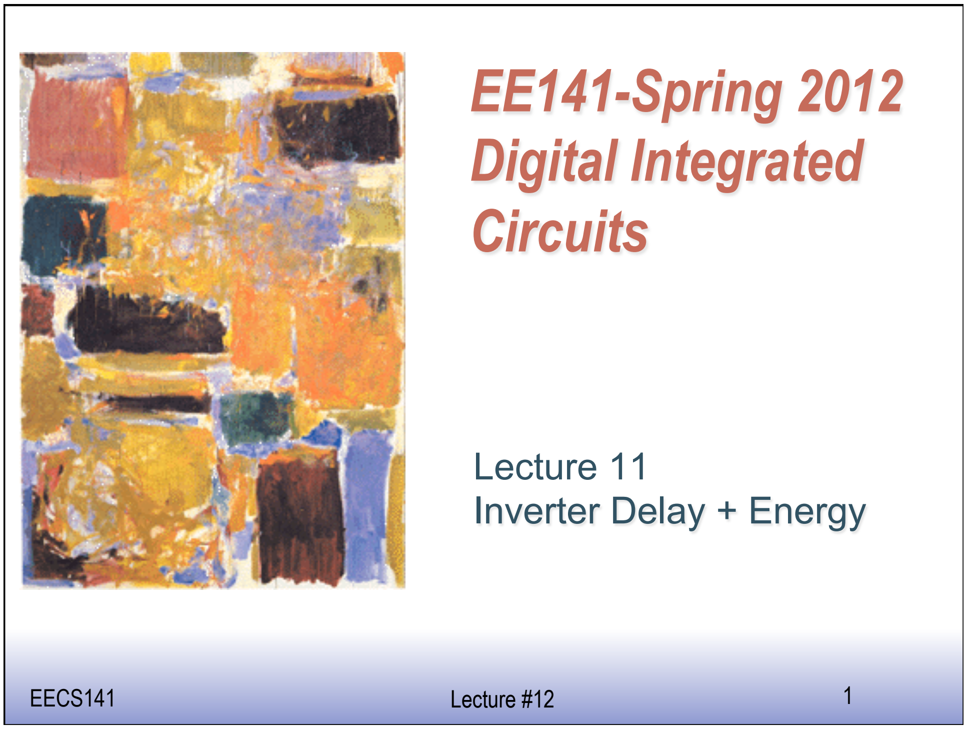
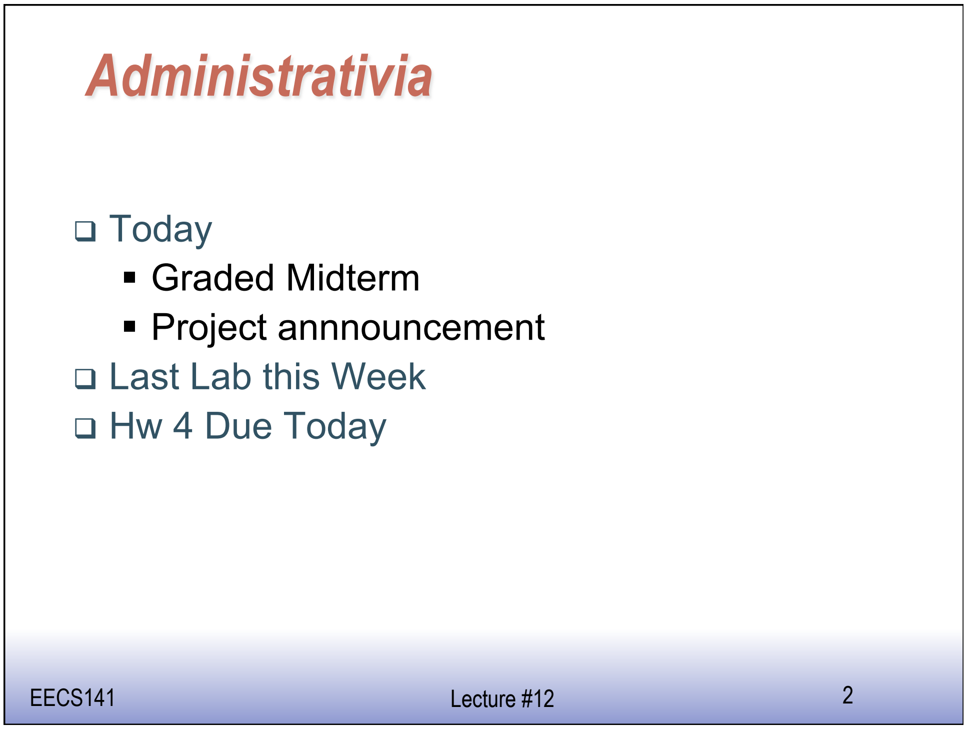


EE141-Spring 2012 Digital Integrated Circuits

Lecture 11
Inverter Delay + Energy

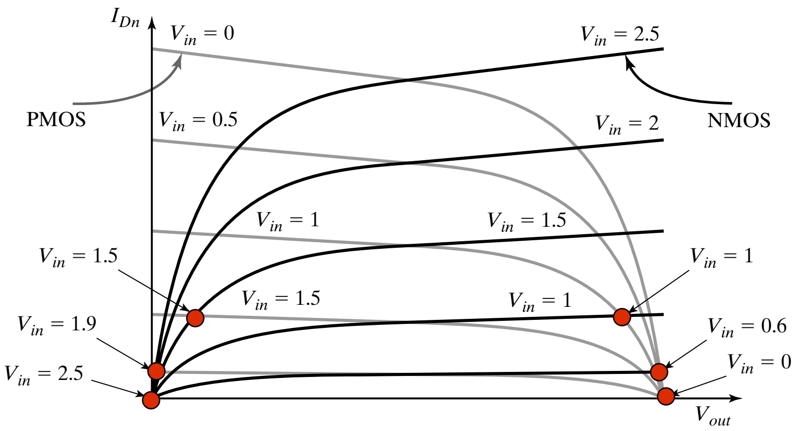


Administrativia

- ❑ Today
  - ▪ Graded Midterm
  - ▪ Project annnouncement
- ❑ Last Lab this Week
- ❑ Hw 4 Due Today

So the key challenge at any operational point is to determine what operation mode each of the transistors is. Because that's going to determine the equations you can write down and what you have to solve.

CMOS Inverter Load Characteristics

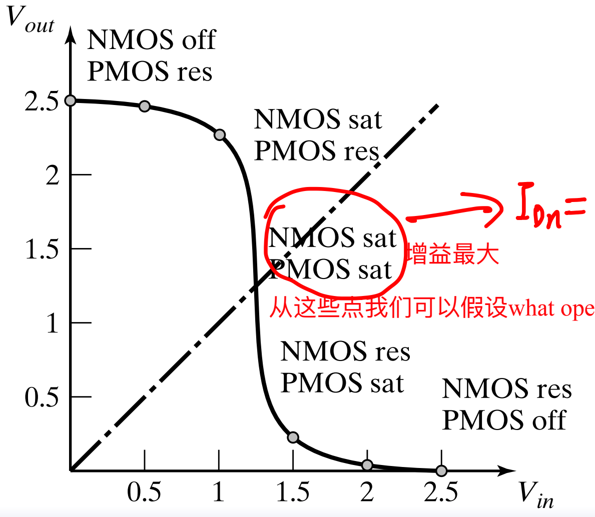


利用这些红点，我们可以从operational perspective看到不同的transition。

EECS141 Lecture #13 3

res=restive=linear

CMOS Inverter VTC



$\longrightarrow I_{Dn} = -I_{Dp}$

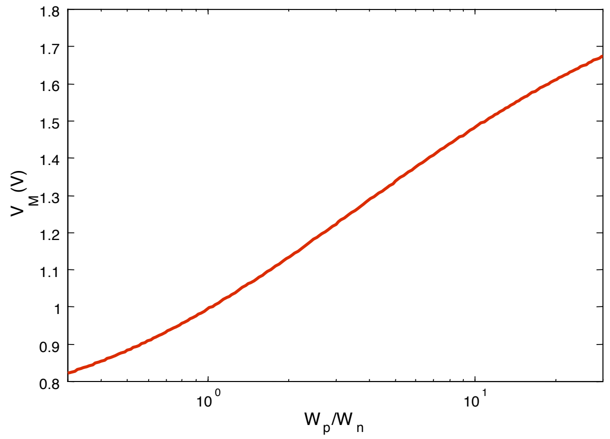
增益最大

写方程，求导数，解方程，得到gain等参数，其它点同理。

从这些点我们可以假设what operational modes of my transistor should be.

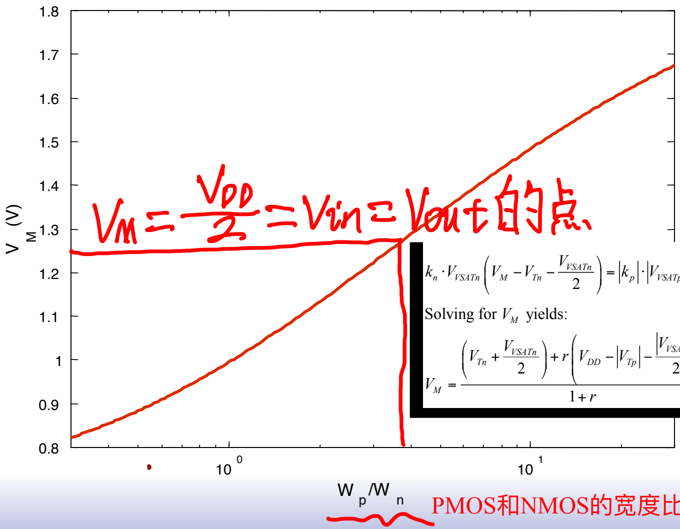
EECS141 Lecture #13 4

Switching Threshold as a Function of Transistor Ratio



$$I_{dn}(V_M) = I_{dp}(V_M)$$

Switching Threshold as a Function of Transistor Ratio



$V_M = \frac{V_{DD}}{2} = V_{in} = V_{out}$的点

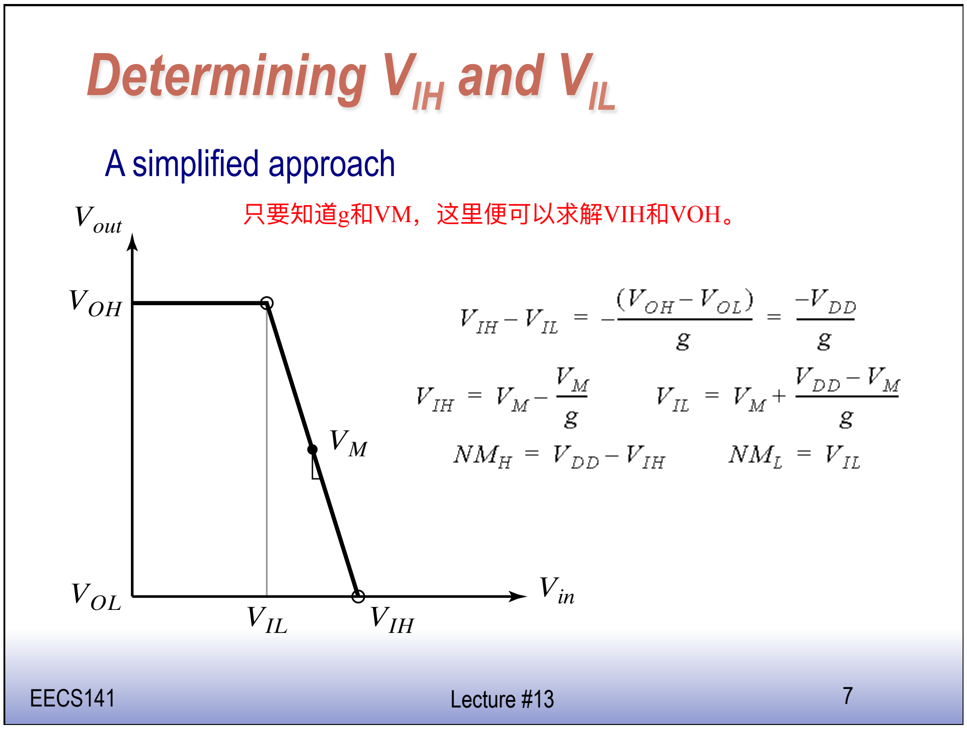
$$k_n \cdot V_{VSATn}\left(V_M - V_{Tn} - \frac{V_{VSATn}}{2}\right) = |k_p| \cdot |V_{VSATp}|\left(V_{DD} - V_M - |V_{Tp}| - \frac{|V_{VSATp}|}{2}\right)$$

Solving for V_M yields:

$$V_M = \frac{\left(V_{Tn} + \frac{V_{VSATn}}{2}\right) + r\left(V_{DD} - |V_{Tp}| - \frac{|V_{VSATp}|}{2}\right)}{1 + r} \text{ with } r = \frac{|k_p| \cdot |V_{VSATp}|}{k_n \cdot V_{VSATn}} = \frac{W_p \upsilon_{satp}}{W_n \upsilon_{satn}}$$

PMOS和NMOS的宽度比

因为从定义求导从而解得VIH和VIL太过复杂，也不利于观察特性，所以需要使用比较简单的模型。



Determining V_{IH} and V_{IL}

A simplified approach

只要知道g和VM，这里便可以求解VIH和VOH。

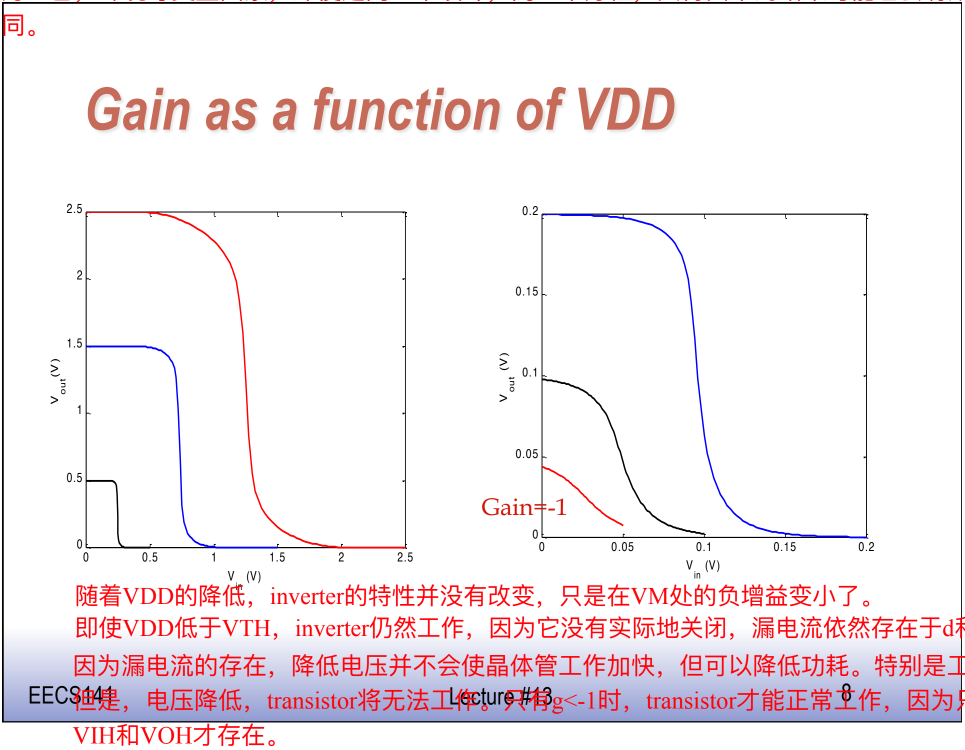
$$V_{IH} - V_{IL} = -\frac{(V_{OH} - V_{OL})}{g} = \frac{-V_{DD}}{g}$$

$$V_{IH} = V_M - \frac{V_M}{g} \qquad V_{IL} = V_M + \frac{V_{DD} - V_M}{g}$$

$$NM_H = V_{DD} - V_{IH} \qquad NM_L = V_{IL}$$

EECS141 Lecture #13 7

因为工艺，环境等变量因素，即使是同一个设计，同一个方程，其得出来的结果可能也会有所不同。



Gain as a function of VDD

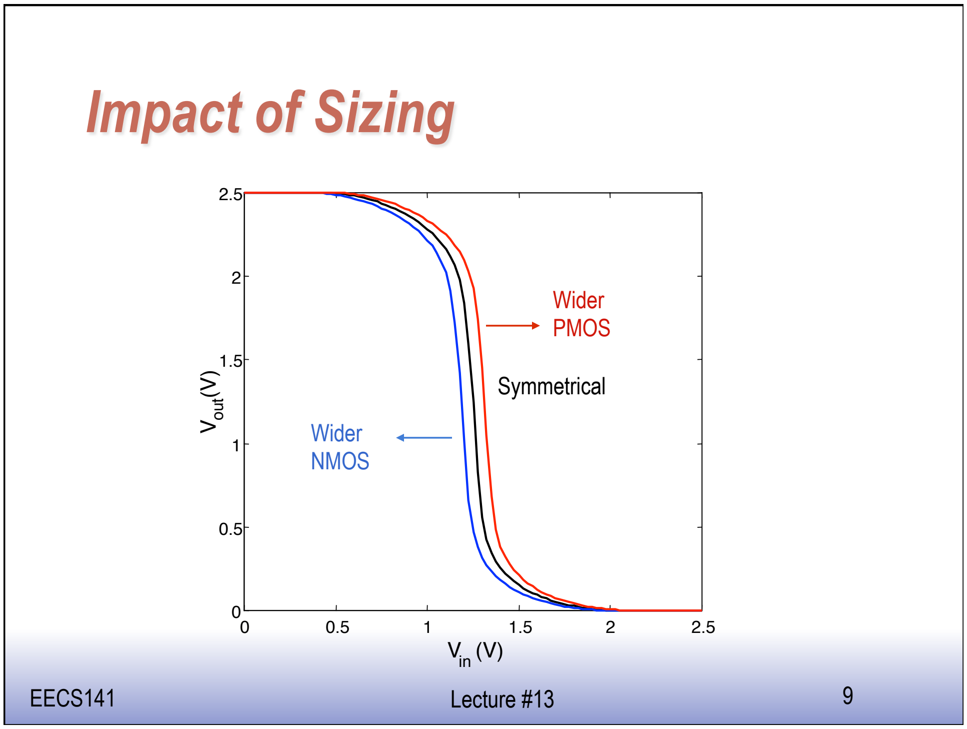
随着VDD的降低，inverter的特性并没有改变，只是在VM处的负增益变小了。

即使VDD低于VTH，inverter仍然工作，因为它没有实际地关闭，漏电流依然存在于d和s之间。

因为漏电流的存在，降低电压并不会使晶体管工作加快，但可以降低功耗。特别是工作在VTH以下。

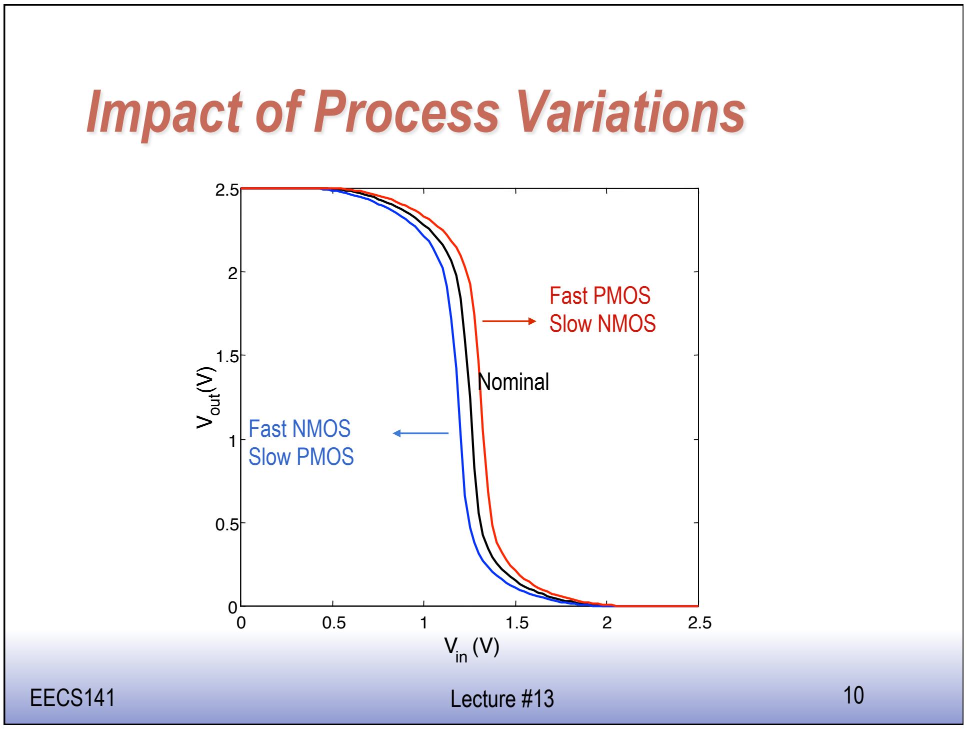
EECS141 Lecture #13 8

但是，电压降低，transistor将无法工作。只有g<-1时，transistor才能正常工作，因为只有这样子，VIH和VOH才存在。



Impact of Sizing

EECS141 Lecture #13 9

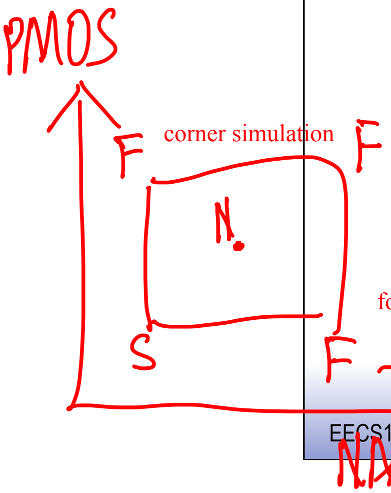


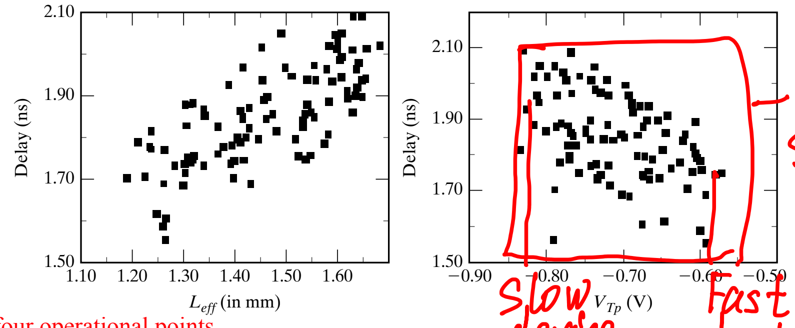
Impact of Process Variations

EECS141 Lecture #13 10

Process Variations

Not all transistors are alike
Impacts parameters such as reliability and performance





corner simulation

slow device

fast device

When doing simulating, not only give a few points about delay which is meaningless. We should give many points like these two picture.

four operational points

Define process corners: SS, FF, SF, FS

F = Fast N = Normal

S = Slow

PMOS is fast means PMOS wider,NMOS is fast means NMOS wider.

不仅仅改变电压，还改变speed和耗电。

MIDTERM 1

PROJECT

SRAM ==static random access memory

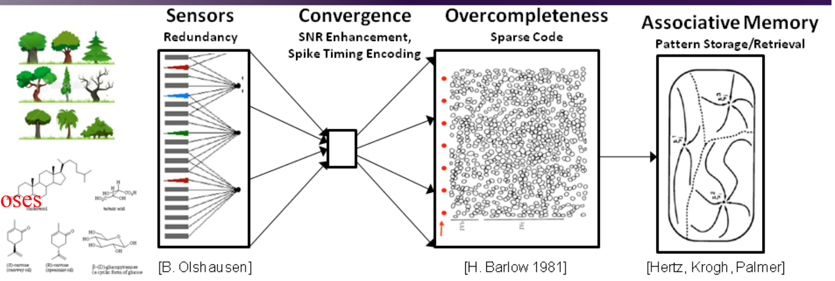
Goal: An Neural Associative Memory

- What is a memory?
  - When applying an address, returns the data stored at that address
- What is an associative memory?
  - When applying data, returns the address at which the data is stored
- What is an associative neural memory?
  - When applying data, returns the address for which the data is the closest match to the applied data

why is neural?
because what it matches is not necessary to be the same things

EECS141 Lecture #13 15

The Sensory Pathway



[B. Olshausen] [H. Barlow 1981] [Hertz, Krogh, Palmer]

These are nodes/noses which have a lot of sensors.

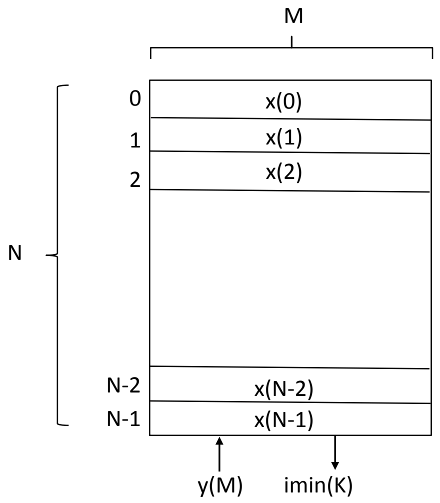
	Sensors (Redundancy)	Convergence	Overcompleteness (Sparse Code)	Associative Memory (Pattern Storage/Retrieval)
Visual	Retina	Ganglion Cells/ LGN	Primary Visual Cortex (V1)	Higher-level Cortex
Olfactory	Olfactory Epithelium (OE)	Olfactory Bulb (OB)	Primary Olfactory Cortex	Higher-level Cortex

EECS141 Lecture #13 16

General Description

- ❑ Given a vector *x*[N] of integers with wordlength M stored in SRAM memory (of size NxM).
- ❑ Realize the following associative function: For an input y, find the value of i (i = 0 … N-1) for which holds that |y-x[i]| is minimum.
- ❑ Such that area/energy is minimized

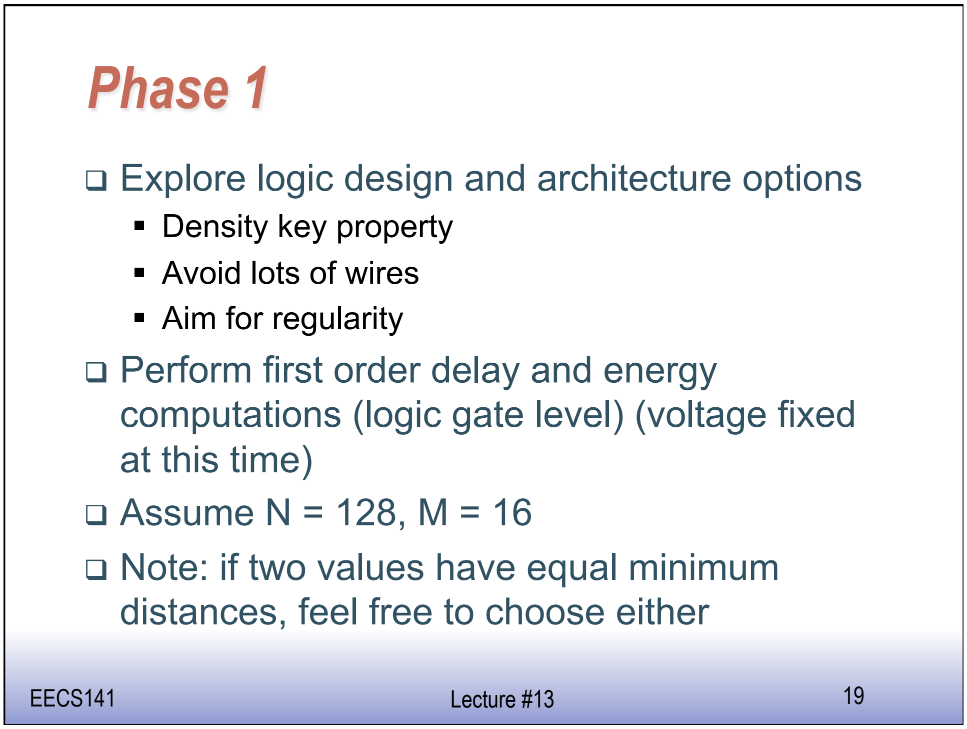
Overall Block Diagram



d[i] = |*x[i]-y*|

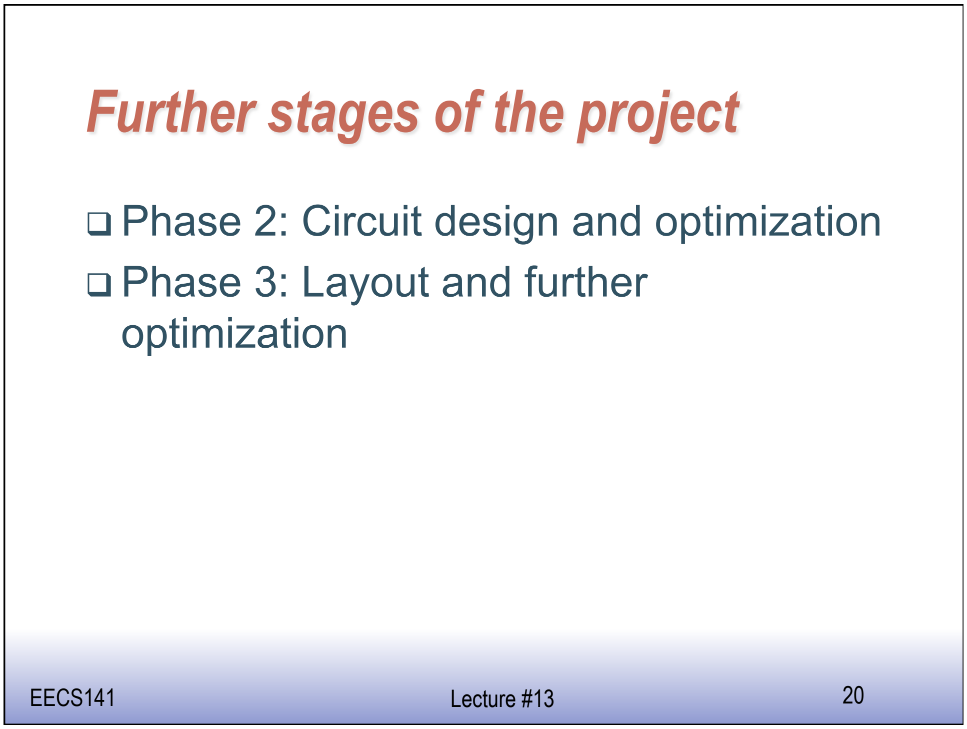
imin: min(d[i]) over all i [0..N-1]

$K = \log_2 (N)$



Phase 1

- ❑ Explore logic design and architecture options
  - Density key property
  - Avoid lots of wires
  - Aim for regularity
- ❑ Perform first order delay and energy computations (logic gate level) (voltage fixed at this time)
- ❑ Assume N = 128, M = 16
- ❑ Note: if two values have equal minimum distances, feel free to choose either



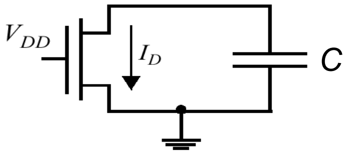
Further stages of the project

- ❑ Phase 2: Circuit design and optimization
- ❑ Phase 3: Layout and further optimization

CMOS Switching Delay (Resistance)

MOS Transistor as a Switch

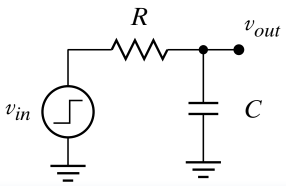
- Discharging a capacitor



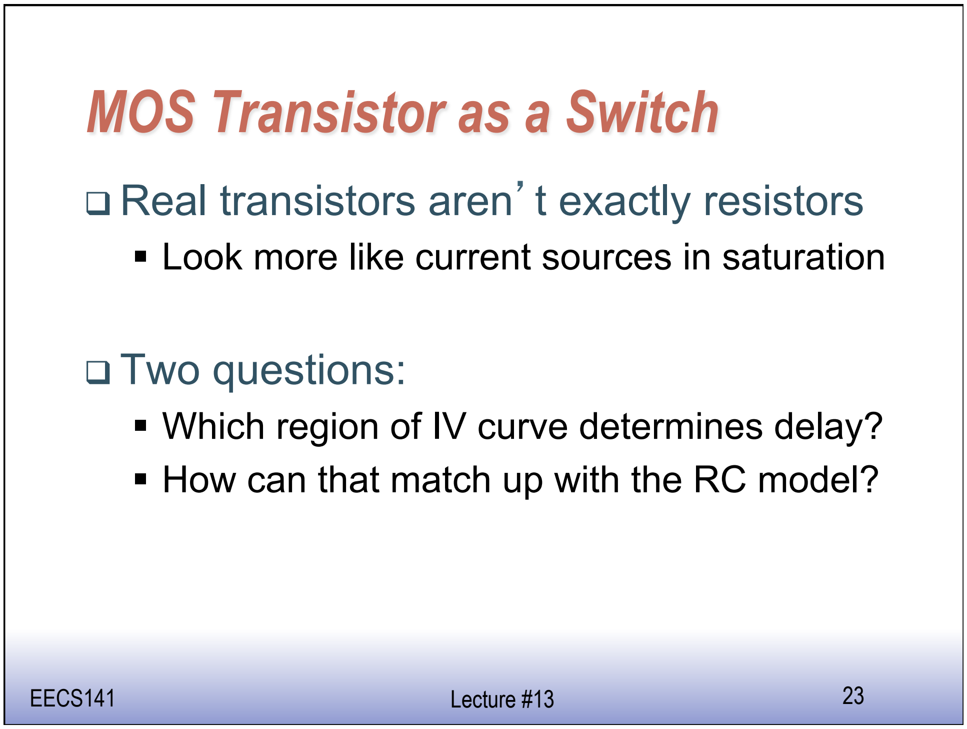
$$i_D = i_D(v_{DS})$$

$$i_D = C\frac{dV_{DS}}{dt}$$

- We modeled this with:

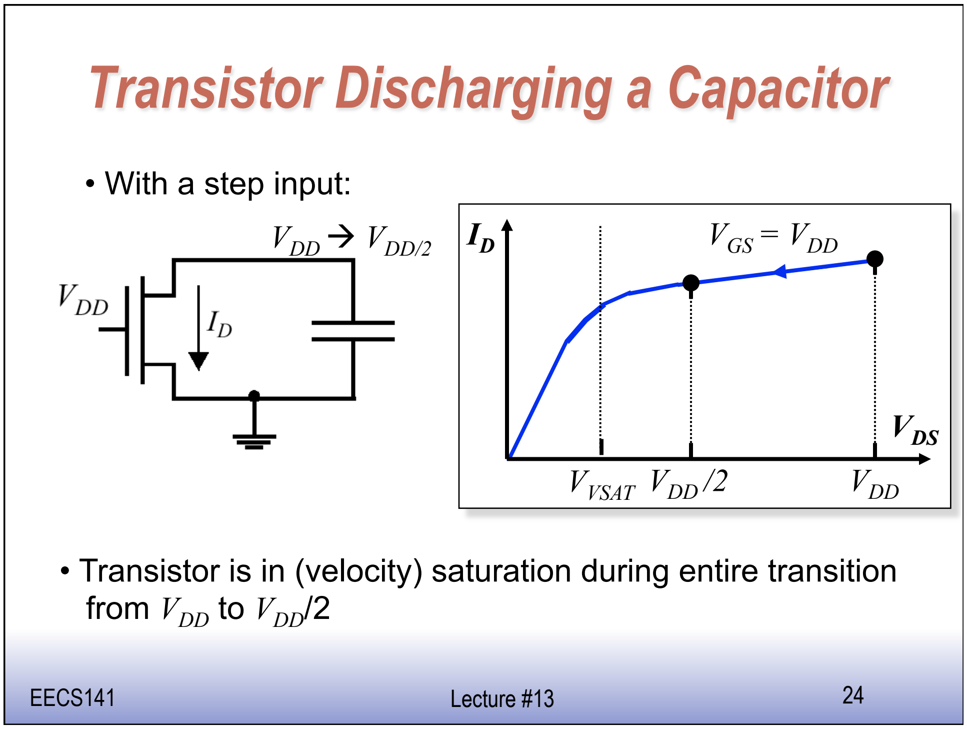


t_p = ln (2) RC



MOS Transistor as a Switch

- ❑ Real transistors aren' t exactly resistors
  - Look more like current sources in saturation
- ❑ Two questions:
  - Which region of IV curve determines delay?
  - How can that match up with the RC model?



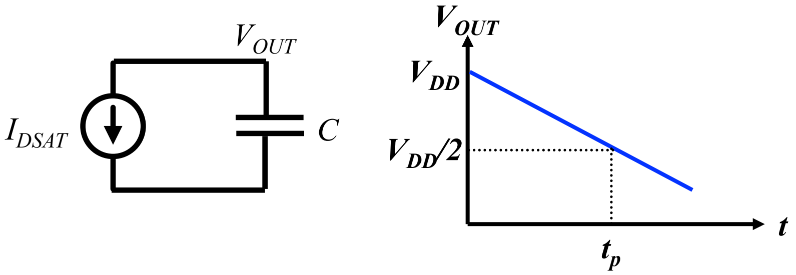
Transistor Discharging a Capacitor

- With a step input:

- Transistor is in (velocity) saturation during entire transition from V_{DD} to $V_{DD}/2$

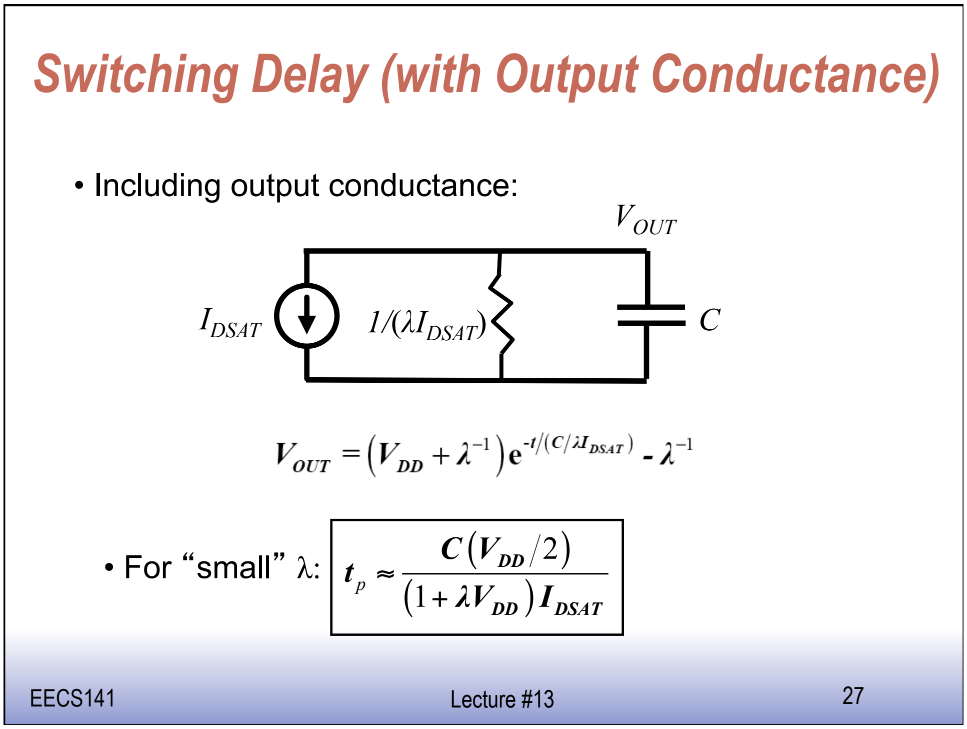
Switching Delay

- In saturation, transistor basically acts like a current source:



$$V_{OUT} = V_{DD} - (I_{DSAT}/C)t \longrightarrow \boxed{t_p = C(V_{DD}/2)/I_{DSAT}}$$

Defining IDSAT



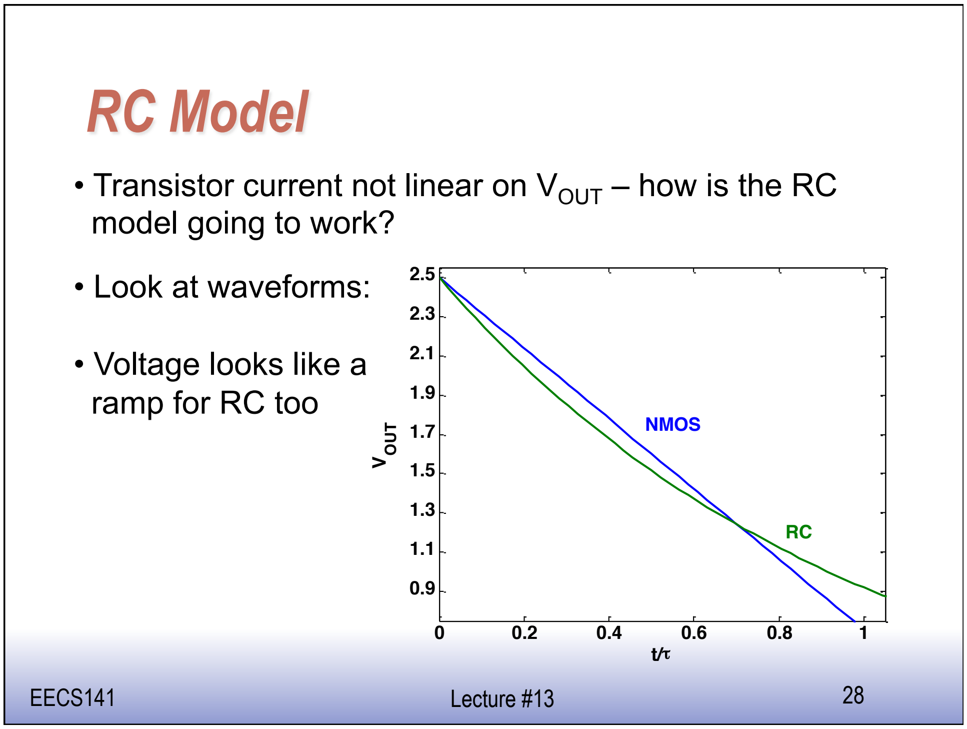
Switching Delay (with Output Conductance)

- Including output conductance:

V_{OUT}
I_{DSAT}
$1/(\lambda I_{DSAT})$
C

$$V_{OUT} = \left(V_{DD} + \lambda^{-1}\right) e^{-t/(C/\lambda I_{DSAT})} - \lambda^{-1}$$

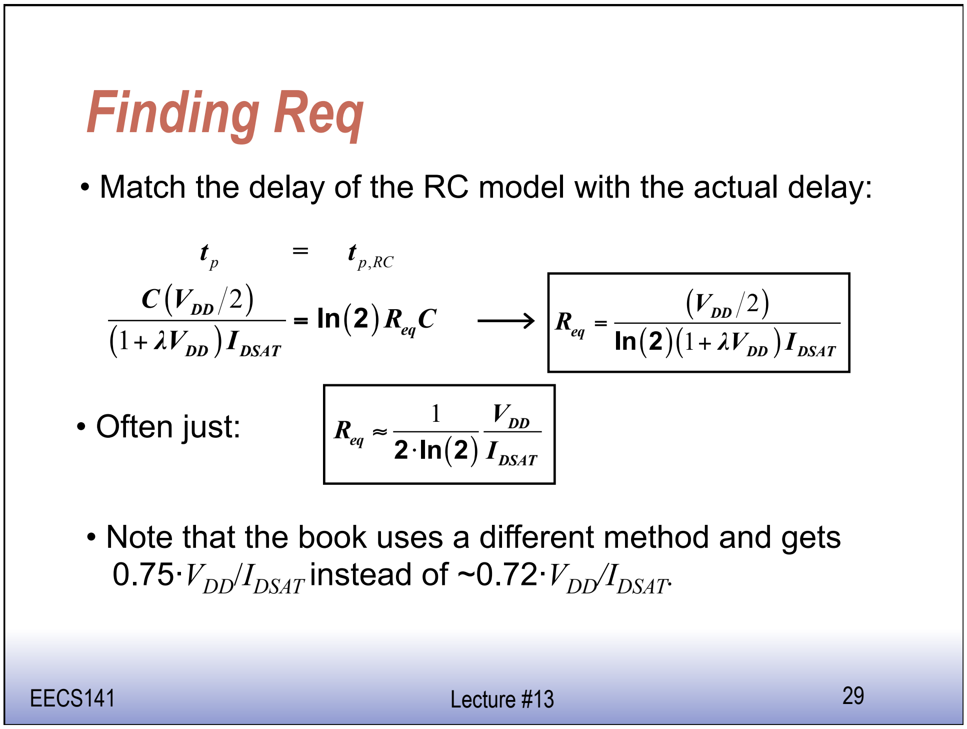
- For "small" λ: $\boxed{t_p \approx \frac{C\left(V_{DD}/2\right)}{\left(1+\lambda V_{DD}\right) I_{DSAT}}}$



RC Model

- Transistor current not linear on V_{OUT} – how is the RC model going to work?
- Look at waveforms:
- Voltage looks like a ramp for RC too

2.5
2.3
2.1
1.9
1.7
1.5
1.3
1.1
0.9
V_{OUT}
0
0.2
0.4
0.6
0.8
1
t/τ
NMOS
RC



Finding Req

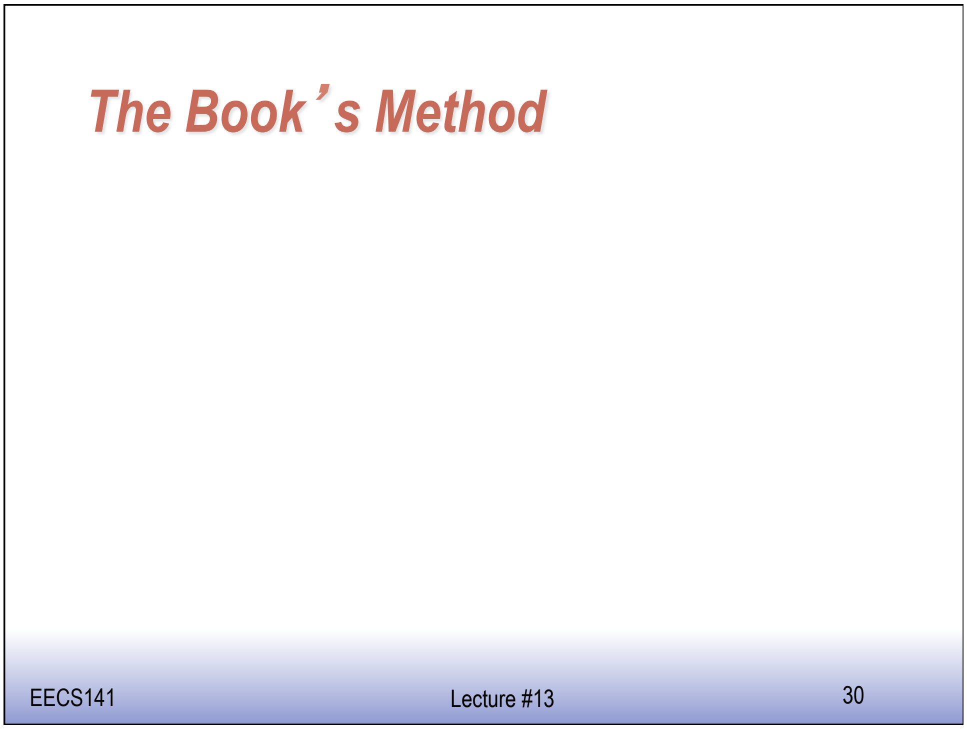
- Match the delay of the RC model with the actual delay:

$$t_p \quad = \quad t_{p,RC}$$

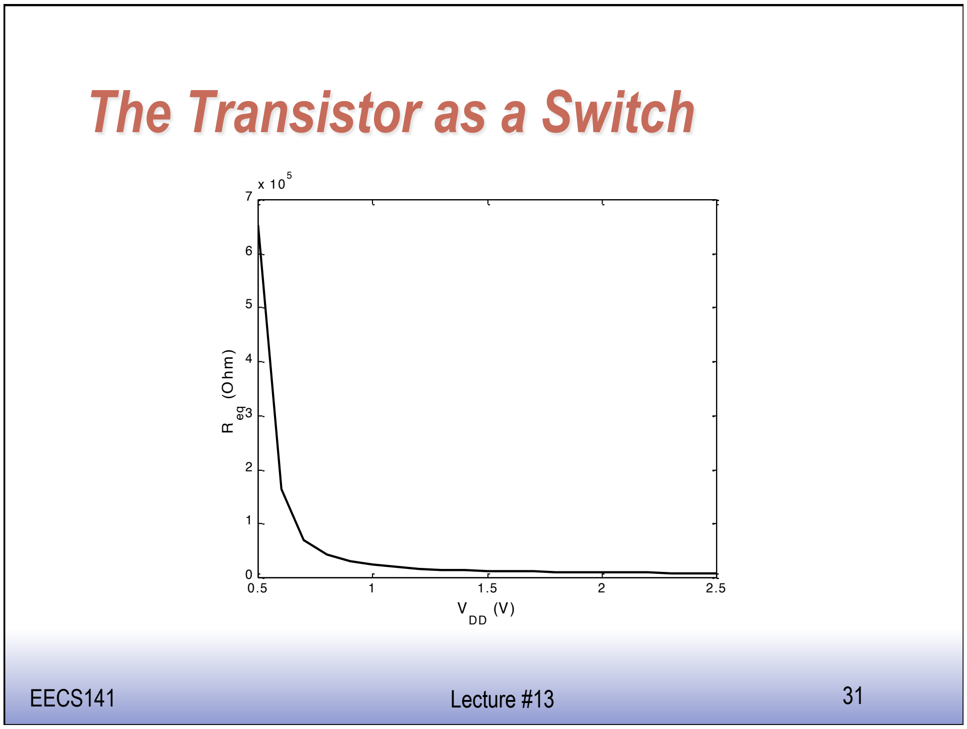
$$\frac{C(V_{DD}/2)}{(1+\lambda V_{DD})I_{DSAT}} = \ln(2)R_{eq}C \longrightarrow \boxed{R_{eq} = \frac{(V_{DD}/2)}{\ln(2)(1+\lambda V_{DD})I_{DSAT}}}$$

- Often just: $\boxed{R_{eq} \approx \frac{1}{2\cdot\ln(2)}\frac{V_{DD}}{I_{DSAT}}}$

- Note that the book uses a different method and gets $0.75\cdot V_{DD}/I_{DSAT}$ instead of ~$0.72\cdot V_{DD}/I_{DSAT}$.

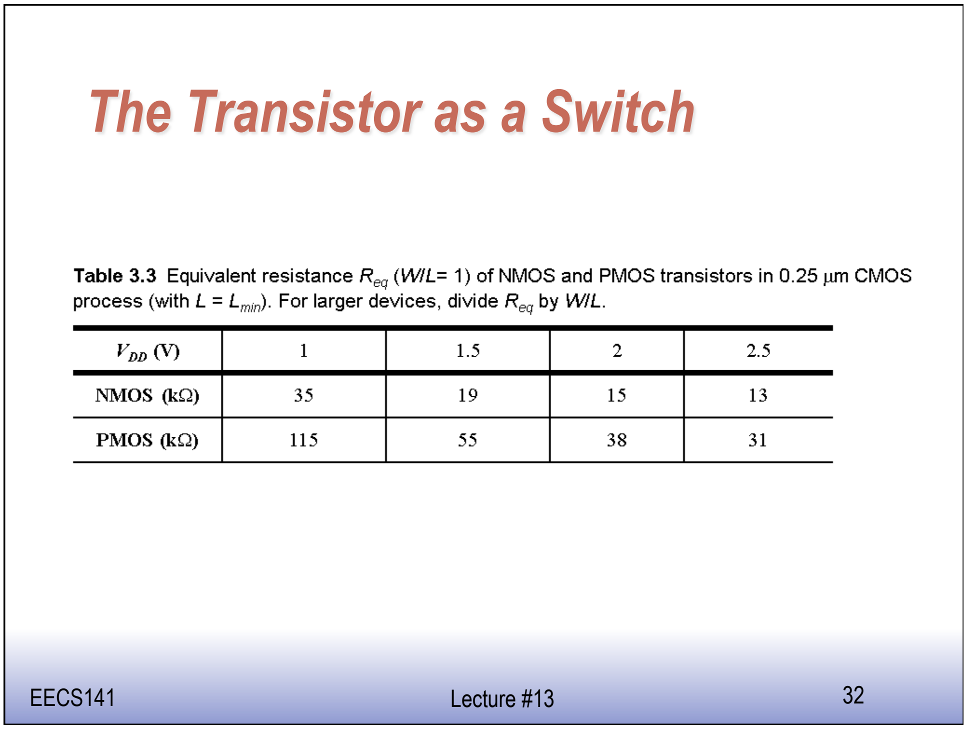


The Book's Method



The Transistor as a Switch

EECS141 Lecture #13 31

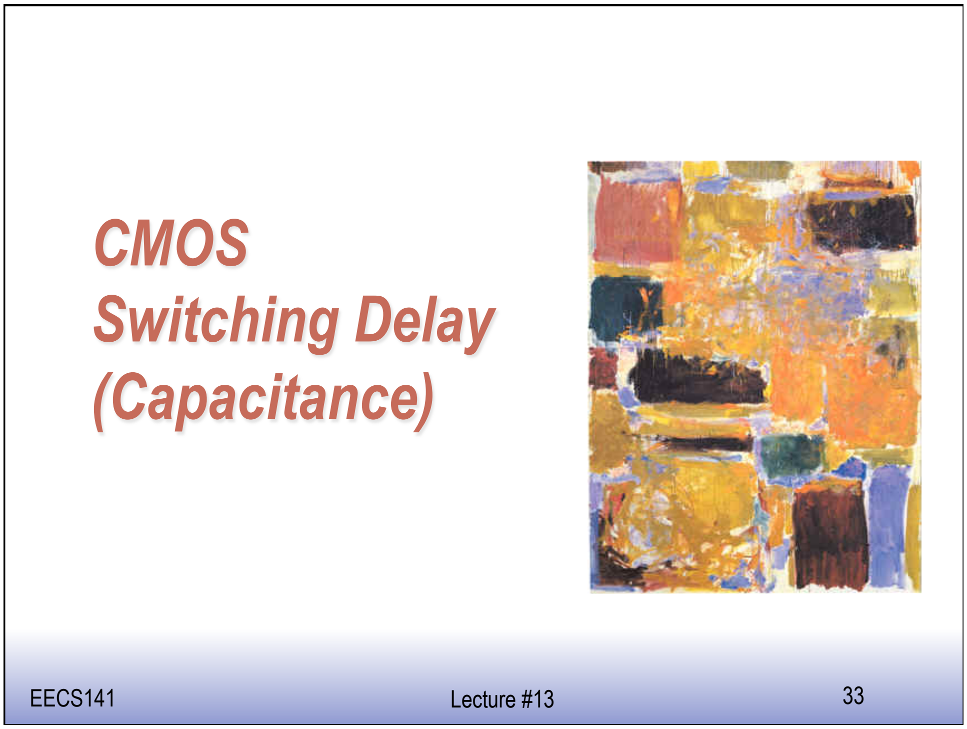


The Transistor as a Switch

Table 3.3 Equivalent resistance R_{eq} (W/L= 1) of NMOS and PMOS transistors in 0.25 μm CMOS process (with $L = L_{min}$). For larger devices, divide R_{eq} by W/L.

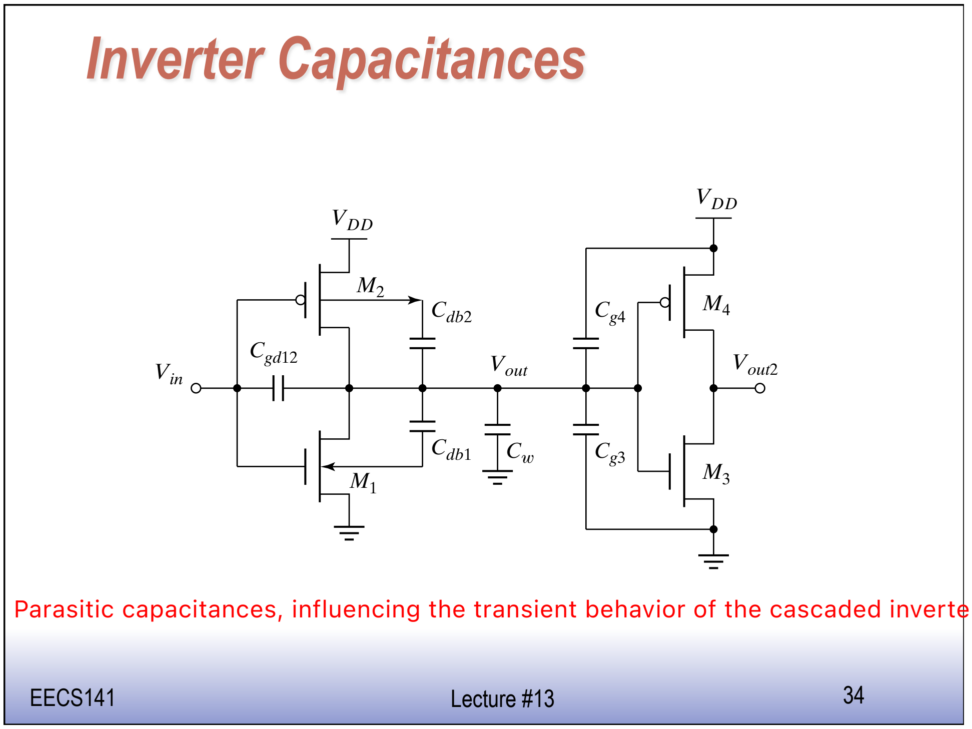
V_{DD} (V)	1	1.5	2	2.5
NMOS (kΩ)	35	19	15	13
PMOS (kΩ)	115	55	38	31

EECS141 Lecture #13 32



CMOS Switching Delay (Capacitance)

To make the analysis tractable, we assume that all capacitances are lumped together into one single capacitor CL , located between Vout and GND.



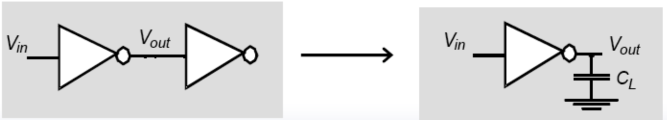
Inverter Capacitances

Parasitic capacitances, influencing the transient behavior of the cascaded inverter pair

This picture includes all the capacitances influencing the transient response of node Vout . It is initially assumed that the input Vin is driven by an ideal voltage source with zero rise and fall times. Accounting only for capacitances connected to the output node, CL breaks down into the following components.

Inverter Capacitance Model

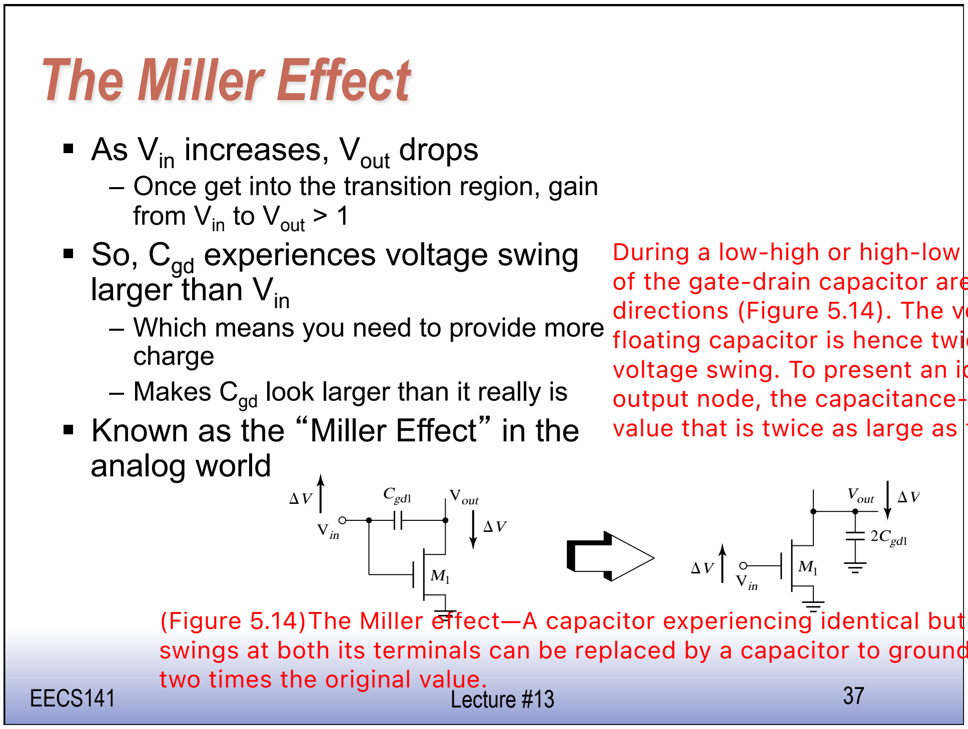
- Capacitance models important for analysis and intuition
  - But often need something simpler to work with
- Simpler model:
  - Lump together as effective linear capacitance to (ac) ground
  - In most processes: $Cg = Cd = 1.5 – 2\text{fF}\cdot W(\mu m)$



Lumping the Caps

M1 and M2 are either in cut-off or in the saturation mode during the first half (up to 50% point) of the output transient. Under these circumstances, the only contributions to Cgd12 are the overlap capacitances of both M1 and M2. The channel capacitance of the MOS transistors does not play a role here, as it is located either completely between gate and bulk (cut-off) or gate and source (saturation) (see Chapter 3).

The lumped capacitor model now requires that this floating gate-drain capacitor be replaced by a capacitance-to-ground.



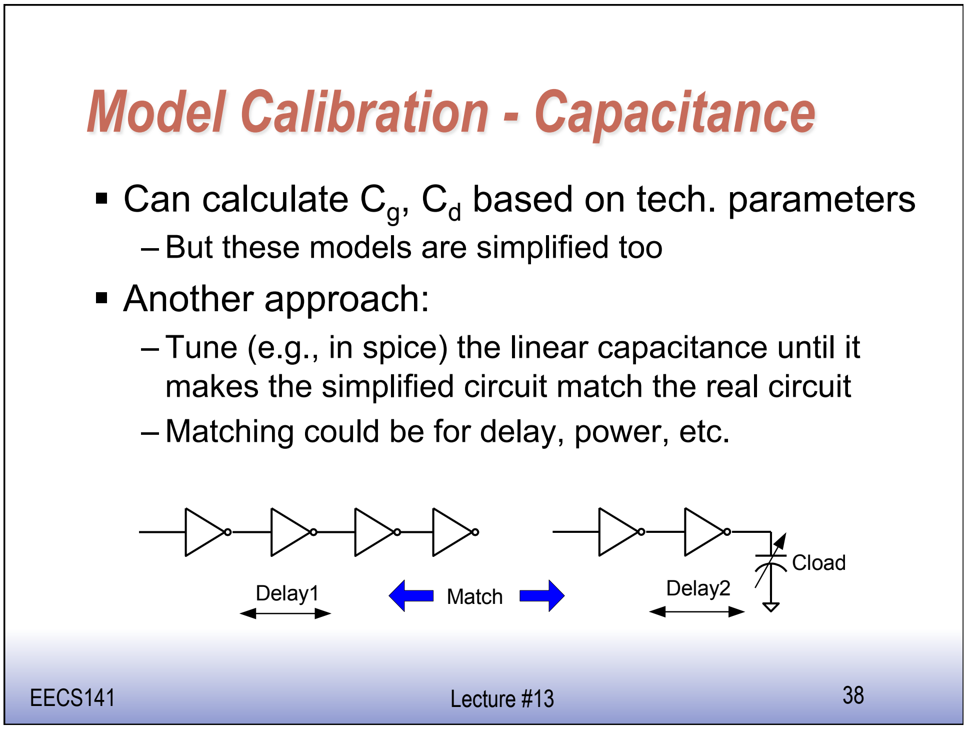
The Miller Effect

- As V_{in} increases, V_{out} drops
  - Once get into the transition region, gain from V_{in} to V_{out} > 1
- So, C_{gd} experiences voltage swing larger than V_{in}
  - Which means you need to provide more charge
  - Makes C_{gd} look larger than it really is
- Known as the "Miller Effect" in the analog world

ΔV
V_{in}
C_{gd1}
V_{out}
ΔV
M_1
V_{out}
ΔV
$2C_{gd1}$
ΔV
V_{in}
M_1

During a low-high or high-low transition, the terminals of the gate-drain capacitor are moving in opposite directions (Figure 5.14). The voltage change over the floating capacitor is hence twice the actual output voltage swing. To present an identical load to the output node, the capacitance-to-ground must have a value that is twice as large as the floating capacitance.

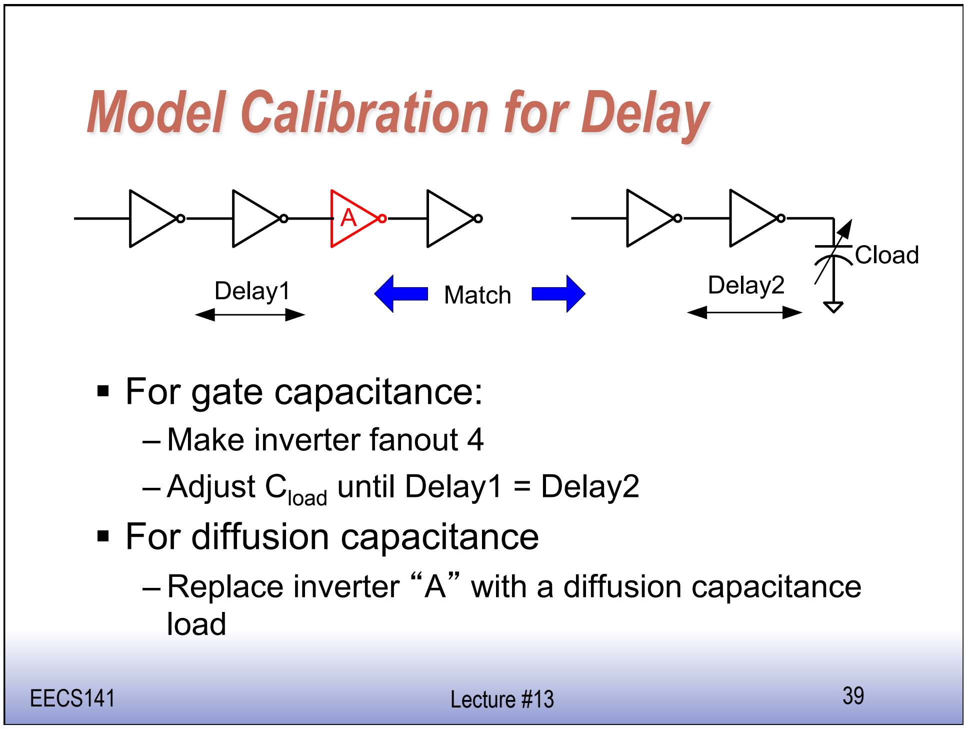
(Figure 5.14)The Miller effect—A capacitor experiencing identical but opposite voltage swings at both its terminals can be replaced by a capacitor to ground, whose value is two times the original value.



Model Calibration - Capacitance

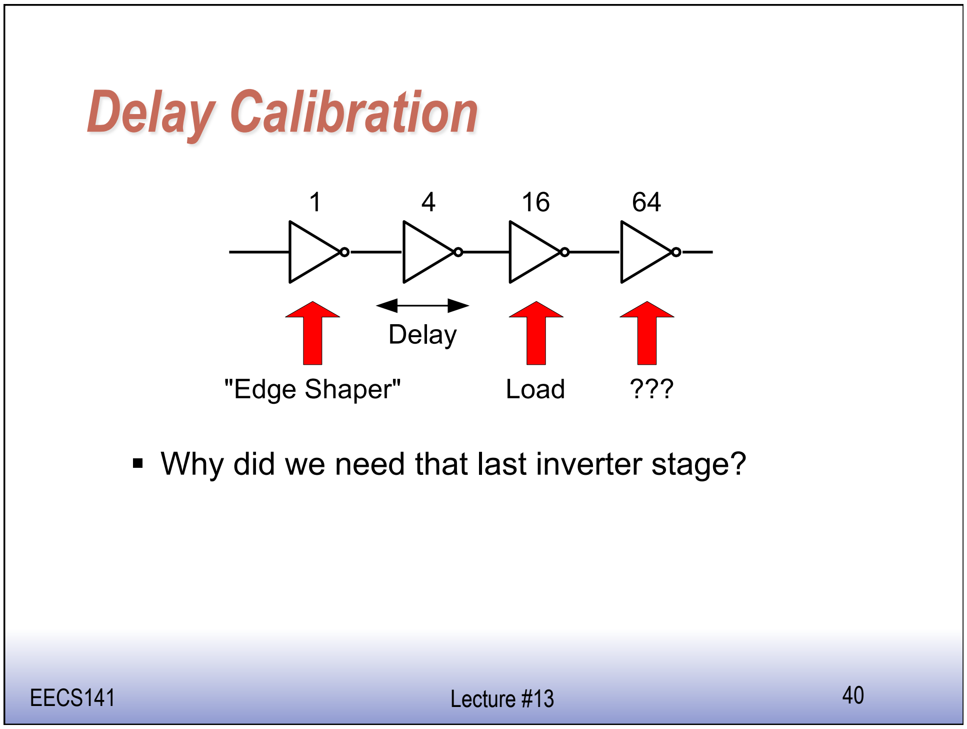
- Can calculate C_g, C_d based on tech. parameters
  - But these models are simplified too
- Another approach:
  - Tune (e.g., in spice) the linear capacitance until it makes the simplified circuit match the real circuit
  - Matching could be for delay, power, etc.

Delay1
Match
Delay2
Cload



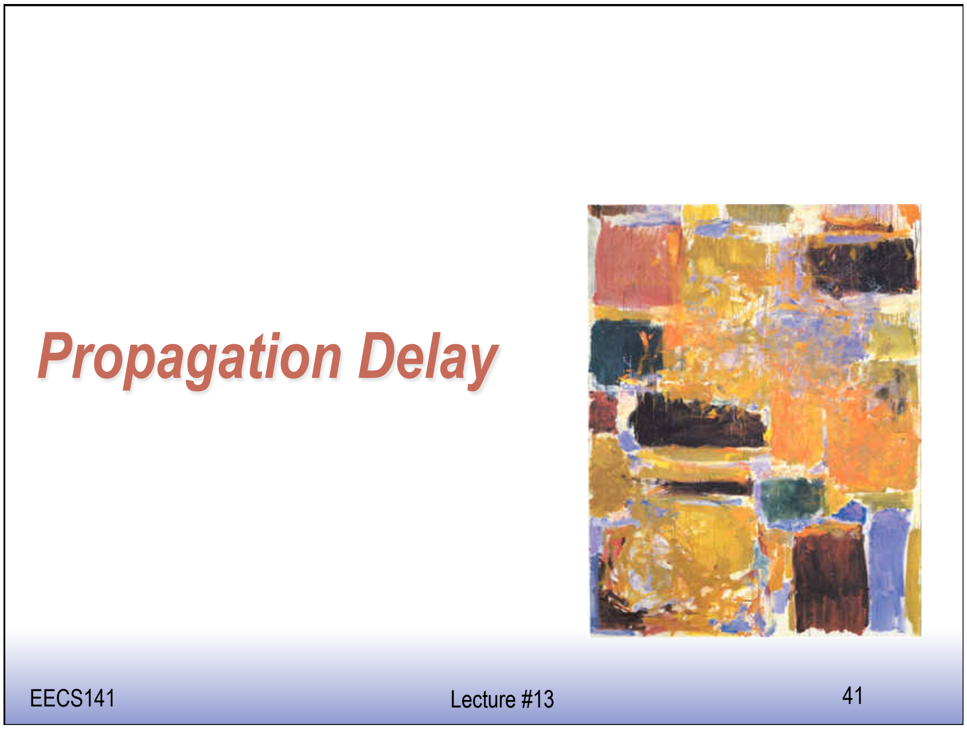
Model Calibration for Delay

- For gate capacitance:
  - Make inverter fanout 4
  - Adjust C_{load} until Delay1 = Delay2
- For diffusion capacitance
  - Replace inverter “A” with a diffusion capacitance load



Delay Calibration

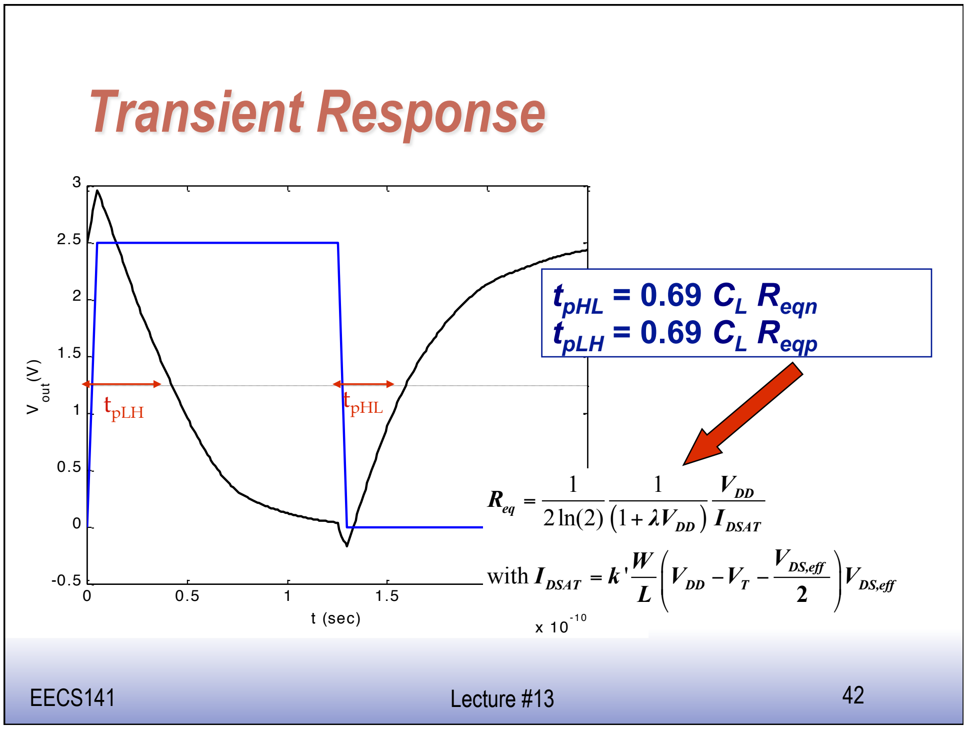
- Why did we need that last inverter stage?



Propagation Delay

EECS141 Lecture #13 41

The voltage-dependencies of the on-resistance and the load capacitor are addressed by replacing both by a constant linear element with a value averaged over the interval of interest.



Transient Response

t_{pHL} = 0.69 C_L R_{eqn}
t_{pLH} = 0.69 C_L R_{eqp}

$$R_{eq} = \frac{1}{2\ln(2)} \frac{1}{(1+\lambda V_{DD})} \frac{V_{DD}}{I_{DSAT}}$$

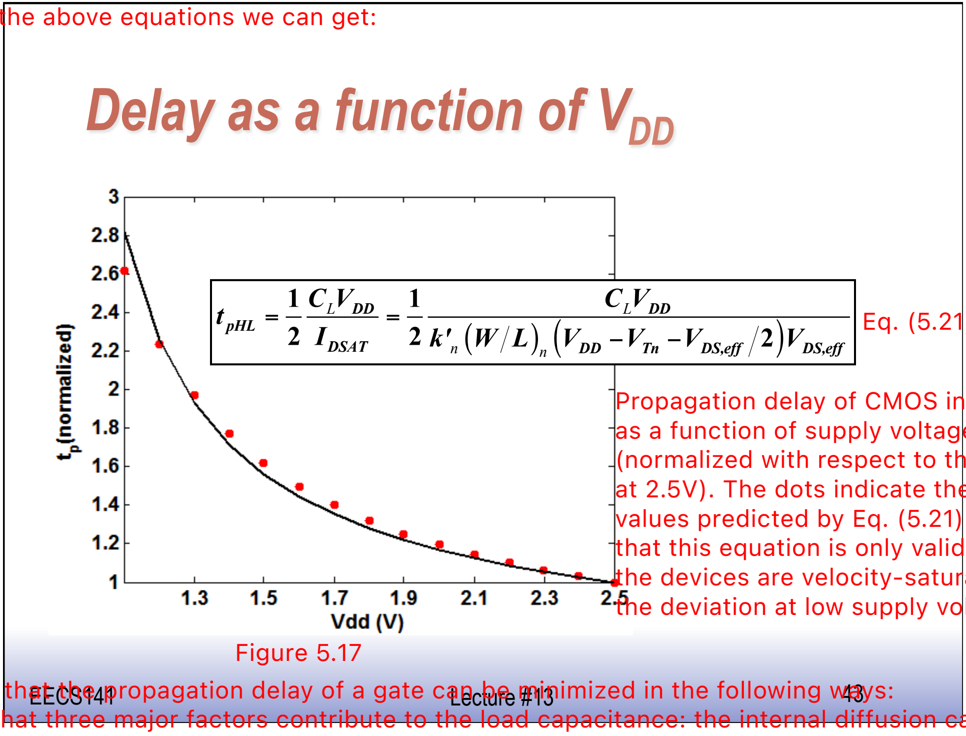
$$\text{with } I_{DSAT} = k'\frac{W}{L}\left(V_{DD} - V_T - \frac{V_{DS,eff}}{2}\right)V_{DS,eff}$$

EECS141 Lecture #13 42

with Reqp the equivalent on-resistance of the PMOS transistor over the interval of interest. This analysis assumes that the equivalent load-capacitance is identical for both the high-to-low and low-to-high transitions.

Very often, it is desirable for a gate to have identical propagation delays for both rising and falling inputs. This condition can be achieved by making the on-resistance of the NMOS and PMOS approximately equal. Remember that this condition is identical to the requirement for a symmetrical VTC.

assuming for the time being that the channel-length modulation factor λ is ignorable, from the above equations we can get:



Delay as a function of V_{DD}

$$t_{pHL} = \frac{1}{2}\frac{C_L V_{DD}}{I_{DSAT}} = \frac{1}{2}\frac{C_L V_{DD}}{k'_n (W/L)_n (V_{DD} - V_{Tn} - V_{DS,eff}/2) V_{DS,eff}}$$ Eq. (5.21)

Propagation delay of CMOS inverter as a function of supply voltage (normalized with respect to the delay at 2.5V). The dots indicate the delay values predicted by Eq. (5.21). Observe that this equation is only valid when the devices are velocity-saturated. Hence, the deviation at low supply voltages.

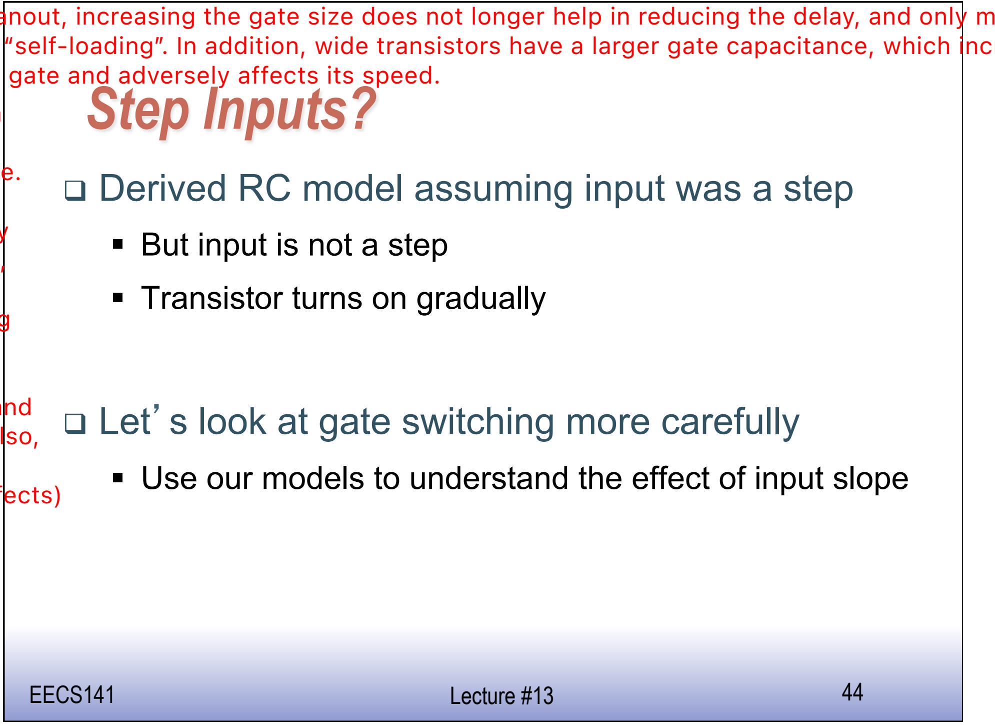
Figure 5.17

EECS141 Lecture #13 43

From the above, we deduce that the propagation delay of a gate can be minimized in the following ways:

1) Reduce CL . Remember that three major factors contribute to the load capacitance: the internal diffusion capacitance of the gate itself, the interconnect capacitance, and the fanout. Careful layout helps to reduce the diffusion and interconnect capacitances. Good design practice requires keeping the drain diffusion areas as small as possible.

2) Increase the W/L ratio of the transistors. This is the most powerful and effective performance optimization tool in the hands of the designer. Proceed however with caution when applying this approach. Increasing the transistor size also raises the diffusion capacitance and hence C L . In fact, once the intrinsic capacitance (i.e. the diffusion capacitance) starts to dominate the extrinsic load formed by wiring and fanout, increasing the gate size does not longer help in reducing the delay, and only makes the gate larger in area. This effect is called "self-loading". In addition, wide transistors have a larger gate capacitance, which increases the fan-out factor of the driving gate and adversely affects its speed.

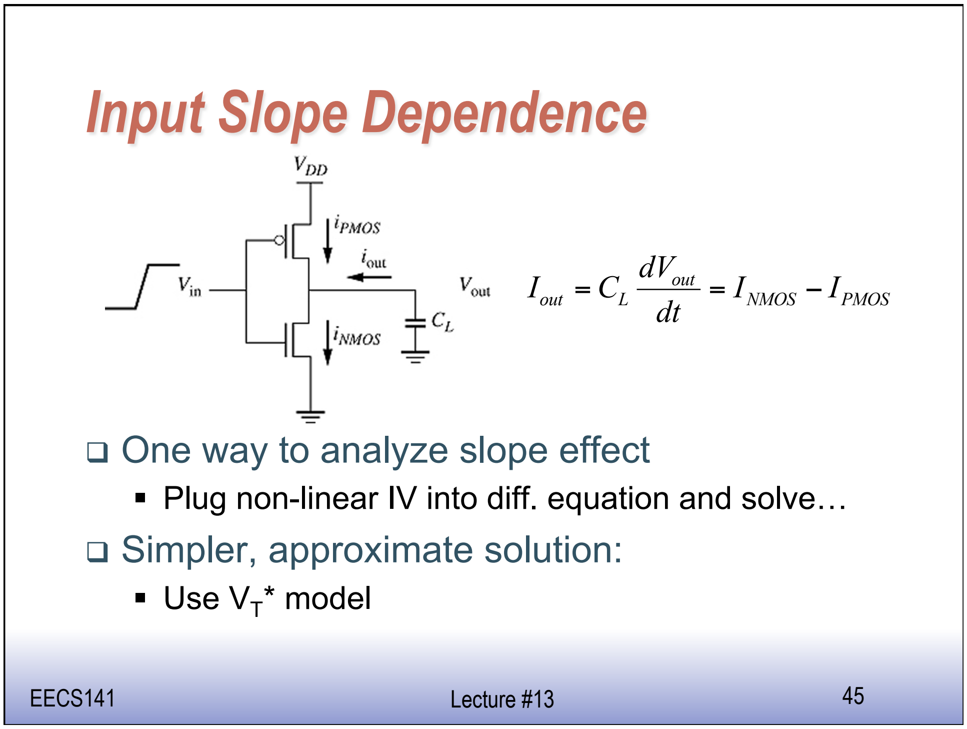


Step Inputs?

- Derived RC model assuming input was a step
  - But input is not a step
  - Transistor turns on gradually
- Let' s look at gate switching more carefully
  - Use our models to understand the effect of input slope

EECS141 Lecture #13 44

in Figure 5.17, the delay of a gate can be modulated by modifying the supply voltage. This flexibility allows the designer to trade-off energy dissipation for performance, as we will see in a later section. However, increasing the supply voltage above a certain level yields only very minimal improvement and hence should be avoided. Also, reliability concerns (oxide breakdown, hot-electron effects) enforce firm upper-bounds on the supply voltage



Input Slope Dependence

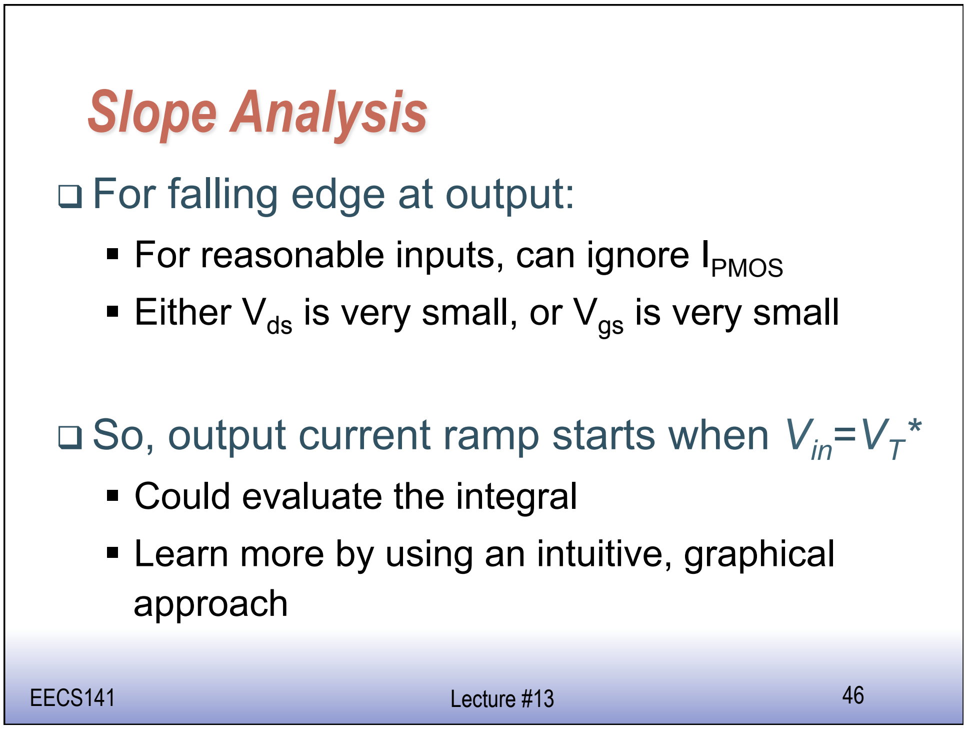
$$I_{out} = C_L \frac{dV_{out}}{dt} = I_{NMOS} - I_{PMOS}$$

❑ One way to analyze slope effect

- Plug non-linear IV into diff. equation and solve…

❑ Simpler, approximate solution:

- Use V_T* model



Slope Analysis

❑ For falling edge at output:

- For reasonable inputs, can ignore I_{PMOS}
- Either V_{ds} is very small, or V_{gs} is very small

❑ So, output current ramp starts when $V_{in}=V_T$*

- Could evaluate the integral
- Learn more by using an intuitive, graphical approach

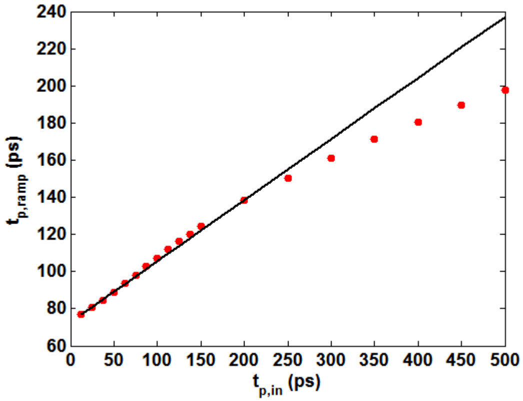
Result Summary

- ❑ For reasonable input slopes:

$$\boxed{t_{p,ramp} = t_{p,step} + \frac{V_T^{\,*}}{V_{DD}} \cdot t_{p,in}}$$

Model vs. Spice Data

- ❑ For reasonable input slope
  - Model matches Spice very well
- ❑ Model breaks with very large t_r
  - Input looks “DC” – traces out VTC
  - Have other problems here anyways
    - – Short-circuit current


$t_{p,ramp}$ (ps)
$t_{p,in}$ (ps)