Lecture 3:

MOS Transistors
Switch and Gate Logic

Mark Horowitz

Modified by Azita Emami

Computer Systems Laboratory

Stanford University

azita@stanford.edu

Overview

Reading

W&E 2.1-2.2 - MOS Transistor Model

(more complex than we need)

W&E 2.4.1 - nMOS like gates

W&E 2.6 - CMOS switches

W & E 1.5, - CMOS gates

Shoji 2.1-2.8 - CMOS gates

Introduction

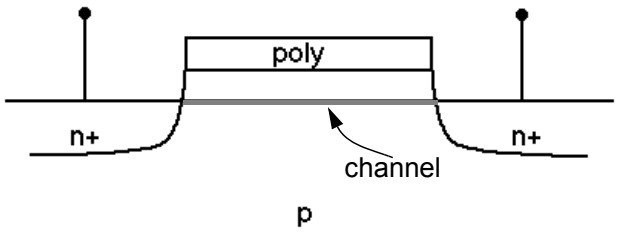
So far we have talked about building logic out of switches, but we were using a very simple model of a transistor. We will again look at building logic from transistors, but this time we will use a more accurate model of a transistor. This model will point out limitation of nMOS switch logic. There is a restricted form of switch logic, called gate logic, that behaves like unidirectional logic functions. Since this is a nice level of abstraction, most CMOS transistors are used to create ‘gates’ that a designer then uses

New Transistor Model

In the first lecture, we approximated a transistor as a simple switch. While this is a good first model for a transistor, we need a better model if we want to understand delay and transistor connection rules in CMOS circuits. While a transistor can be viewed as a switch, it is a switch with some interesting properties. A transistor is really a non-linear device where the output current is a function of the size of the device, and voltages on its terminals.

Luckily, for the stuff that we do, we don't need to use the real i-V curve of a transistor. We can approximate it as a voltage controlled resistor. But to understand what value resistance we should use and how it will change with technology, it is important to get a feeling for transistor operation. The notes have only a quick review. W&E 2.1 - 2.2 has more details if you need/want them.

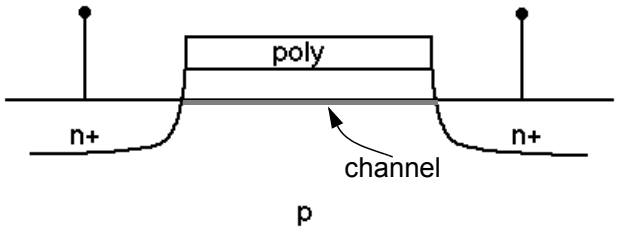
Transistors



For nMOS transistors

- Raising the gate voltage attracts electrons to form a thin n-region under the gate. This n-region is called the channel, and forms a bridge from between the two n+ region, and allows current to flow. If the channel is not present, the two n+ regions are separated by two back to back diodes, which blocks current flow. This induced n-region forms a resistor, the more carriers in the channel, the lower the resistance between the source and drain

+ Transistor i-V Curve



- There is a lot of semiconductor physics, that I will skip (bandgaps, fermi energies …).
- What basically happens is that the poly - oxide - silicon sandwich under the gate poly is a capacitor. To increase the gate voltage, I need to add positive charge to the poly, and negative charge to the silicon. At first the negative charge comes from pushing away the holes in the channel (leaving the negatively charged ionized acceptor atoms). After some point (called the threshold voltage) a channel of mobile electrons forms.

+ Transistor i-V cont'd

Electron charge in the channel can be easily determined:

$$Q_{mobile} = C_{ox}(V_{gs} - V_{th}) = \frac{\varepsilon \times Area \times (V_{gs} - V_{th})}{t_{ox}}$$

The (mobile) electrons in the channel will move if a voltage is applied

- Voltage applied between source and drain
- Follows Ohm's Law (i =V/R)

  Implies that the carrier velocity is proportional to E-field (V/L)
- The mobility of the electrons (μ_n) relates the E-field (V/L) to velocity
- The current is the charge/length * velocity.

The relation between E-field and carrier speed is only true at low field levels. There is a speed limit for carriers in silicon, so at high fields you get less current than you might expect.

$$i_{ds} = \frac{\mu_n \varepsilon \times Area \times (V_{gs} - V_{th})}{t_{ox} L} \frac{V_{ds}}{L} = \frac{W}{L} \times \frac{\mu_n \varepsilon (V_{gs} - V_{th})}{t_{ox}} V_{ds}$$

+ i-V cont'd

$$i_{ds} = \frac{W}{L} \times \frac{\mu_n \varepsilon (V_{gs} - V_{th})}{t_{ox}} V_{ds}$$

- The value of the current is proportional to the gate to source voltage (remember how the source is defined) minus threshold voltage, V_{th}

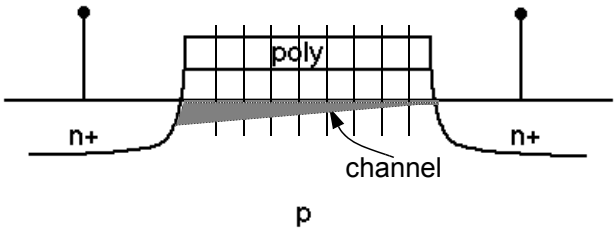
  Since (V_{gs} - V_{th}) set the number of carriers in the channel
- Current inversely proportional to the oxide thickness.
- Current proportional to width (width of the diffusion), inversely proportional to length (width of the poly)

So

- Resistance of transistor is proportional to length and inversely proportional to width

Unfortunately this derivation is missing something… What?

+ i - V Current Saturation



When Vds is not zero, the number of carriers in the channel is not constant either, since the voltage of the channel is changing. Closer to the drain there are less electrons, so the resistance will be higher.

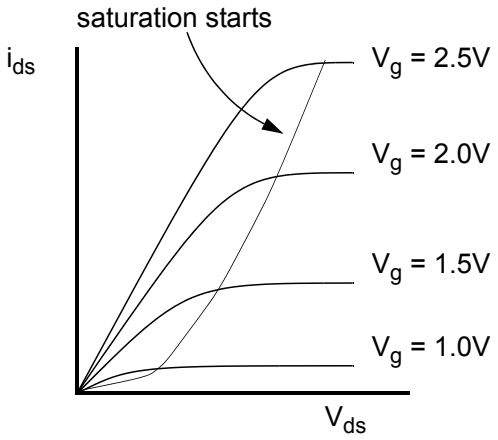
To solve for the i - V curve, break the transistor into a number of small transistors in series. Each little transistor will have a very small V_{ds}, so the previous formula will hold. This gives the quadratic i - V curve.

When V_g - V_d < V_{th}, the model breaks down, and current no longer depends on V_{ds}

Notice that the current through each transistor must be the same, since otherwise it would accumulate charge.

This model breaks down when there is not enough charge to support the needed current. Without velocity saturation this happens when the channel charge becomes 0 (eq. shown). With velocity sat, it occurs earlier.

+ Ideal Quadratic nMOS i-V



Note that the i-V with Vds is the same as our initial formula if we use the average value of source/drain voltages.
When the transistor saturates, the formula does not hold, and the current remain constant.

In the linear region $i_{ds} = \frac{W}{L} \times \frac{\mu_n \varepsilon}{t_{ox}} \left(V_{gs} - V_{th} - \frac{V_{ds}}{2} \right) V_{ds}$

In the saturation region $i_{ds} = \frac{W}{L} \times \frac{\mu_n \varepsilon}{2 t_{ox}} (V_{gs} - V_{th})^2$

+ Velocity Saturation

Most models (book and notes) give a quadratic model of a MOS transistor

- $I_{ds} = K (V_{gs} - V_{th})^2$
  - Larger voltage gate to source increases carriers in the channel, which increases the current, and the larger voltage drain to source means that the carriers move faster. Thus the current increases quadratically with voltage

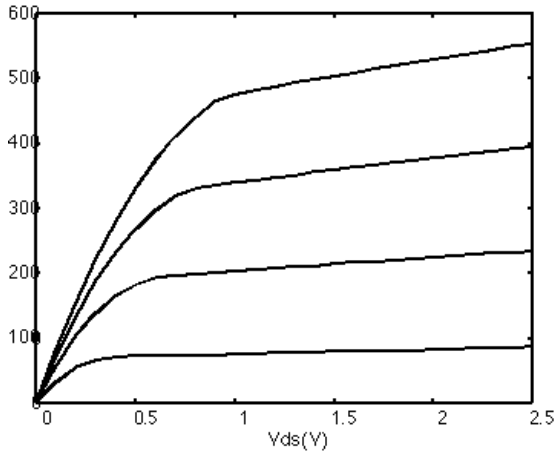
In advanced CMOS devices (L < 1μ) the electric fields are so large that the carriers are moving as fast as they can. Increasing the lateral field does not make them move faster. As a result, the current is not quadratic on voltage.

- $I_{ds} = K (V_{gs} - V_{th})^{1.3}$
  - Larger voltage increases the channel charge, but that is about all. The effective change in velocity is small

Bottom Line:

If the i-V curve of the transistor is important to you, you should use a good model of the transistor. Usually given by a simulation model (SPICE)

+ Real nMOS i-V Curve



W= 1μm L= 0.35μm NFET

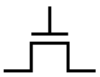
I_{DSS} = 550μA at 2.5V

K = 140μA/V²

C_G = 2fF

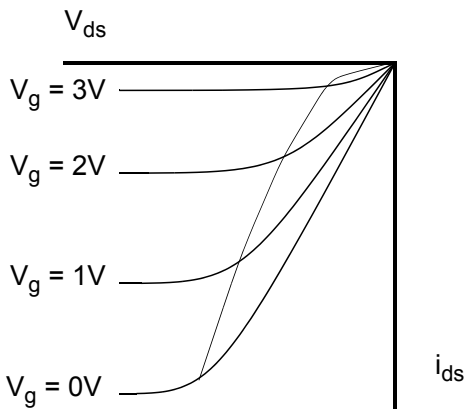
C_S = 1fF

R_{ON} = 1/(280μA/V) ~ 3.5KΩ



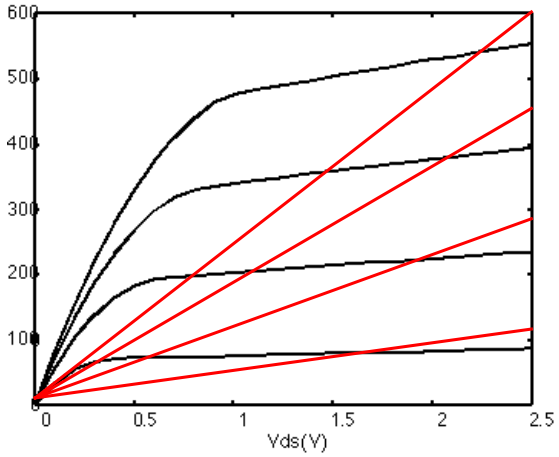
+ Current - Voltage Curves

Divided into two regions, depending on whether the drain is ‘connected’ to the channel (remember this is a pMOS device with its source at 5V):



nMOS Approximation

We will approximate the transistor as a voltage variable resistor



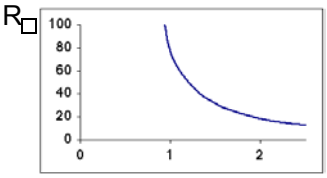
This approximation is ok for timing estimates, but not for analog circuits. Resistor values are set to match real nMOS iV curves, not quadratic model.

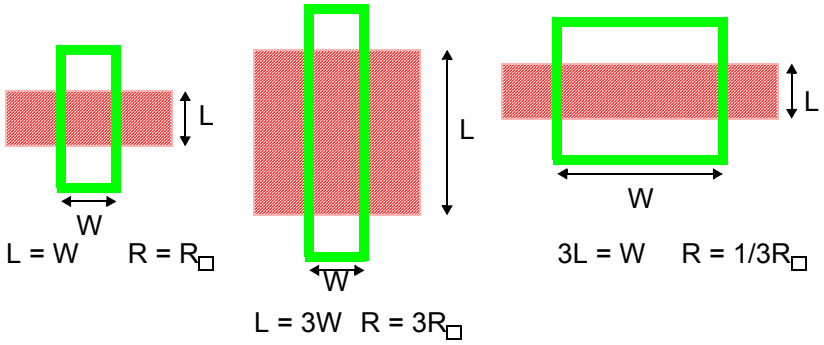
Electrical Model

Resistance is proportional to L/W (number of squares)

The resistance / square of a transistor is also inversely proportional to (Vgs -Vth)

At Vgs = Vdd, R is about 10K/sq





- Aside - Size Terminology

How to refer to the size of a device?

1. Number of squares - L/W
2. Width of the device - W/L

The second method is more common in industry, and is the method that I will use in the class. Larger transistors mean more current, not more resistance. As we will see later, almost all transistors have minimum length, so often people refer to transistor size solely by its width.

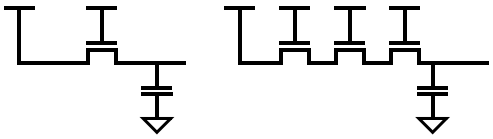
nMOS Switch Logic

With this new model of an nMOS transistor, we can see some limitations of nMOS switch logic.

A high output of switch logic is a degraded signal; it is the voltage on the gate minus a threshold voltage. This is because there must be a V_{th} between the gate and the source for the transistor to conduct. Since the input (drain) is equal to the gate for logic 1, the transistor turns off when the output becomes $V_{gate} - V_{th}$.

Note that the output voltage does not depend on the number of switch transistors that the signal travels through. It only depends on the gate voltage of those switches. The output voltage will be set by the lowest gate voltage on any of the switch transistors it passes through.

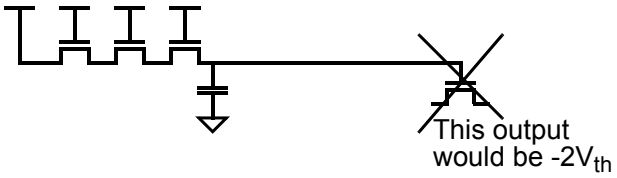
You need to be a little careful about the value of the threshold voltage. It depends on the voltage of the source (back gate effect). While it is 0.5V when the source is gnded, it can be 0.8V when the source is at 2V. This makes the degraded levels even worse



Output of both circuits is Vdd -Vth

nMOS Switch Logic

If you connect this degraded output to the gate of another nMOS switch, you would get an output that is degraded by 2 V_{th}. This may be too low to detect as a high output (remember we need to provide signals where the digital abstraction). In fact in many design styles, no degraded levels are allowed. We will see later in this lecture how to build switches that don't degrade the high level.

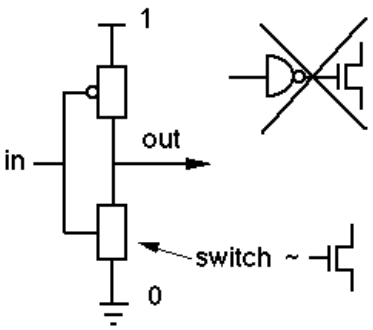


Passing a logic 0 is much easier, since then the transistor is always on (V_{gs} = Vdd). nMOS devices don't degrade low levels.

Note: nMOS switch logic has two limitations. It can't invert signals, and it's outputs can not be used to drive the gates of switches directly. To do useful stuff we clearly need (at least) inverters.

nMOS Inverters

Need to build an inverter without using of a switch with an inverted control line which does not exist in nMOS:

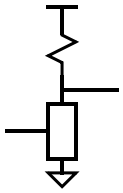


And need to get rid of the degraded high output.

The solution is to forget about building the inverted switch - use a device that is always weakly on.

nMOS Inverters

An inverter consists of a switch and a resistor; for this to work the resistance of the switch transistor must be much lower than the resistance of the resistor (depletion transistor). When the input is low, the switch transistor is off and the resistor pulls up the output to Vdd; when the input is high, the switch transistor fights the pullup resistor and the output falls close to Gnd.



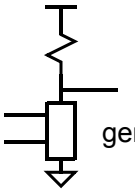
Ratio Rule: The resistance of the pulldown must be 4 times lower than the resistance of the pullup to guarantee a good low level (close to 0V).[1]

1. Transistors are not linear resistors, so you can't find the output voltage from the standard voltage divider equation. See W&E 2.4 for a discussion of a similar problem

+General nMOS Gates

Consist of an nMOS switch network between the output and Gnd

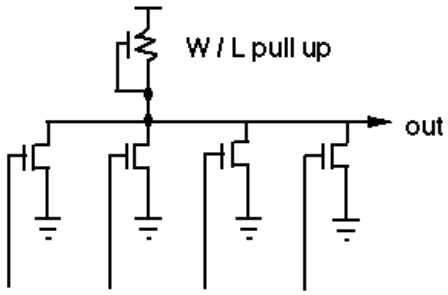
- Uses a default pullup device
- Need to make the pullup device weak enough



Since the output is low when the switch function connects, the logic function is the complement of the function of the switch network.

- The total resistance of any path through the pulldown network must be less then 1/4 the resistance of the pullup device

+In nMOS NORs are Nice



$W / L_{pd} \geq 4\ W / L_{pull\ up}$

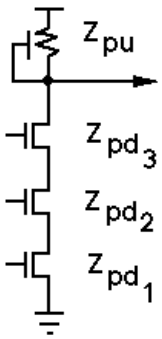
(each transistor)

W / L does not depend on the number of inputs!

For NOR gates, the size of the transistors is independent of the number of inputs. If each pulldown transistor was 12:2, then the pullup would be 3:2, for a 2 input NOR and a 8 input NOR.

Notice also that the pullup resistance and the pulldown resistance (when only one transistor is on, which is the worst case) does not depend on the number of inputs the gate has.

+NAND are Not



$Z = L / W$

$Z \geq 4 \cdot \sum_{i=1}^{n} Z_{pdi}$

Large n (# of inputs) $\Rightarrow$

1) wide pull down transistors

or

2) long pullup transistors

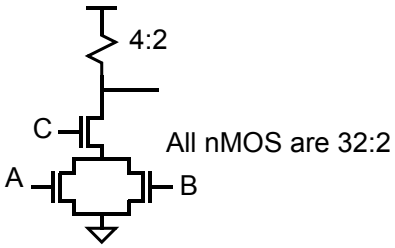
So if the pulldown transistors were each 8:2 devices, and there were 3 of them, the effective pulldown resistance would be 3 times the resistance of each transistor, or 3/4 of a square. For the ratio rule, the pullup device would have to be 3 squares (2:6). And this gate would be slow since it has a high resistance.

Limited series stack to around 3 transistors, to reduce this problem. In nMOS, large fanin gates are always NOR gates.

+Complex Gate Example

Look at building $\overline{(A + B)\, C}$

- Switch network function is (A + B) C



- Since all pulldown paths have two series device, each device must be 8 times as wide as the pullup, (1/8 the resistance)

nMOS Summary

nMOS switch logic (many switch networks, output always driven)

- Degraded outputs
- Control must come from gates

nMOS gates (one switch network to Gnd, default pullup)

- Dense (n+1 transistors for an n-input gate) and fast
- Complete (NANDs and NORs)

But scaling killed it

- Scaling increases the number of gates, and decreases the resistance of the transistors
- Power went through the roof

  Every gate with a low output dissipates power

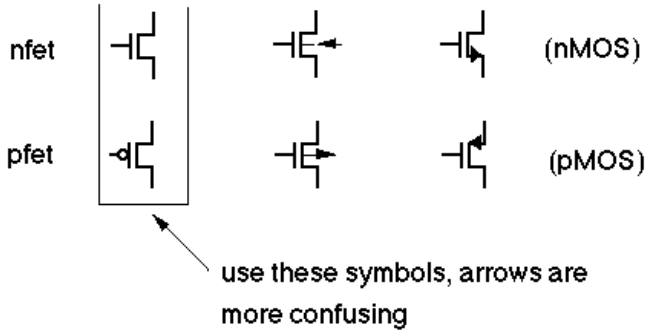
  Chips in the early 80's dissipated watts

CMOS Technology

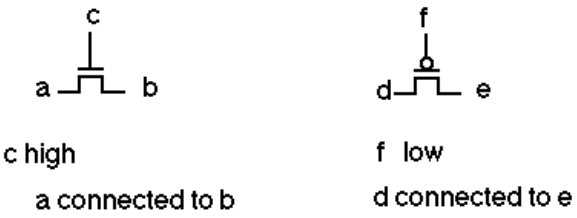
Complementary MOS

Have both nMOS and pMOS devices

There are many way to draw nMOS and pMOS devices



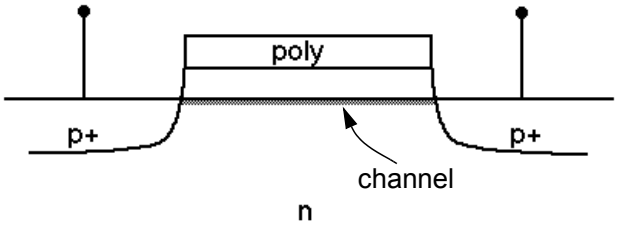
Complementary Transistors



That is the reason for the ‘o’ on the gate of the pMOS device

Like nMOS transistors, pMOS transistors have a threshold voltage, but for a pMOS device it is negative. pMOS devices turns on when the gate is LOWER than the source by more than the threshold voltage. And for pMOS devices the source is the diffusion terminal with the HIGHER voltage on it.

pMOS Transistors

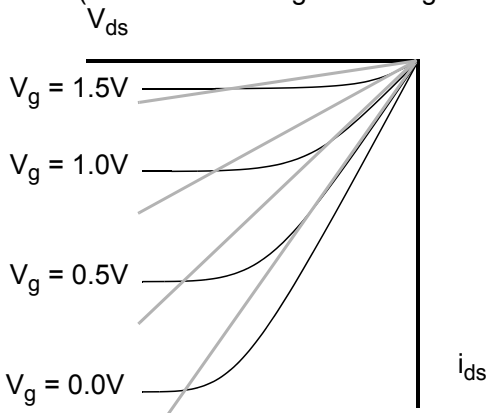


Are like nMOS, except the sign of all the voltages is reversed

- Channel carriers have + charge. Attracted by a negative voltage
- Lowering the gate voltage attracts holes to form a thin p-region. This region allows holes to flow between the two p+ regions. When the channel is not formed, the p+ regions are again isolated by two back to back diodes.

pMOS i - V

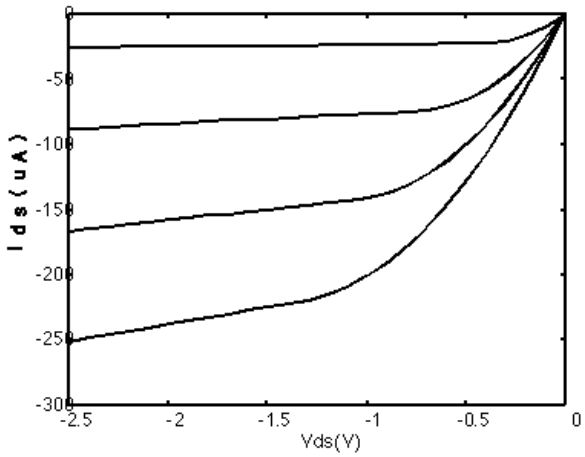
Like nMOS but rotated (current and voltages are negative):



Again use a voltage variable resistor approximation.

Current is about 1/2 nMOS current since the holes are slower than electrons (mobility is smaller)

+ Real pMOS i-V Curve



Transistor is 1μ wide, 0.35μ long

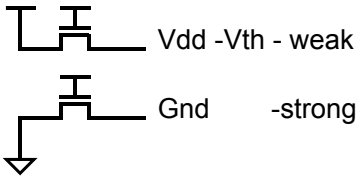
CMOS Transistor Switches

In CMOS you have a richer set of transistor switches

nMOS

connected when gate is high

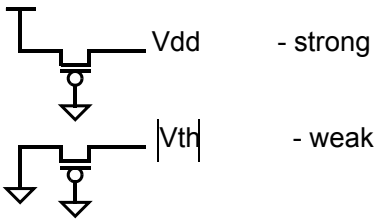
high output is degraded



pMOS

connected when gate is low

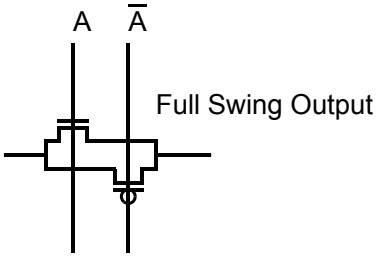
low output is degraded



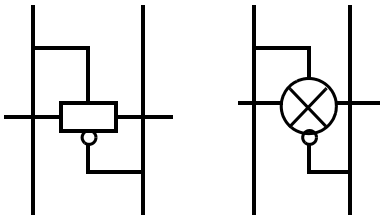
Transmission Gates

By using both nMOS and pMOS neither output is degraded

But you need the true and complement of the control signal



Other symbols used



Using transmission gates, you don't have degraded levels

- Difference between gates and switch logic gets less clear
- Simpler design issues, since there is only one type of signal

So for our CMOS designs you are not allowed to have degraded levels

In CMOS designs, gates are really a kind of switch logic, where inputs can only go to the control terminals of switches, and not source/drains

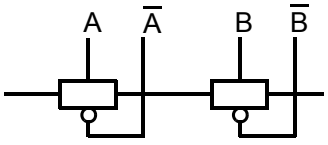
CMOS Switch Networks

In general one needs to use full CMOS transmission gates[1]

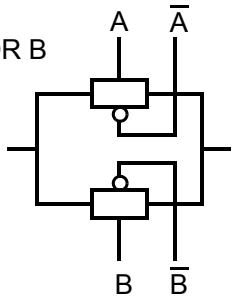
- Two control lines per switch
- No degraded levels

Examples:

A AND B



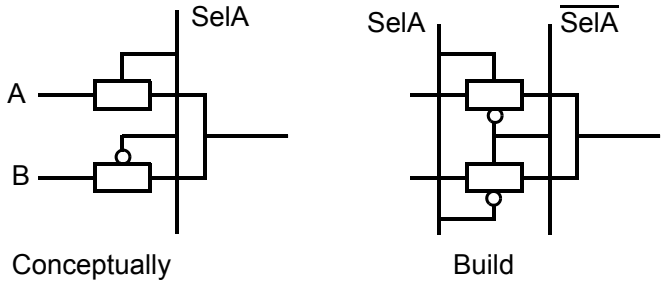
A OR B



1. If the switch network only connects to a constant (Vdd or Gnd) then you don't need both transistors. Connections to Vdd only need pMOS, and connections to Gnd only need nMOS.

Switch Logic

Example: 2-1 Mux:



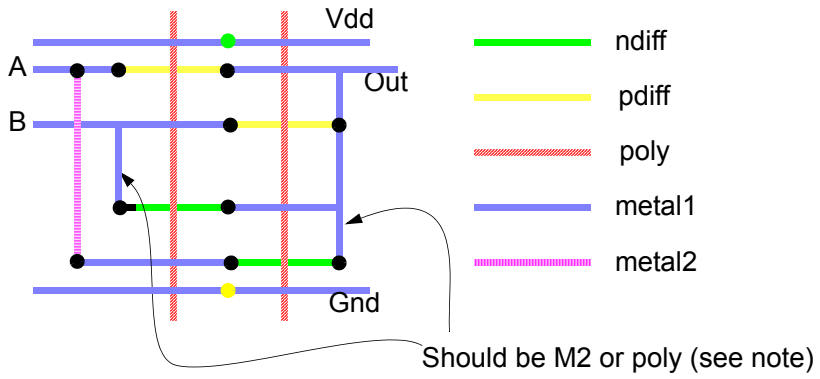
Conceptually Build

CMOS switch logic need a large number of control wires

- Each control is needed in true and complement form
- For 2-1 Mux this works out well, but for a 3-1 mux, this means 6 ctrls
  SelA, SelB, SelC and their complements

2-1 Mux Stick Diagram

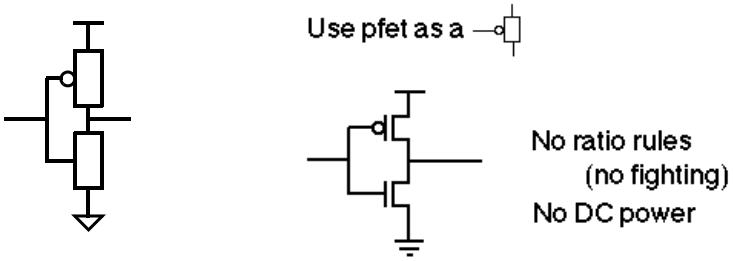
Follow the basic rules from last lecture.



Note that using M1 for the vertical sections is really cheating since then I use M1 in both the vertical and horizontal directions. This is generally bad since it blocks other horizontal M1 from routing in this area. But in this case there are no other M1 lines so it is ok.

CMOS Inverter

Two simple switch networks, one to Vdd and the other to Gnd

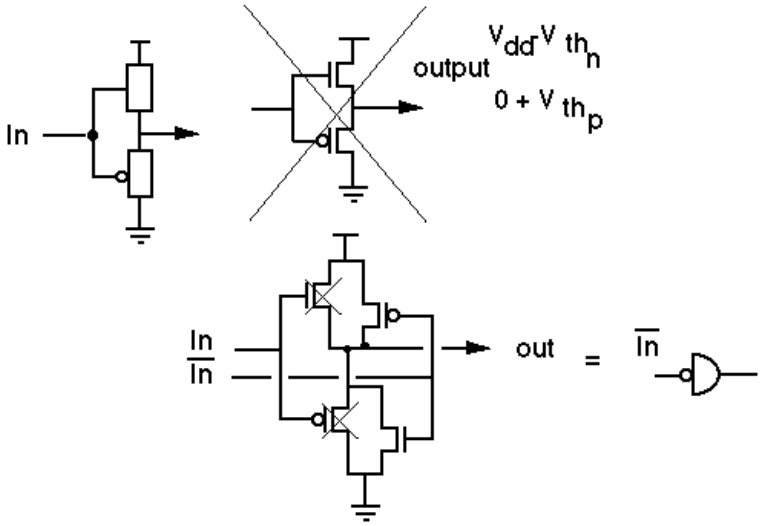


- Can simplify the switches because they connect to constants
- The CMOS inverter does not dissipate DC power since either the path to Vdd or Gnd is off.

  You can build large transistors without worrying about power.
- Gate is more complex than nMOS, you need to drive both transistors

CMOS Buffer

Can build any gate using switches, even a buffer



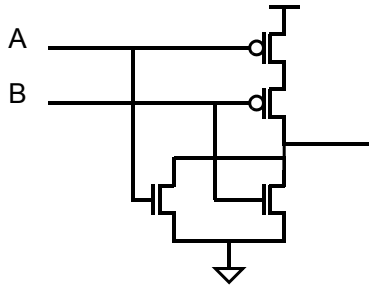
CMOS Gates

Since we have both type of switches, just route the correct supply to the output:

- NOR

  Output is low when either A or B is high

  Output is high when A and B are both low

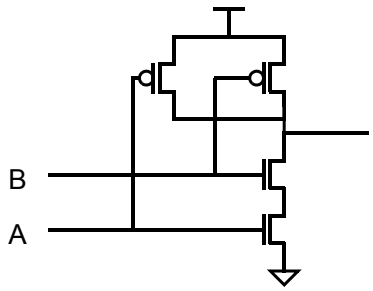


NAND Gate

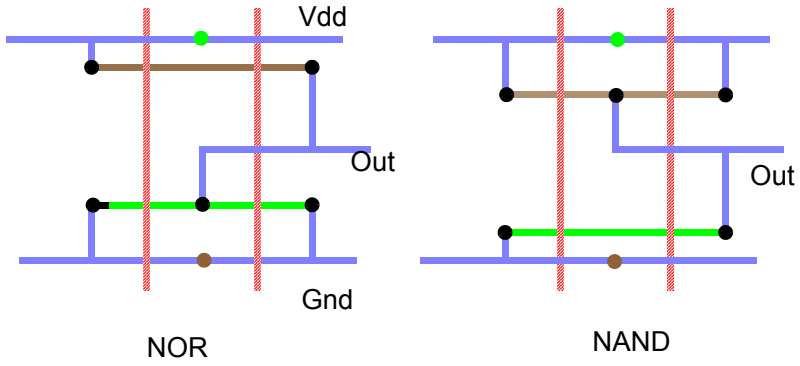
- NAND

  Output is low when A and B are both high

  Output is high when either A or B is low



NAND and NOR Stick Diagram



CMOS Gates

Large Fanin gates are hard to do

- In nMOS, NOR gates were good for large fanin gates
  - No series devices
  - More inputs just mean more parallel transistors
- In CMOS large fanin implies series devices
  - NOR - series pMOS pullup
  - NAND - series nMOS pulldown
- Series devices are slower
  - Resistance of series transistors add
  - Speed set by RC
- For speed reasons, fanin of CMOS is limited to about 3 series devices