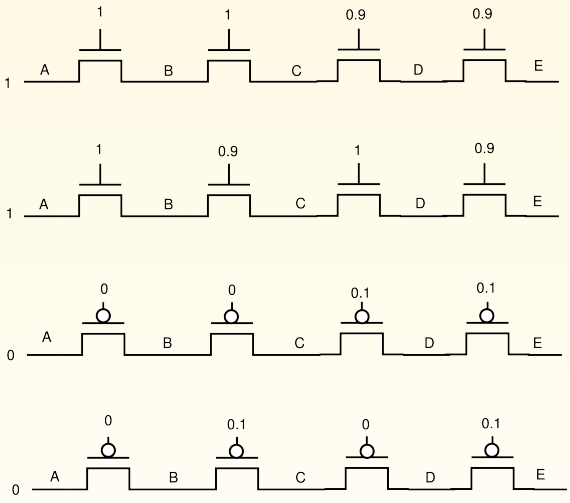
$V_{dd} = 5V$, $V_{tn} = 1V$ and $|V_{tp}| = 0.7V$



Example 2

Assume: initial voltage of 0.5V on all the internal nodes

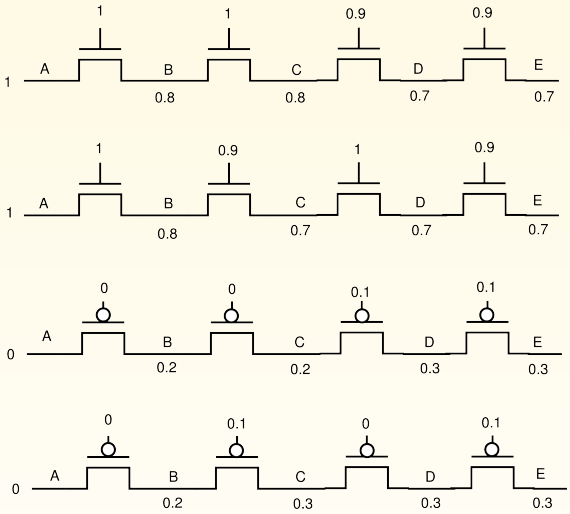
$V_{dd} = 1.0V$, $V_{t_n} = 0.2V$ and $|V_{t_p}| = 0.2V$



Example 2, Cont'd

Assume: initial voltage of 0.5V on all the internal nodes

$V_{dd} = 1.0V$, $V_{t_n} = 0.2V$ and $|V_{t_p}| = 0.2V$



Effective Resistance

Resistance of a bar of uniform material

- $R = \frac{\rho \times L}{A} = \left(\frac{\rho}{t}\right)\left(\frac{L}{W}\right)$
  - where ρ = resistivity of the material
  - A = cross-section of the resistor
  - t, W = thickness, width of the material
- The channel resistance of a MOS transistor in the linear region, $R_c = k\left(\frac{L}{W}\right)$,
  - where $k = \frac{1}{\mu C_{ox}(V_{gs} - V_t)}$

Resistance values depend on the technology

- Obtain the information from the technology files
- **Sheet resistance ($\Omega/\square$)**
  - Lowest for metal, increases for poly, active, highest for Well
- **Contact (via) resistance** becomes more important as processes scale down
- **Channel (turned-on transistor)** on the order of 1000 $\Omega/\square$

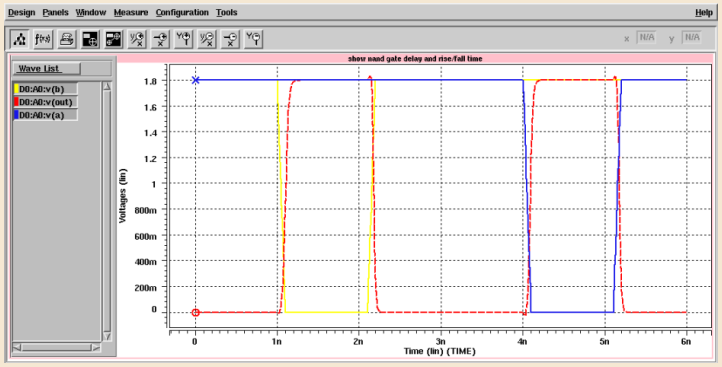
Example of Process Parameters and Simulation

Example: TSMC 0.18μ process

http://www.europractice-ic.com/technologies_TSMC.php

Look at one process

Example of SPICE simulation



Resistance on a Turned-On Transistor

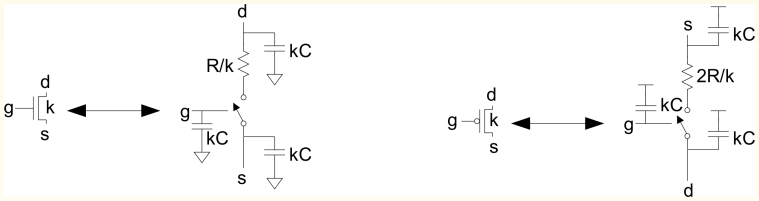
Circuit delay depends on resistance and capacitance – delay depends on RC

- Need to deal with the resistance of conducting transistors and interconnects
- Shockley models have limited value for obtaining resistance
  - 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
  - $I_{ds} = \frac{V_{ds}}{R}$
  - R averaged across switching of digital gate
- Too inaccurate to predict current at any given time, but good enough to predict RC delay
- More accurate values of delay obtained from detailed design using the tools

RC Delay Model

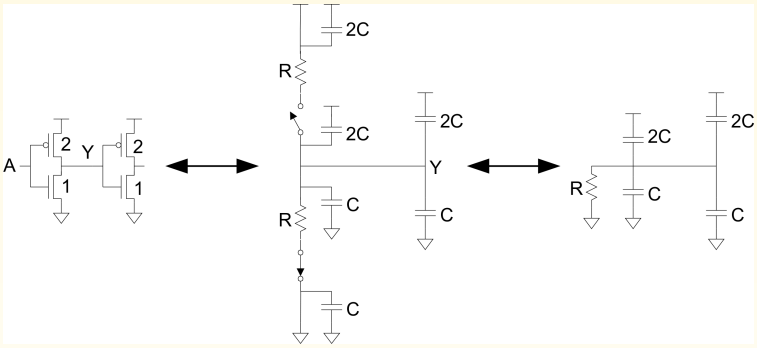
Use equivalent circuits for MOS transistors

- 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



Inverter Delay Estimate

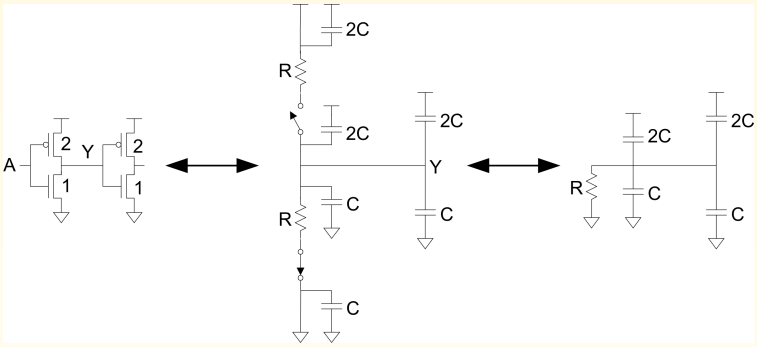
Estimate the delay of a fanout-of-1 inverter



d = 6RC

Inverter Delay Estimate

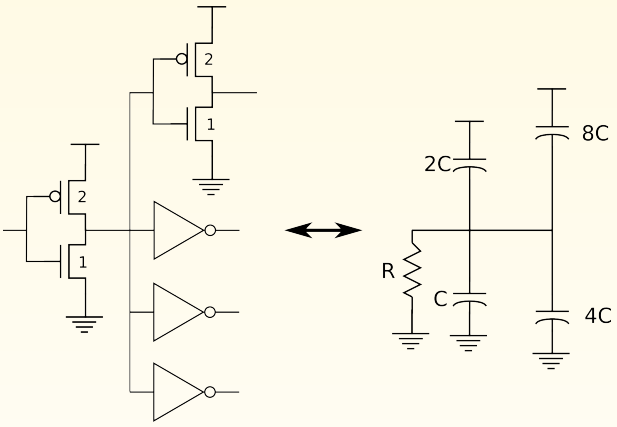
Estimate the delay of a fanout-of-1 inverter



d = 6RC

Inverter Delay Estimate, Cont'd

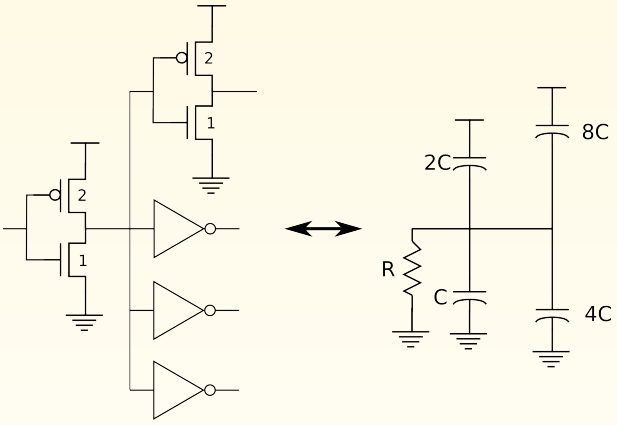
Estimate the delay of an inverter driving 4 identical inverters –
Fanout-of-4 (FO4) delay
An important abstraction at higher levels of the design



d = 15RC

Inverter Delay Estimate, Cont'd

Estimate the delay of an inverter driving 4 identical inverters –
Fanout-of-4 (FO4) delay
An important abstraction at higher levels of the design



d = 15RC