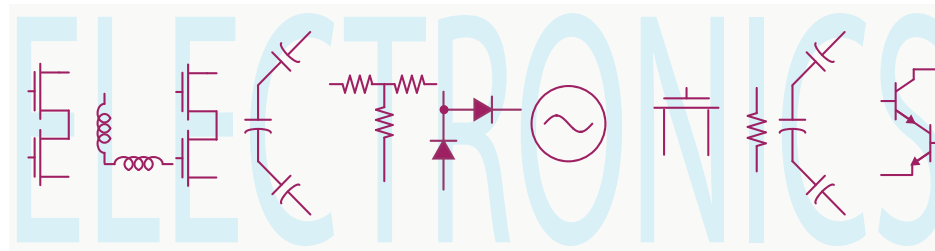


EE 42/100

Lecture 23: CMOS Transistors and Logic Gates



Rev A 4/15/2012 (10:39 AM)

Prof. Ali M. Niknejad

University of California, Berkeley

Copyright © 2012 by Ali M. Niknejad

Logic Levels

- In a real circuit, the actual voltage waveforms take on a continuous range of values. In order for our digital abstraction to work, we have to define the valid range of logic levels that we will interpret as a zero or one.
- Consider one gate driving another gate. For the driver gate, we call the V_{OH} the smallest valid “high” output. Likewise, we call V_{OL} the largest valid “low” output.
- For the receiver gate, we call the smallest valid “high” input level V_{IH} and the largest valid “low” input level V_{IL} . Any input outside this range are in the forbidden zone and the output is indeterminate if the inputs fall in this range.

Noise Margin

- Notice that if $V_{OL} < V_{IL}$ and $V_{OH} > V_{IH}$ we have some *noise margin* in our circuit. In other words, the output of one gate can vary by as much as $NM_H = V_{OH} - V_{IH}$ and the data will still be valid on the high side. Likewise, the low side noise margin is $NM_L = V_{IL} - V_{OL}$.
- This is an important concept because it means that digital logic gates “clean up” their inputs and allow cascading of gates to occur without ever worrying about the NM decreasing. In fact, quite the opposite happens.

DC Transfer Characteristic

- The transfer characteristics for an ideal logic inverter and a real inverter are shown below. Note that for an ideal inverter there's a threshold voltage at the midpoint. If the signal is below the midpoint, the output saturates to the supply. If it's above, it saturates to the ground potential.
- A real curve has a transition region in which the output voltage is not well defined (high or low). This is the forbidden region for the gate.
- The V_{OH} and V_{OL} are defined by the points where dV_o/dV_i is equal to -1 . This is done to maximize the noise margin of the circuit.

CMOS Transistors

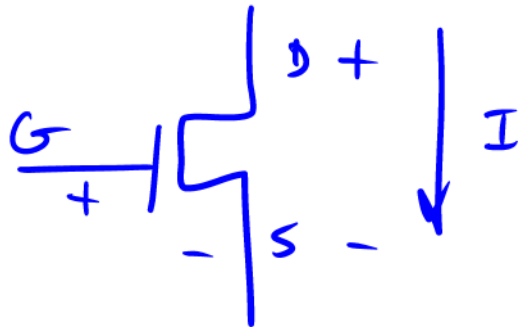
- The vast majority of logic gates today are built from silicon CMOS transistors. CMOS transistors have been the driving force of electronic miniaturization and power scaling for the past few decades.
- CMOS transistors are used to make logic gates, memory cells, and buffers and amplifiers.
- The level of integration has been growing exponentially for the past four decades and today you build a chip with 1 billion transistors!

Fabrication

- CMOS stands for Complementary MOS, and MOS stands for Metal-Oxide-Silicon. A MOS sandwich is the basic structure which forms a capacitor between the gate and the body of the transistor. The structure is fabricated by lithographically defining an opening in the silicon where the native SiO_2 oxide is grown. A p-type silicon is usually used for the body.
- To complete the transistor, two new terminals are added, the source and drain. The source and drain are grown by doping openings in the silicon adjacent to the gate. Under normal operating conditions, the body is biased at the most negative potential (ground), and that means that the source/drain junctions are reverse biased and isolated from the body and from each other.

nMOS Operation

- Ignoring the gate terminal for now, since the source/drain form two back to back diodes, no current can flow through the device.
- If a positive voltage is applied at the gate terminal, it begins to attract negative charge carriers to the surface. It's a capacitor after all and in equilibrium we would expect the positive charge on the gate to be balanced by negative charge at the surface of silicon. Thus we can say that a "channel" begins to form due to the field-effect of the gate terminal.
- If the voltage is made sufficiently large, the channel becomes conductive enough that current can flow from the source to the drain. Thus the gate voltage modulates the conductivity of the channel.
- The device is called a FET, or field-effect transistor.

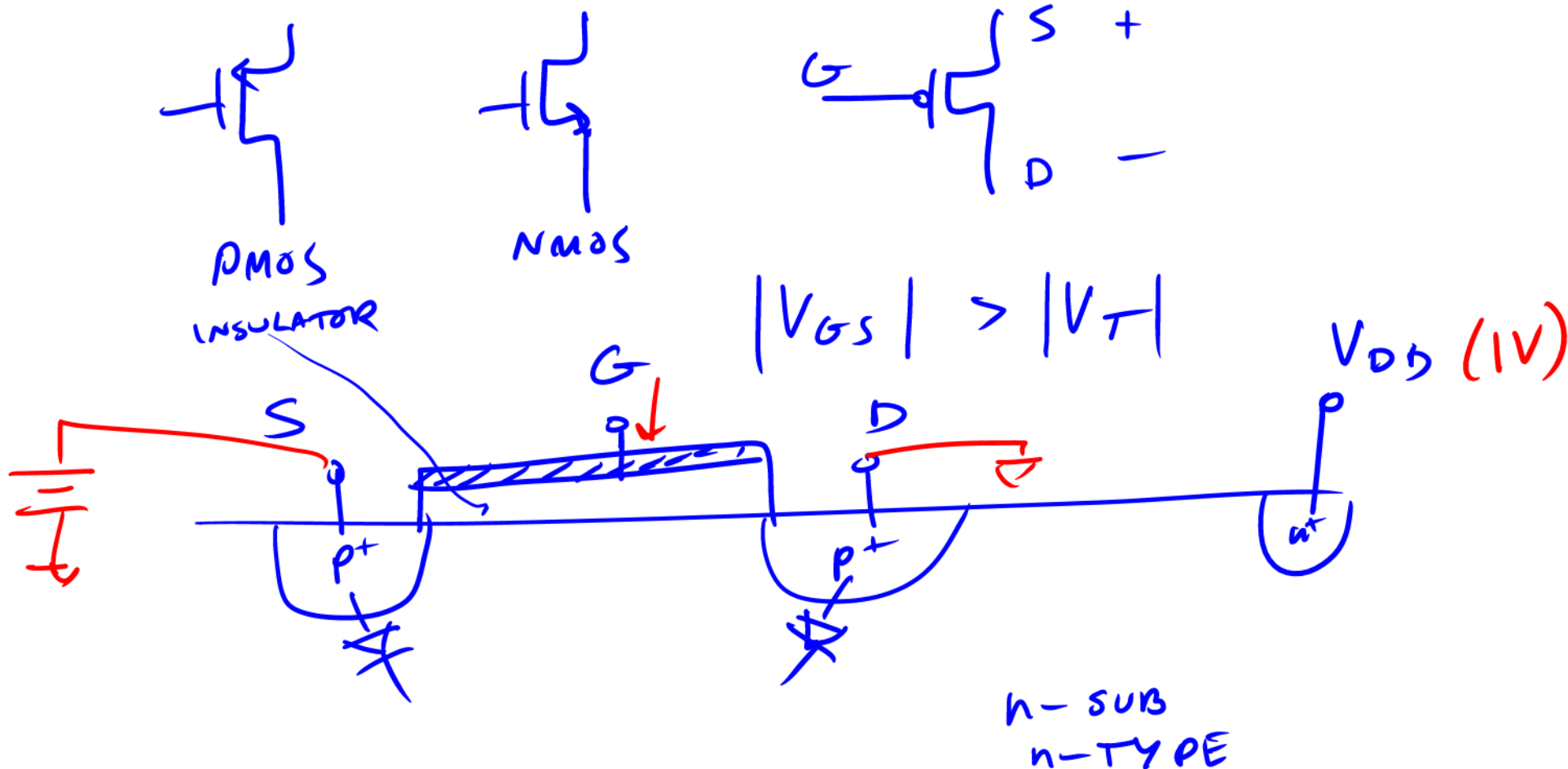


$V_{GS} \geq V_T \Rightarrow$ CHANNEL
CONDUCTS BTWN
 $D \rightarrow S$

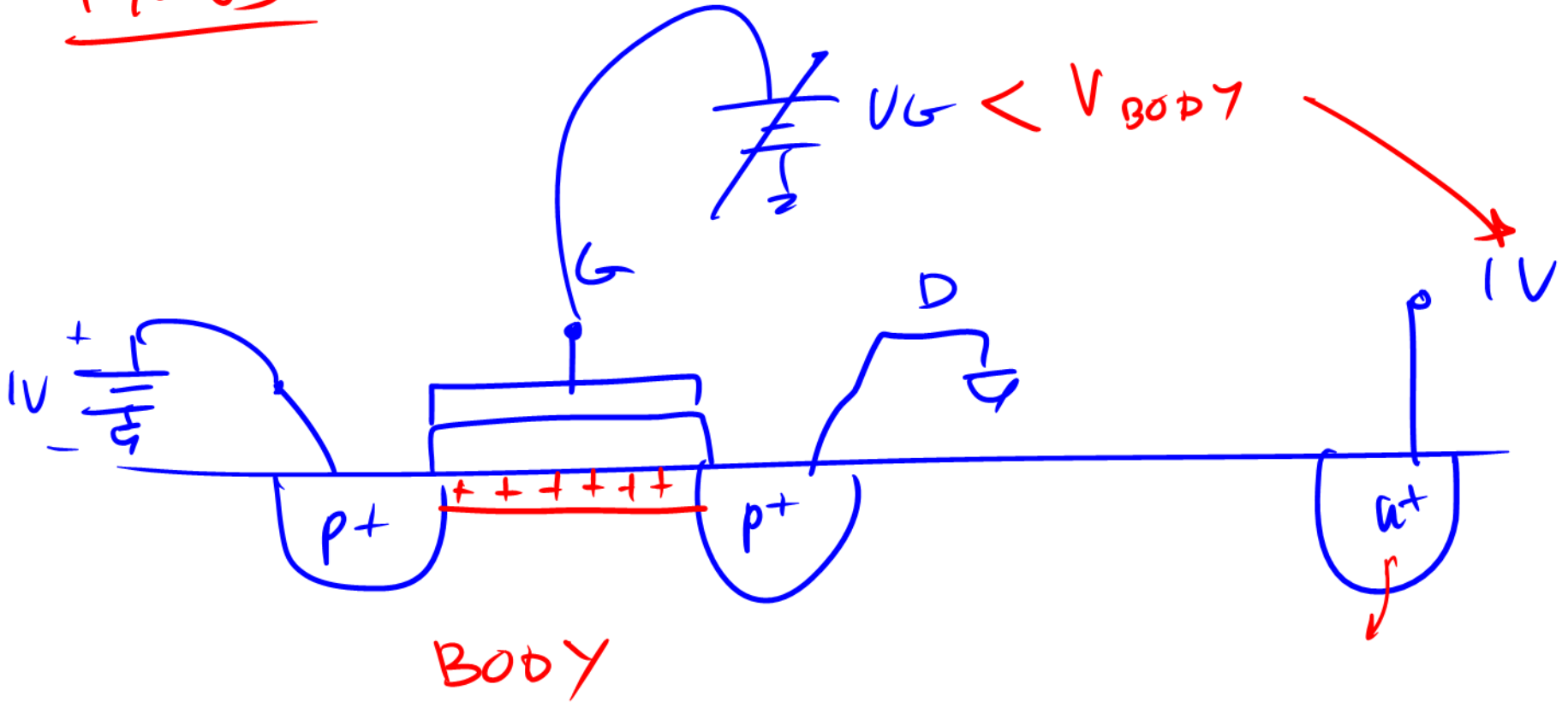
D & S ARE SYMMETRIC

pMOS Operation

- The pMOS device has exactly the opposite or complementary behavior. It's grown in an n-type body and the source and drain are p-type. To get current to flow through the device, therefore, requires positive charges to accumulate at the surface. To do this a negative voltage is applied to the gate terminal.
- The voltage for which conduction begins is called the threshold voltage, V_{TH} .



PMOS



CHANNEL FORMS WHEN

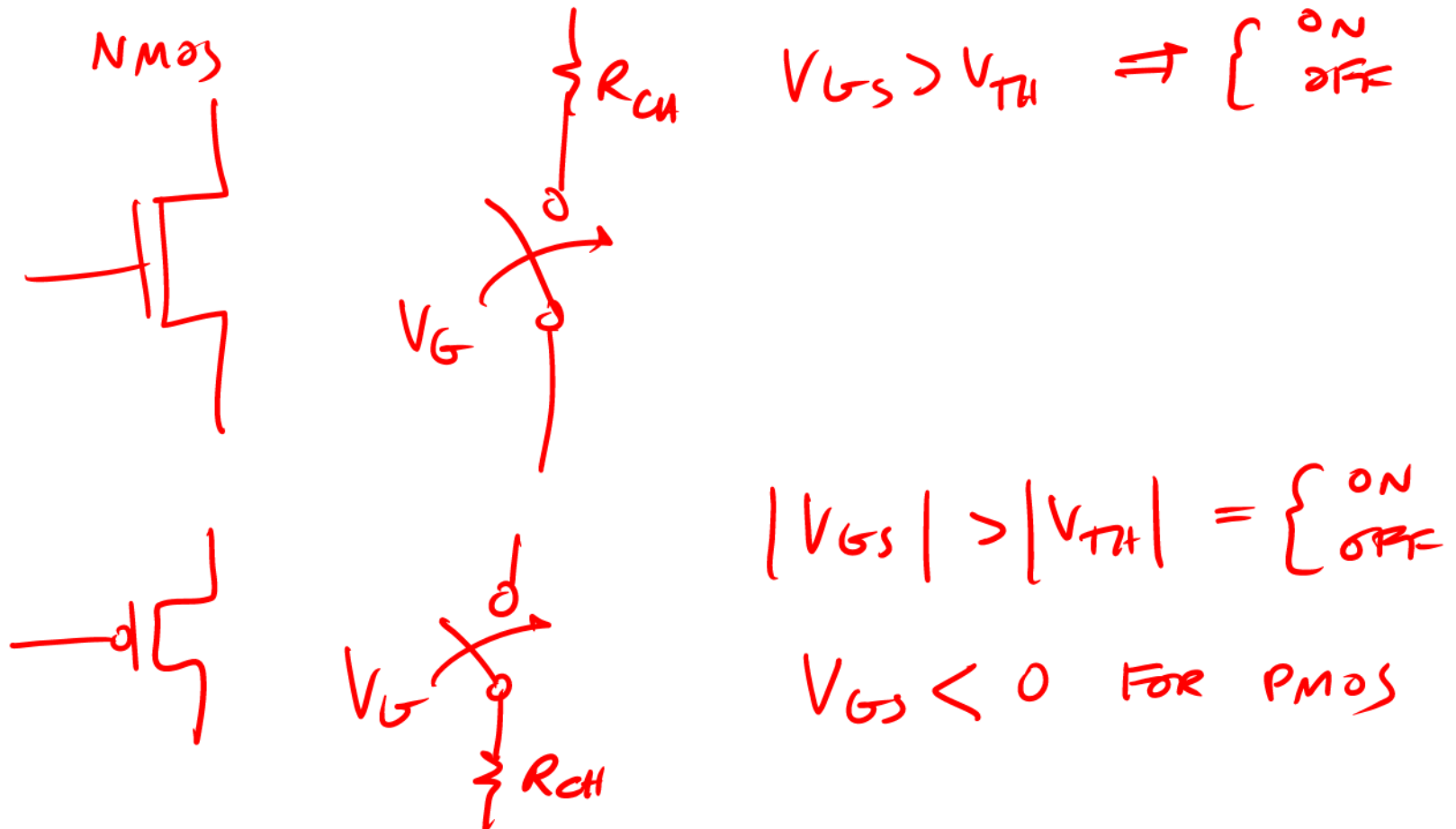
$$V_{G,B} = -V_T$$

$$V_G - V_B = -V_T$$
$$V_G = V_B - V_T$$

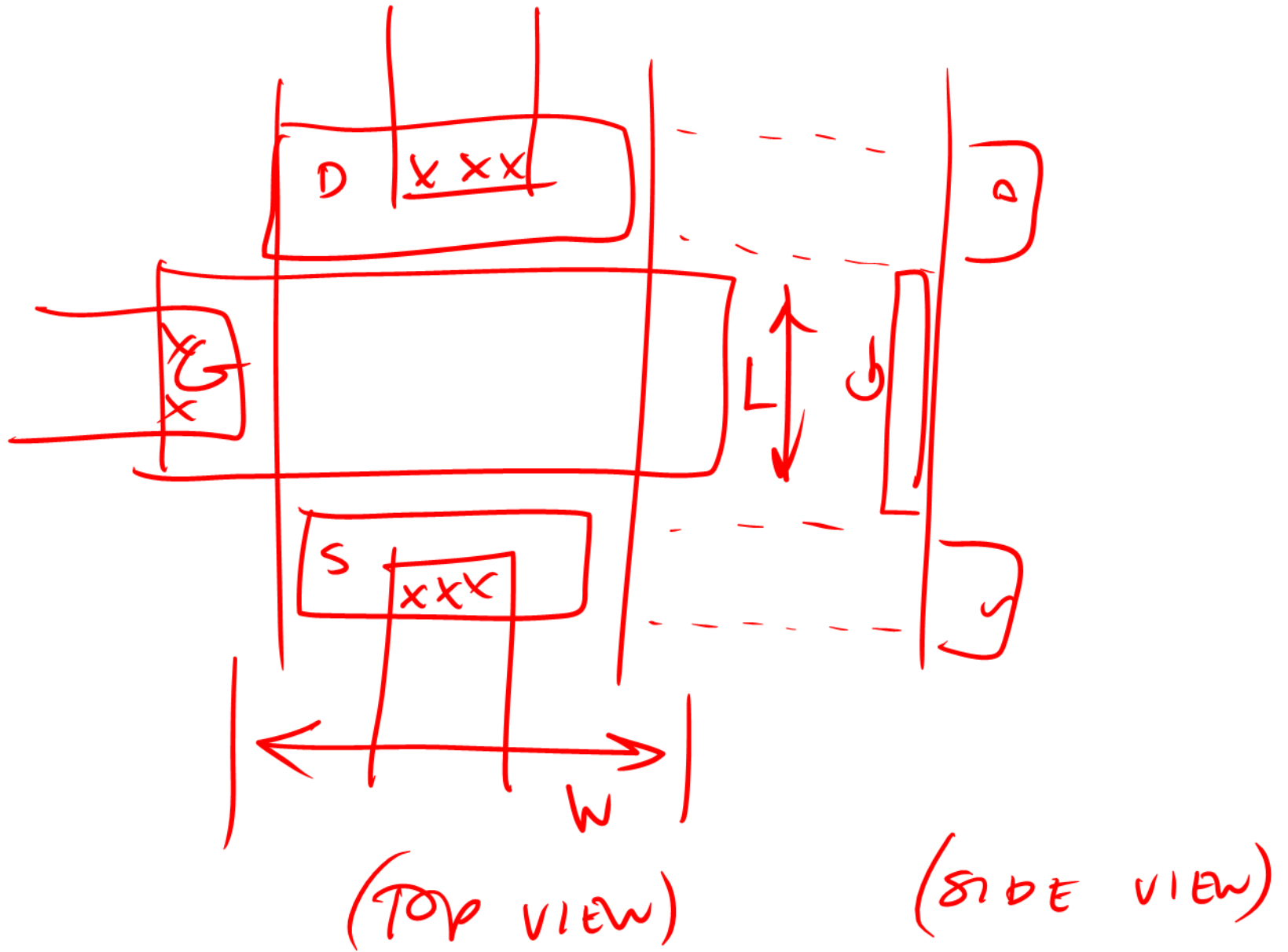
n-TYPE
SUP

Switch Model

- The easiest way to describe a FET transistor is to model it as a switch. In the nMOS device, when a voltage $V_G > V_{TH}$ is applied to the gate terminal, current can flow through the device. It does have “on-resistance” but we can usually ignore that. The gate terminal is insulated from the transistor (it’s like a capacitor) and so no DC current flows into the gate.
- When the gate voltage is too low to form a channel, $V_G < V_{TH}$, virtually no current flows and we model the device as an open circuit.



TOP VIEW



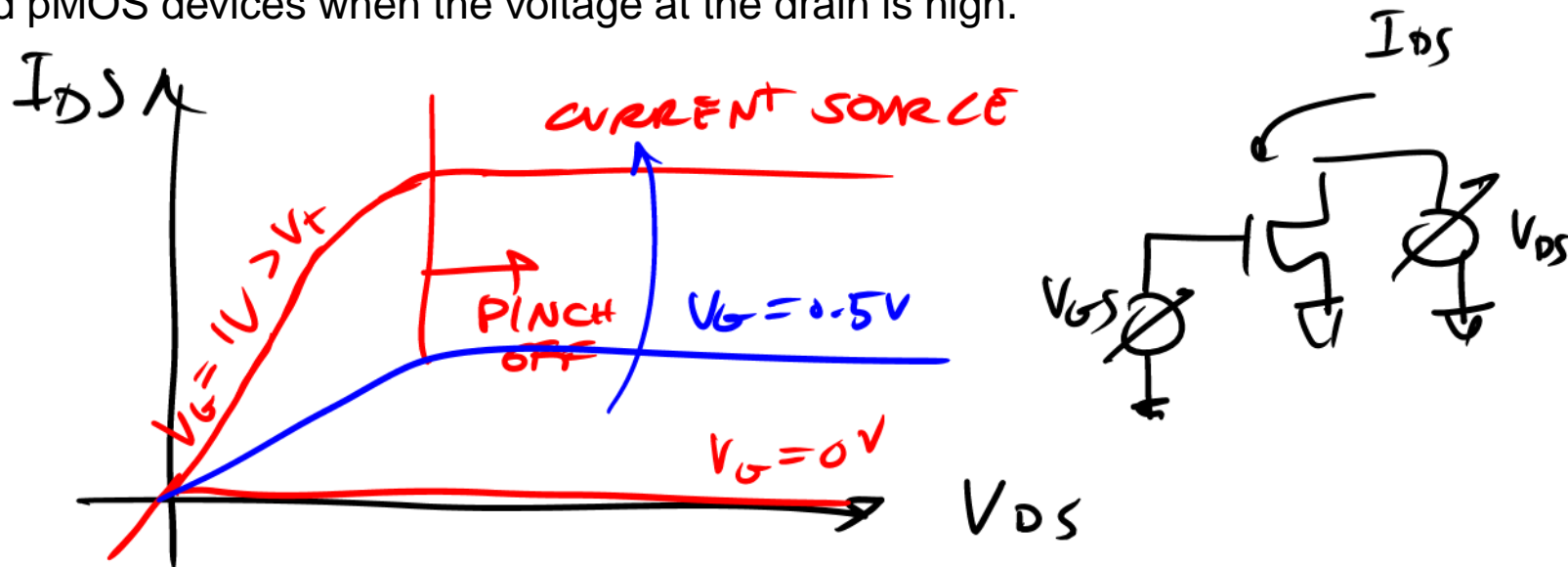
Limit to Switch Operation

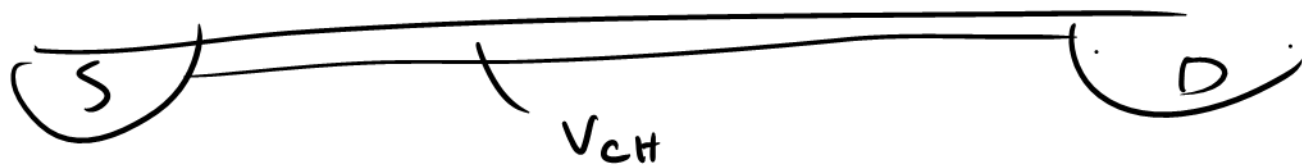
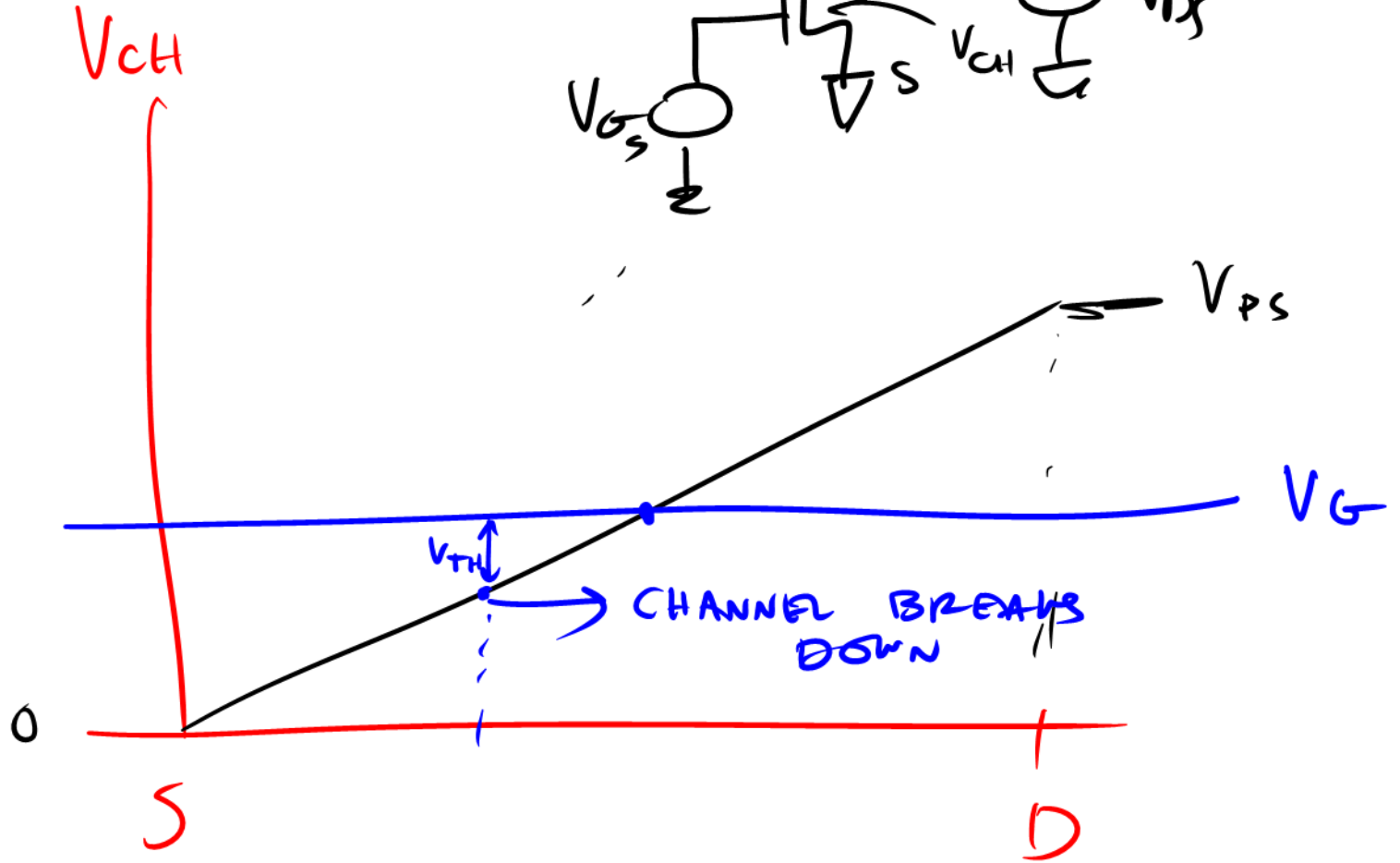
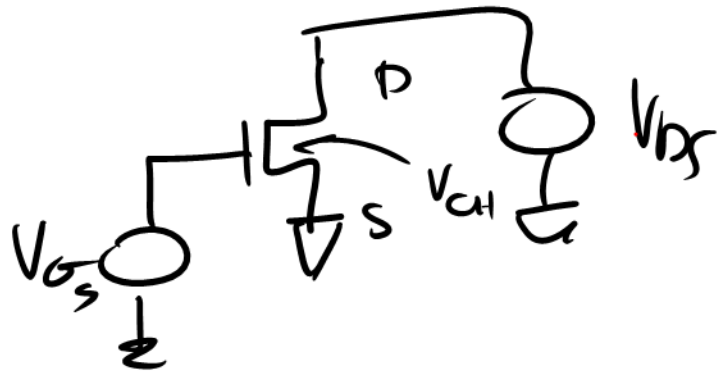
- In the nMOS device, if the drain voltage is made too positive, then the field between the gate and the channel near the drain gets too low and the channel begins to disappear near the drain end. Under these conditions the switch model breaks down.
- To keep this from happening, we like to keep the drain voltage low

$$V_G - V_D > V_{TH}$$

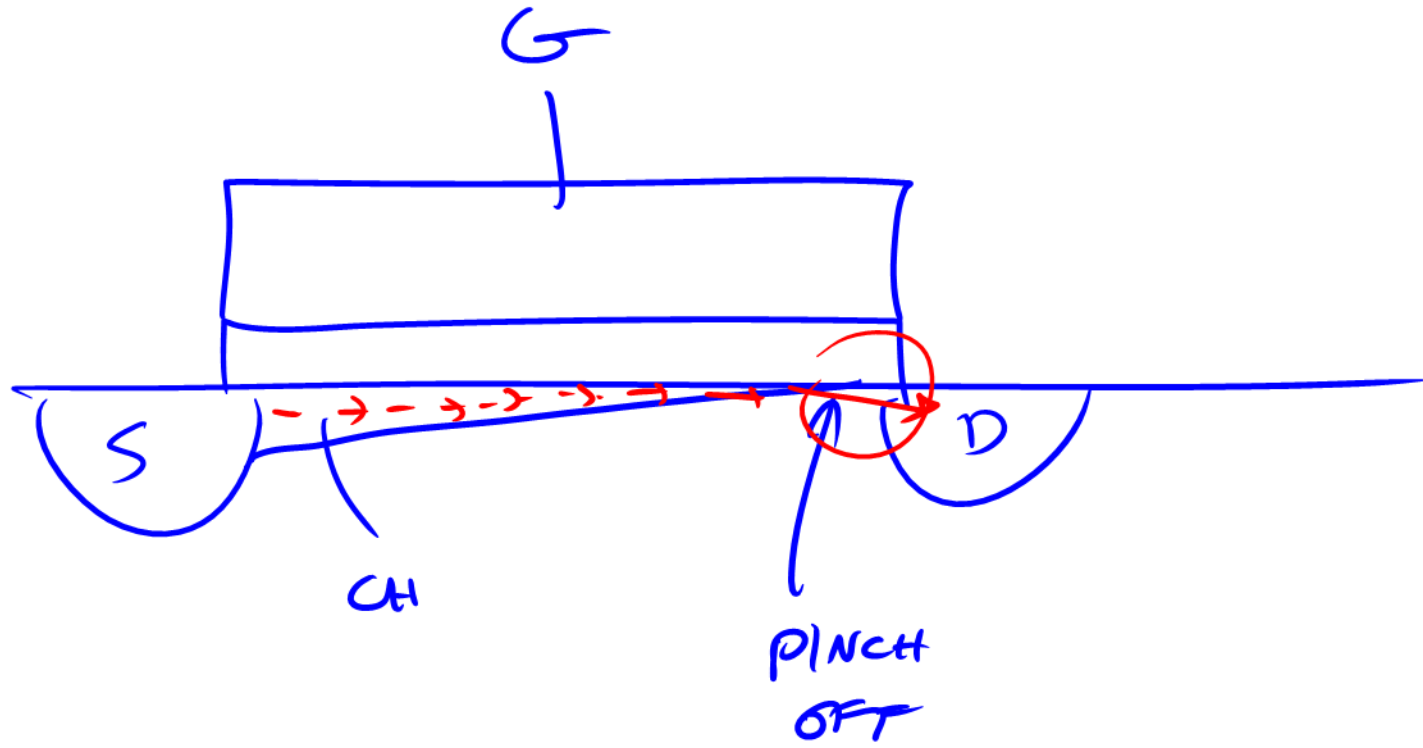
$$V_D < V_G - V_{TH}$$

- Exactly the opposite is true for the pMOS. Therefore to operate the devices as good switches, we should employ nMOS devices when the voltage on the drain is low and pMOS devices when the voltage at the drain is high.

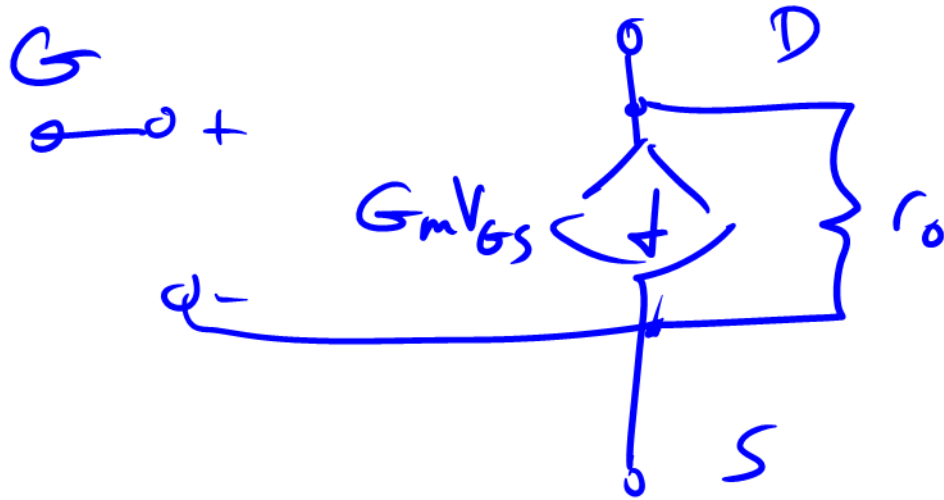




PINCH OFF



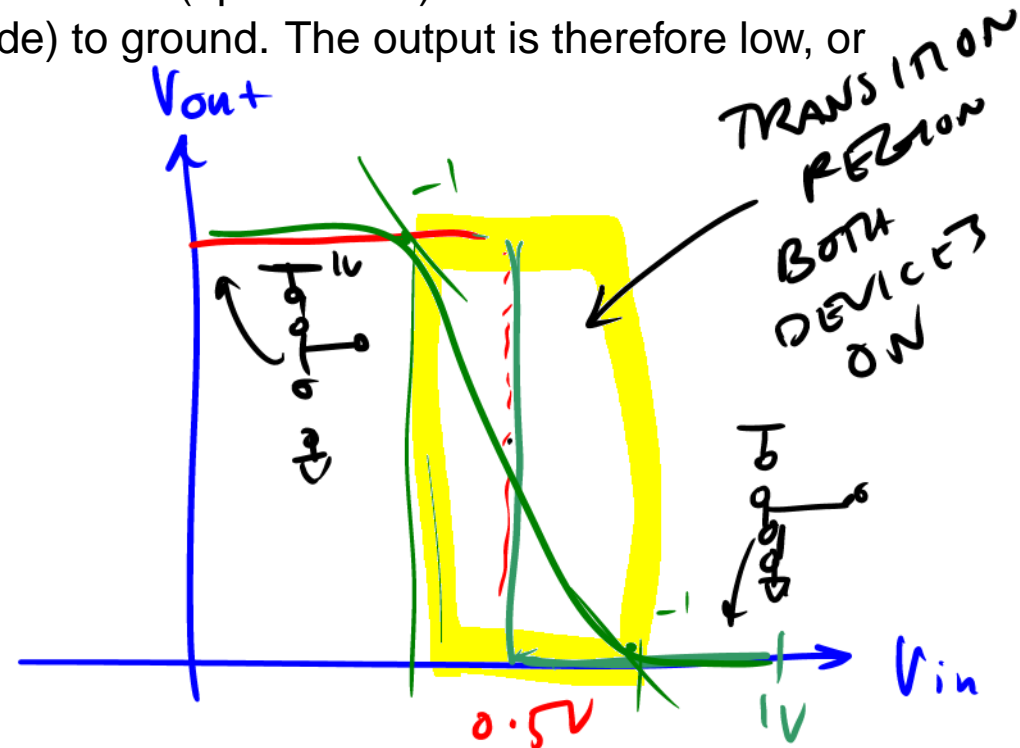
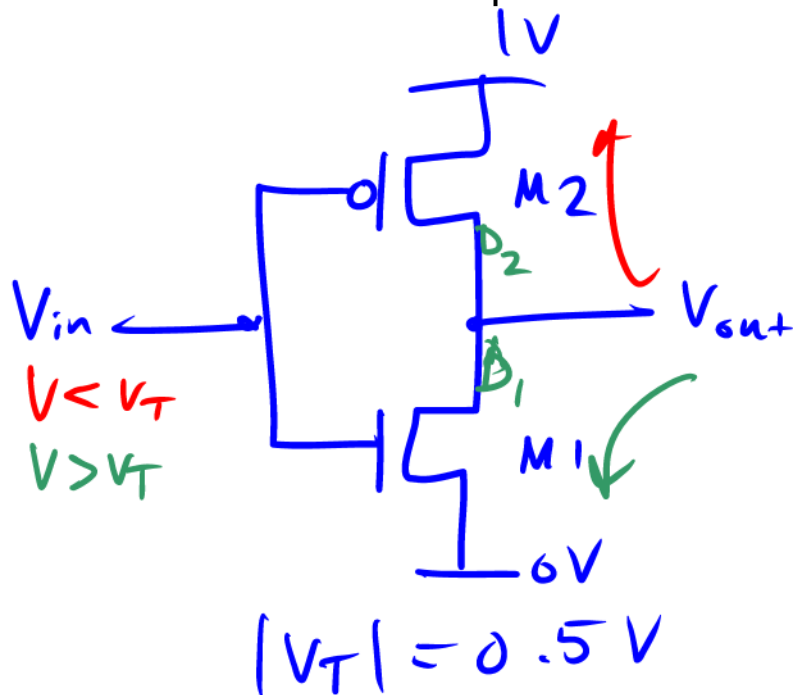
TRANSISTOR IN SATURATION



$$\underline{V_C < V_S}$$

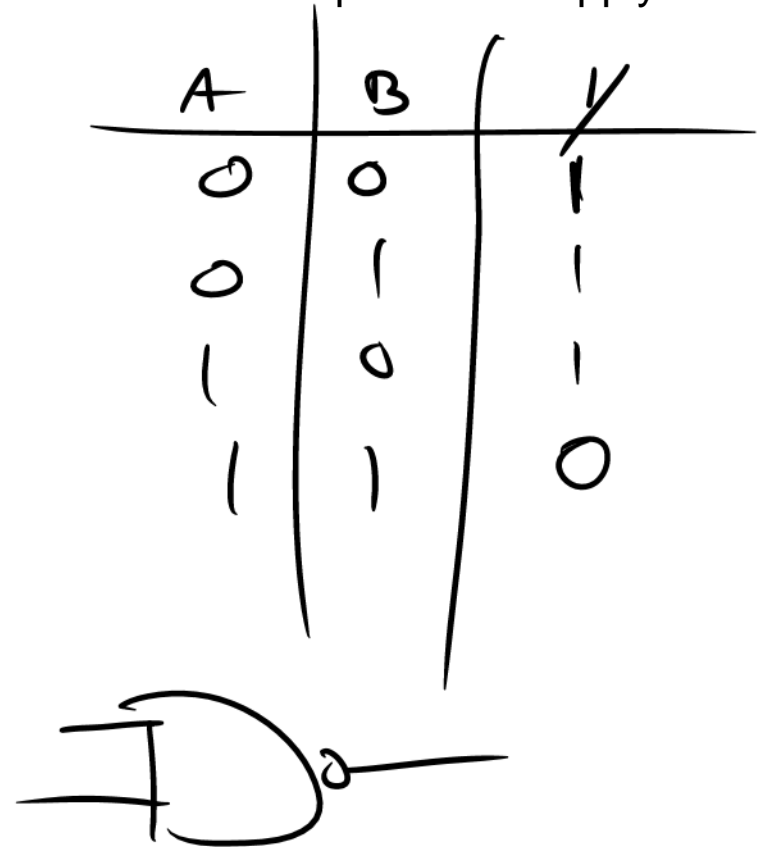
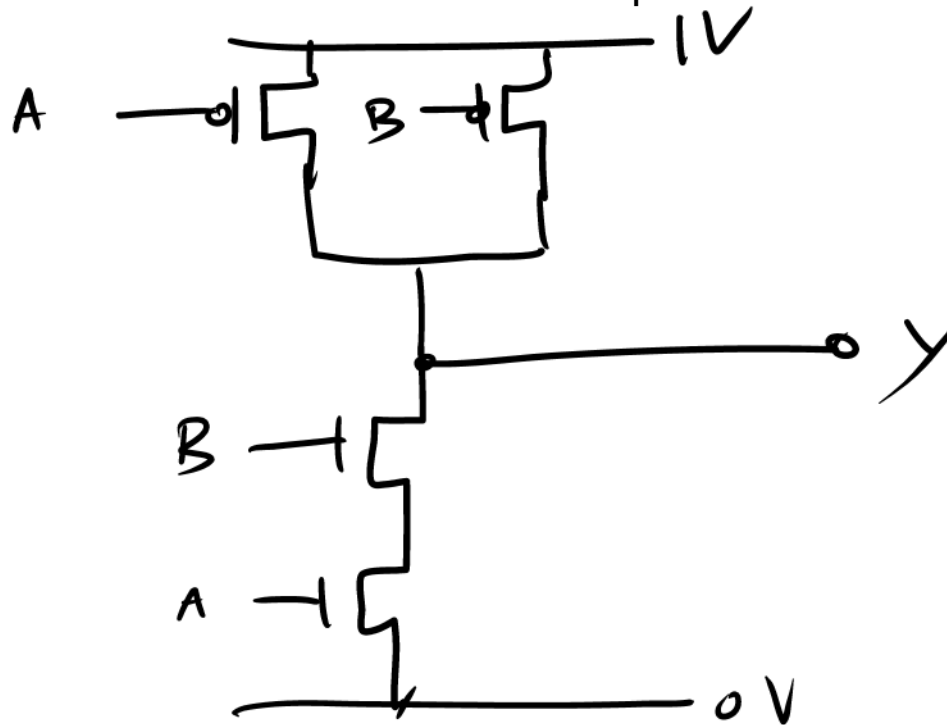
CMOS Not Gate – Inverter

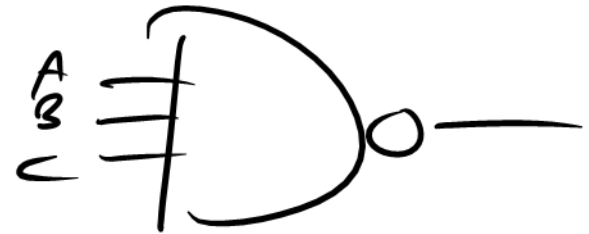
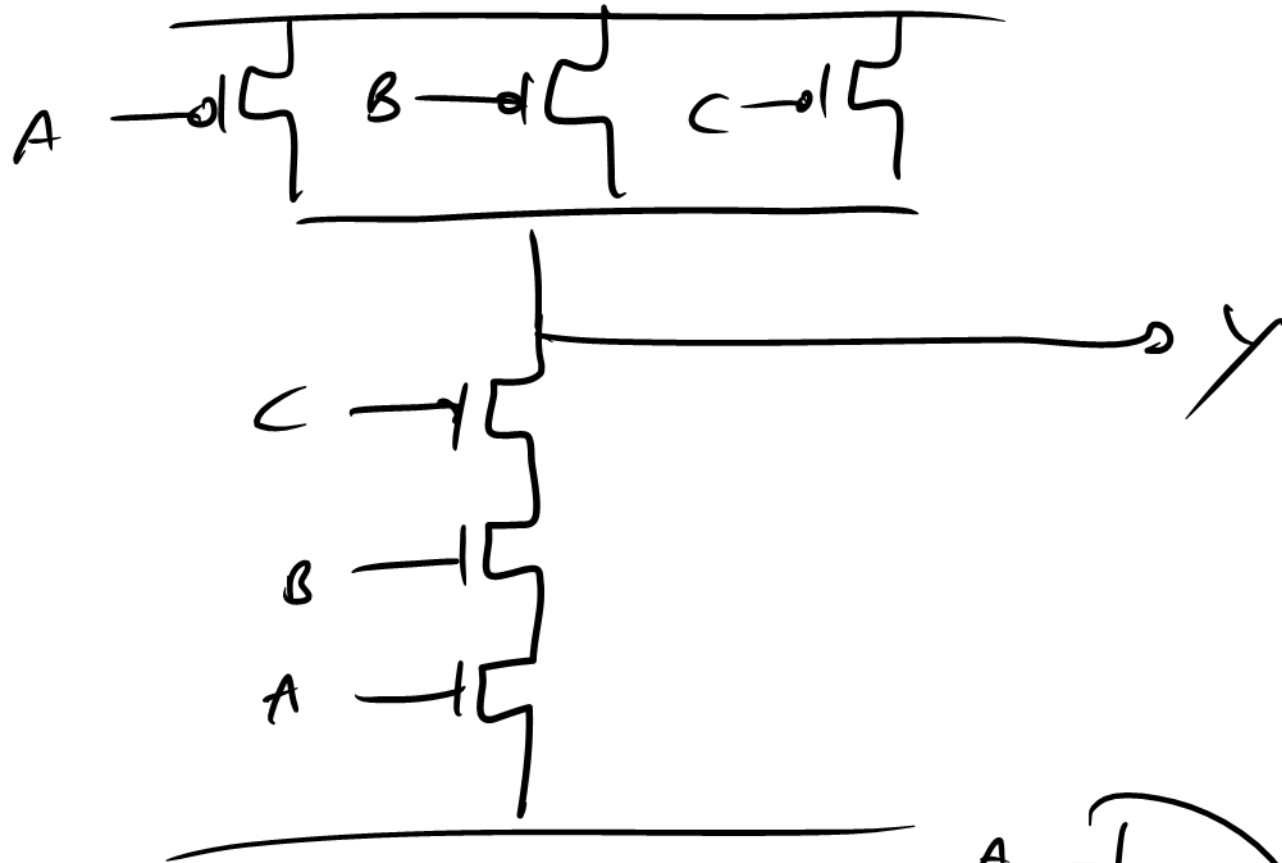
- The CMOS inverter is one of the most important circuit building blocks. The gates and drains of two transistor are connected together. The nMOS is on the bottom and the pMOS is on top. The sources are connected to ground and the supply voltage.
- When a low input is applied, the nMOS is off (open circuit) but the pMOS is on and shorts the output node (drain node) to the supply. The output is therefore high, or inverse of the input.
- When a high input is applied, the pMOS is off (open circuit) but the nMOS is on and shorts the output node (drain node) to ground. The output is therefore low, or inverse of the input.



CMOS NAND Gate

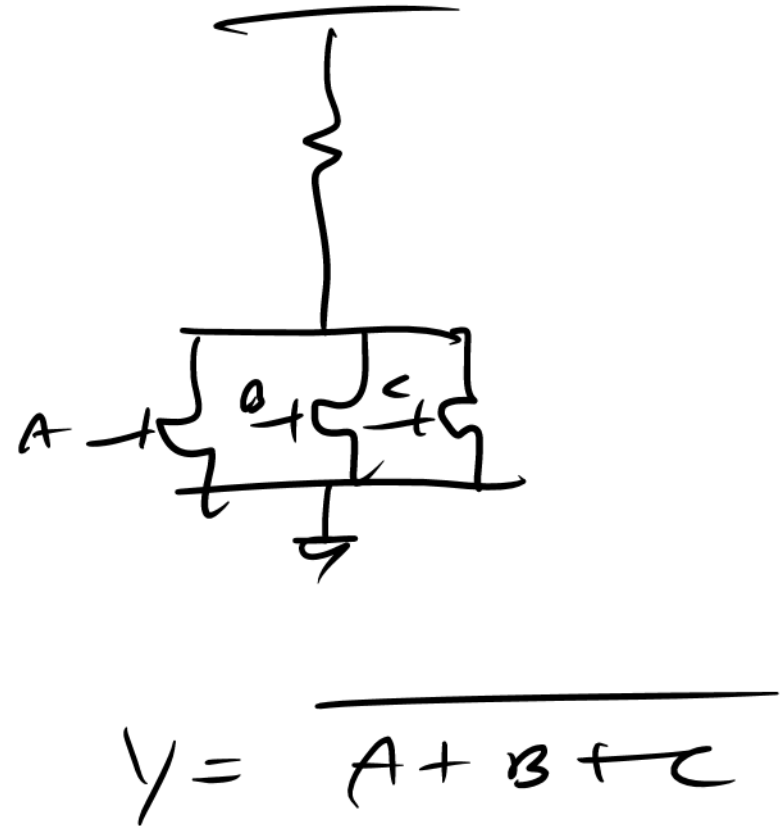
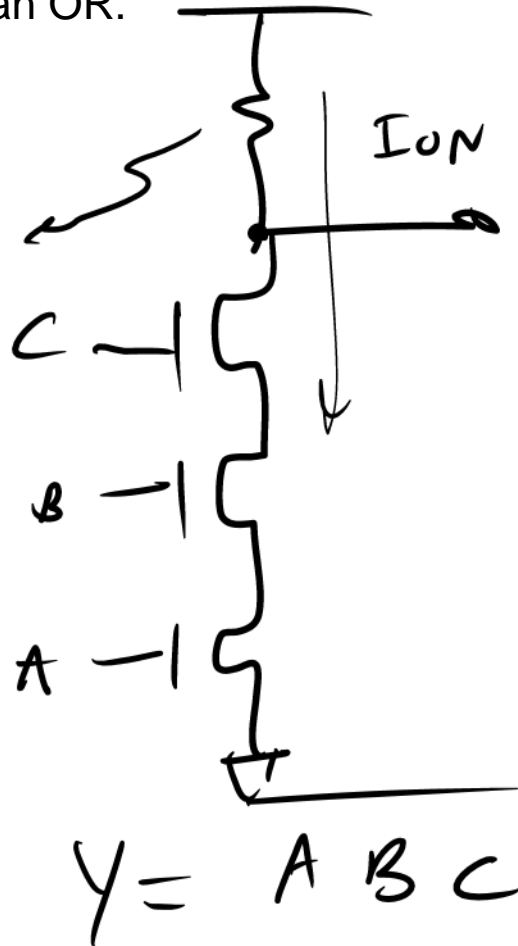
- Consider the circuit below. There are two nMOS transistors in series at the bottom and two parallel pMOS transistors on top.
- When both inputs are high, the nMOS transistors both turn on and short the output to ground.
- When only one input is high, though, there is no conducting path to ground. But then one of the pMOS transistors turns on and shorts the output to the supply. The same is true when both inputs are low.





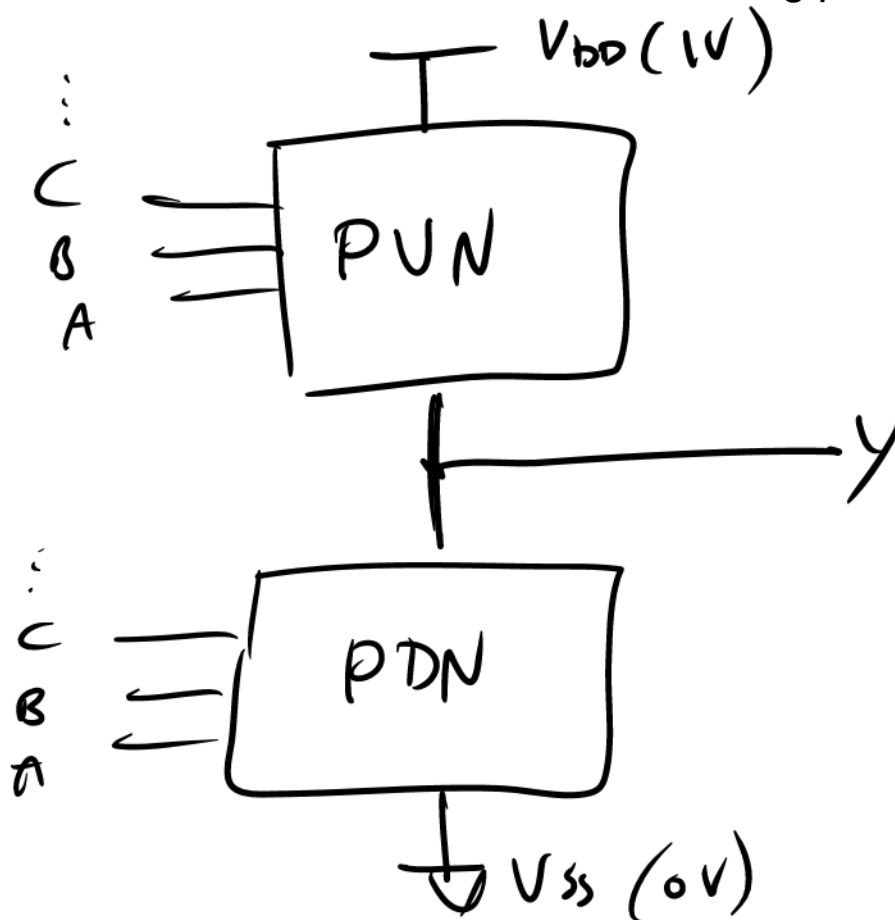
Series and Parallel Connections

- We can generalize the results of the previous examine and note that whenever we put transistors in series, they function as an AND. When in parallel, they function as an OR.



Pull Up and Pull Down Networks

- CMOS logic gates work by having a pull-up and pull-down network. These networks should always complement each other, otherwise a contention would occur and the output would go into an undefined state. When the output is low, the pull-down network should turn on (and pull-up should turn off) and provide a conducting path to ground.
- When the output is high, the pull-up network should turn on and the pull-down network should turn off and a conducting path should be present to the supply.



$$Y = PUN(A, B, C, \dots)$$

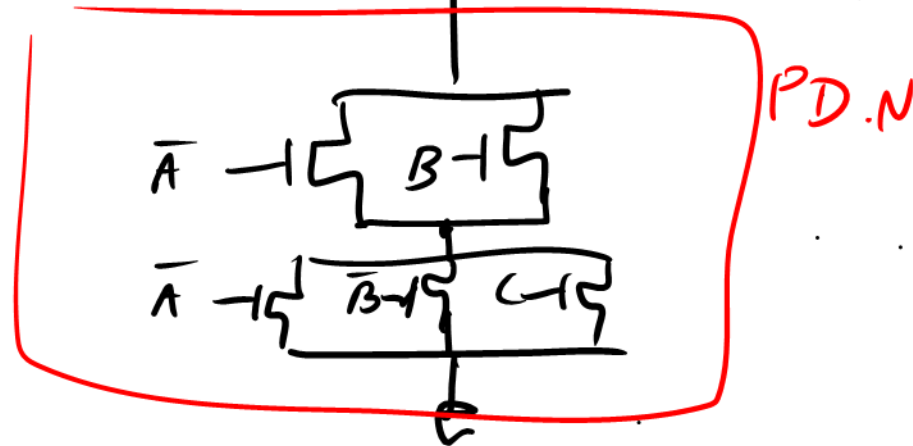
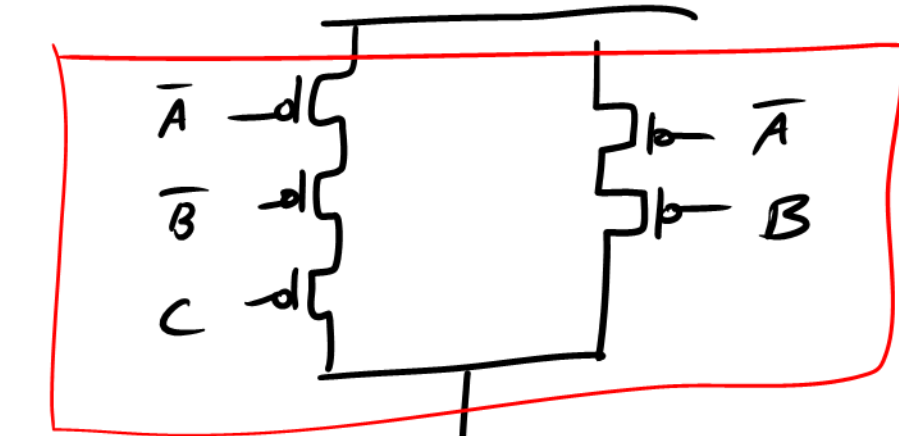
$$PDN(A, B, C, \dots) = \overline{PUN(A, B, C, \dots)}$$

EX

$$y = A B \bar{C} + A \bar{B}$$

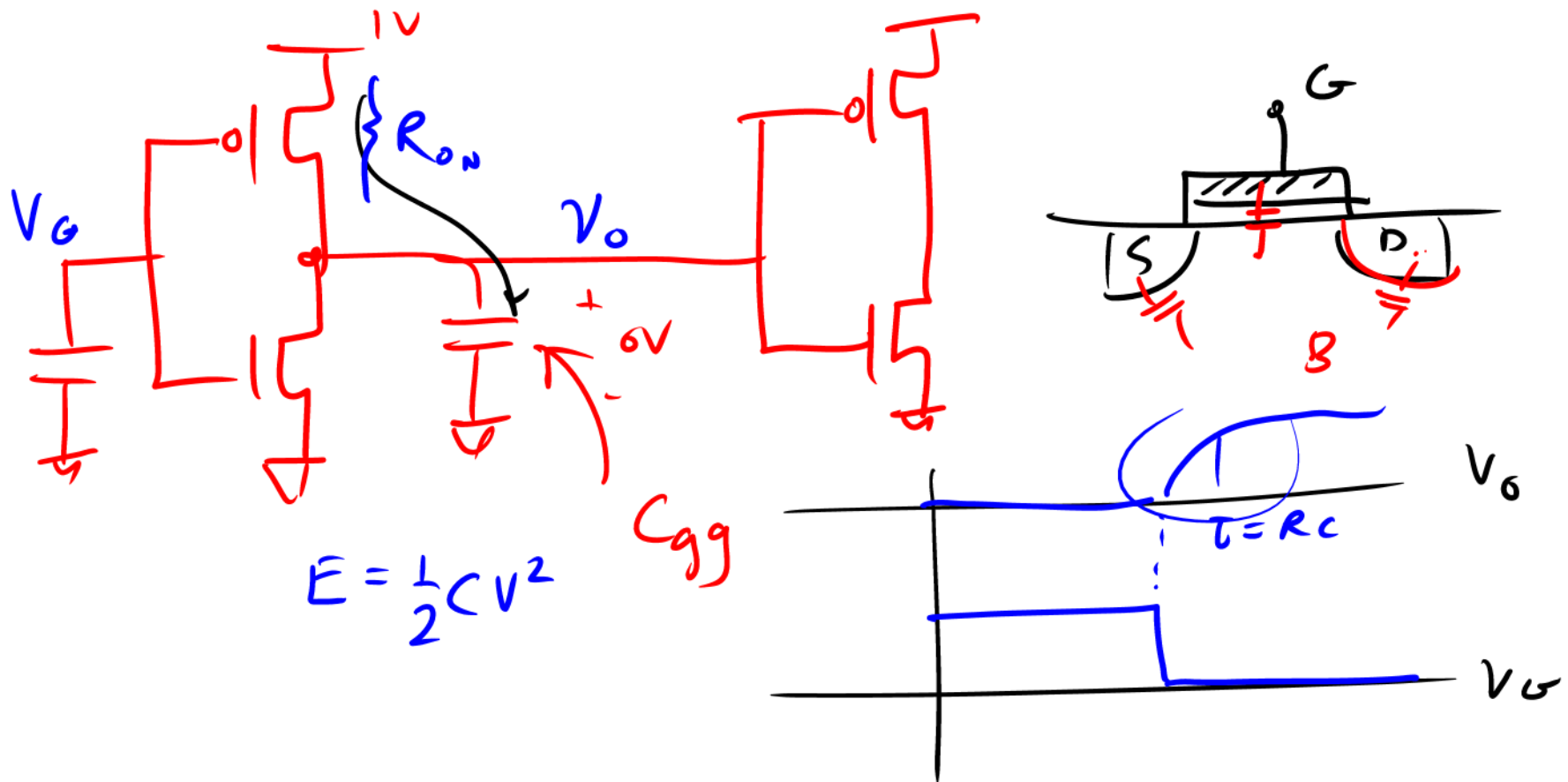
$$\bar{y} = \overline{(A B \bar{C})} \cdot \overline{(A \bar{B})} = (\bar{A} + \bar{B} + C) \cdot (\bar{A} + B)$$

PUN = y



CMOS Power Consumption

- We can estimate the power consumption of CMOS by noting that under ideal conditions no DC power consumption occurs. That's because of the complementary nature of the pull-up and pull-down networks.
- The output node of a CMOS logic gate has capacitance, though, which is from the gates of all the devices loading the output (the load capacitance) as well as the internal capacitance of the output itself (the drain nodes).



CMOS Power Consumption

- We know that to charge the capacitance to a voltage V_{DD} requires charge $q = C_L V_{DD}$ and an energy $E = qV_{DD} = C_L V_{DD}^2$ is drawn from the supply.
- Since this capacitance is charged and discharged during an active cycle, the net energy drawn from supply and then return to ground. The total power consumption is

$$P = E/T = C_L V_{DD}^2 f \alpha$$

- In the above equation, α is the activity factor (switching rate) and it's a statistical quantity because it depends on how often the inputs transition.
- The above equation has been the guiding light for CMOS transistor design for the past several decades. Voltage levels have scaled from tens of volts down to 1V.
- As the devices have gotten smaller and smaller, the C_L term has scaled but the sheer number of transistors has gone up and switching speeds have gone up $f \sim 1$ GHz today. Just imagine 10 million transistors switching, each with 3fF of capacitance, at a rate of 1 GHz. That's a whopping 30W of power!
- Today instead of making transistors run faster, it's better to use parallelism to run multiple cores at lower speeds but yet try to get the same amount of work done.