EE141-Spring 2012 Digital Integrated Circuits

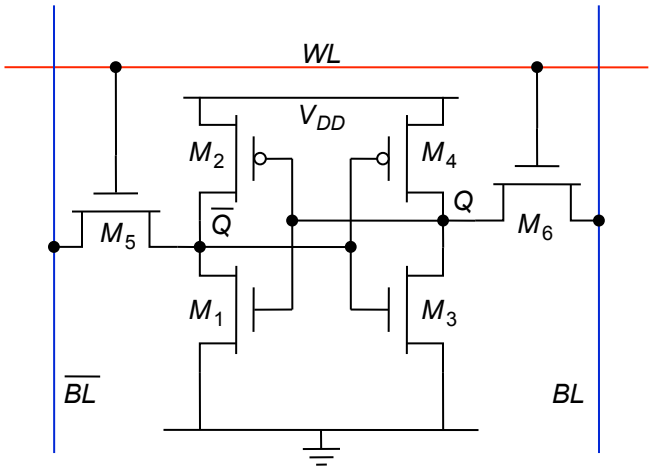
Lecture 15
Inverter Delay + Energy

Administrativia

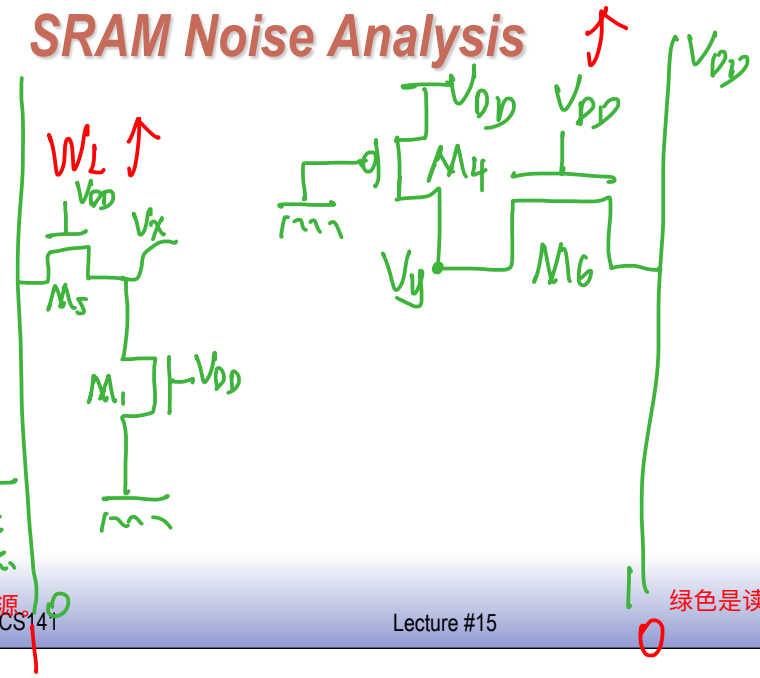
- Midterm 2 on Friday March 23.
  - Topics: Everything including arithmetic.
  - Memory NOT included.
- Project phase 1 due March 21.
- Hw 6 due today.
- Hw 7 will be posted today but will only be due after spring break.

Memory - SRAM

SRAM Noise Margin Analysis



SRAM Noise Analysis



在读取数值的过程中，为了减少Vx，M1/M5必须保持在一个适当的比例，而且这个值越大越好。

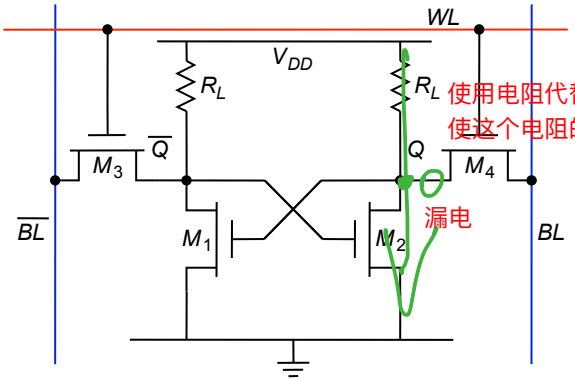
这个电容相当于一个直流电源。

绿色是读取操作，红色是写入操作

在写入数值的过程中，为了让Vx打过VM，M1/M5必须保持在一个适当的比例，而且这个值越小越好。当然，我们也可以适当调整M2/M6的比例，让Vy的值小于VM，而且这个值是越小越好。因为写和读，对比例值要求的矛盾，因此SRAM cell往往是不对称的，一个是通过size NMOS来达成目的，一个是通过size PMOS来达成目的。

为什么在读取值时的预充电是使两条线的值变为1，因为这样可以把所有工作都交给NMOS来做，而NMOS的工作效率比PMOS高，可以有效减少面积。

Resistive Pull-Ups

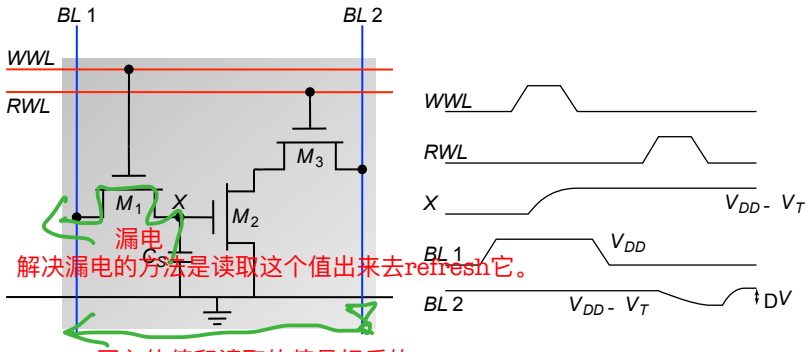


使用电阻代替PMOS也可以做到相同的操作。不过要尽量使这个电阻的值大，这样可以减少漏电。

Static power dissipation -- Want R_L large
Bit lines precharged to V_{DD} to address t_p problem

Memory - DRAM

3-Transistor DRAM Cell



漏电

解决漏电的方法是读取这个值出来去refresh它。

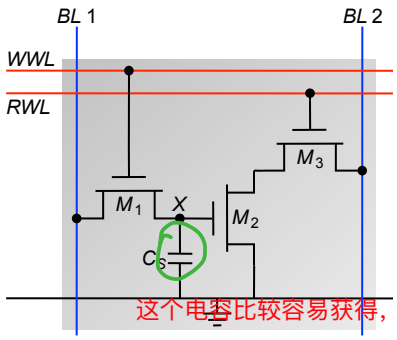
写入的值和读取的值是相反的。

No constraints on device ratios

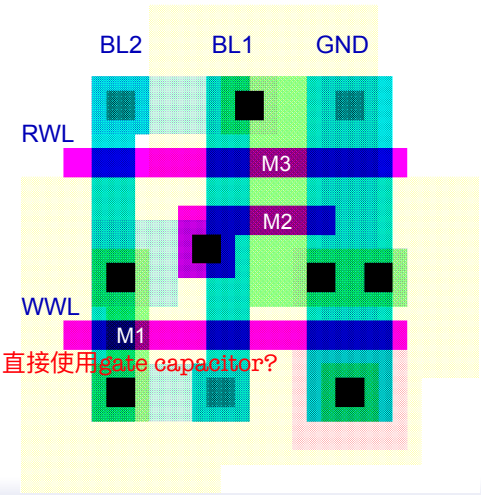
Reads are non-destructive 如果不漏电的话，读取里面的数值不会破坏被存储的数值。

Value stored at node X when writing a “1” = $V_{WWL}-V_{Tn}$

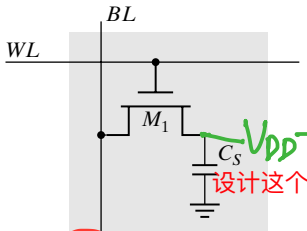
3T DRAM Layout



这个电容比较容易获得，直接使用gate capacitor?



1-Transistor DRAM Cell



$V_{DD}-V_{TH}$ (当存储1时)

设计这个很困难。

这个会很大，因为线电容和lump capacitor的关系。

因为读取数值会破坏DRAM Cell存储的值，因此在读取出来以后又要写回去。

When DRAM Cell stores 0. If we read it , we should precharge C_{BL},and 它存储的电荷等于 $V_{DD}C_{BL}$

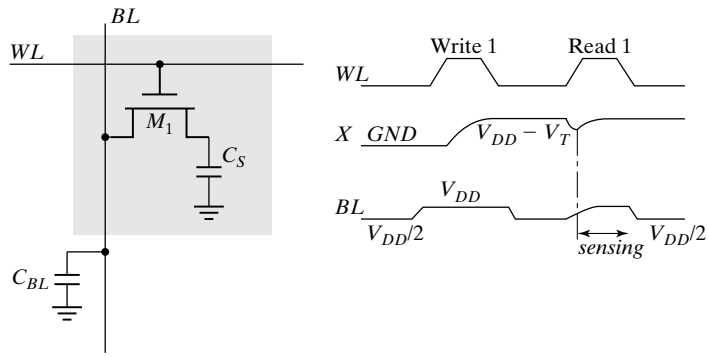
如果打开M1，那么此时的总电荷为 $V_{DD}C_{BL}+0\cdot C_S=V_X(C_{BL}+C_S)$

因此，新的电压为 $V_X=\frac{C_{BL}}{C_{BL}+C_S}V_{DD}$

$\Delta V=\frac{C_S}{C_{BL}+C_S}V_{DD}$ 尽可能大

DRAM需要一个sense amplifier，否则它不会工作。理由和SRAM cell一样。

1-Transistor DRAM Cell



Write: C_S is charged or discharged by asserting WL and BL.
Read: Charge redistribution takes places between bit line and storage capacitance

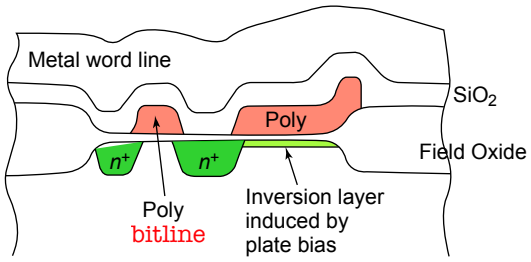
$$\Delta V = V_{BL} - V_{PRE} = V_{BIT} - V_{PRE} \frac{C_S}{C_S + C_{BL}}$$

Voltage swing is small; typically around 250 mV.

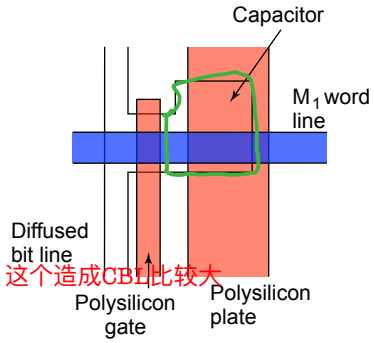
voltage is very small。

关键挑战是如何create capacitance，太小，会难以读取数据。

1T DRAM Cell



Cross-section

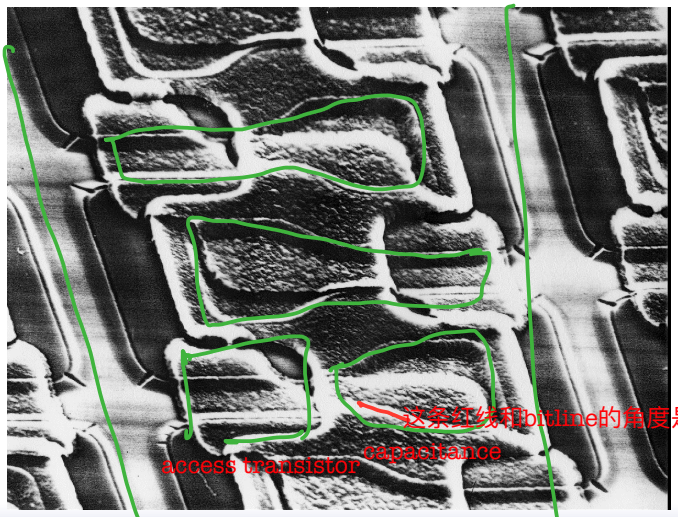


Layout

Uses Polysilicon-Diffusion Capacitance
Expensive in Area

Micrograph of 1T DRAM

为了使使用面积更有效率，同一列的DRAM Cell呈z形摆放。

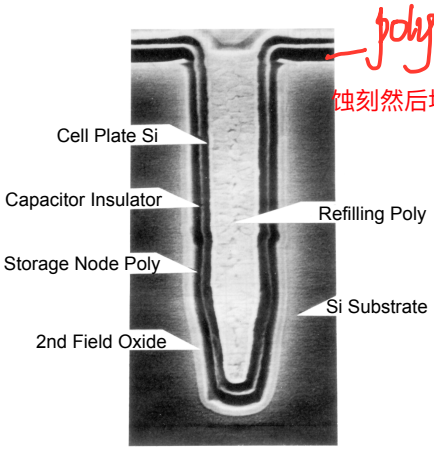


Advanced 1T DRAM Cells

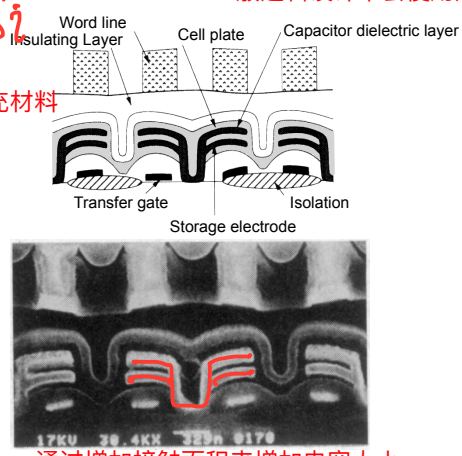
这个方法可以加大DRAM Cell所需要的电容，但是一般逻辑设计不会使用这个方法。

poly Si

蚀刻然后填充材料



Trench Cell



通过增加接触面积来增加电容大小。

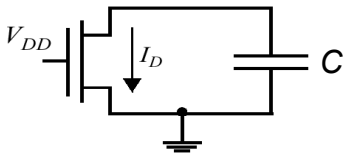
Stacked-capacitor Cell

CMOS Switching Delay

MOS Transistor as a Switch

transistor and capacitors 都是非线性的，因此我们需要找到一些等效的东西来代替它们。

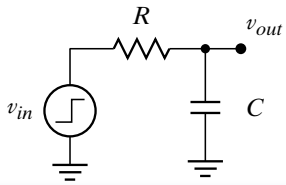
- Discharging a capacitor



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

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

- We modeled this with:



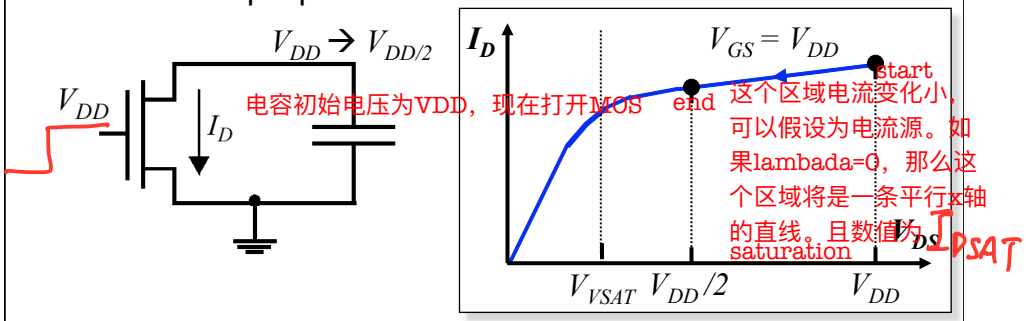
t_p = ln (2) RC

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:

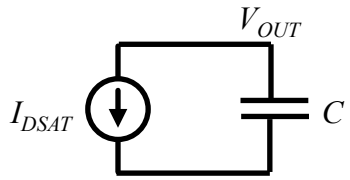


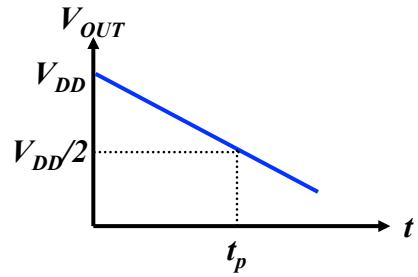
- Transistor is in (velocity) saturation during entire transition from V_{DD} to $V_{DD}/2$

经过上面saturation简化得到：

Switching Delay

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





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

$$= \frac{C\Delta V}{I_{DSAT}}$$

Defining IDSAT

$$I_{DSAT} = [I_D(V_{DD}) + I_D(\tfrac{V_{DD}}{2})]/2$$

$$\Downarrow$$

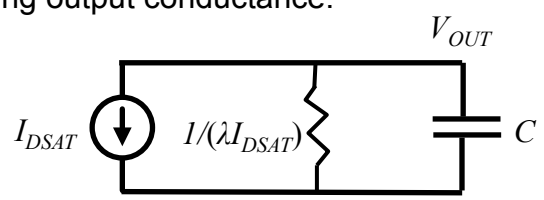
$$\frac{1}{2}\left(I_{DSAT}(1+\lambda V_{DD}) + I_{DSAT}(1+\lambda \tfrac{V_{DD}}{2})\right)$$

$$= I_{DSA}(1+\tfrac{3\lambda}{4}V_{DD})$$

Switching Delay (with Output Conductance)

$V_{GS} = V_{DD}$ (always)

- Including output conductance:



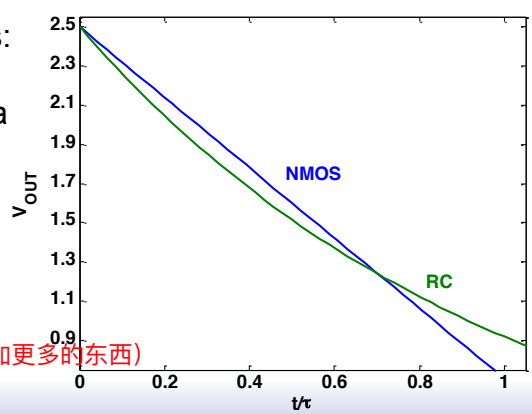
$$V_{OUT} = \left(V_{DD} + \lambda^{-1}\right) e^{-t/(C/\lambda I_{DSAT})} - \lambda^{-1}$$

- For "small" λ: $t_p \approx \dfrac{C(V_{DD}/2)}{(1+\lambda V_{DD}) I_{DSAT}}$ $= C\Delta V$

事实上是平入

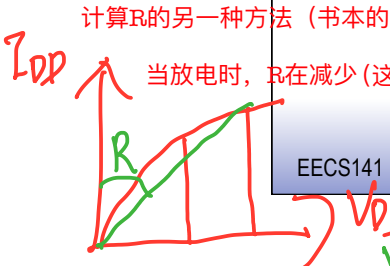
RC Model

- 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



计算R的另一种方法（书本的方法）：

当放电时，R在减少（这个方法更好，可以添加更多的东西）



$$R = \frac{V}{I} = \frac{V_{DD}}{I_{DSAT}(1+\lambda V_{DS})} = \left[R(V_{DD}) + R\left(\frac{V_{DD}}{2}\right)\right]/2$$

Finding Req

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

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

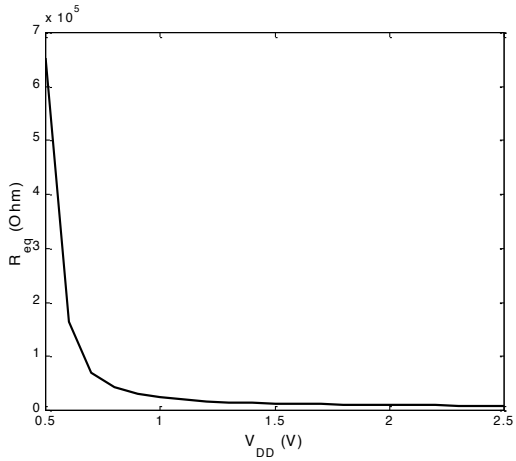
$$\frac{C(V_{DD}/2)}{(1+\lambda V_{DD})I_{DSAT}} = \ln(2)R_{eq}C \longrightarrow \boxed{R_{eq} = \frac{(V_{DD}/2)}{\ln(2)(1+\lambda V_{DD})I_{DSAT}}}$$

- 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 ~$0.72\cdot V_{DD}/I_{DSAT}$.
（书本的方法）

The Book's Method

The Transistor as a Switch



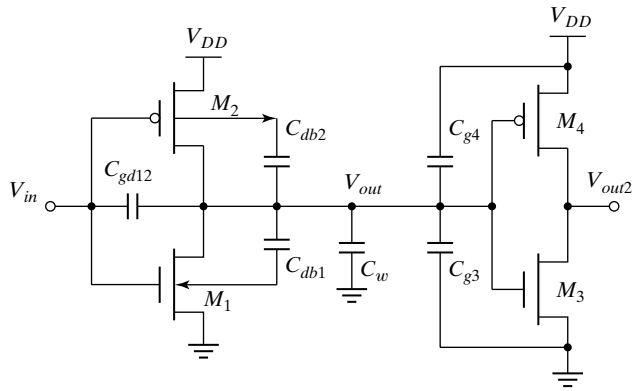
The Transistor as a Switch

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

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

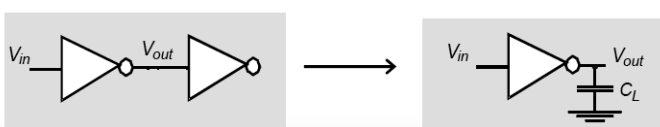


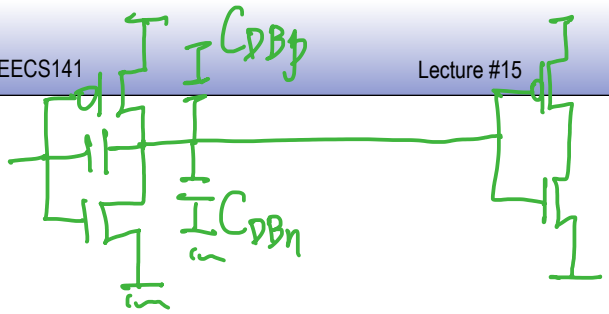
Parasitic capacitances, influencing the transient behavior of the cascaded inverter pair

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.

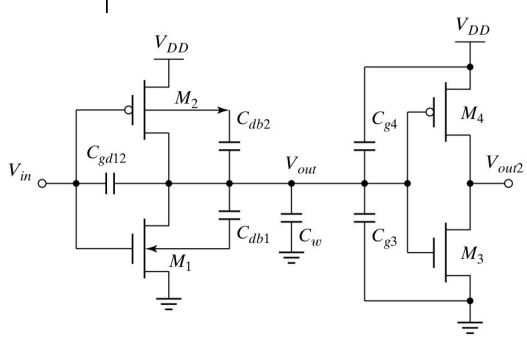
Inverter Capacitance Model

- 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 – 2fF·W(μm)





Lumping the Caps

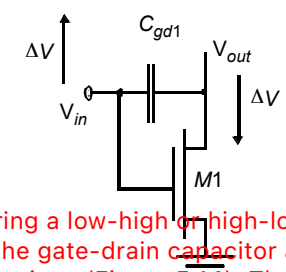


M1 and M2 are either in cut-off or in the saturation mode during the first half (up to 50% point) of the output transient. Under these circumstances, the only contributions to Cgd12 are the overlap capacitances of both M1 and M2. The channel capacitance of the MOS transistors does not play a role here,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

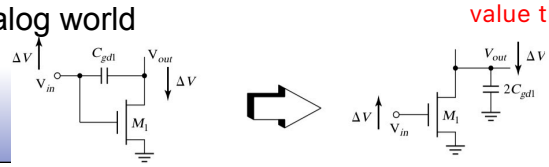
The Miller Effect

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

- 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



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.



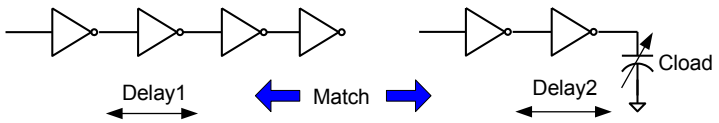
(Figure 5.14)The Miller effect—A capacitor experiencing identical but opposite voltage swings at both its terminals can be replaced by a capacitor to ground, whose value is two times the original value.

$C_L = C_{DB(n,p)} + 2C_{ov(n,p)} + C_w$

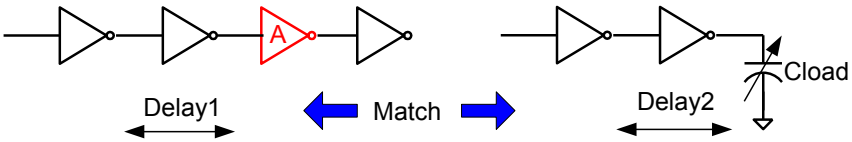
ov = overlap

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.

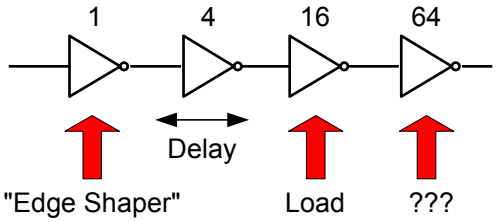


Model Calibration for Delay



- 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

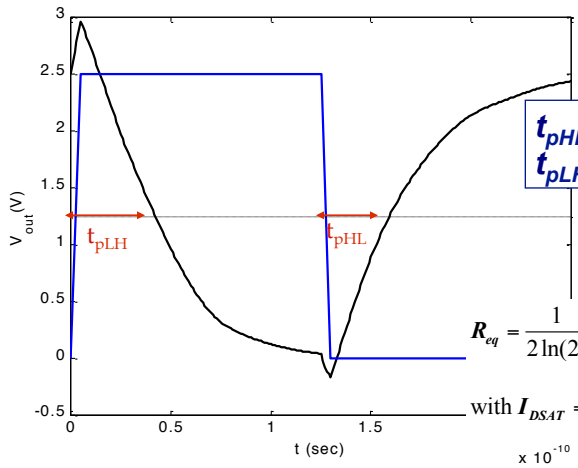


- Why did we need that last inverter stage?

Propagation Delay

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



$$t_{pHL} = 0.69\, C_L\, R_{eqn}$$
$$t_{pLH} = 0.69\, C_L\, R_{eqp}$$

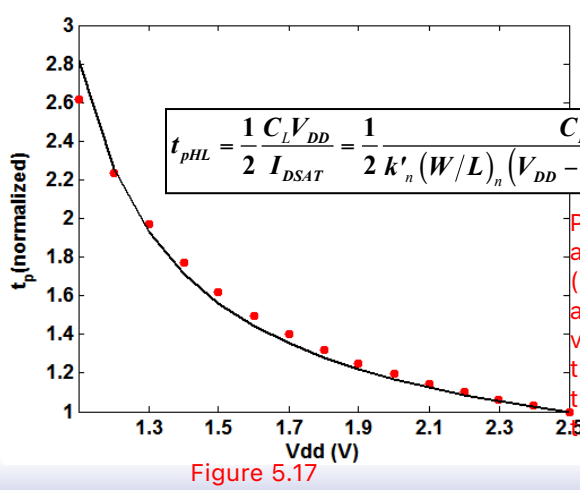
$$R_{eq} = \frac{1}{2\ln(2)} \frac{1}{(1+\lambda V_{DD})} \frac{V_{DD}}{I_{DSAT}}$$

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

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.

Delay as a function of V_{DD}



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

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

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.

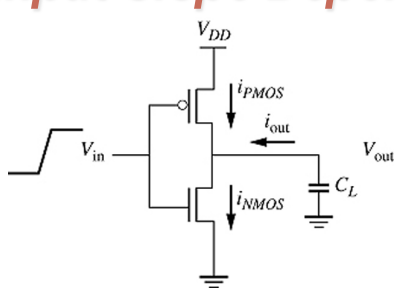
2) Increase the W/L ratio of the transistors. This is the most powerful and effective performance optimization tool in the hands of the designer. Proceed however with caution when applying this approach. Increasing the transistor size also raises the diffusion capacitance and hence C L . In fact, once 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.

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

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

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

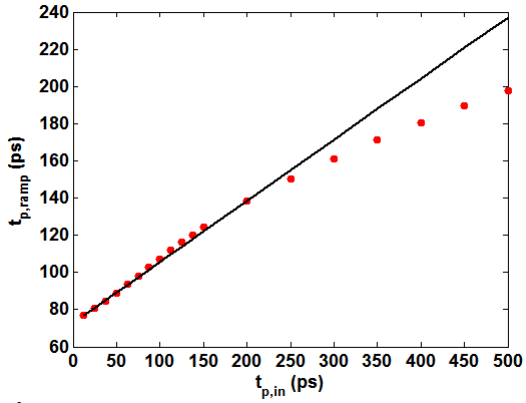
Result Summary

- For reasonable input slopes:

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

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

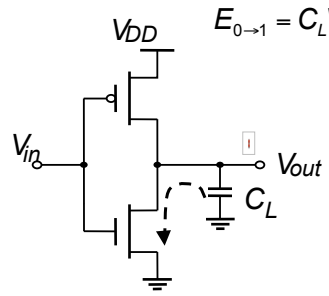


CMOS Power Dissipation

Where Does Power Go in CMOS?

- Switching power
  - Charging/discharging capacitors
- Leakage power
  - Transistors are imperfect switches
- Short-circuit power
  - Both pull-up and pull-down on during transition
- Static currents
  - Biasing currents, in e.g. analog, memory

Dynamic Power Consumption



$$E_{0\to1} = C_L V_{DD}^2$$

$$E_R = \frac{1}{2} C_L V_{DD}^2$$

$$E_C = \frac{1}{2} C_L V_{DD}^2$$

- One half of the energy from the supply is consumed in the pull-up network, one half is stored on C_L
- Energy from C_L is dumped during the 1→0 transition

Dynamic Power Consumption

Power = Energy/transition • Transition rate

$$= C_L V_{DD}^2 \bullet f_{0\to1}$$

$$= C_L V_{DD}^2 \bullet f \bullet \alpha_{0\to1}$$

$$= C_{switched} V_{DD}^2 \bullet f$$

- Power dissipation is data dependent – depends on the switching probability
- Switched capacitance $C_{switched} = C_L \bullet \alpha_{0\to1}$

Transition Activity and Power

- Energy consumed in N cycles, E_N:

$$E_N = C_L \bullet V_{DD}^2 \bullet n_{0\to1}$$

$n_{0\to1}$ – number of 0→1 transitions in N cycles

$$P_{avg} = \lim_{N\to\infty} \frac{E_N}{N} \cdot f = \left(\lim_{N\to\infty} \frac{n_{0\to1}}{N} \right) \cdot C_L \cdot V_{DD}^2 \cdot f$$

$$\alpha_{0\to1} = \lim_{N\to\infty} \frac{n_{0\to1}}{N}$$

$$\boxed{P_{avg} = \alpha_{0\to1} \cdot C_L \cdot V_{DD}^2 \cdot f}$$