CS/EE 5710/6710

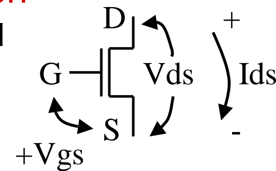
MOS Transistor Models

Electrical Effects

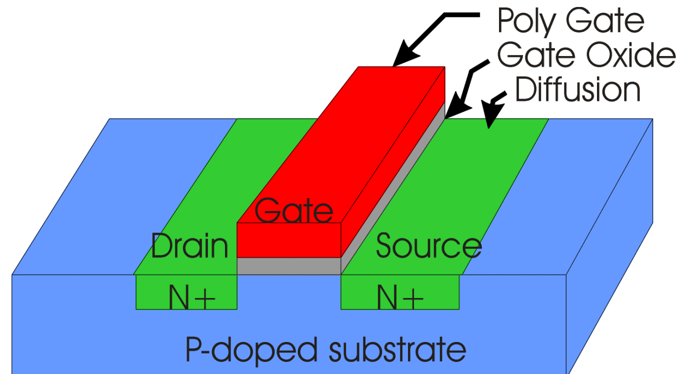
Propagation Delay

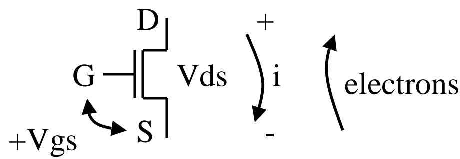
Transistor Characteristics

- Three conduction characteristics
  - Cutoff Region
    - No inversion layer in channel
    - Ids = 0
  - Nonsaturated, or linear region
    - Weak inversion of the channel
    - Ids depends on Vgs and Vds
  - Saturated region
    - Strong inversion of channel
    - Ids is independent of Vds
  - As an aside, at very high drain voltages:
    - "avalanche breakdown" or "punch through"
    - Gate has no control of Ids…

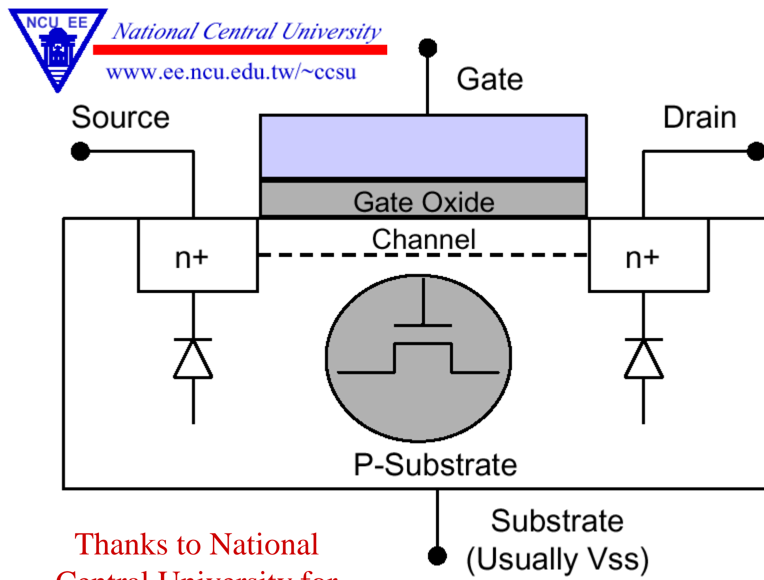


N-type Transistor

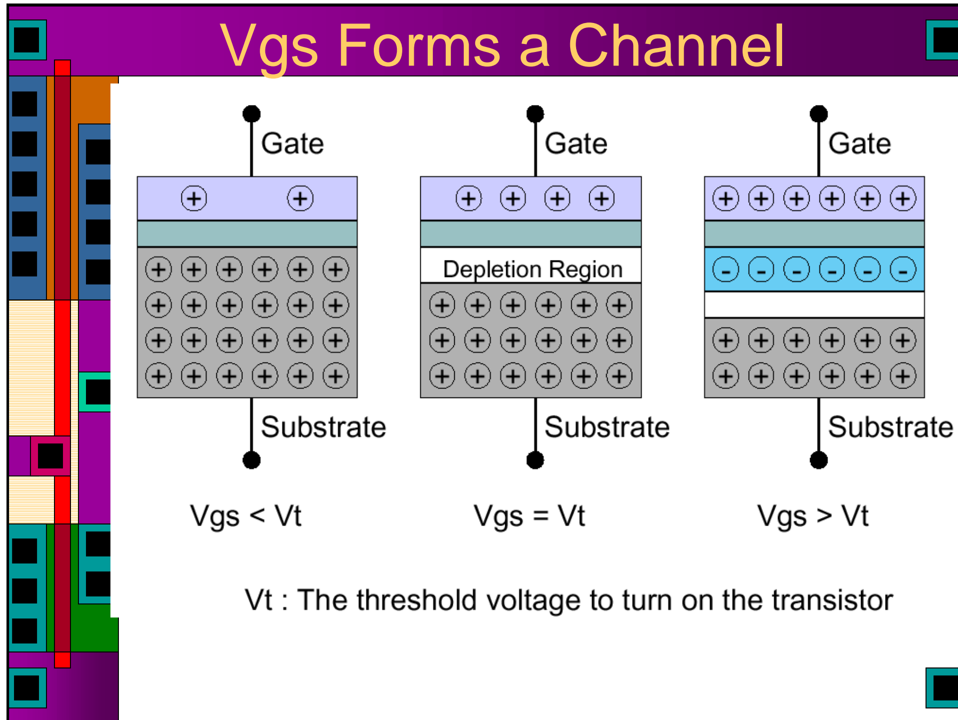




Another Cutaway View



Thanks to National Central University for Some images

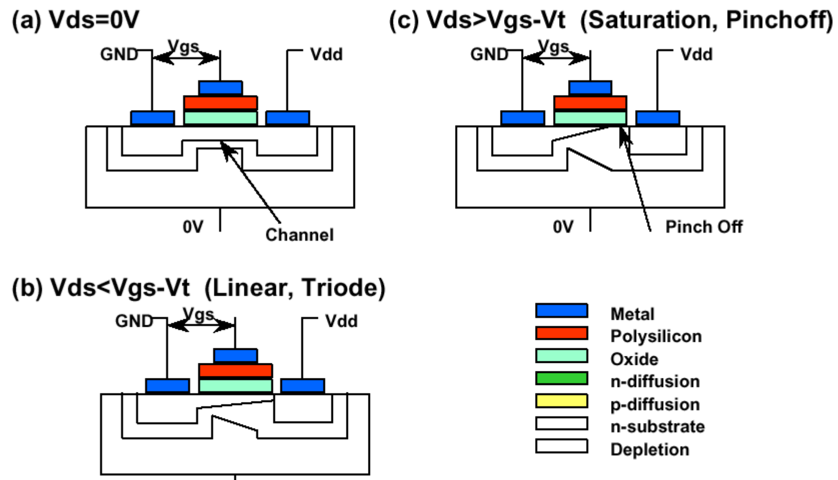


Basic N-Type MOS Transistor

- Conditions for the regions of operation
  - Cutoff: If Vgs < Vt, then Ids is essentially 0
    - Vt is the “Threshold Voltage”
  - Linear: If Vds < (Vgs – Vt) then Ids depends on both Vgs and Vds
    - Channel becomes deeper as Vgs goes up
  - Saturated: If (Vgs – Vt) < Vds then Ids is essentially constant (Saturated)

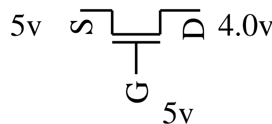
D G S +Vgs Vds + Ids - electrons

Visualizing the Channel

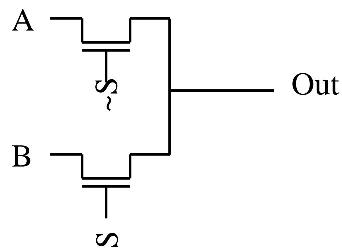


"Saturated" Transistor

- In the 0 < (Vgs – Vt) < Vds case
  - Ids Current is effectively constant
  - Channel is "pinched off" and conduction is accomplished by drift of carriers
  - Voltage across pinched off channel (I.e. Vds) is fixed at Vgs – Vt
    - This is why you don't use an N-type to pass 1's!
    - High voltage is degraded by Vt
    - Iv Vt is 1.0v, 5v in one side, 4.0v out the other

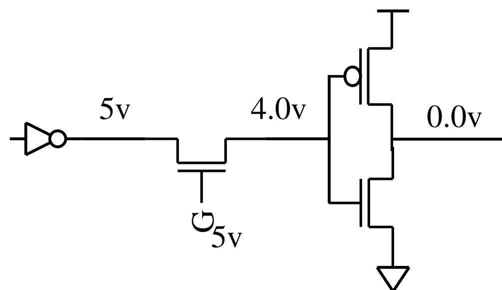


Aside: N-type Pass Transistors



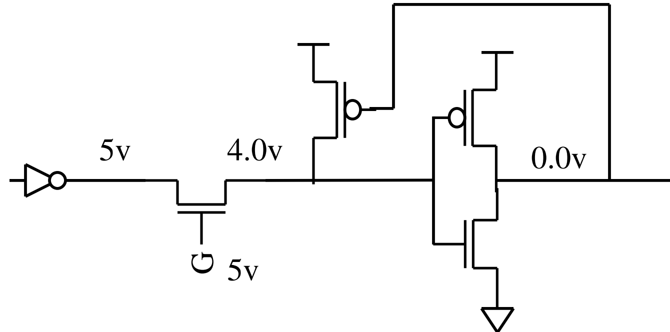
- If it weren't for the threshold drop, N-type pass transistors (without the P-type transmission gate) would be nice
  - 2-way Mux Example…

N-type Pass Transistors



- One one hand, the degraded high voltage from the pass transistor will be restored by the inverter
- On the other hand, the P-device may not turn off completely resulting in extra power being used

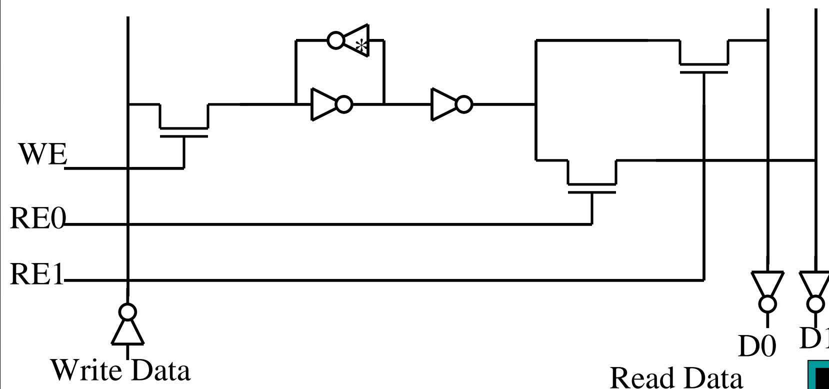
N-type Pass Transistors

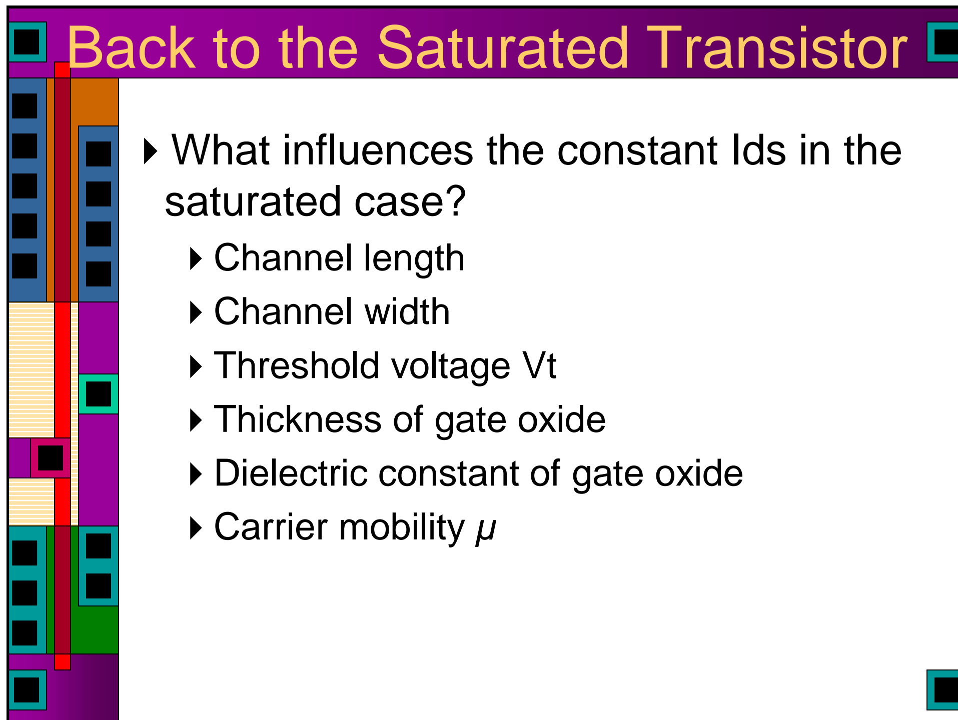


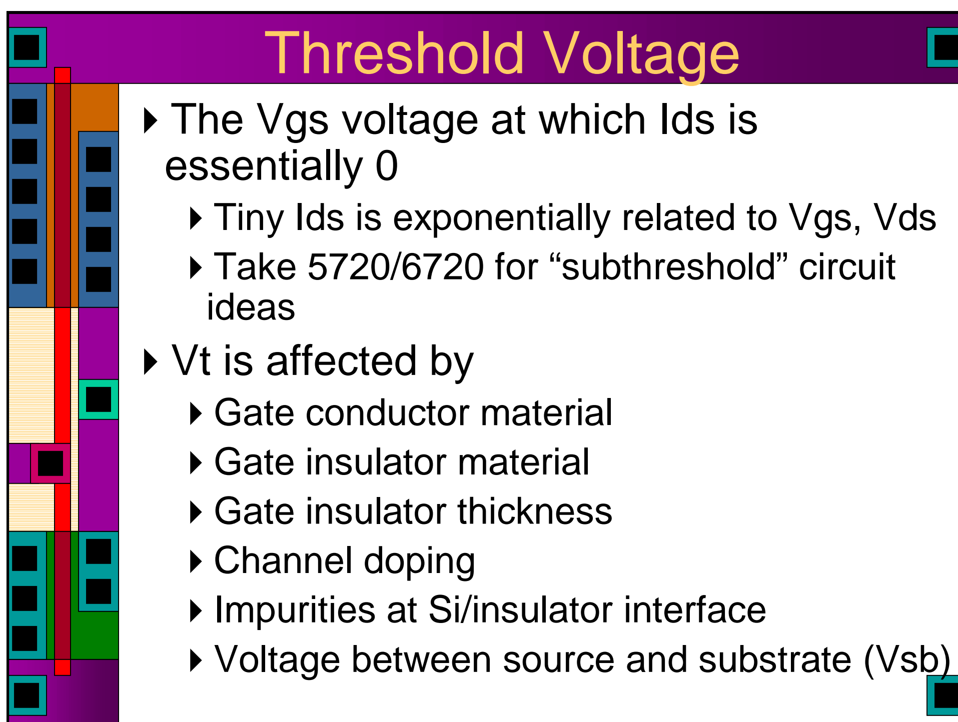
- Another option is a "keeper" transistor fed back from the output
  - This pulls the internal node high when the output is 0
  - But is disconnected when output is high
- Make sure the size is right…

N-type Pass Transistors

- In practice, they are used fairly often, but be aware of what you're doing
  - For example, read/write circuits in a Register File

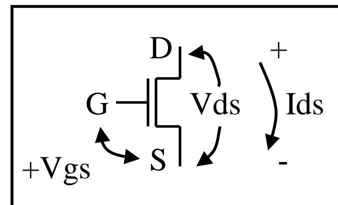



Back to the Saturated Transistor
What influences the constant Ids in the saturated case?
Channel length
Channel width
Threshold voltage Vt
Thickness of gate oxide
Dielectric constant of gate oxide
Carrier mobility μ

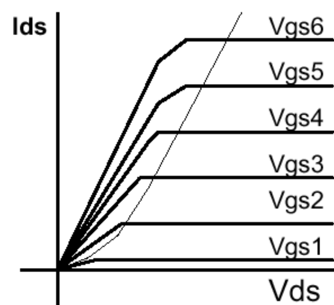

Threshold Voltage
The Vgs voltage at which Ids is essentially 0
Tiny Ids is exponentially related to Vgs, Vds
Take 5720/6720 for "subthreshold" circuit ideas
Vt is affected by
Gate conductor material
Gate insulator material
Gate insulator thickness
Channel doping
Impurities at Si/insulator interface
Voltage between source and substrate (Vsb)

Basic DC Equations for Ids

- Cutoff Region
  - Vgs < Vt, Ids = 0
- Linear Region
  - 0 < Vds < (Vgs – Vt)
    Ids = β[(Vgs – Vt)Vds – Vds²/2]
  - Note that this is only "linear" if Vds²/2 is very small, I.e. Vds << Vgs –Vt
- Saturated Region
  - 0 < (Vgs – Vt) < Vds
    Ids = β[(Vgs – Vt)²/2]



Ids Curves



$$\beta = \frac{\mu\varepsilon}{t_{ox}}\left(\frac{W}{L}\right) = \mu C_{ox}\left(\frac{W}{L}\right)$$

Cutoff Region

$$V_{gs} < V_t$$

$$I_{ds} = 0$$

Triode (Linear) Region

$$V_{gs} - V_t > V_{ds} > 0$$

$$I_{ds} = \beta\left[(V_{gs} - V_t)V_{ds} - \frac{V_{ds}^2}{2}\right]$$

Saturation Region

$$V_{gs} - V_t > V_{ds} > 0$$

$$I_{ds} = \beta\frac{(V_{gs} - V_t)^2}{2}$$

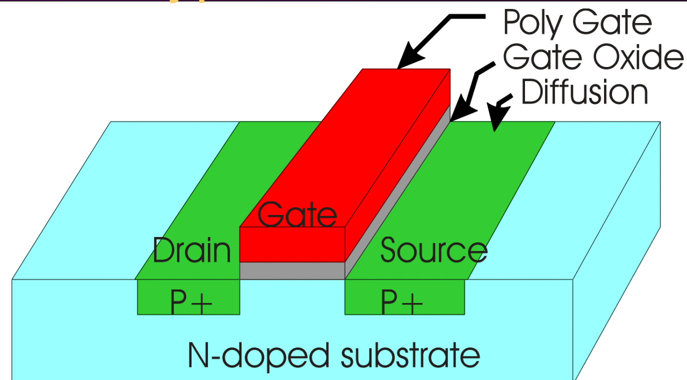
Transistor Gain

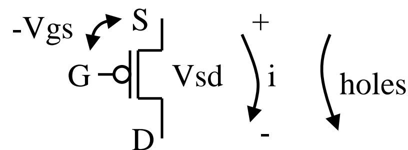
- β is the MOS transistor gain factor
- $\beta = (\mu\varepsilon/t_{ox})(W/L)$

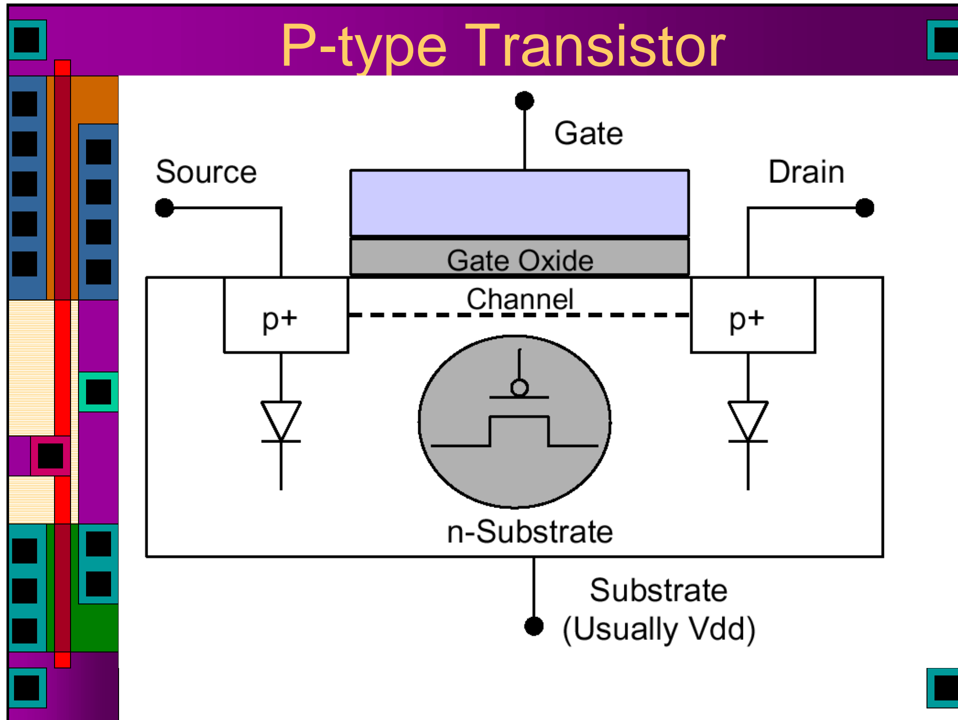
  Process-dependent: $(\mu\varepsilon/t_{ox})$ — Layout dependent: (W/L)

  - μ = mobility of carriers
    - Note that N-type is twice as good as P-type
  - ε = permittivity of gate insulator
  - T_{ox} = thickness of gate oxide
  - Book calls $(\mu\varepsilon/t_{ox}) = k'$
- Increase W/L to increase gain

P-type Transistor







P-type Transistors

- Source is Vdd instead of GND
  - Vsg = (Vdd - Vin), Vsd = (Vdd -Vout), Vt is negative
- Cutoff: (Vdd-Vin) < -Vt, Ids=0
- Linear Region
  - (Vdd-Vout) < (Vdd - Vin + Vt)

  $Ids = \beta[(Vdd-Vin+Vt)(Vdd-Vout) - (Vdd-Vout)^2/2]$
- Saturated Region
  - ((Vdd - Vin) + Vt) < (Vdd - Vout)

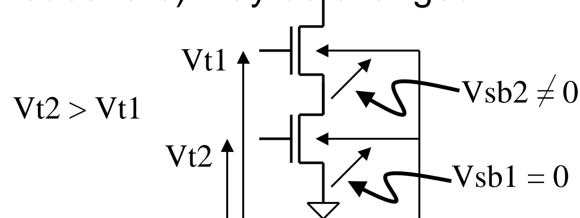
  $Ids = \beta[(Vdd - Vin + Vt)^2/2]$

-Vgs
S
G
Vsd
+
i
-
D

2nd Order Effect: Body Effect

- A second order effect that raises Vt
- Recall that Vt is affected by Vsb (voltage between source and substrate
  - Normally this is constant because of common substrate
  - But, when transistors are in series, Vsb (Vs – Vsubstrate) may be changed

Vt2 > Vt1



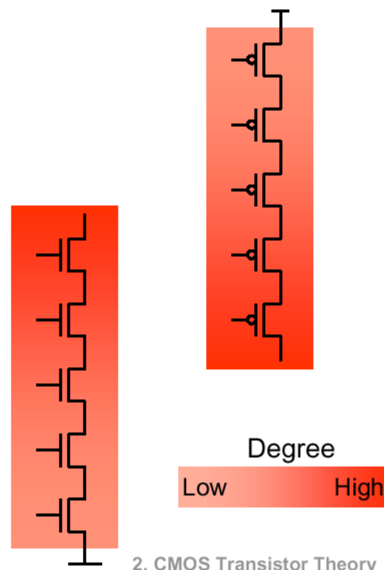
2nd Order Effect: Body Effect

- **Body Effect -**
  Vt is a function of voltage between source and substract

$$V_t = V_{to} + \gamma[\sqrt{(2\varphi_b + |V_{sb}|)} + 2\sqrt{\varphi_b}]$$

$$\varphi_b = \frac{kT}{q}\ln\left(\frac{N_A}{N_i}\right)$$

$$\gamma = \frac{t_{ox}}{\varepsilon_{ox}}\sqrt{2q\varepsilon_{si}N_A} = \frac{1}{C_{ox}}\sqrt{2q\varepsilon_{si}N_A}$$



Degree

Low High

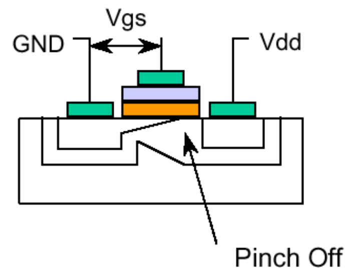
2nd Order Effect

- **Channel Length Modulation -**
  Channel length is a function of Vds. When Vds increase, the depletion region of the pinch off at drain shorten the channel length.

$$L_{eff} = L = L_{short}$$

$$L_{short} = \sqrt{2\frac{\varepsilon_{si}}{qN_A}(V_{ds} - (V_{gs} - V_t))}$$

$$Ids = \frac{kW}{2L}(V_{gs} - V_t)^2(1 + \lambda V_{ds})$$



2nd Order Effect

- **Mobility Variation -**
  The mobility of the carrier decreases when the carrier density increases. Therefore, when Vgs is large. The density of the carrier in the channel increases. As a result, the mobility decreases.

$$\mu = \frac{Average_carrier_drift_velocity(V)}{Electrical_Frield(E)}$$

$$\mu_n = 600\, cm^2 / V \cdot \sec$$

$$\mu_p = 250\, cm^2 / V \cdot \sec$$

2nd Order Effect

- **Fowler-Nordheim Tunneling**
  When the gate oxide is very thin, a current can flow from gate to source by electron tunneling through the gate oxide.

$$I_{FN} = C_1 W L E_{ox}^2 e^{\frac{-E_0}{E_{ox}}}$$

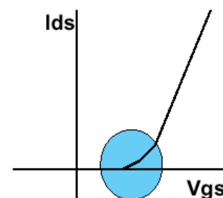
$$E_{ox} = \frac{Vgs}{t_{ox}}$$

- **Drain Punchthrough**
  When the drain voltage is high enough, the depletion region around the drain may extend to the source. Thus, causing current to flow irrespective of the gate voltage.

2nd Order Effect

- **Impact Ionization - Hot Electrons**
  When the source-drain electric field is too large, the electron speed will be high enough to break the electron-hole pair. Moreover, the electrons will penerate the gate oxide, causing a gate current.

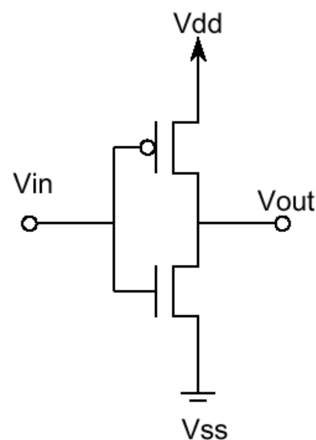
- **Subthreshold Region**
  The cutoff region is also referred to as the subthreshold region, where Ids increase exponentially with Vds and Vgs.

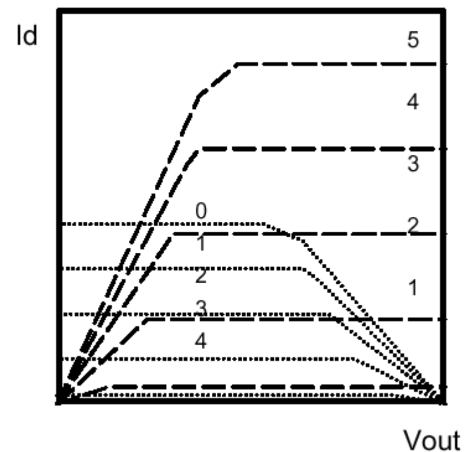


Inverter Switching Point

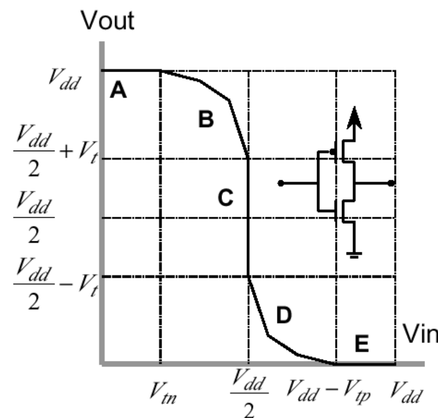
- Inverter switching point is determined by ratio of βn/βp
  - If βn/βp = 1, then switching point is Vdd/2
- If W/L of both N and P transistors are equal
  - Then βn/βp = μ_n / μ_p = electron mobility / hole mobility
  - This ratio is usually between 2 and 3
  - Means ratio of W_{ptree}/W_{ntree} needs to be between 2 and 3 for βn/βp = 1
  - For this class, we'll use $W_{ptree}/W_{ntree} = 2$

Inverter Switching Point





Inverter Operating Regions



Region	NMOS	PMOS
A	Off	-
B	Sat	Linear
C	Sat	Sat
D	Linear	Sat
E	-	Off

Gate Sizes

- Assume minimum inverter is Wp/Wn = 2/1 (L = Lmin, Wn = Wmin, Wp = 2Wn)
  - This becomes a 1x inverter
- To drive larger capacitive loads, you need more gain, more Ids
  - Increase widths to get 2x inverter
  - Wp/Wn is still 2/1, but Wp and Wn are double the size
  - For most gates, diminishing returns after about 4x size

Inverter β Ratios

Vdd

$\frac{W_p}{L_p}$

Vin

Vout

$\frac{W_n}{L_n}$

Vss

Vout

V_{dd}

$\frac{V_{dd}}{2}+V_t$

$\frac{V_{dd}}{2}$

$\frac{V_{dd}}{2}-V_t$

$\beta_n < \beta_p$

$\beta_n = \beta_p$

$\beta_n > \beta_p$

Vin

V_{tn} $\frac{V_{dd}}{2}$ $V_{dd}-V_{tp}$ V_{dd}

Inverter Noise Margin

$$NM_H = V_{OH} - V_{IH}$$
$$NM_L = V_{IL} - V_{OL}$$

Vout

V_{OH}

V_{OL}

Vin

V_{IL} V_{IH}

V_{OH}

NM_H

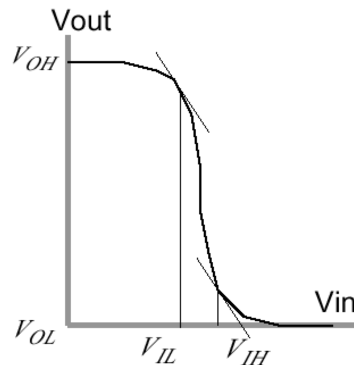
V_{IH}

V_{IL}

NM_L

V_{OL}

Inverter Noise Margin



If $\beta_n = \beta_p$ and $V_{tn} = V_{tp}$

$$V_{IH} = \frac{1}{8}(5V_{dd} - 2V_t)$$

$$V_{OH} = V_{dd}$$

$$NM_H = \frac{1}{8}(3V_{dd} + 2V_t)$$

$$V_{IL} = \frac{1}{8}(3V_{dd} + 2V_t)$$

$$V_{OH} = 0$$

$$NM_H = \frac{1}{8}(3V_{dd} + 2V_t)$$

Performance Estimation

- First we need to have a model for resistance and capacitance
  - Delays are caused (to first order) by RC delays charging and discharging capaciters
- All these layers on the chip have R and C associated with them
- Mostly this is handled in the Spectre simulator
  - But it’s good to have an idea what’s going on

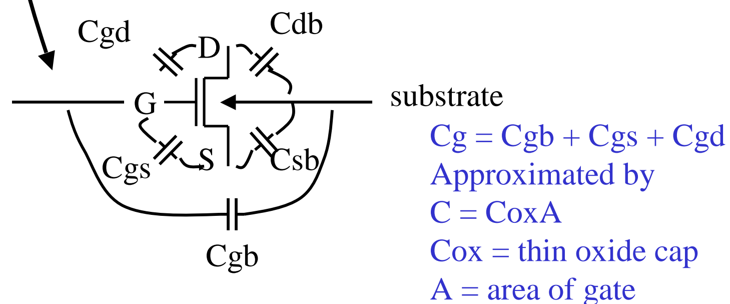
Resistance

- R = (ρ/t)(L/W) = Rs(L/W)
  - ρ = resistivity of the material
  - t = thickness
  - Rs I= sheet resistance in Ω/square
- Typical values of Rs

	Min	Typ	Max
M3	0.03	0.04	0.05
M1, M2	0.05	0.07	0.1
Poly	15	20	30
Silicide	2	3	6
Diffusion	10	25	100
Nwell	1k	2k	5k

Capacitance

- Three main forms:
  - Gate capacitance (gate of transistor)
  - Diffusion capacitance (drain regions)
  - Routing capacitance (metal, etc.)



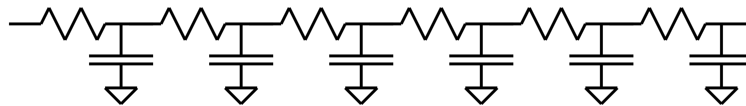
Cg = Cgb + Cgs + Cgd
Approximated by
C = CoxA
Cox = thin oxide cap
A = area of gate

Routing Capacitance

- First order effect is layer->substrate
  - Approximate using parallel plate model
  - $C = (\varepsilon/t)A$
    - ε = permittivity of insulator
    - t = thickness of insulator
    - A = area
  - Fringing fields increase effective area
- Capacitance between layers becomes very complex!
  - Crosstalk issues…

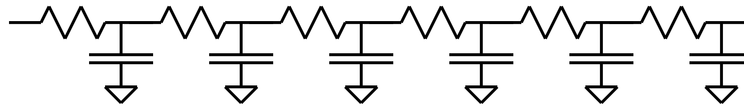
Distributed RC on Wires

- Wires look like distributed RC delays
  - Long resistive wires can look like transmission lines
  - Inserting buffers can really help delay



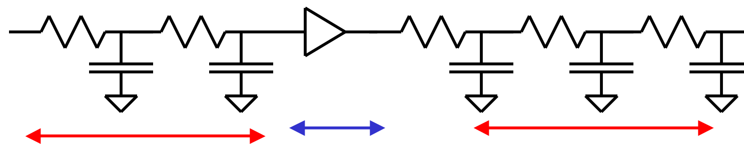
- $T_n = RCn(n+1)/2$
- $T = RCL^2/2$ as the number of segments becomes large
  - R = resistance per unit length
  - C = capacitance per unit length
  - L = length of wire

RC Wire Delay Example



- $R = 20\Omega/sq$
- $C = 4 \times 10^{-4}$ pF/um
- L = 2mm
- $T = 4 \times 10^{-15} (2000)^2$ s
  - delay = 16 ns

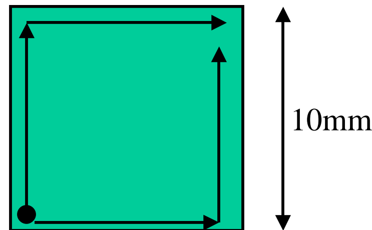
RC Wire/Buffer Delay Example



- Now split into 2 1mm segments with a buffer
- $T = 2 \times (4 \times 10^{-15} (1000)^2) +$ Tbuf
  = 8ns + Tbuf
- Assuming Tbuf is less than 8ns (which it will be), the split wire is a win

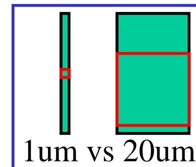
Another Example: Clock

- 50pF clock load distributed across 10mm chip in 1um metal
  - Clock length = 20mm
  - R = 0.05Ω/sq, C = 50pF/20mm
  - $T = (RC/2)L^2 = (6.25X10^{-17})(20{,}000)^2$
    $= 25ns$

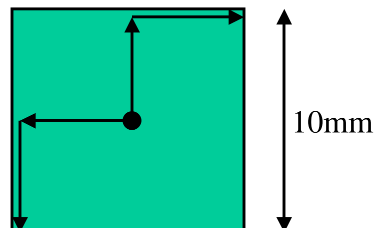


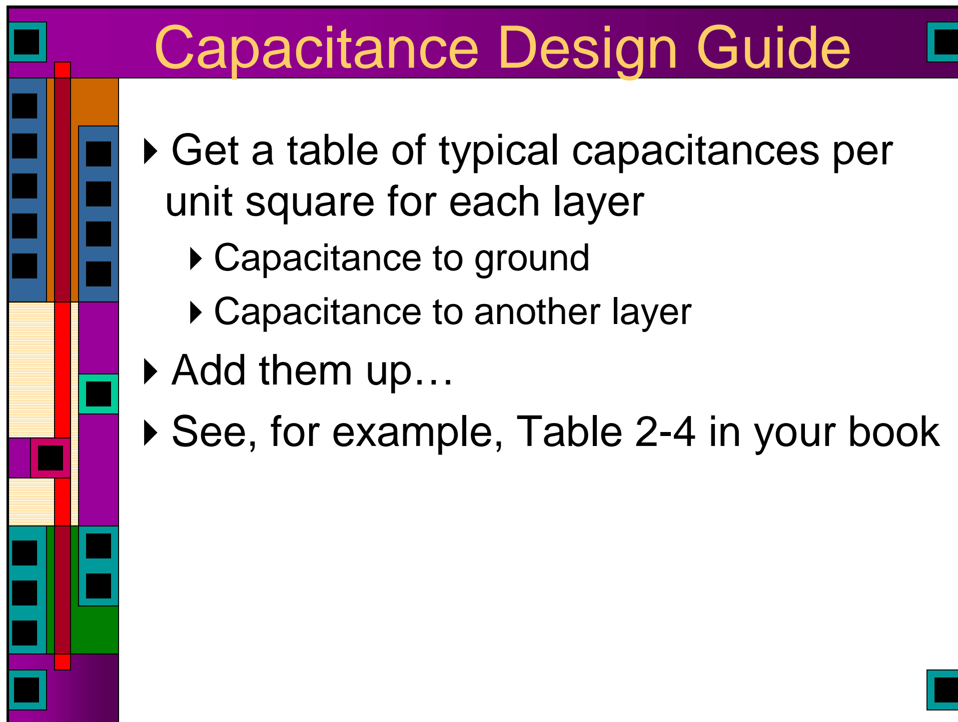
Different Distribution Scheme

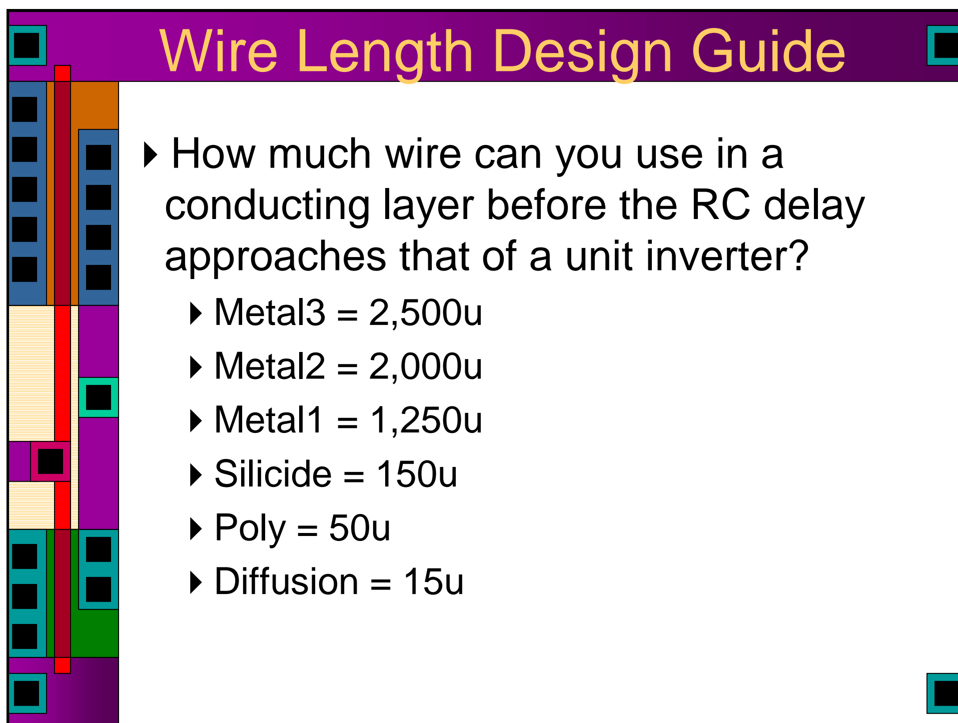
- Put clock driver in the middle of the chip
- Widen clock line to 20um wires



  - Clock length = 10mm
  - R = 0.05Ω/sq, C = 50pF/20mm
  - $T = (RC/2)L^2 = (0.31X10^{-17})(10{,}000)^2$
    $= 0.31ns$
  - Reduces R by a factor of 20, L by 2
  - Increases C a tiny bit

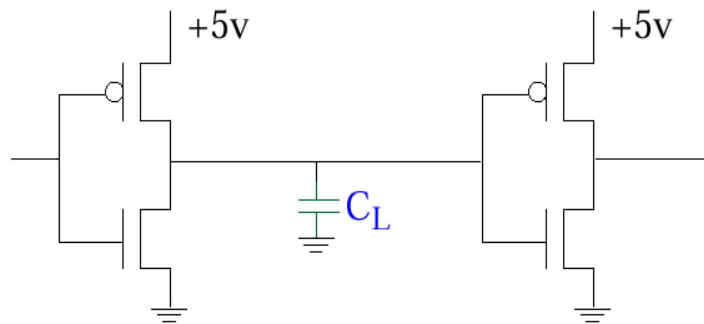



Capacitance Design Guide
Get a table of typical capacitances per unit square for each layer
Capacitance to ground
Capacitance to another layer
Add them up…
See, for example, Table 2-4 in your book

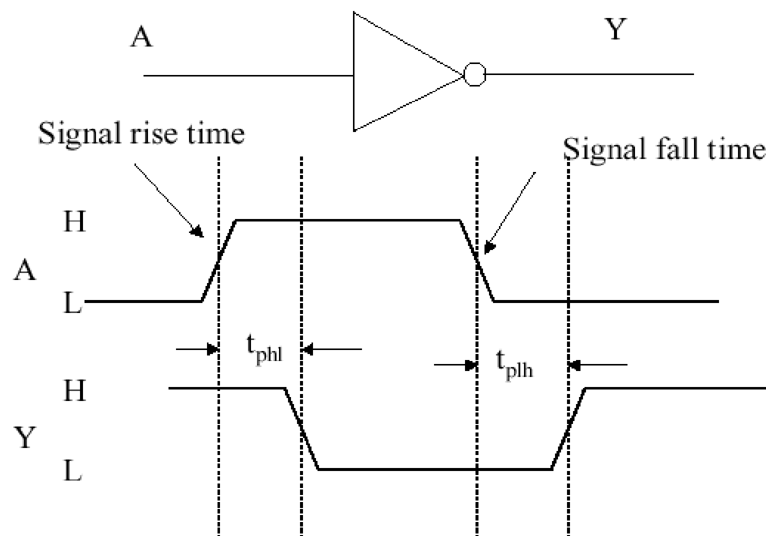

Wire Length Design Guide
How much wire can you use in a conducting layer before the RC delay approaches that of a unit inverter?
Metal3 = 2,500u
Metal2 = 2,000u
Metal1 = 1,250u
Silicide = 150u
Poly = 50u
Diffusion = 15u

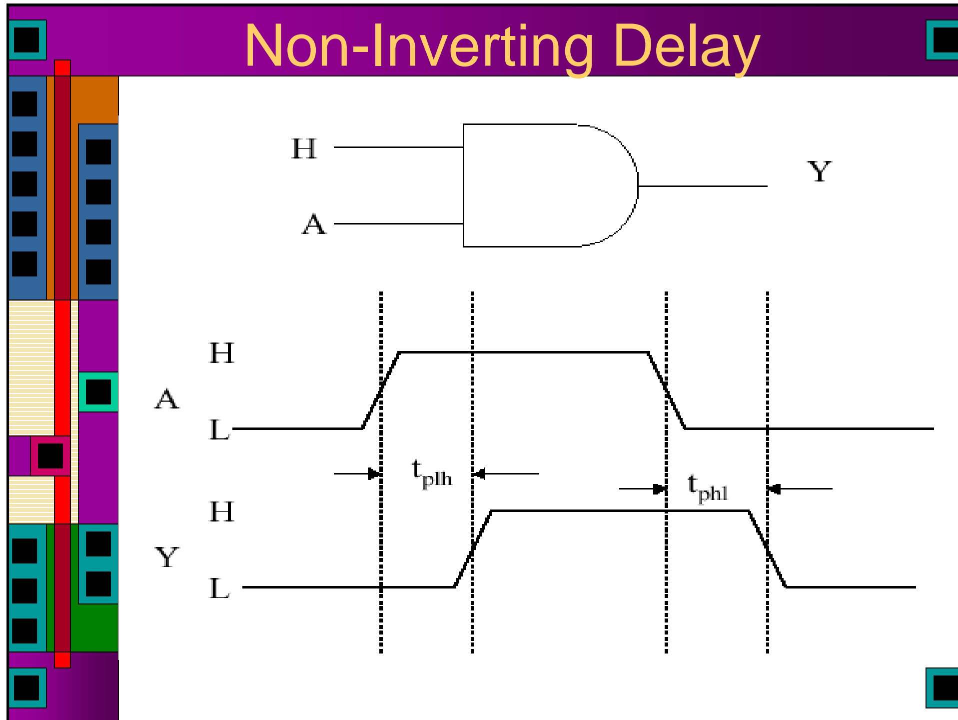
Propagation Delay

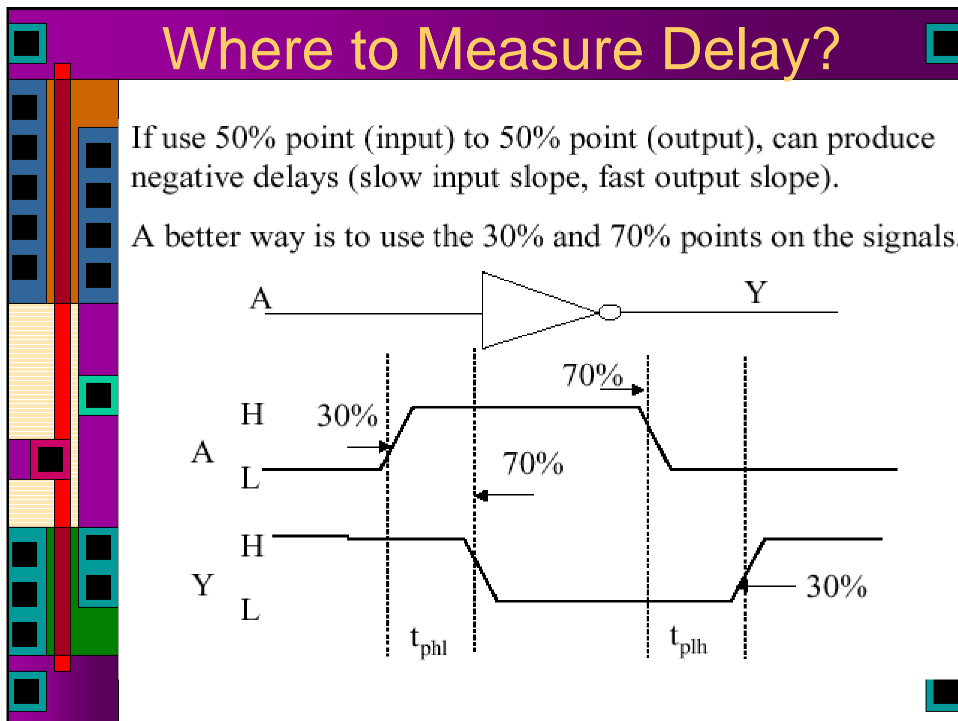
- Recall that it takes time to charge capacitors
- Recall that the gate of a transistor looks like a capacitor
- Wires have resistance and capacitance also!

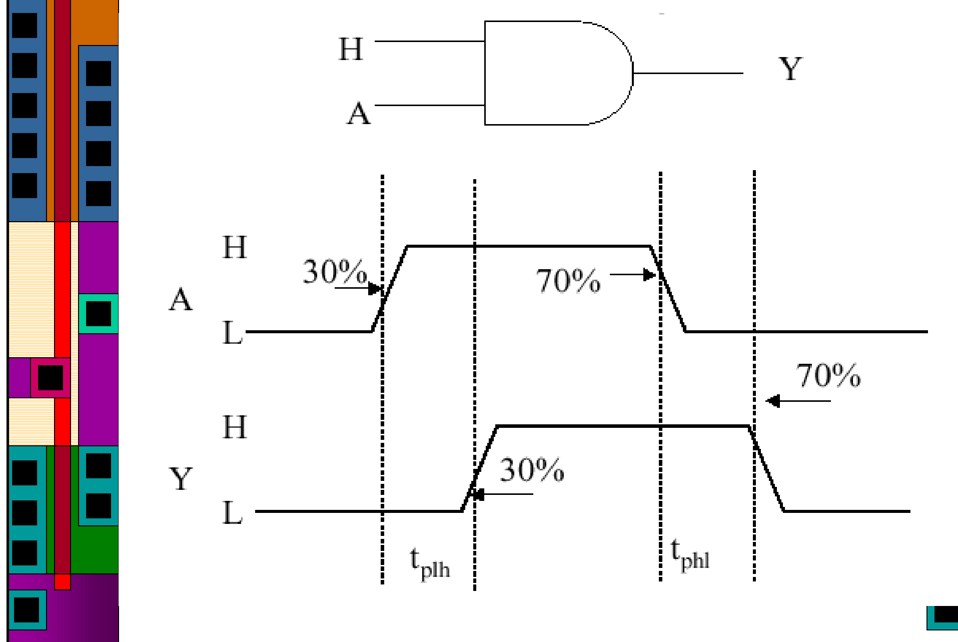


Inverting Propagation Delay




Non-Inverting Delay
H
A
Y
H
A
L
t_{plh}
t_{phl}
H
Y
L


Where to Measure Delay?
If use 50% point (input) to 50% point (output), can produce negative delays (slow input slope, fast output slope).
A better way is to use the 30% and 70% points on the signals.
A
Y
70%
H
30%
A
L
70%
H
Y
L
30%
t_{phl}
t_{plh}

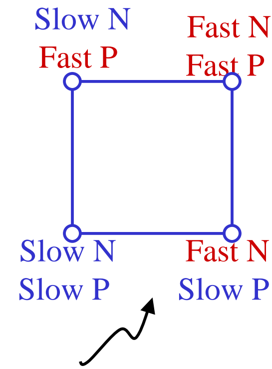


What Affects Gate Delay?

- Environment
  - Increasing Vdd decreases delay
  - Decreasing temperature decreases delay
  - Fabrication effects, fast/slow devices
- Usually measure delay for at least three cases:
  - Best - high Vdd, low temp, fast N, Fast P
  - Worst - low Vdd, high temp, slow N, Slow P
  - Typical - typ Vdd, room temp (25C), typ N, typ P

Process Corners

- When parts are specified, under what operating conditions?
- Temp: three ranges
  - Commercial: 0 C to 70 C
  - Industrial: -40 C to 85 C
  - Military: -55 C to 125 C
- Vdd: Should vary ± 10%
  - 4.5 to 5.5v for example
- Process variation:
  - Each transistor type can be slow or fast



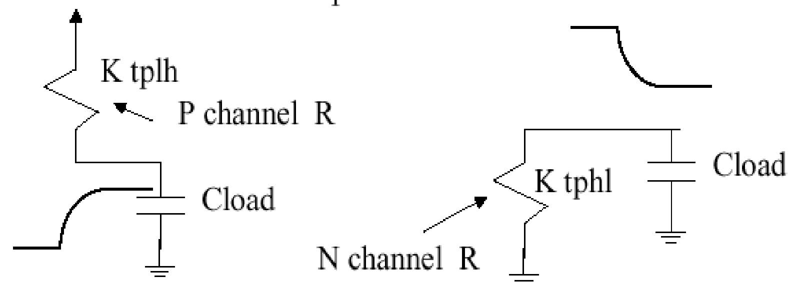
What Else Affects Gate Delay?

Input slew and output load both effect timing. For a FIXED input slope, FIXED environment, a simple timing model is:

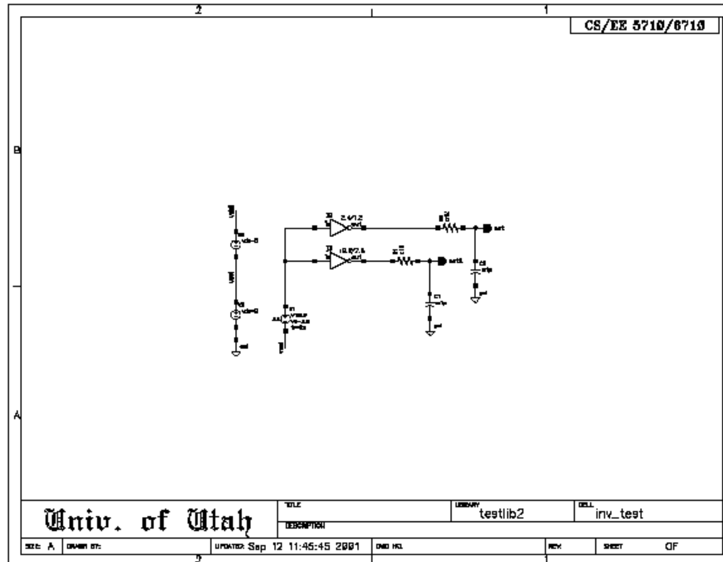
delay = Tnoload + K* Cload

Tnoload is the delay of the gate with no external load.

K is different for TPLH, TPHL since it represents the channel resistance. Same equation is used for Slew values.

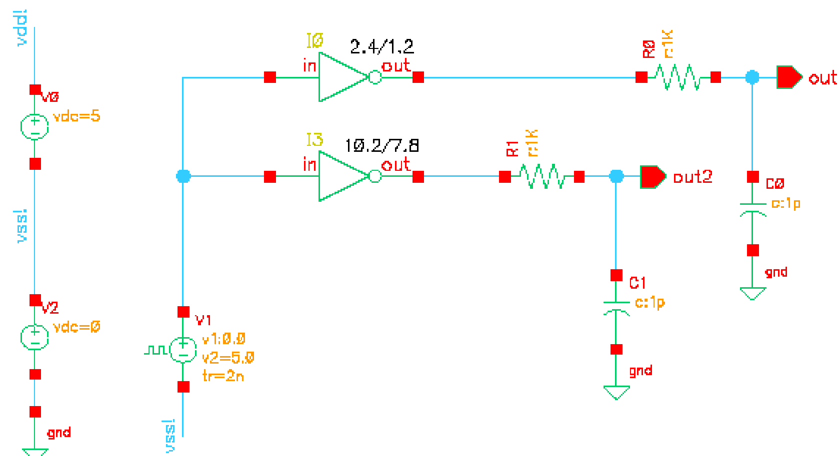


Inv_Test Schematic

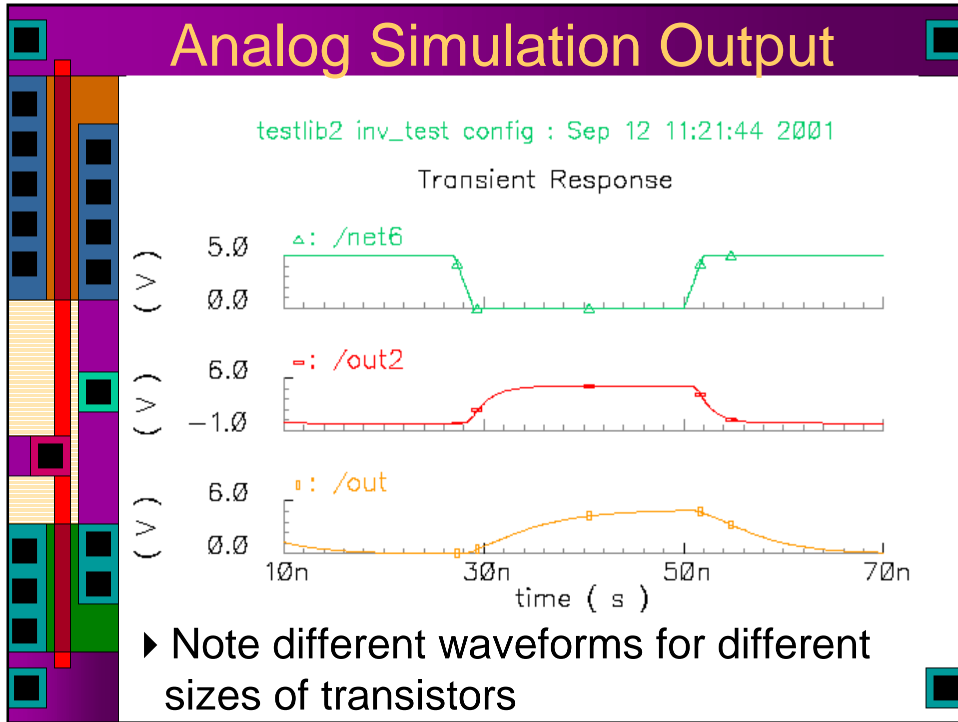


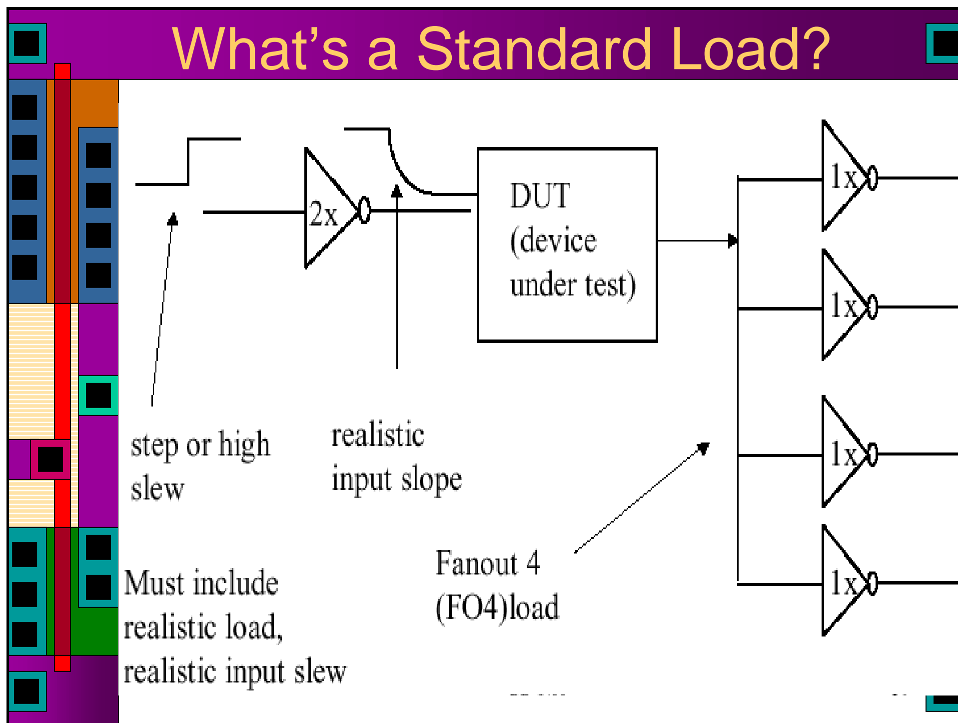
- From CAD #2

Closeup of Inv-Test



- Note the sizes I used for this example…


Analog Simulation Output
testlib2 inv_test config : Sep 12 11:21:44 2001
Transient Response
5.0
0.0
/net6
6.0
-1.0
/out2
6.0
0.0
/out
10n
30n
50n
70n
time (s)
Note different waveforms for different sizes of transistors


What's a Standard Load?
2x
DUT (device under test)
1x
1x
1x
1x
step or high slew
realistic input slope
Fanout 4 (FO4)load
Must include realistic load, realistic input slew

What About Gates in Series

- Basically we want every gate to have the delay of a “standard inverter”
  - Standard inverter starts with 2/1 P/N ratio
- Gates in series? Sum the conductance to get the series conductance
- $\beta n\text{-eff} = 1/(1/\beta 1 + 1/\beta 2 + 1/\beta 3)$
  - $\beta n\text{-eff} = \beta n/3$
- Effect is like increasing L by 3
  - Compensate by increasing W by 3