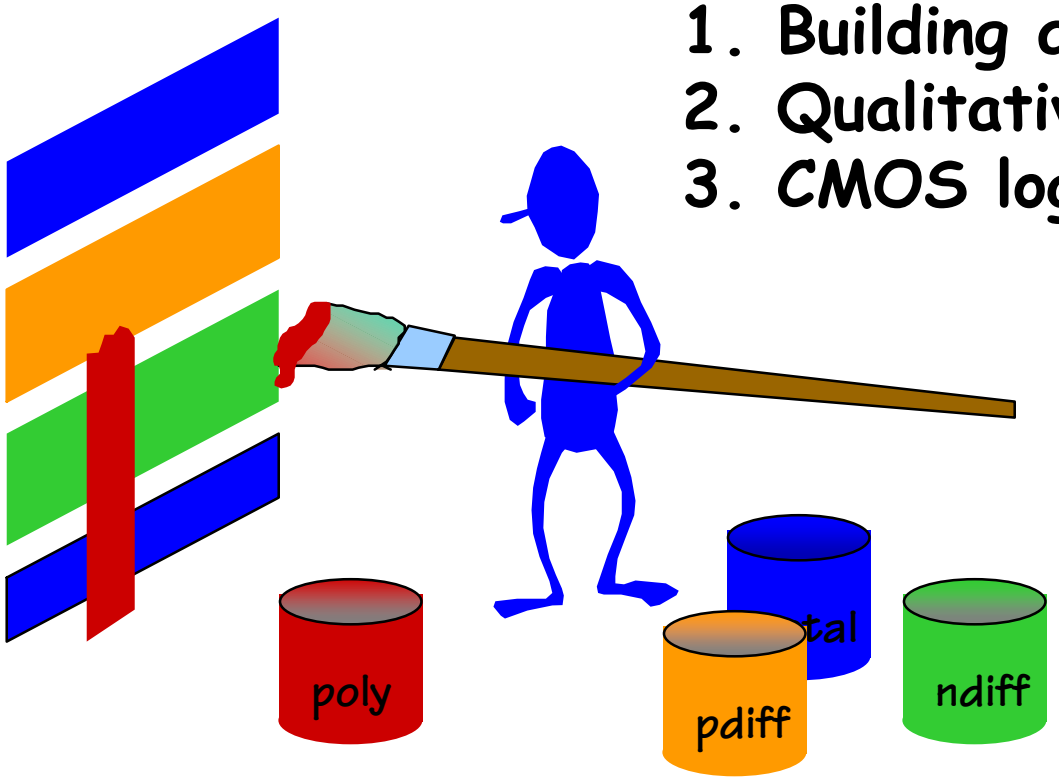
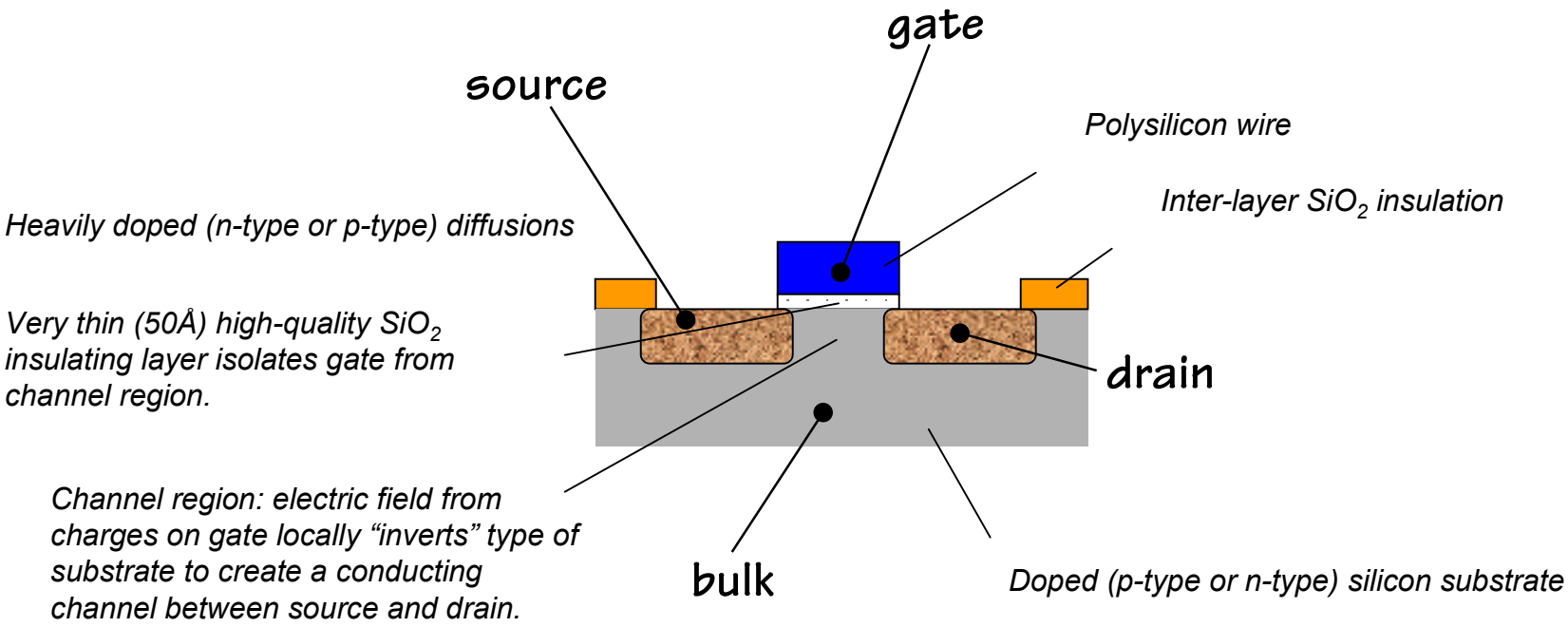
CMOS Technology



1. Building a MOSFET
2. Qualitative MOSFET model
3. CMOS logic gates

Handouts: Lecture Slides

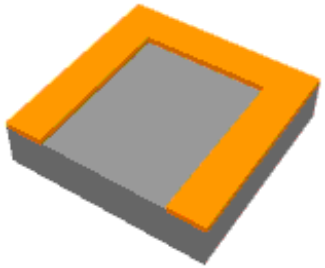
MOSFETS: Gain & non-linearity



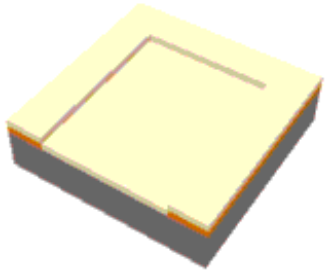
MOSFETs (metal-oxide-semiconductor field-effect transistors) are four-terminal voltage-controlled switches. Current flows between the diffusion terminals if the voltage on the gate terminal is large enough to create a conducting "channel", otherwise the mosfet is off and the diffusion terminals are not connected.

Why are MOS devices King?

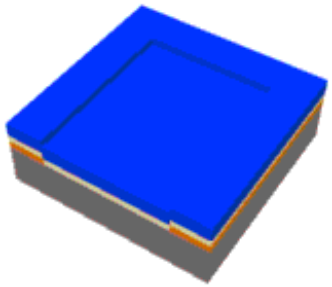
Advantage #1: Self-Aligning Gates



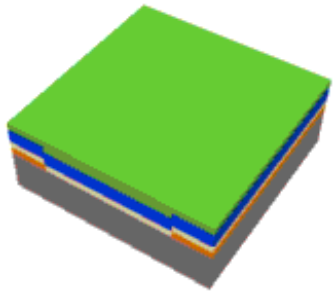
1. We first etch out regions where we want to place our MOSFETs.



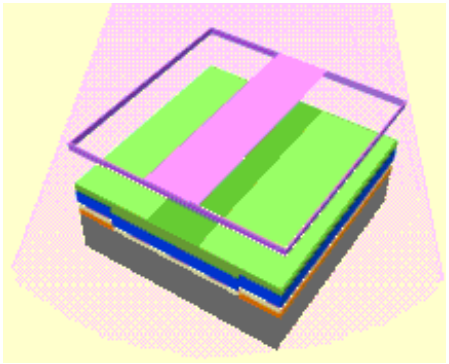
2. Then cover the whole chip surface with a thin insulating oxide layer.



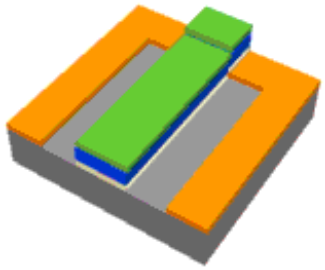
3. Cover that surface with a conducting "gate" material.



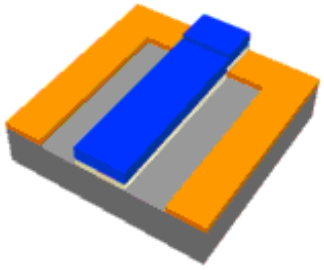
4. Then, we cover it with a "photo-sensitive" compound.



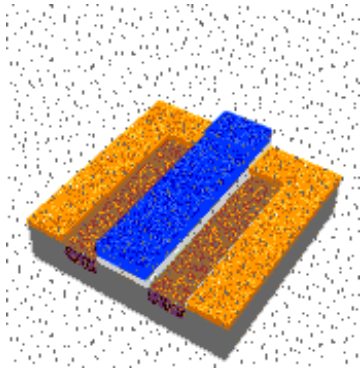
5. We expose the chip to an optical "mask".



6. We etch away the exposed materials.



7. Remove the photoresist.

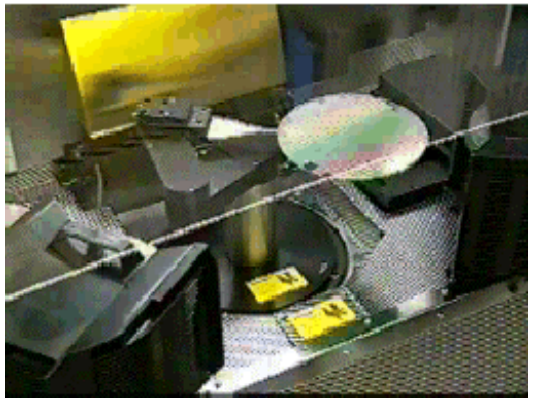


8. Bombard the substrate with ions. The gate protects the region beneath it from implantation.

Let's build a MOSFET

Start with a 500μ slice of a silicon ingot that has been doped with an acceptor (typically boron) to increase the concentration of holes to $10^{14}/cm^3$ - $10^{18}/cm^3$. At room temperature, all the dopants in this *p-type* material are ionized, turning the silicon into a *semiconductor*.

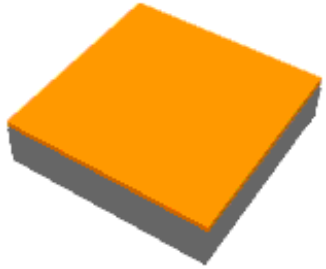
We'll build many copies of the same circuit onto a single wafer. Only a certain percentage of the chips will work; those that work will run at different speeds. The yield decreases as the size of the chips increases and the feature size decreases.



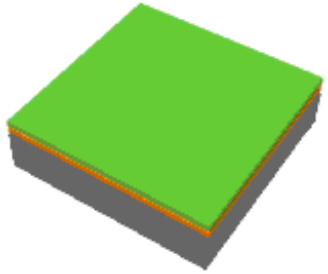
Wafers are processed by automated fabrication lines. To minimize the chance of contaminants ruining a process step, great care is taken to maintain a meticulously clean environment. So put on your bunny suits and let's begin…

Creating patterns on the wafer

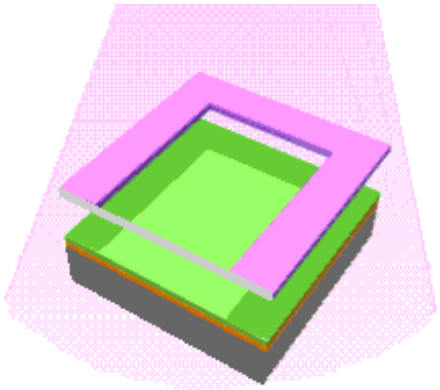
Images from: http://www.intel.com/education/chips/index.htm



A "thick" (0.4u) layer of SiO_2 is formed by oxidizing the surface of the wafer with wet oxygen (we rust it!). The SiO_2 will erve as insulation between the conductive substrate and subsequent conductive layers we'll build on top of the oxide.

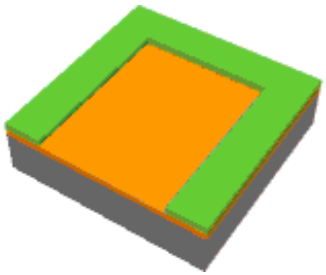


Now we'll form a pattern in the SiO_2 using a mask & etch process. First the wafer is coated with a layer of photoresist. Photoresist becomes soluable when exposed to ultraviolet light…

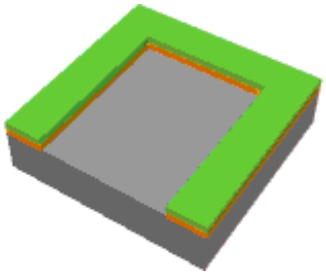


Using a mask to protect parts of the wafer, we'll expose those portions of the wafer where we want to remove the photoresist. We'll use different masks when creating each of the different structures on the wafer.

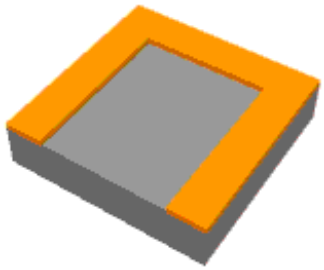
The etching process



The exposed photoresist is removed with a solvent. The unexposed photoresist remains, masking portions of the underlying SiO_2 layer.

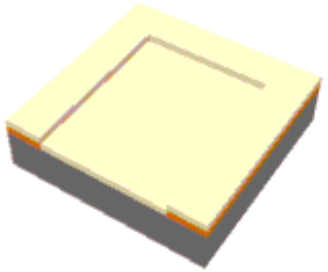


A chemical etch is then used to remove the revealed silicon dioxide.

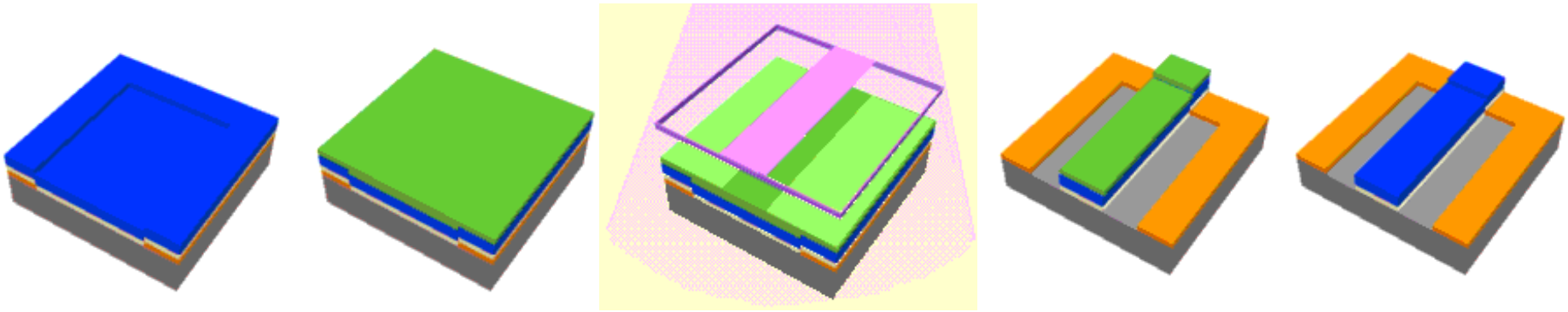


Finally, the remaining photoresist is removed with a different solvent and we're left with pattern of insulating SiO_2 on top of exposed p-type substrate.

Gate oxide & polysilicon

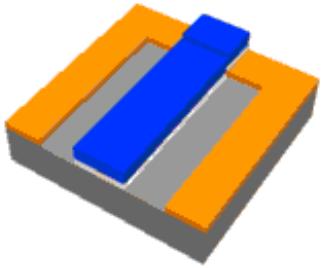


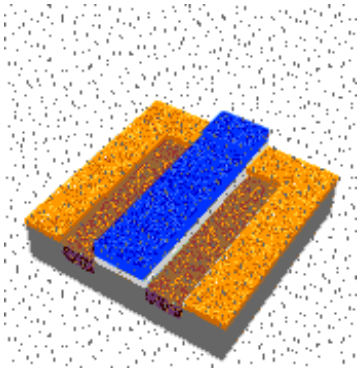
Now a "thin" (100 Å) layer of SiO_2, called *gate oxide*, is grown on the surface. The gate oxide needs to be of high quality: uniform thickness, no defects! The thinner the oxide, the more oomph the FET will have (we'll see why soon) but the harder it is to make it defect-free. Coming soon to a fab near you: 50 Å gate oxide...

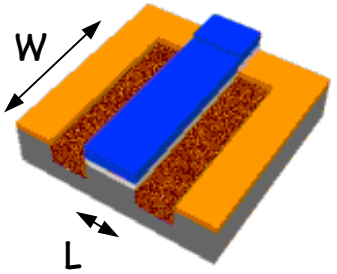


On top of the thin oxide a 0.7u thick layer of polycrystalline silicon, called *polysilicon* or *poly* for short, is deposited by CVD. The poly layer is patterned and plasma etched (thin ox not covered by poly is etched away too!) exposing the surface where the source and drain junctions will be formed. Poly has a high sheet resistance of 20 Ω/sq which can be reduced by adding a layer of a silicided refractory metal such titanium ($TiSi_2$), tantalum ($TaSi_2$) or molybdenum ($MoSi_2$) => 1, 3 or 5 Ω/sq.

Source/drain diffusions



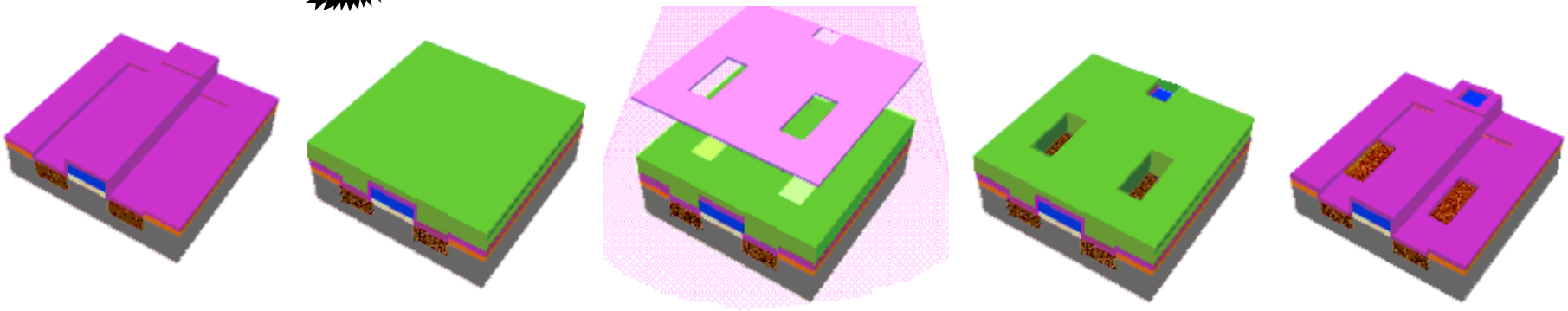




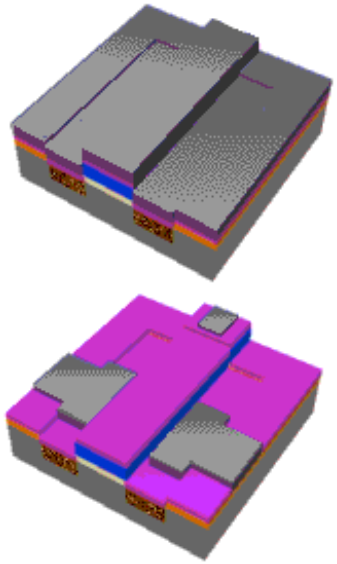
Donor implants are used to create self-aligned MOSFET source/drain diffusions and substrate contacts. Usually As is preferred to obtain shallow N-type diffusions and minimal lateral diffusion. High doses are needed to make low resistance (25 Ω/sq) diffusion wires. Afterwards a short thermal annealing step is performed to repair surface damage caused by the implantation.

This completes the construction of the MOSFET itself. Now we'll add the metal wiring layers…

Metal interconnect

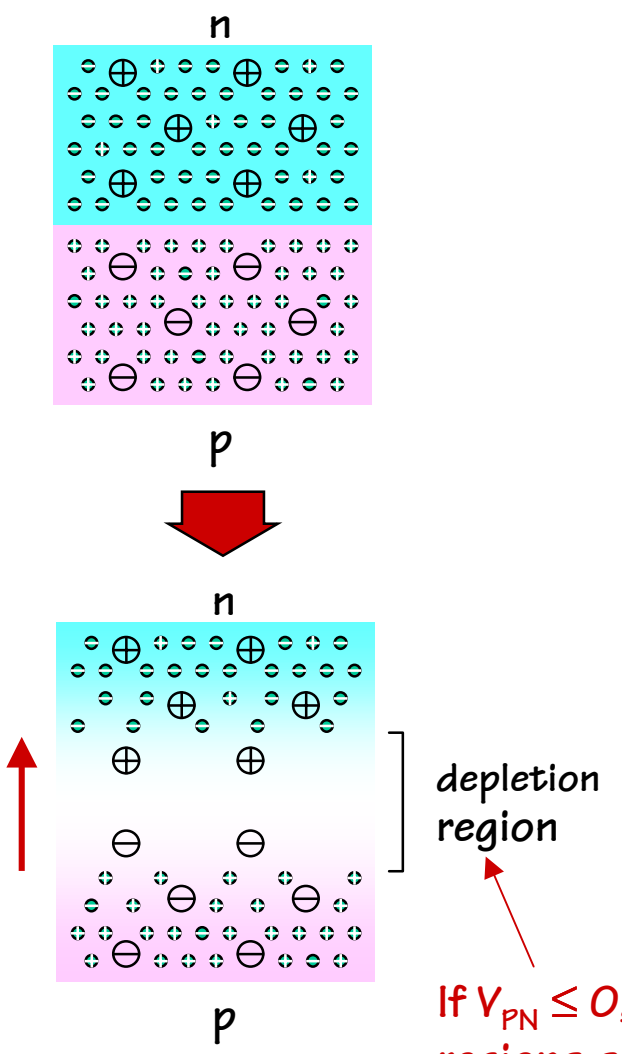


Each level of metal interconnect proceeds in two major steps. The first, shown in the sequence above, deposits an insulating layer of SiO_2 and forms vias (holes) in the oxide to the layer(s) below. In this example, vias are made to the source/drain diffusions and the polysilicon gate.



In the second major step, shown to the left, the aluminum is deposited, patterned, then etched to form low-resistance (.07 Ω/sq) interconnect. With planarization (a mechanical polishing step that creates a flat surface), multiple levels of metal interconnect are possible -- 3 to 6 layers are common in today's processes.

Advantage #2: Junctions as Insulation

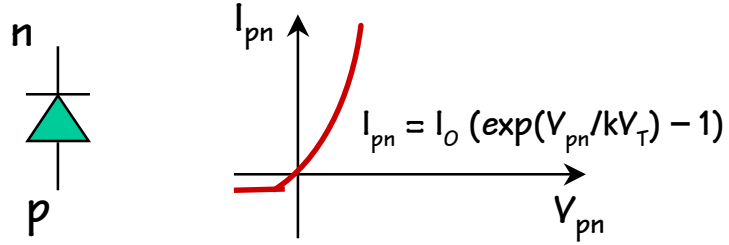


Once the two materials are in contact, the mobile carriers move:

> <u>diffusion</u> of holes from P to N and electrons from N to P ⇒ depletion of majority carriers in boundary region.
>
> <u>drift</u> of majority carriers due to E field formed by fixed ions ⇒ acts in opposite direction of diffusion

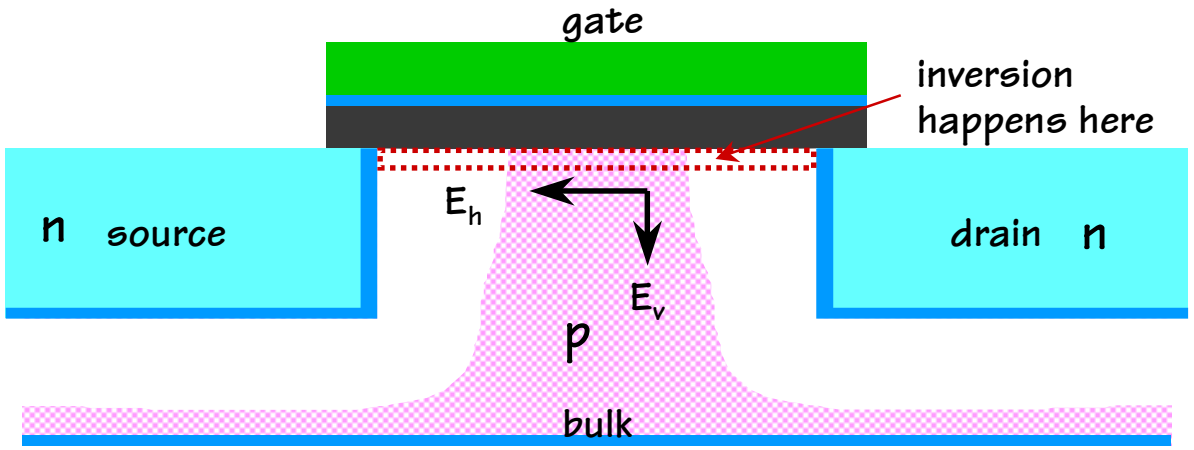
At equilibrium, the sum of the drift currents = sum of the diffusion currents. A depletion region is formed with a voltage across it due to induced field. At room temp, with doping concentrations of $10^{15}/cm^3$, this voltage is 0.6v. The net result is a diode:

If $V_{PN} \leq 0$, the two regions are electrically isolated



Advantage #3: FETs as switches

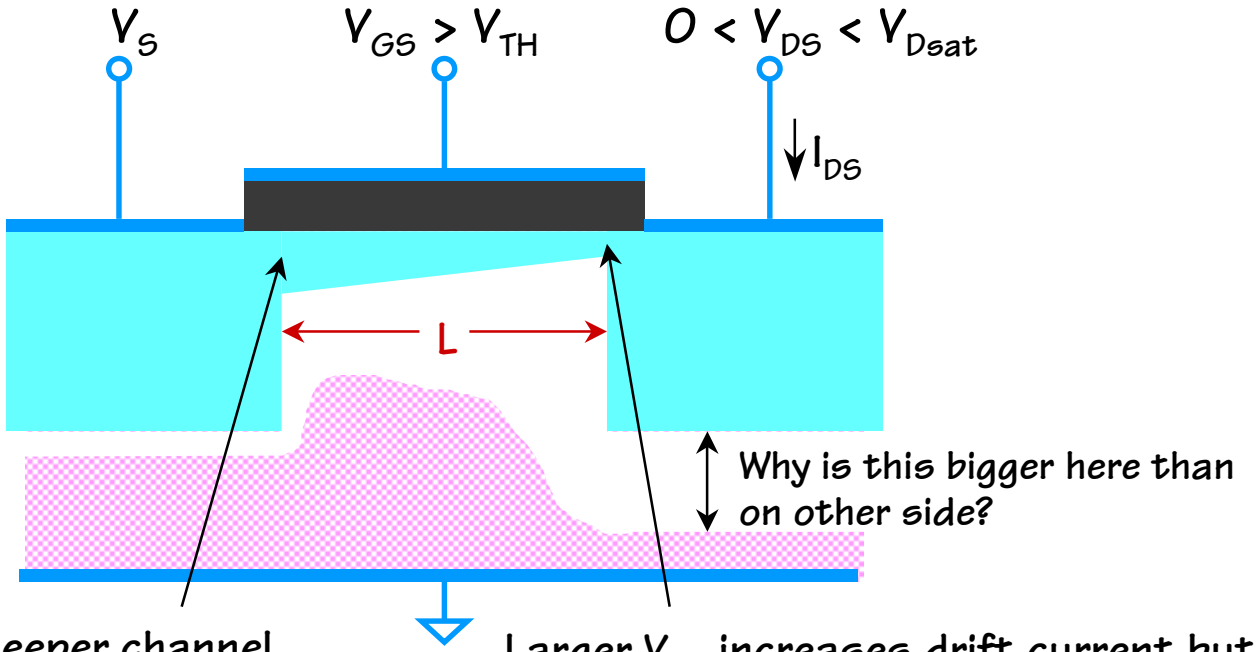
The four terminals of a Field Effect Transistor (gate, source, drain and bulk) connect to conducting surfaces that generate a complicated set of electric fields in the channel region which depend on the relative voltages of each terminal.



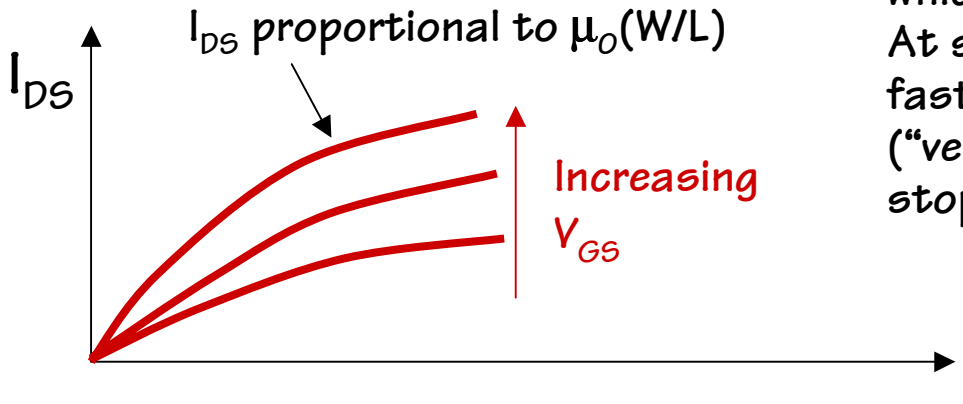
INVERSION:
A sufficiently strong vertical field will attract enough electrons to the surface to create a conducting n-type channel between the source and drain.

CONDUCTION:
If a channel exists, a horizontal field will cause a drift current from the drain to the source.

"Linear" operating region

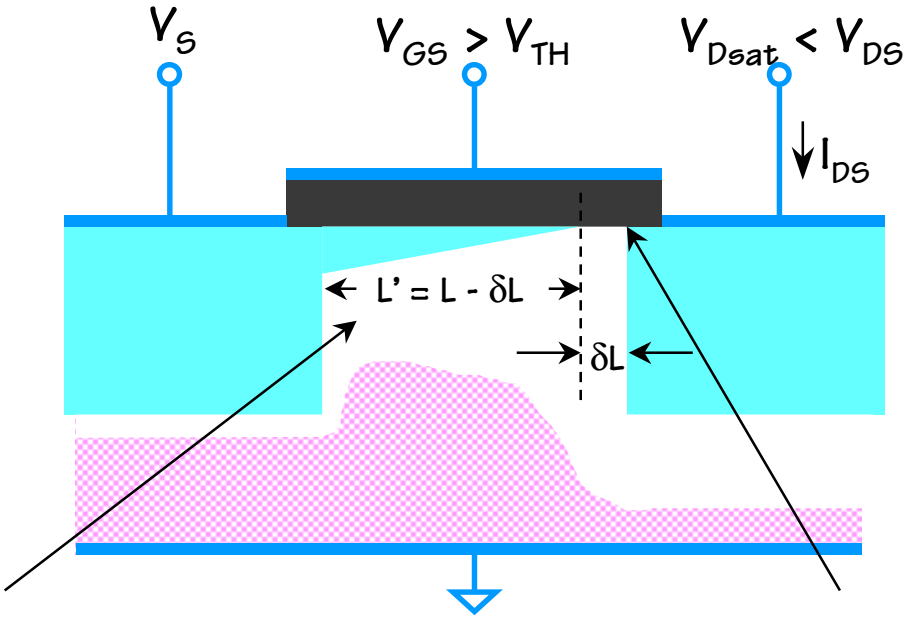


Larger V_{GS} creates deeper channel which increases I_{DS}



Larger V_{DS} increases drift current but also reduces vertical field component which in turn makes channel less deep. At some point, electrons are traveling as fast as possible through the channel ("velocity saturation") and the current stops growing linearly.

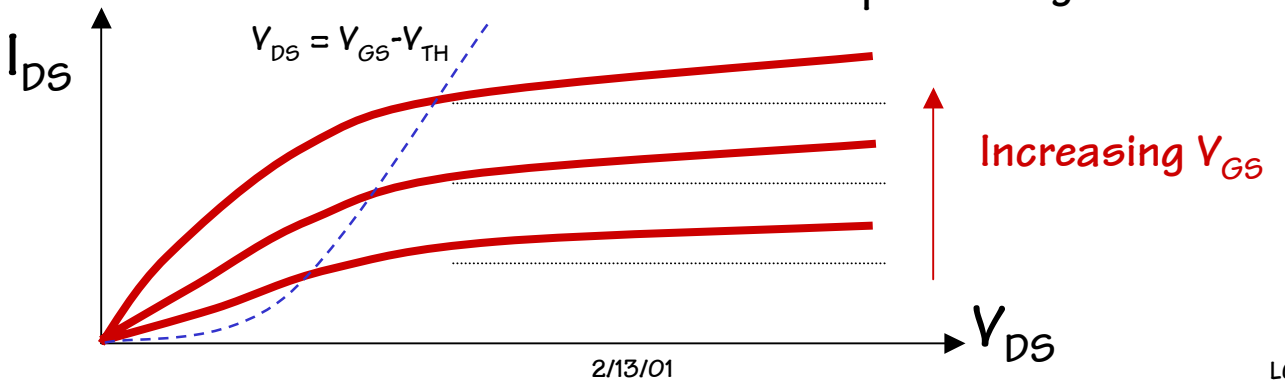
Saturated operating region



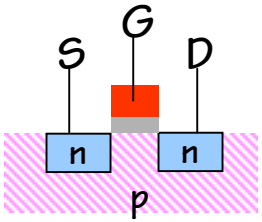
$V_{Dsat} \equiv V_{GS}-V_{TH}$

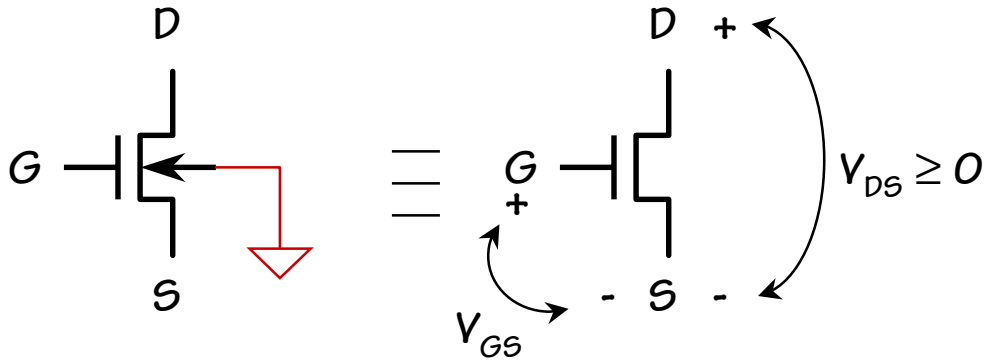
This looks just like a fet with a channel length of L' < L. Shorter L' implies greater I_{DS}. As V_{DS} increases, δL gets larger.

When $V_{DS} = V_{GS}-V_{TH}$ the vertical field component is reduced and the channel is pinched-off. Electrons just keep traveling across depletion region...



NFET Summary





Operating regions:

cut-off:

$V_{GS} < V_{TH}$ ← 0.8V

linear:

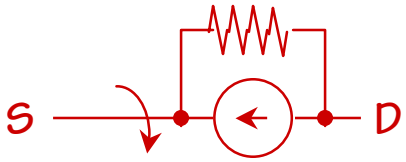
$V_{GS} \geq V_{TH}$

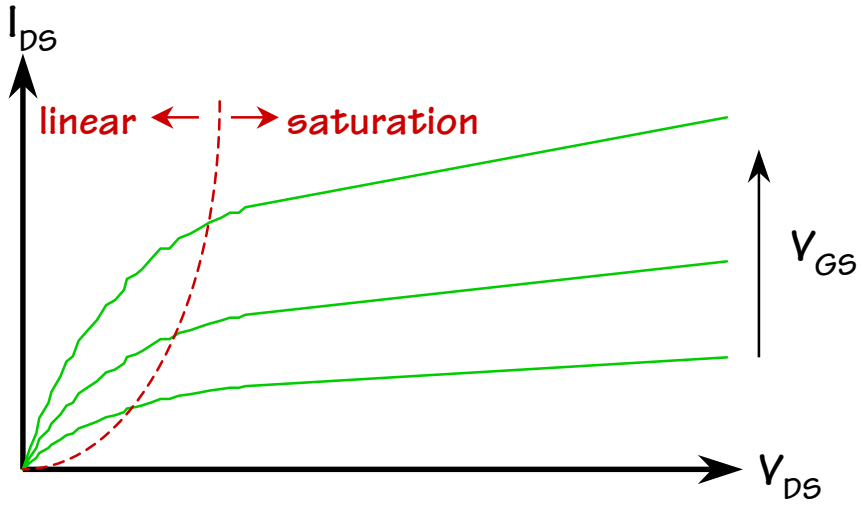
$V_{DS} < V_{Dsat}$ ← $V_{GS} - V_{TH}$

saturation:

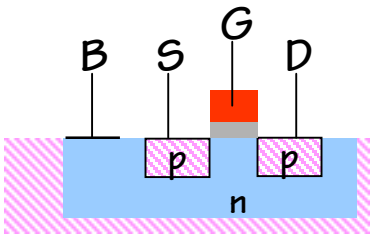
$V_{GS} \geq V_{TH}$

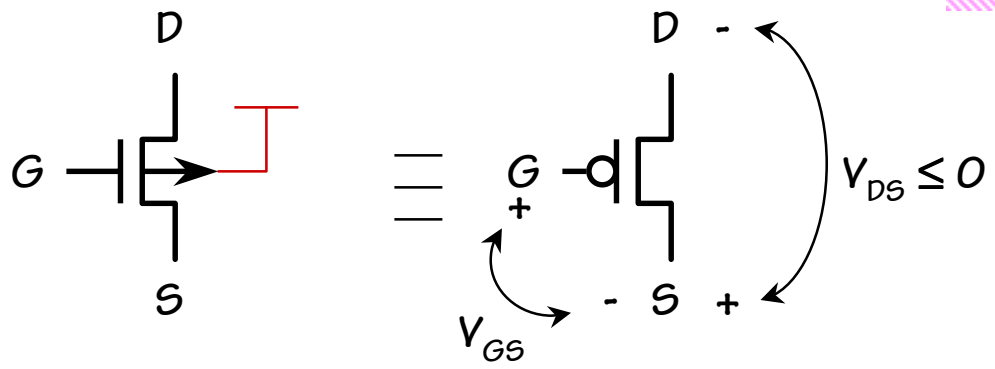
$V_{DS} \geq V_{Dsat}$





PFET Summary





Operating regions:

cut-off:

$V_{GS} > V_{TH}$ ← −0.8V

linear:

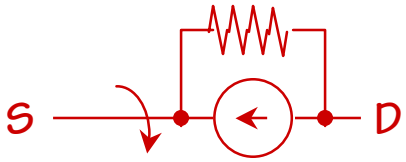
$V_{GS} \leq V_{TH}$

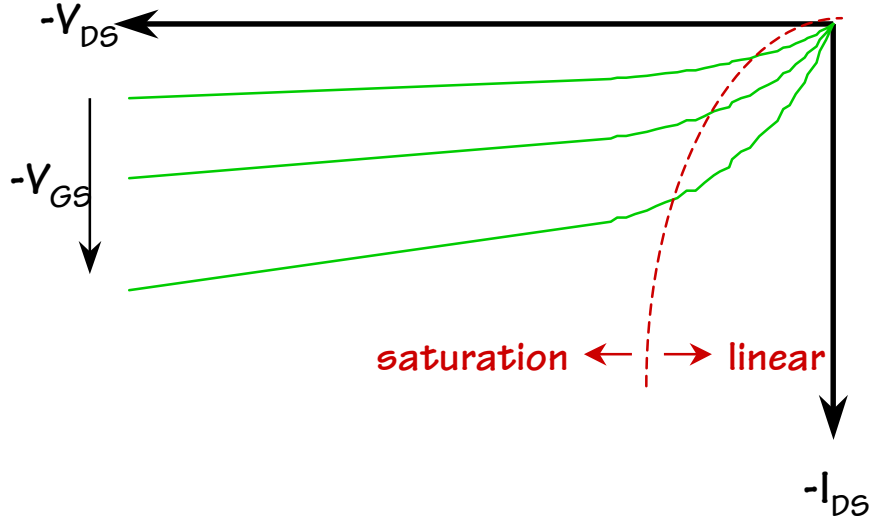
$V_{DS} > V_{Dsat}$ ← $V_{GS} - V_{TH}$

saturation:

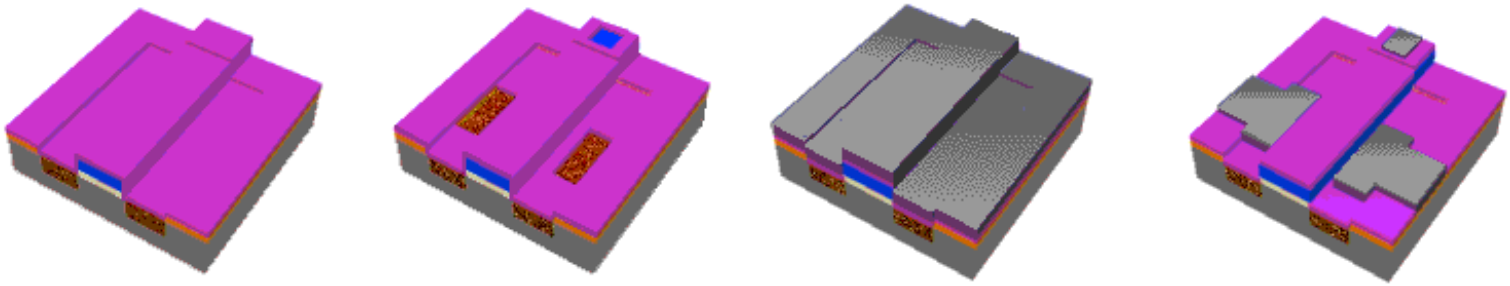
$V_{GS} \leq V_{TH}$

$V_{DS} \leq V_{Dsat}$





Wires: metal interconnect



Deposit SiO_2 insulation

Etch openings for vias/contacts

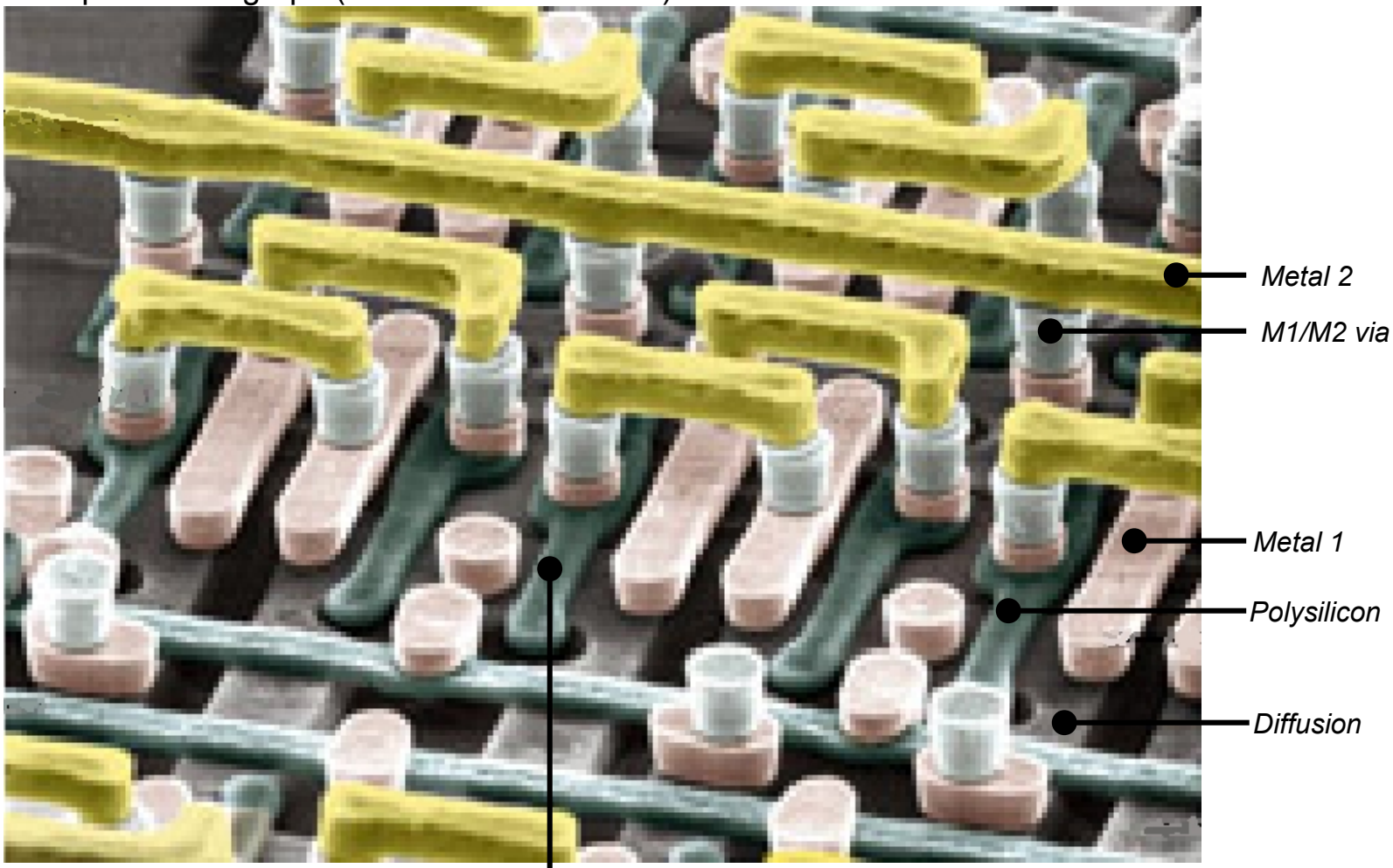
Deposit Al/Cu conductor

Etch away Al/Cu leaving wires

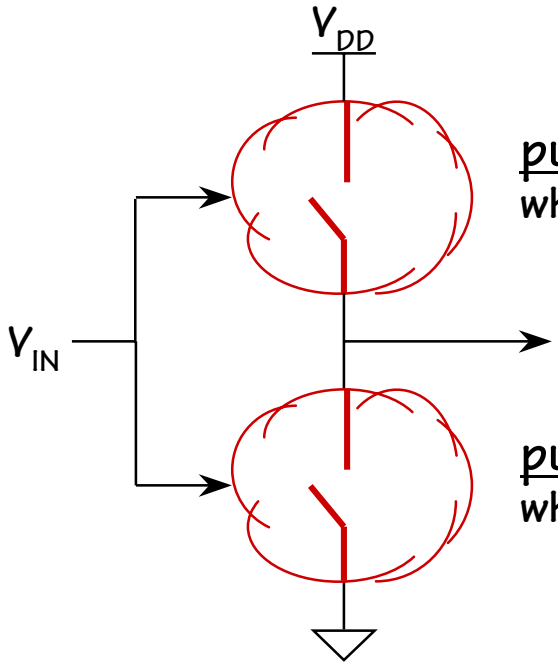
After a layer of SiO_2 insulation has been deposited, aluminum or copper is deposited, patterned, then etched to form low-resistance (.07 Ω/sq) interconnect. With planarization of the SiO_2 (a mechanical polishing step that creates a flat surface), multiple levels of metal interconnect are possible -- 3 to 6 layers are common in today's processes.

Multiple interconnect layers

IBM photomicrograph (Si has been removed!)



Finally... logic gates!



Inverter recipe:

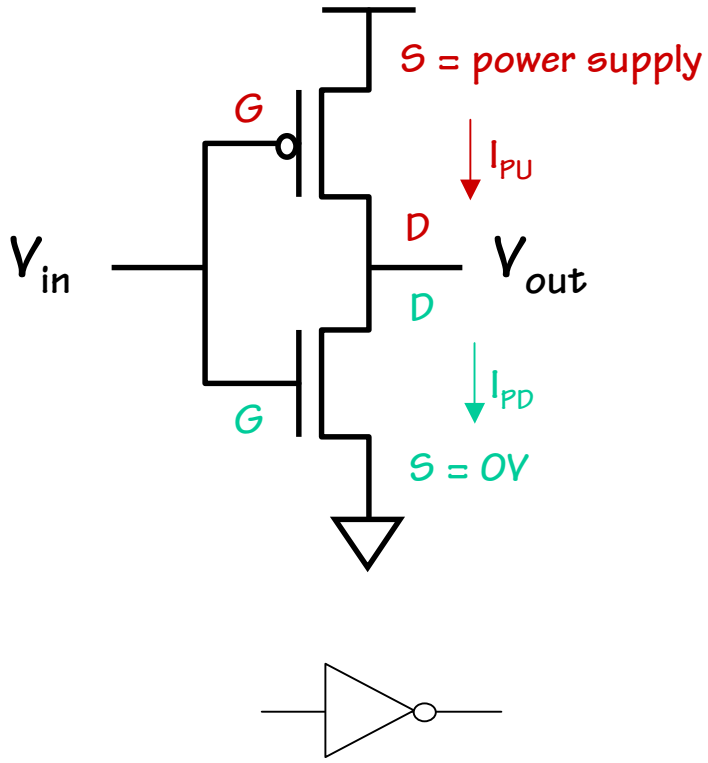
pullup: make this connection when V_{IN} near 0 so that $V_{OUT} = V_{DD}$

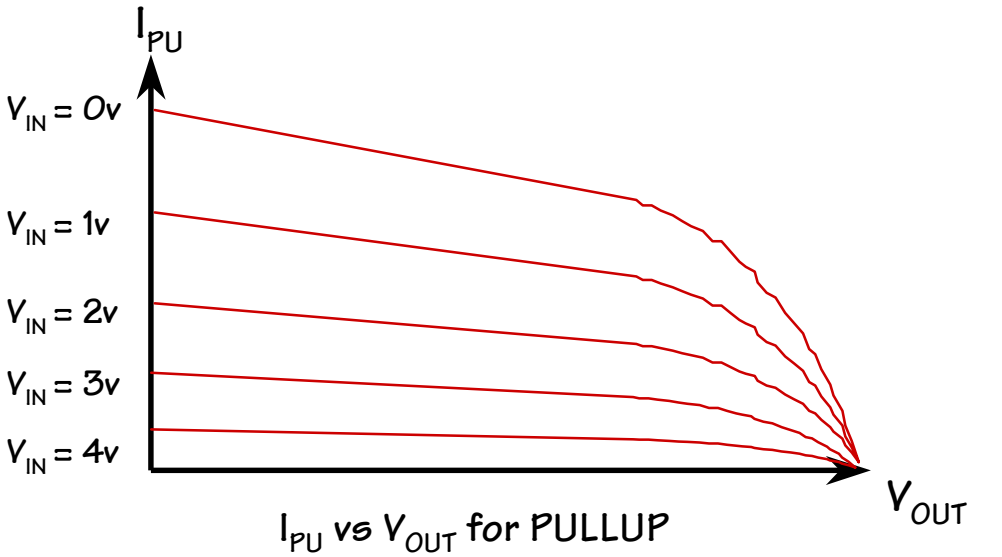
pulldown: make this connection when V_{IN} near V_{DD} so that $V_{OUT} = 0$

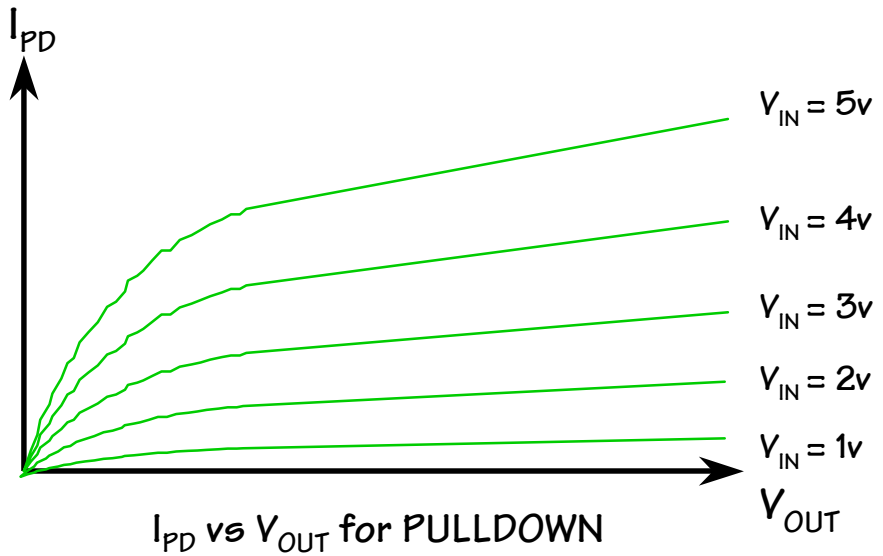
Why choose a binary (e.g., "0" and "1") logic system:

- one power supply => low impedance source for 2 levels
- receiving devices have a simple job => only make one decision
- no DC power if connections not "made" at same time
- Boolean logic has been around a long time

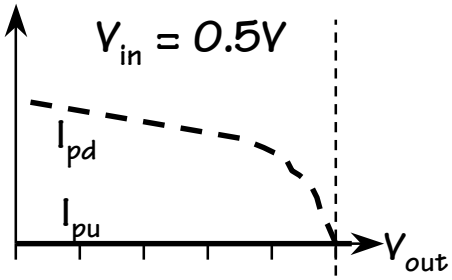
CMOS Inverter

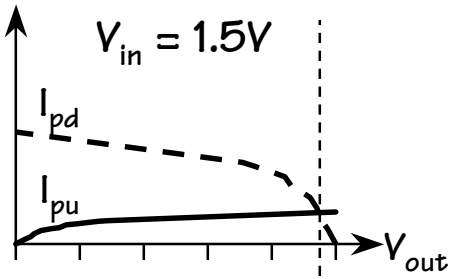


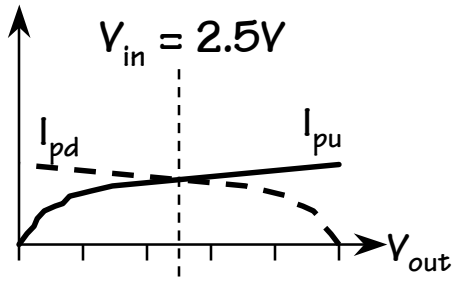


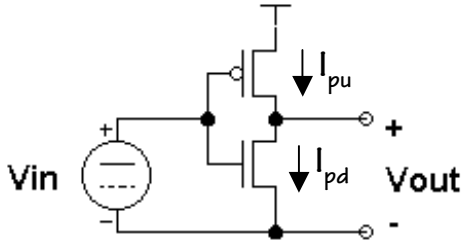


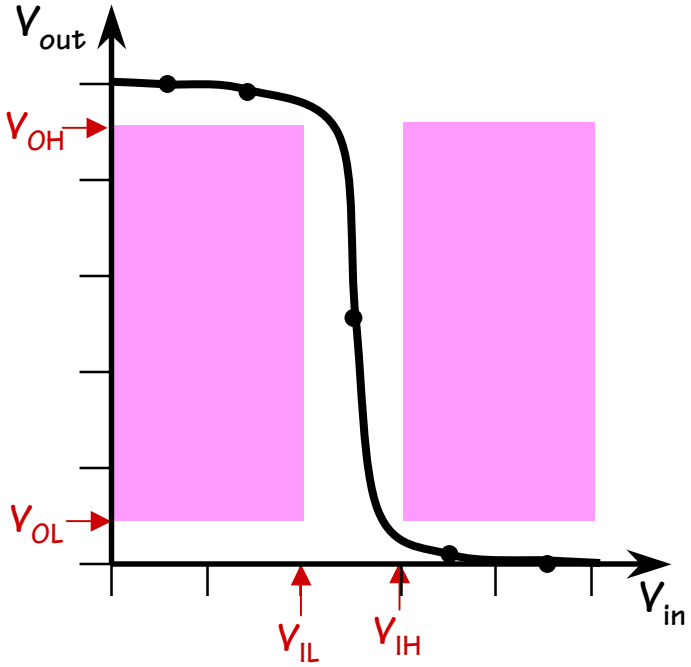
CMOS Inverter VTC



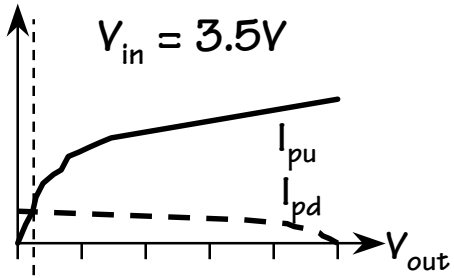


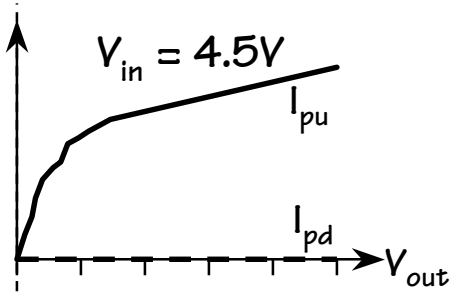






Steady state reached when V_{out} reaches value where $I_{pu} = I_{pd}$.

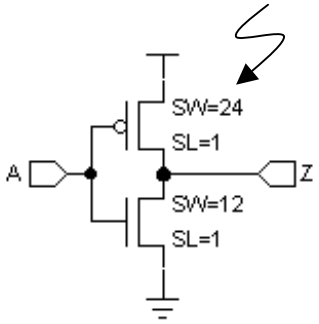




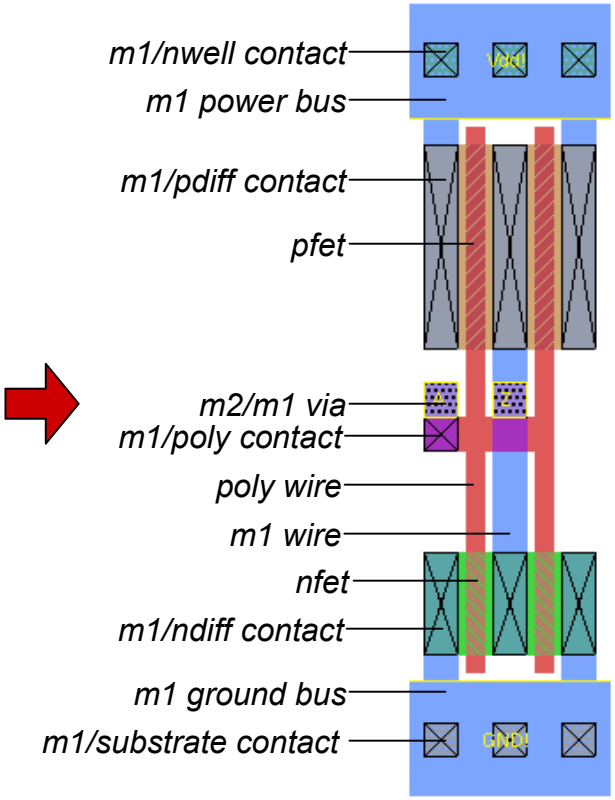
When both fets are saturated, small changes in V_{in} produce large changes in V_{out}

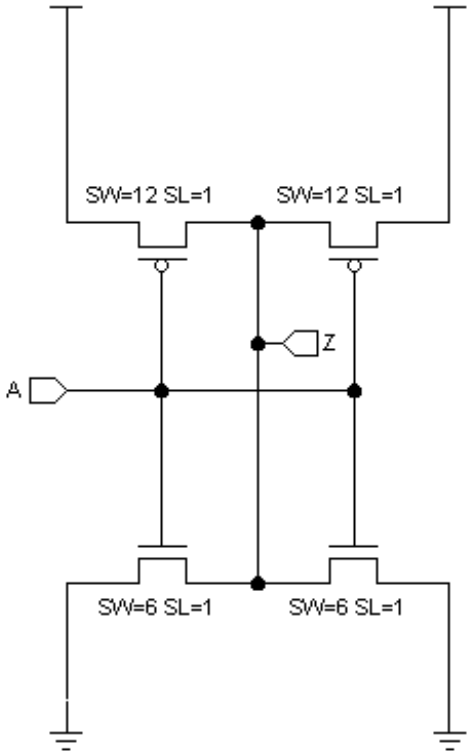
Standard Cell Layout for Inverter

SW: "scaled width" used in Process-independent design



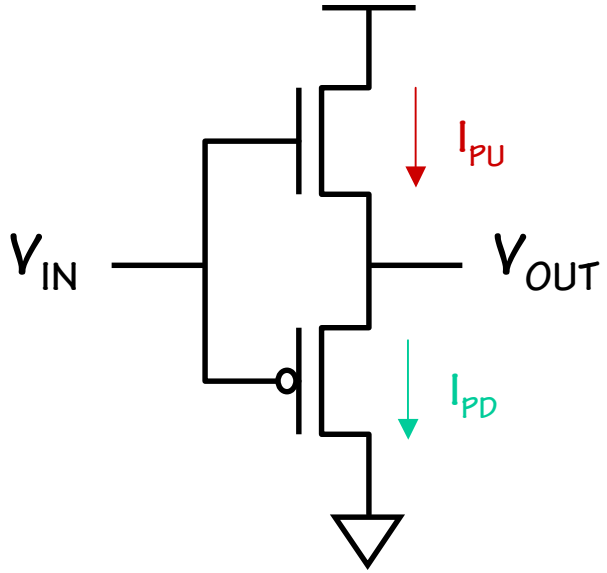
Use two narrow mosfets in parallel instead of one wide mosfet



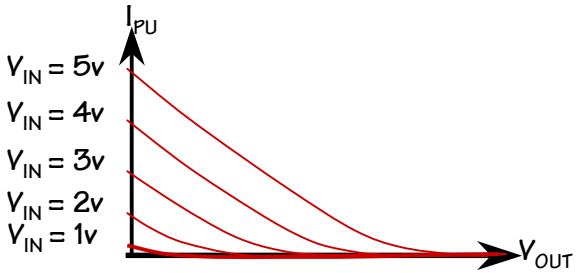


Physical design of a CMOS gate is represented by a mask layout showing where material on each layer (ndiff, pdiff, poly, m1, m2, …) should be placed on the silicon wafer. Each manufacturing process has a set of *design rules* that determine minimum widths, spacings, overlaps, etc.

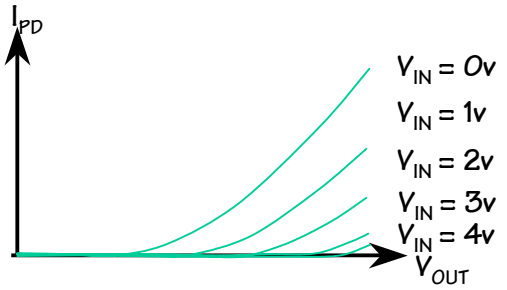
Can we build a CMOS buffer?



Not like this! Here's why:

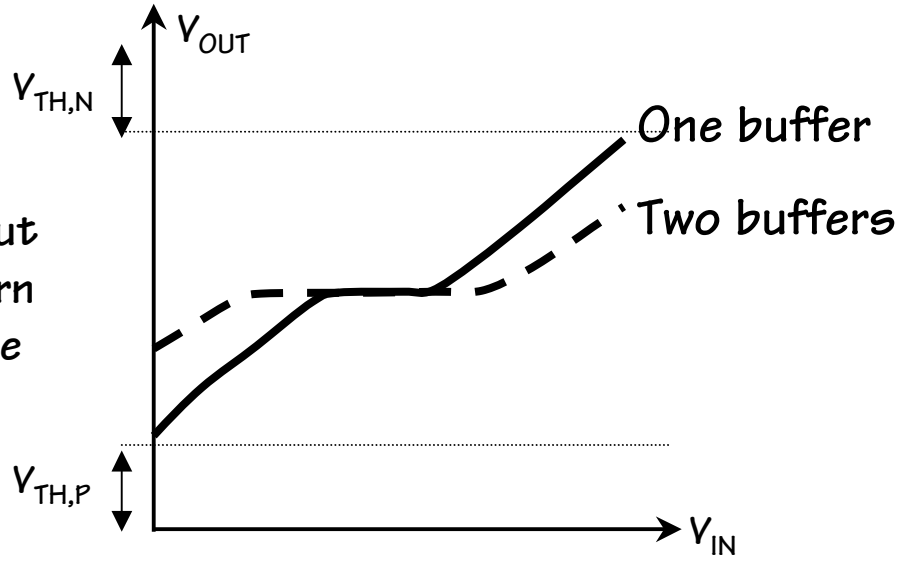


I_{PU} vs V_{OUT} for PULLUP



I_{PD} vs V_{OUT} for PULLUP

Nfets turn off when V_{GS} falls below the threshold voltage V_{TH}. So, even if the input voltage is, say, 5V, then the pullup will turn off when V_{OUT} reaches $5V - V_{TH} = 4.2V$. The pulldown will also turn off before V_{OUT} reaches 0V.



Complementary pullups and pulldowns

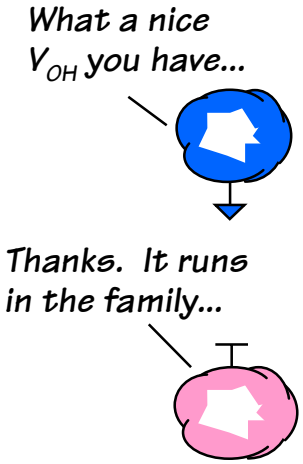
Now you know what the "C" in CMOS stands for!

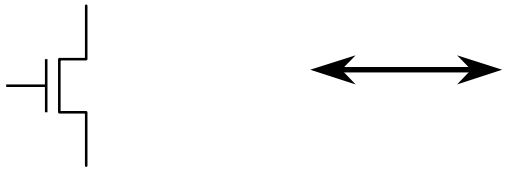
We want *complementary* pullup and pulldown logic, i.e., the pulldown should be "on" when the pullup is "off" and vice versa.

pullup	pulldown	F(A_1,...,An)
on	off	driven "1"
off	on	driven "0"
on	on	driven "X"
off	off	no connection

Since there's plenty of capacitance on the output node, when the output becomes disconnected it "remembers" its previous voltage -- at least for a while. The "memory" is the load capacitor's charge. Leakage currents will cause eventual decay of the charge (that's why DRAMs need to be refreshed!).

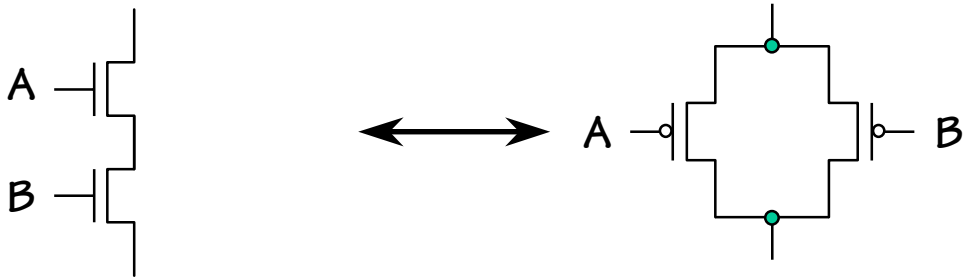
CMOS complements





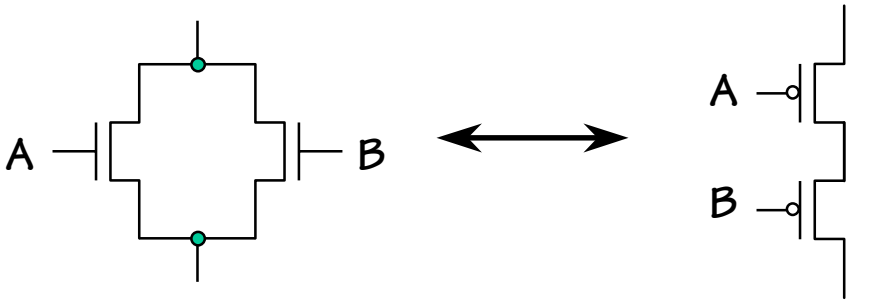
conducts when V_{GS} is high

conducts when V_{GS} is low



conducts when A is high <u>and</u> B is high: $A \cdot B$

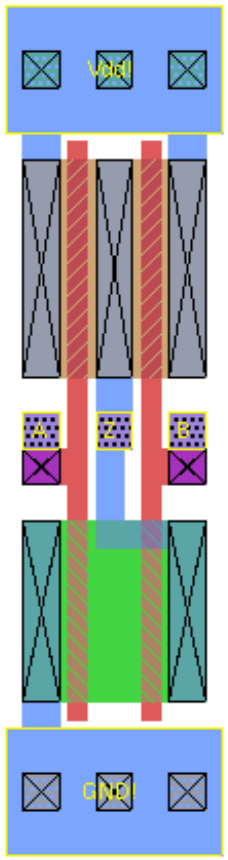
conducts when A is <u>low</u> <u>or</u> B is low: $\overline{A}+\overline{B} = \overline{A \cdot B}$

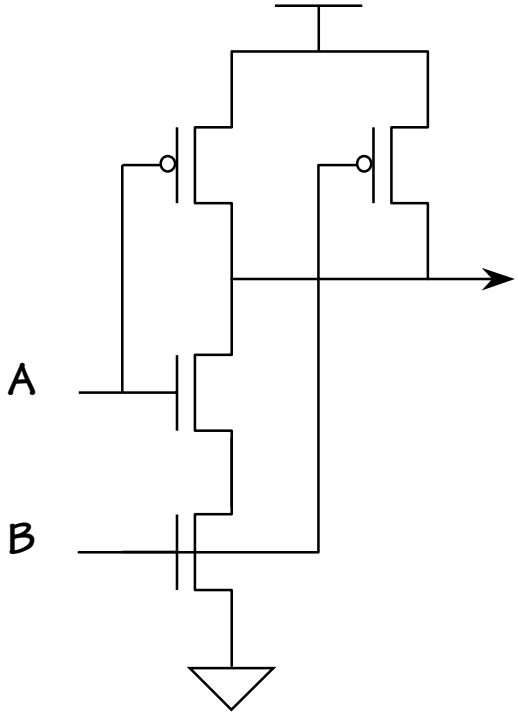


conducts when A is high <u>or</u> B is high: $A+B$

conducts when A is <u>low</u> <u>and</u> B is low: $\overline{A} \cdot \overline{B} = \overline{A+B}$

A pop quiz!



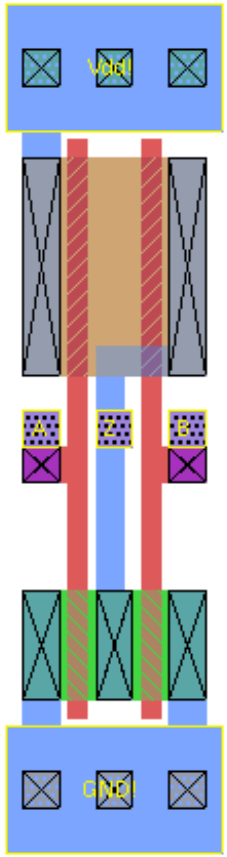


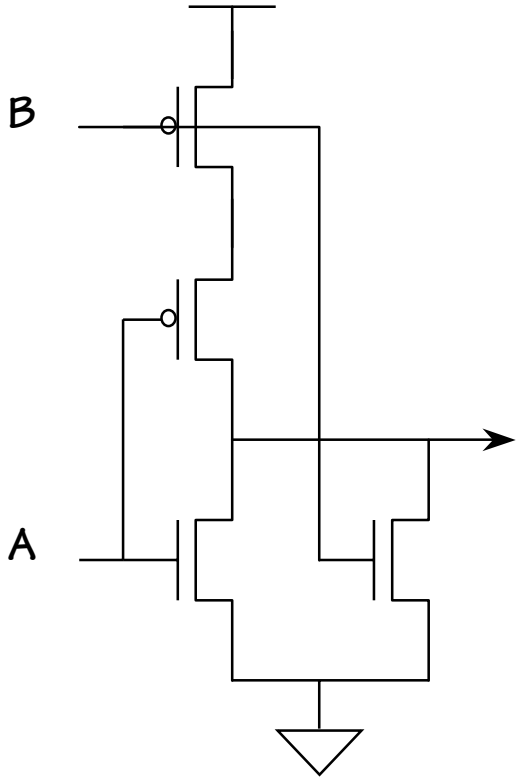
What function does this gate compute?

A	B	C
0	0	1
0	1	1
1	0	1
1	1	0

NAND

Here's another…





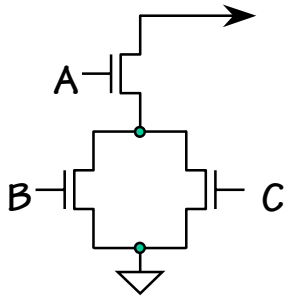
What function does this gate compute?

A	B	C
0	0	1
0	1	0
1	0	0
1	1	0

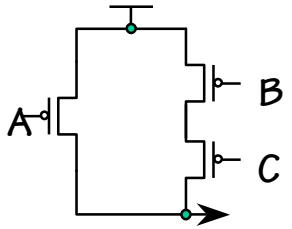
NOR

General CMOS gate recipe

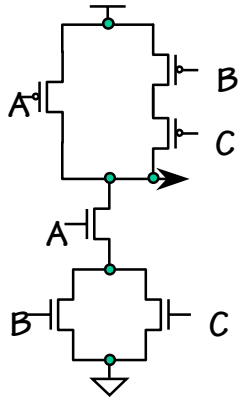
Step 1. Figure out pulldown network that does what you want, *e.g.*, F = A*(B+C)
(What combination of inputs generates a low output)



Step 2. Walk the hierarchy replacing nfets with pfets, series subnets with parallel subnets, and parallel subnets with series subnets



Step 3. Combine pfet pullup network from Step 2 with nfet pulldown network from Step 1 to form fully-complementary CMOS gate.



But isn't it hard to wire it all up?

Summary

- MOSFET features
  - PN junctions provide electrical isolation
  - Self-aligned diffusions are easy to make reliably
  - Switch-like behavior controlled by V_{GS}
  - Shrinking geometries improves performance
- CMOS features
  - CMOS logic is "naturally" inverting: "1" inputs lead to "0" outputs
  - Good noise margins because
    - $V_{OL} = 0$, $V_{OH} = V_{DD}$
    - complementary logic has high gain
  - No static power dissipation
  - Pullup/pulldown "chains" make for good mask layouts
- Next time: timing, converting functionality to logic