Lecture 2
CMOS Technology

Comp Eng 303

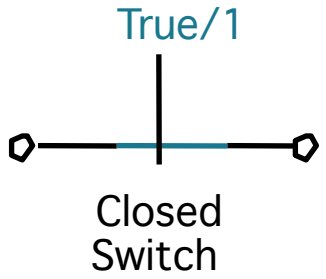
Advanced Digital Design

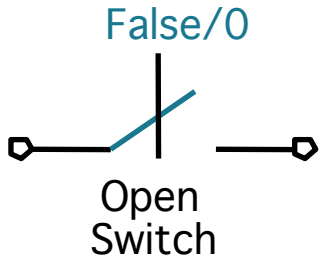
Representations of Digital Design: Transistor Switch Model

Normally Open

A switch connects two points under control signal.
when the control signal is 0 (false), the switch is open
when it is 1 (true), the switch is closed

NMOS:



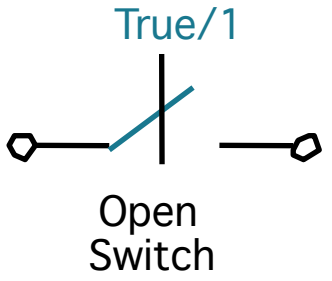


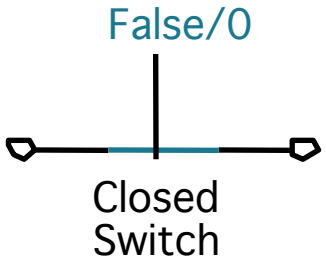
Representations of Digital Design: Transistor Switch Model

Normally Closed

when control is 1 (true), switch is open
when control is 0 (false), switch is closed

PMOS:





CMOS

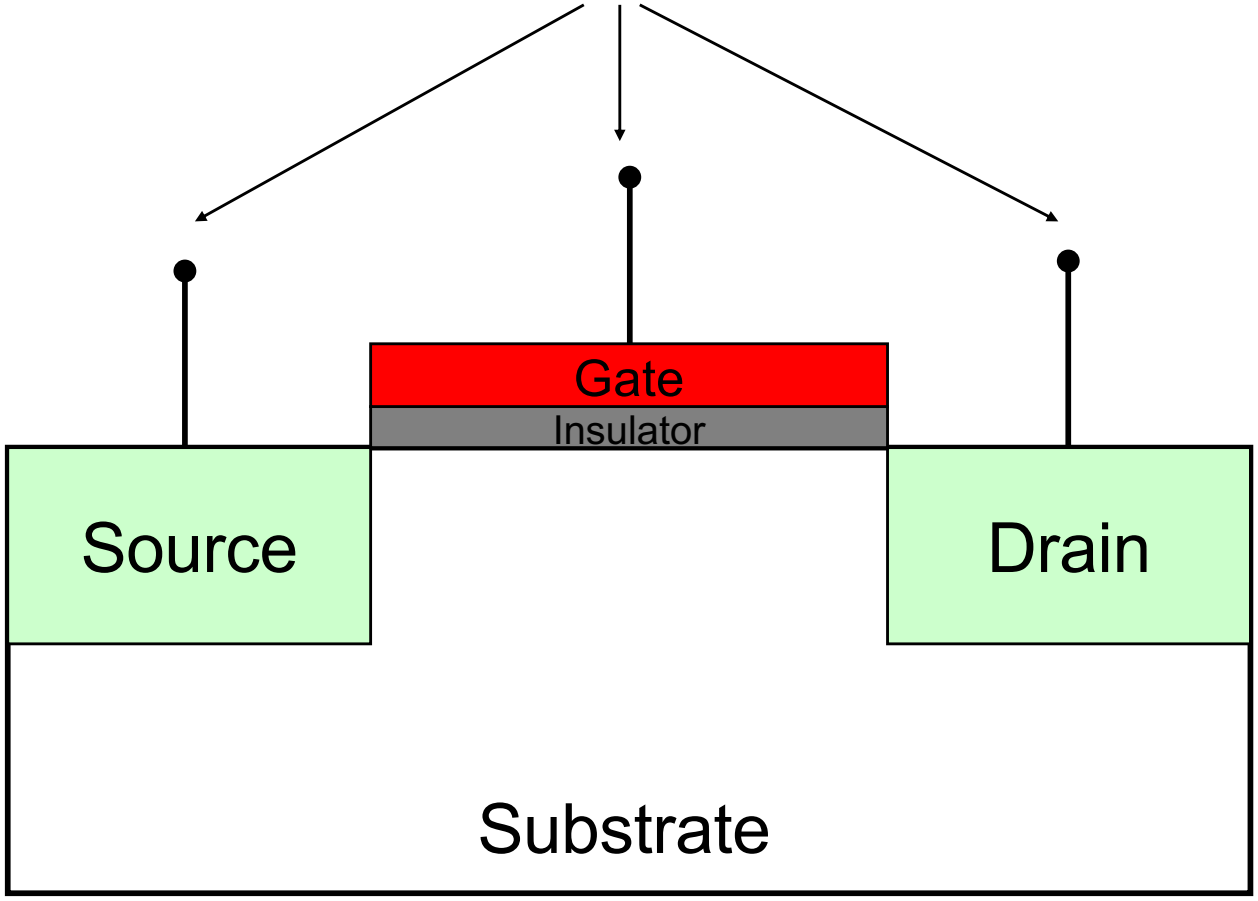
- **M**etal **O**xide **S**ilicon
  - Name is originated from the three fundamental materials that are used to make transistors
  - Silicon, Metal (Cu), Silicondioxide (SiO2)
  - Silicon enriched in Positive and Negative carriers
- **C**omplimentary MOS
  - Two symmetric types of transistors combined in pairs

Modeling CMOS

- Normally open and normally closed switch models are very useful for modeling CMOS circuits
  - Behavior of NMOS and PMOS transistors follow the same principle

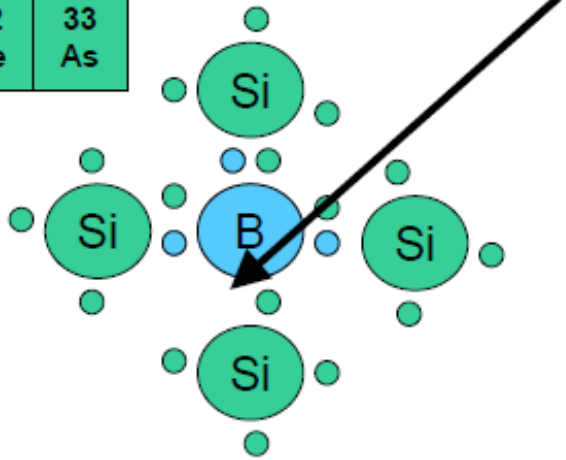
Modeling CMOS

Three main terminals where voltage inputs can be applied



Modeling PMOS

III	IV	V
5 B	6 C	7 N
13 Al	14 Si	15 P
31 Ga	32 Ge	33 As



Boron

Si is replaced with an atom from column III (B)

Now there is a missing electron called a hole

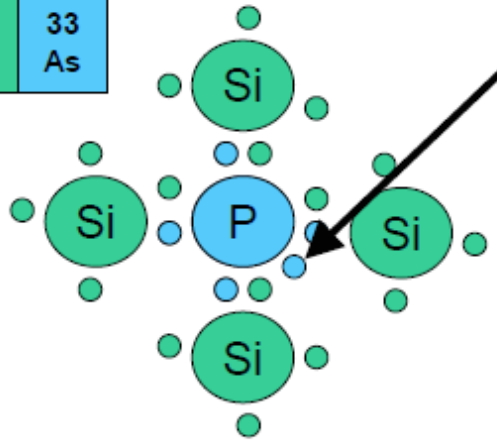
The dopant is called an acceptor

Si doped with acceptors is known as p-type

- Extra holes (effective positive charge)

Modeling NMOS

III	IV	V
5 B	6 C	7 N
13 Al	14 Si	15 P
31 Ga	32 Ge	33 As



Replace Si with an atom from column V (P, As)

Now there is one extra electron

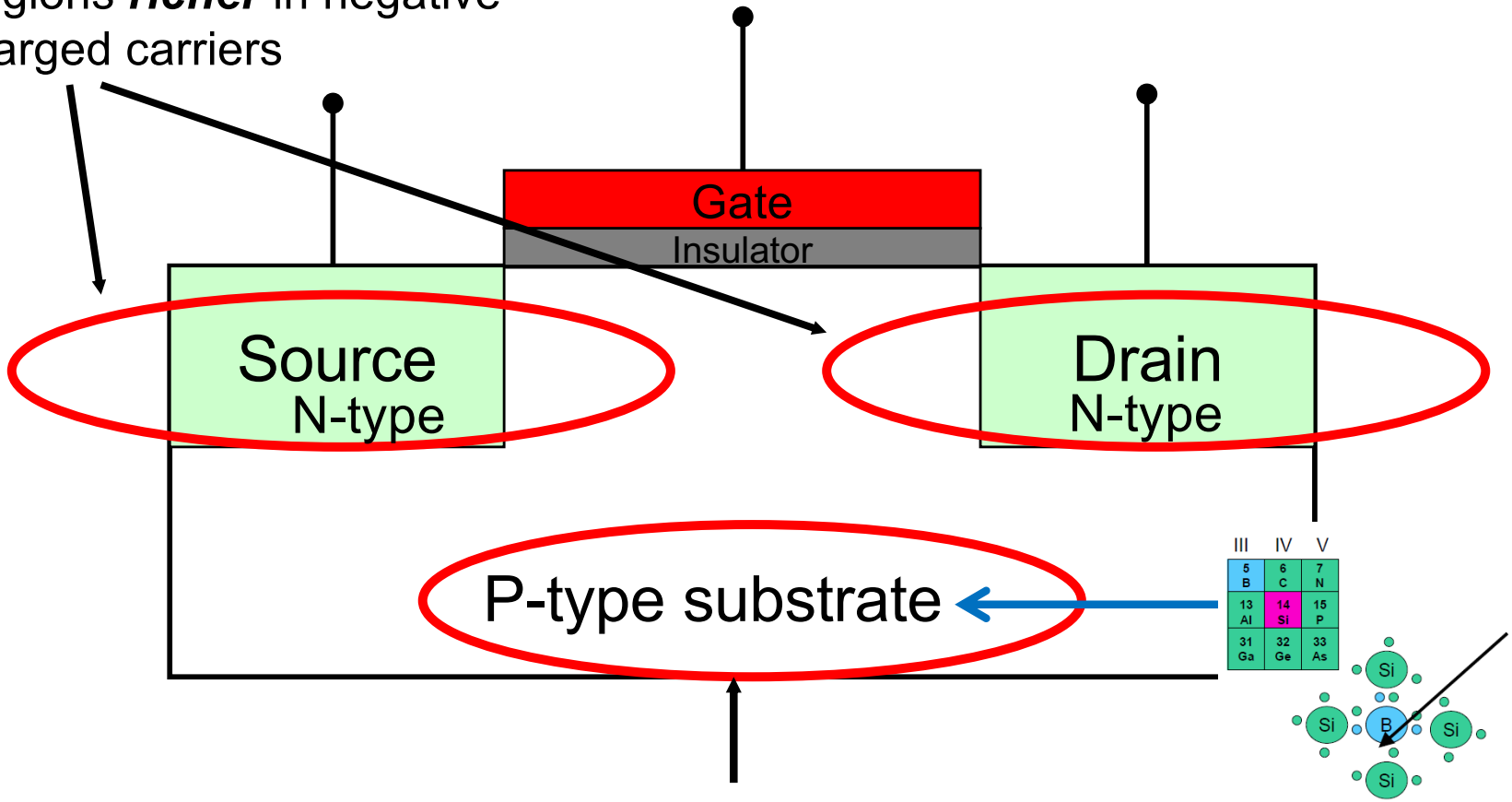
The dopant is called a donor

Si doped with donors is known as n-type

P: Phosphorus
As: Arsenic

Modeling NMOS

Regions ***richer*** in negative charged carriers

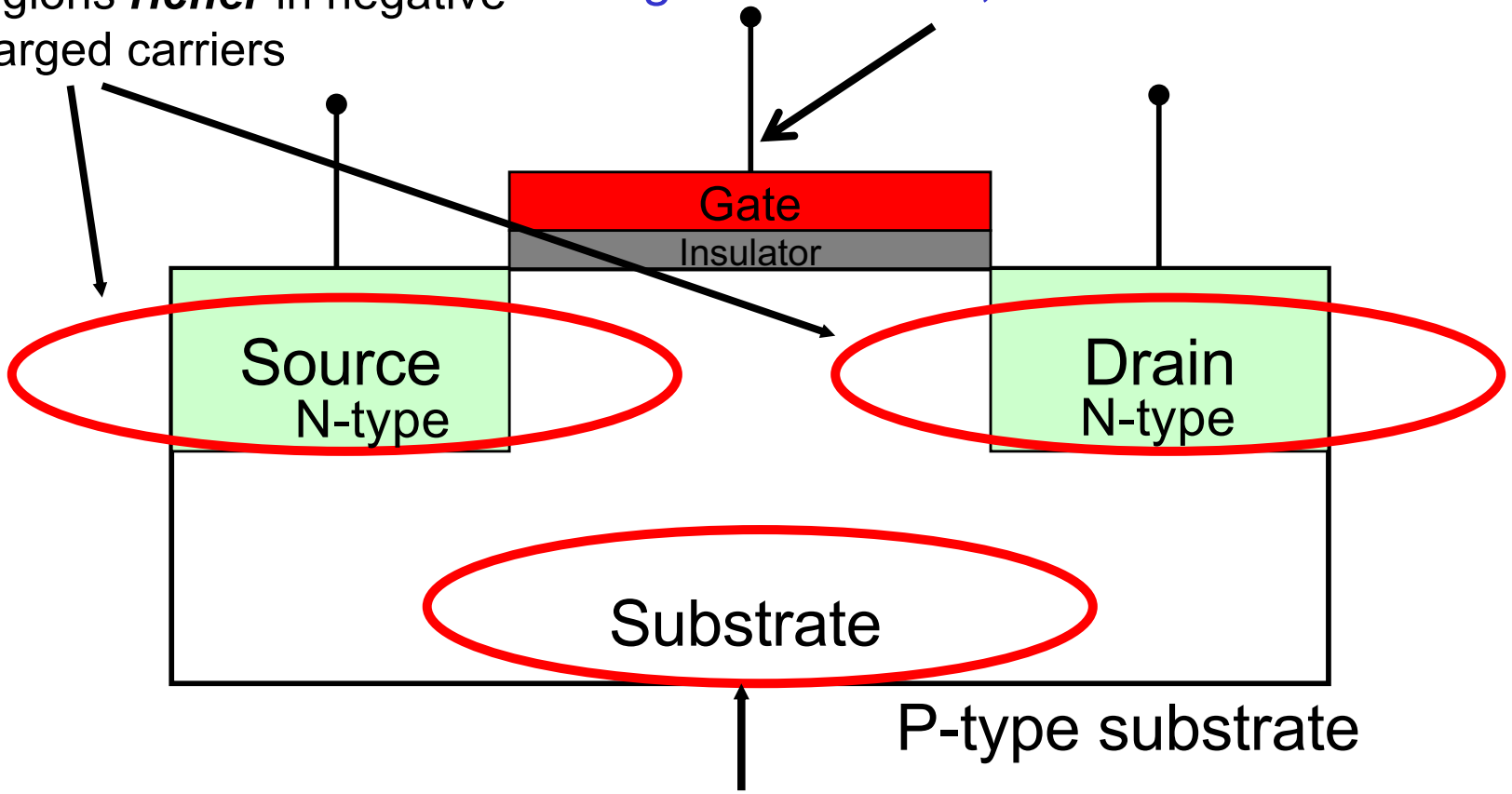


Region ***poorer*** in negative charged carriers (richer in positive charge carriers (holes))

Modeling NMOS

At low gate voltage, depletion region under gate is formed, transistor is not conductive (off)

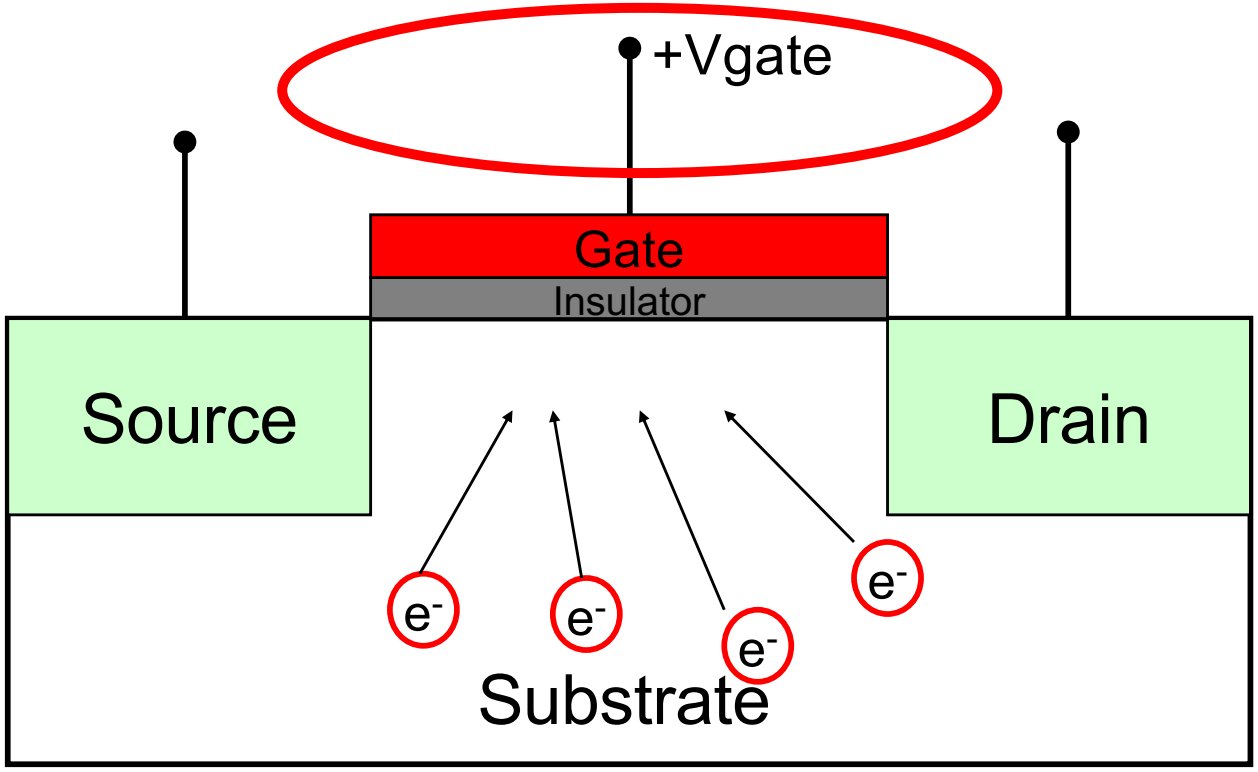
Regions ***richer*** in negative charged carriers



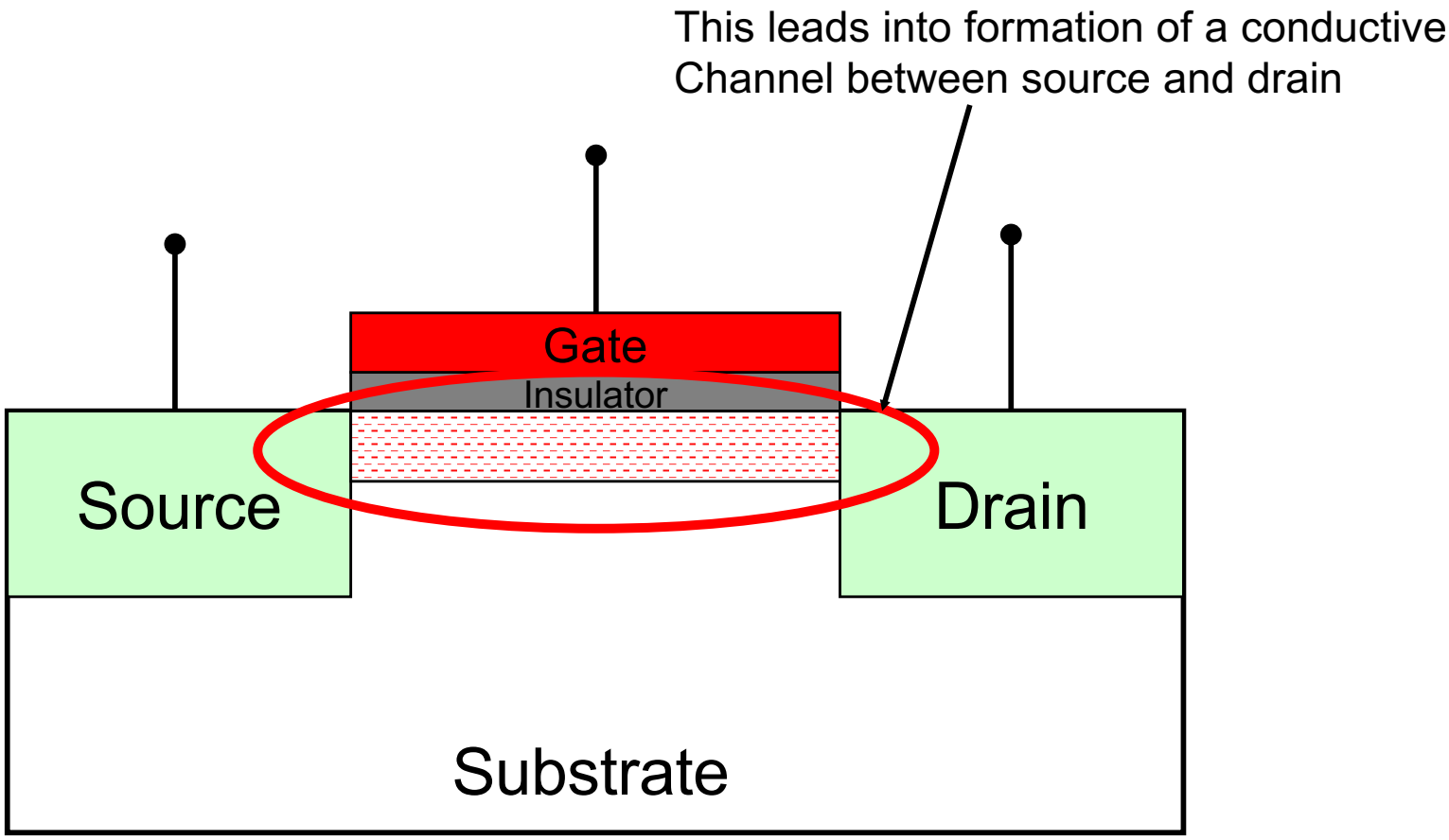
Region ***poorer*** in negative charged carriers (richer in positive charge carriers (holes))

Modeling NMOS

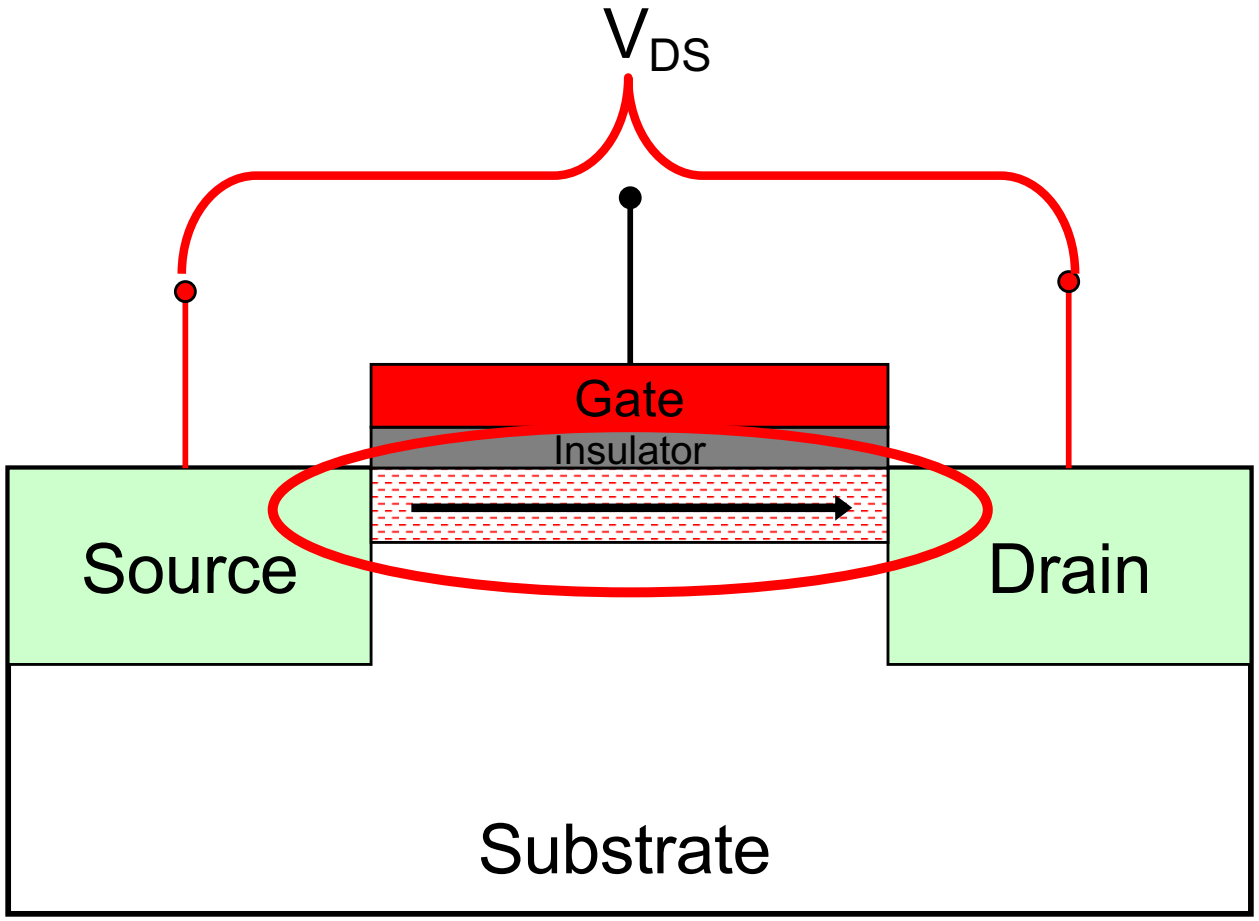
Positive voltage applied here **attracts** electrons
from the substrate upwards towards the region right beneath the Gate



Modeling NMOS



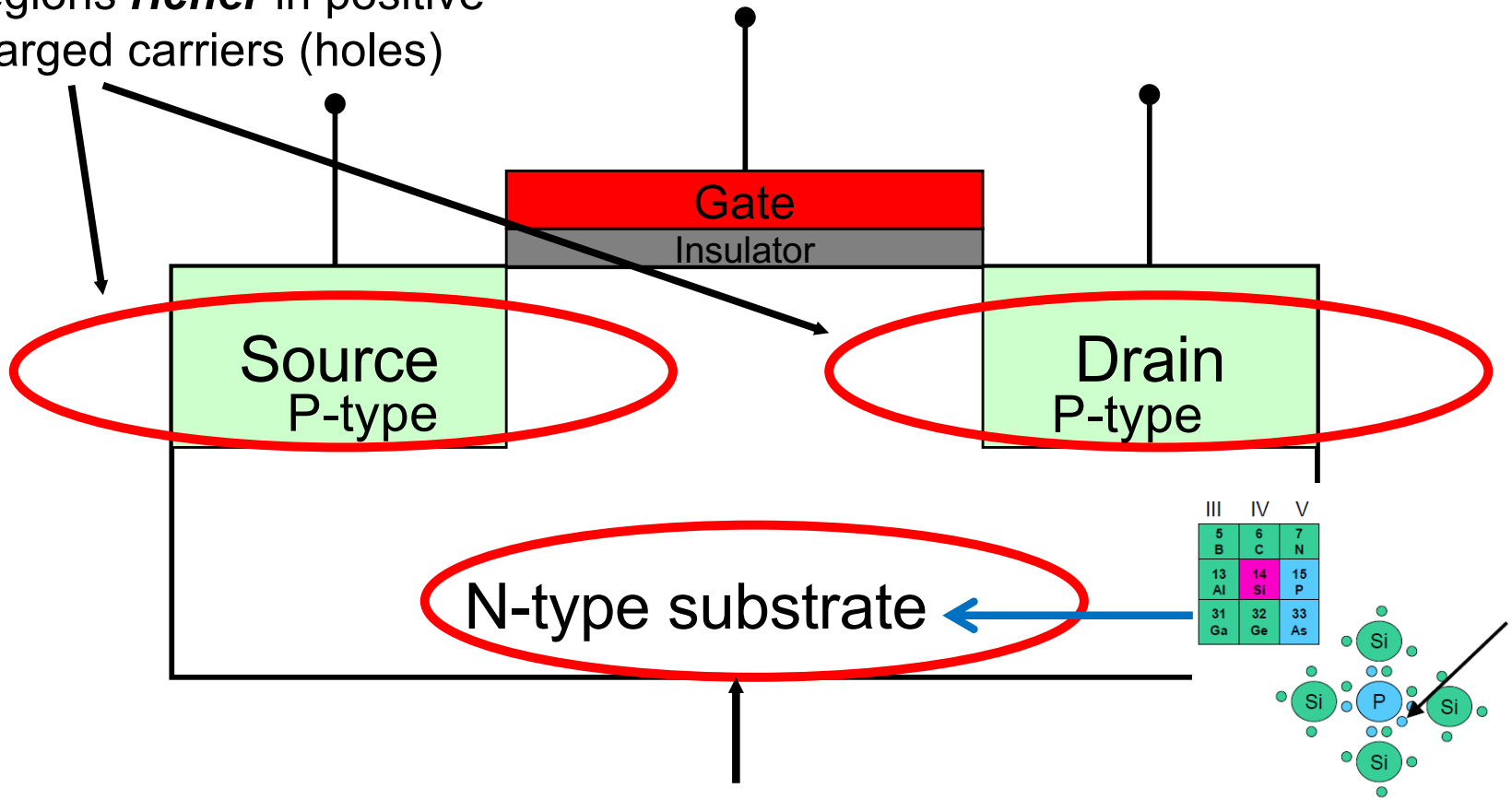
Modeling NMOS



- Voltage difference between drain and source causes electrons to flow from Source to Drain
- Transistor is on (Vgate is high)

Modeling PMOS

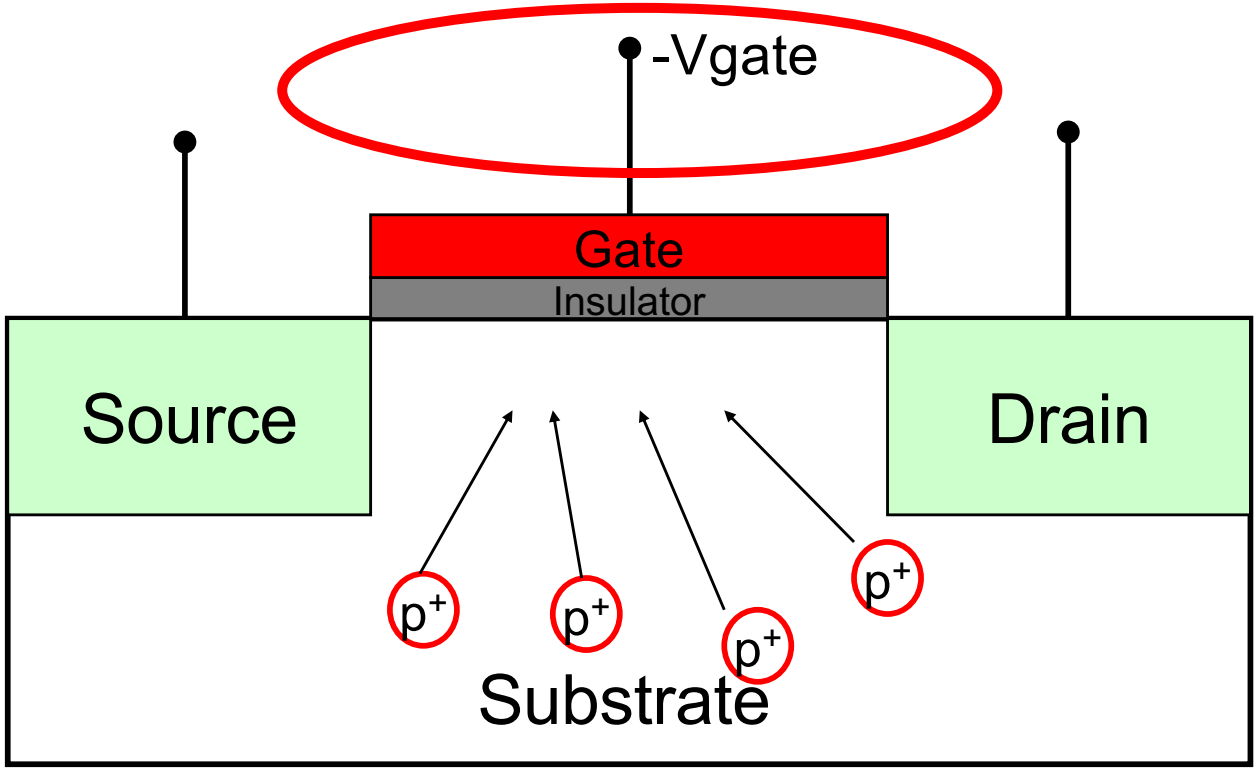
Regions ***richer*** in positive charged carriers (holes)



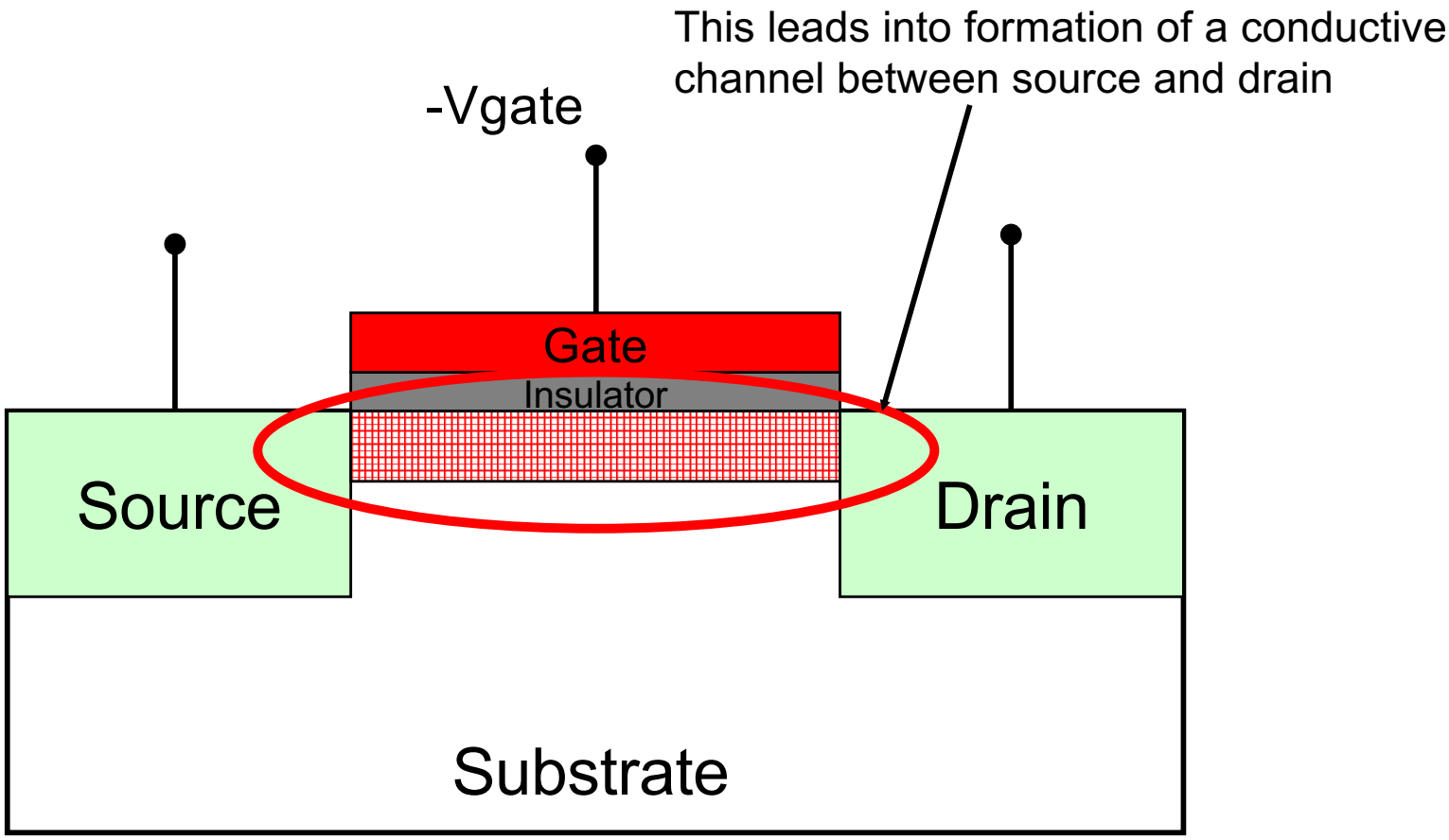
Region ***poorer*** in positive charged carriers (richer in negative charge carriers (electrones))

Modeling PMOS

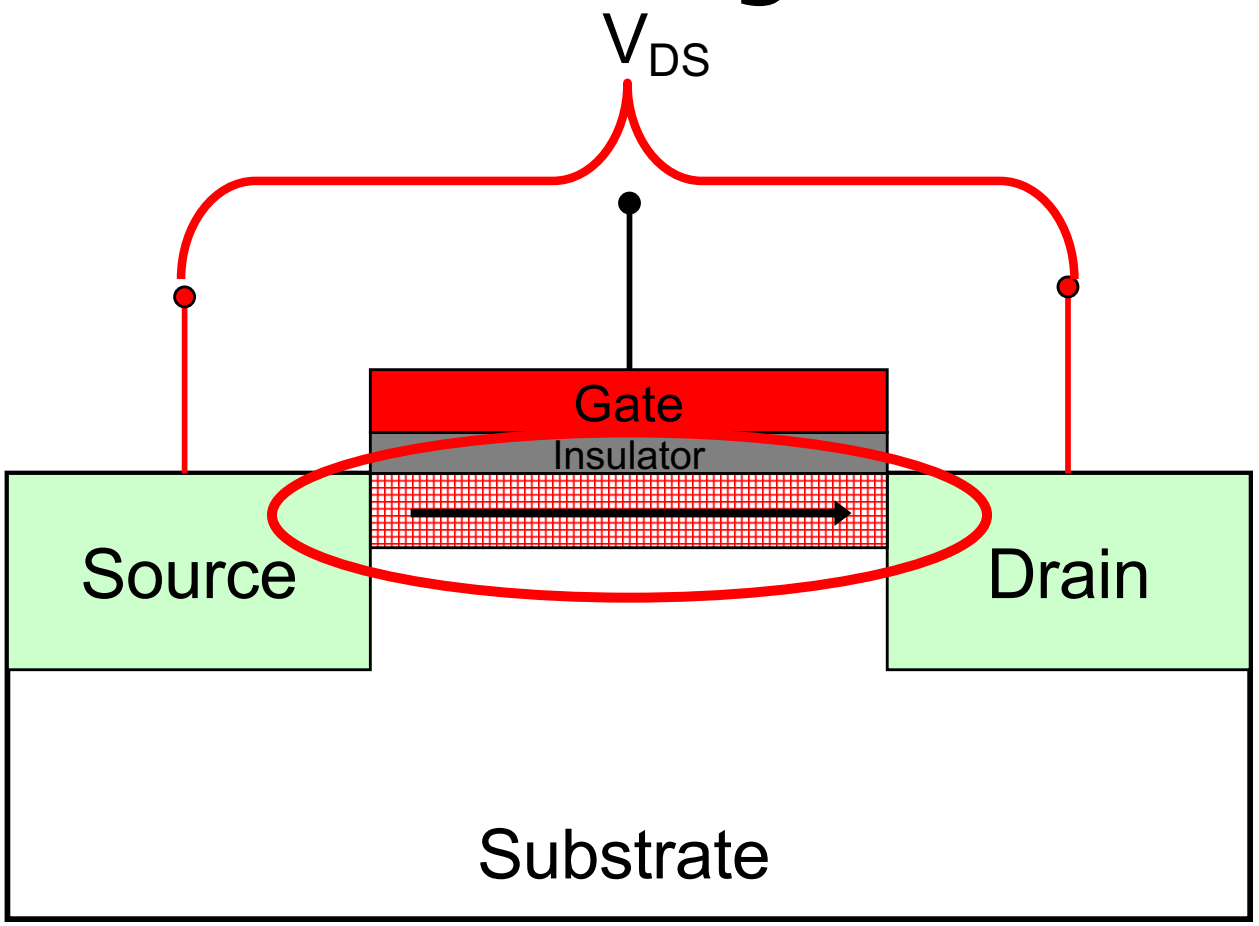
Negative voltage applied here **attracts** positive charge carriers from the substrate upwards towards the region right beneath the Gate



Modeling PMOS

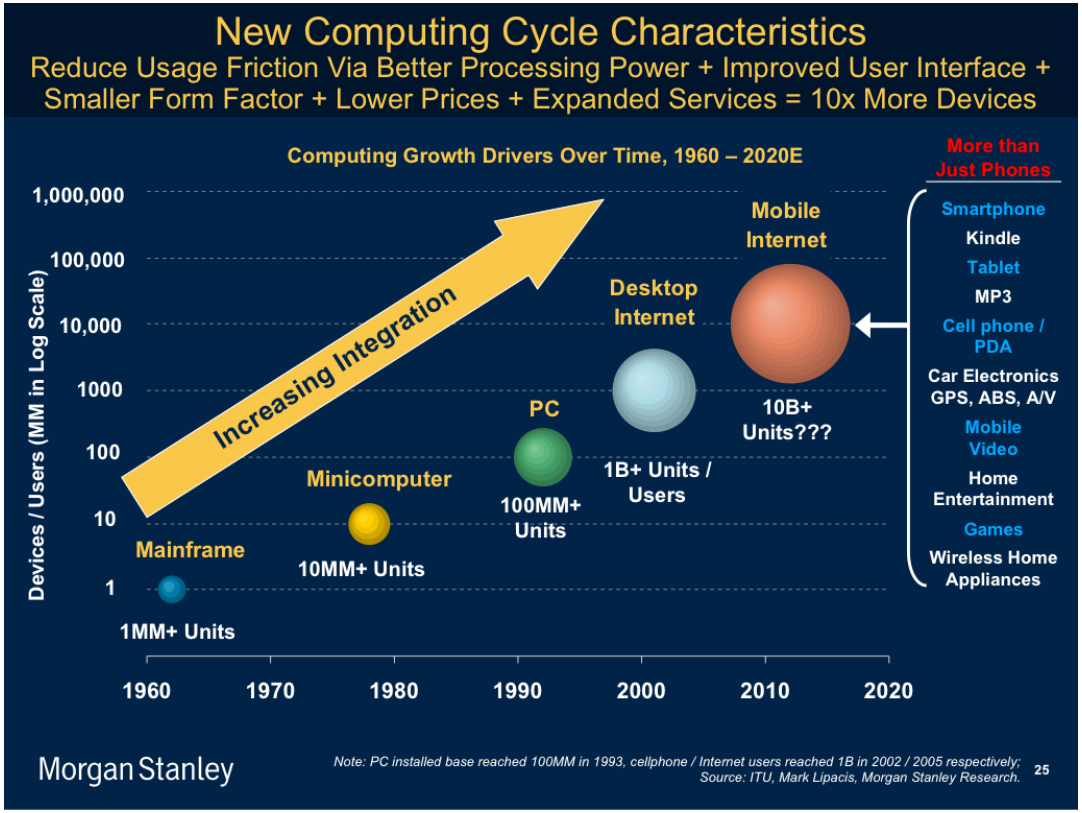


Modeling PMOS



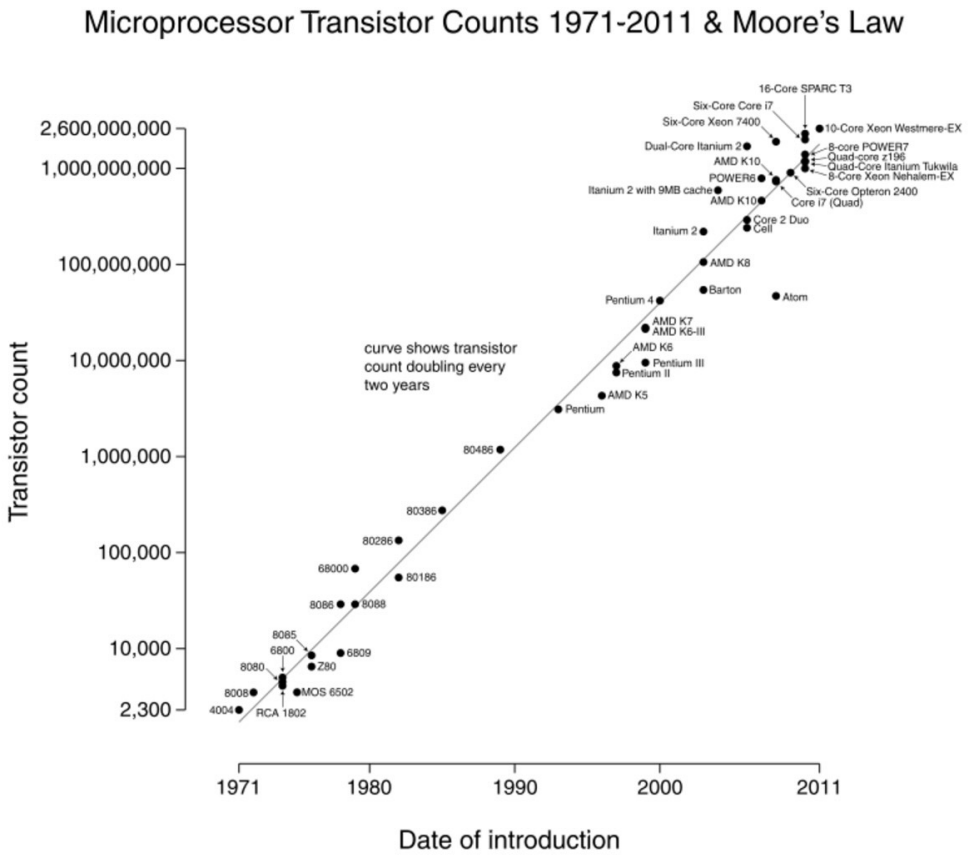
- Voltage difference between drain and source causes positive charge carriers (holes) to flow from Source to Drain
- Transistor is on (-Vgate: means gate has lower voltage than Source and Drain)

Historical Growth of Electronics



- History: 10X growth in each generation of electronic devices

Semiconductor Technology Development



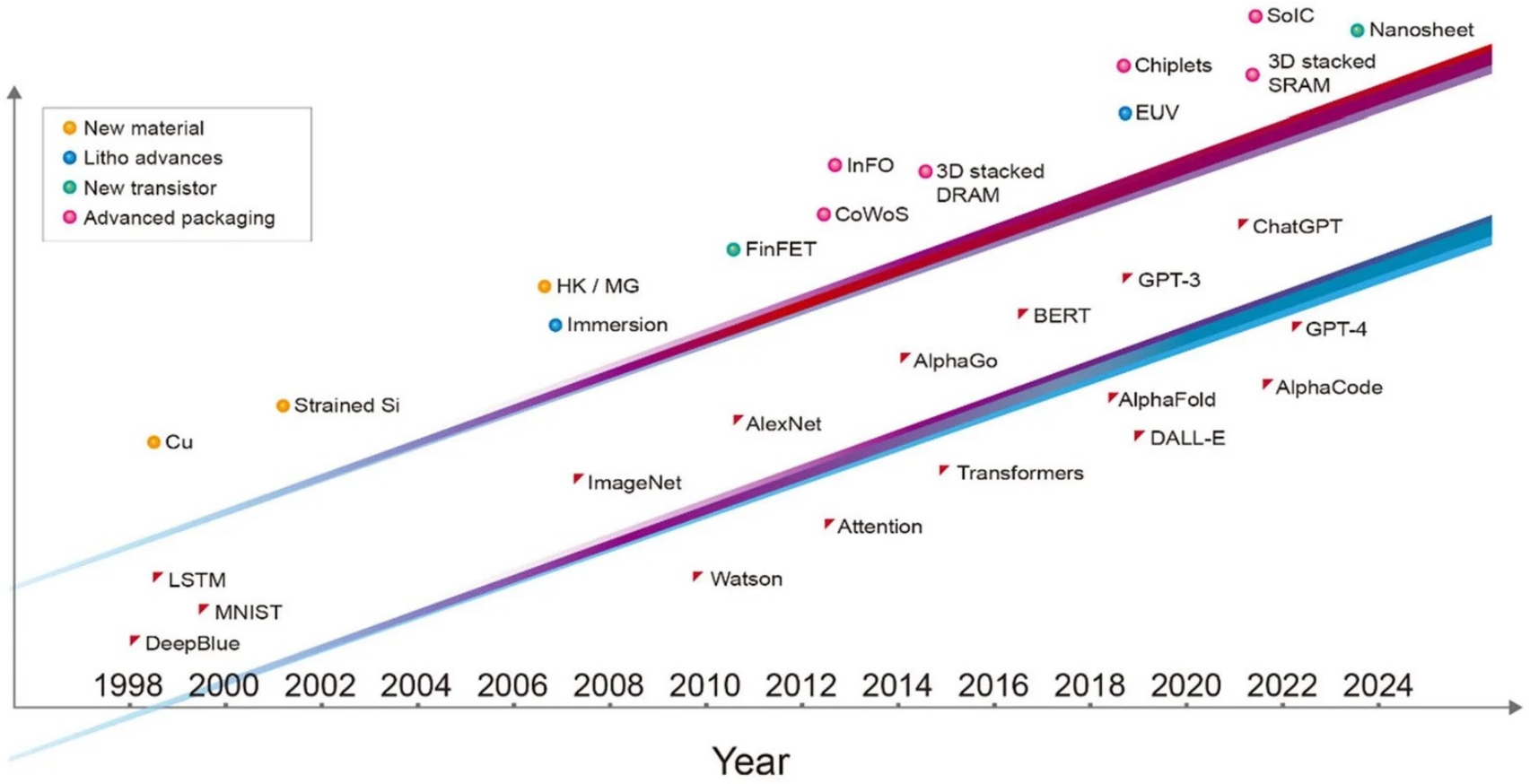
- Moore's law: Transistor count double every 18 months
  - Slow down since 2013 and expect to double every 3 years
- Multiple billions of transistors on a chip today

https://www.youtube.com/watch?v=Z7M8etXUEUU

Semiconductor Trends

- Dennard's Scaling (1974)
  - voltage and current should be proportional to the linear dimensions of a transistor
  - As transistors shrink, power remains proportional to the area of the transistor
  - Power density (power/area) remains constant
  - There are sources of power consumption beyond the switching activity
    - Resulted in a "Power Wall" that has limited practical processor frequency to around 4 GHz since 2006

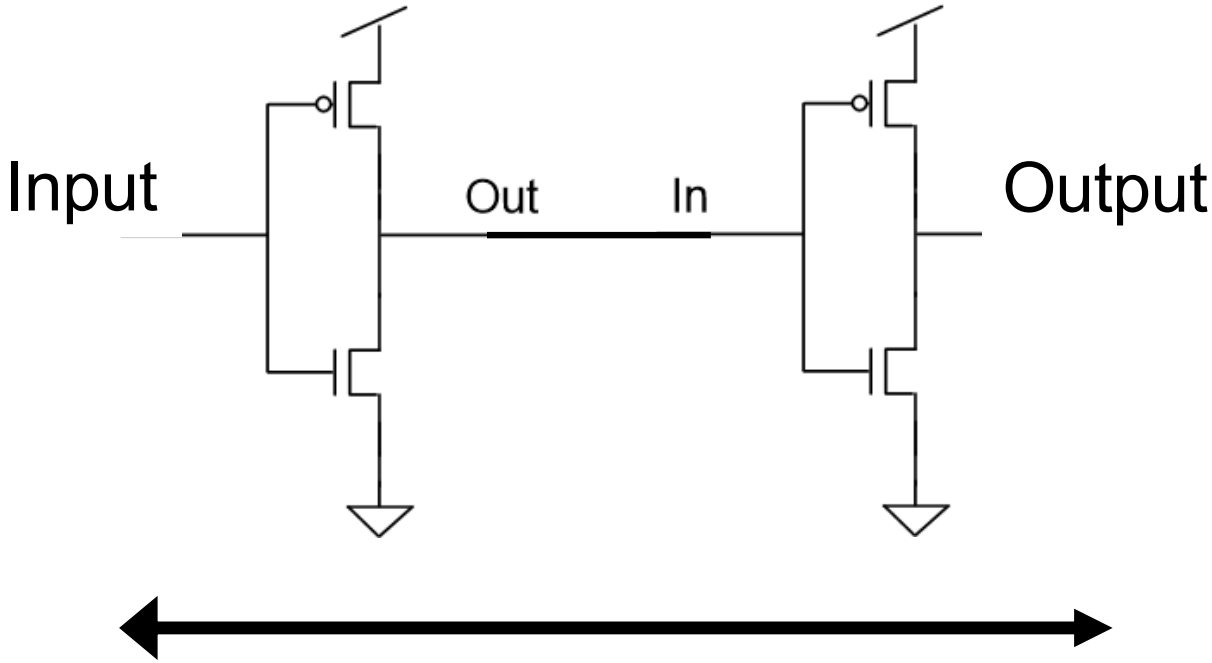
Semiconductor Trends



Advances in semiconductor technology [top line]—including new materials, advances in lithography, new types of transistors, and advanced packaging—have driven the development of more capable AI systems [bottom line].

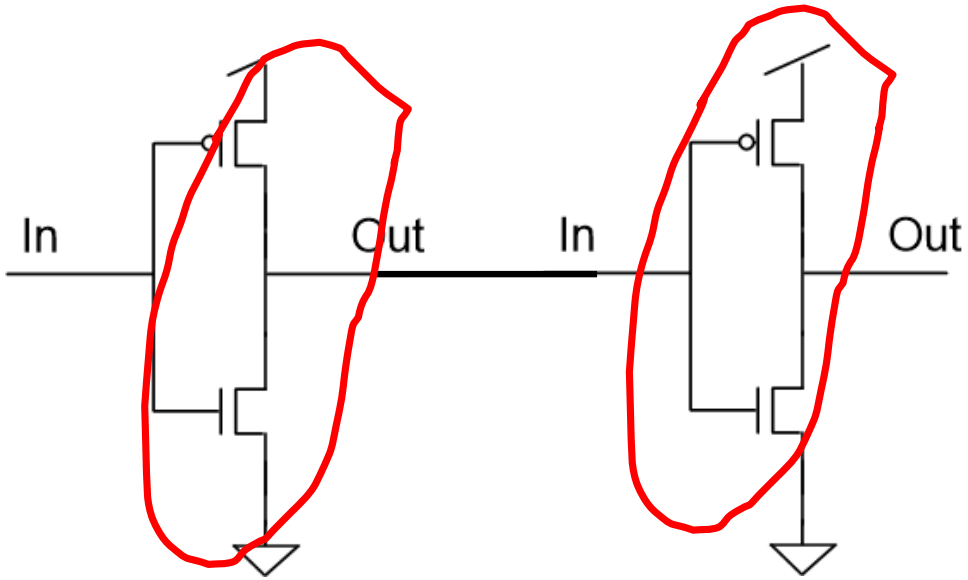
IEEE Spectrum, March 2024

What determines the performance of NMOS/PMOS?

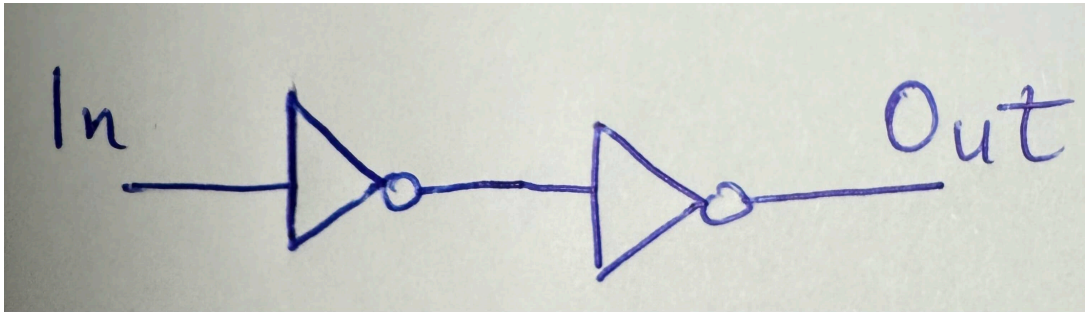


How fast will I observe the consequence of a switch/change in digital value/state of the Input after it propagates through one (or multiple stages of CMOS transistors/gates)?

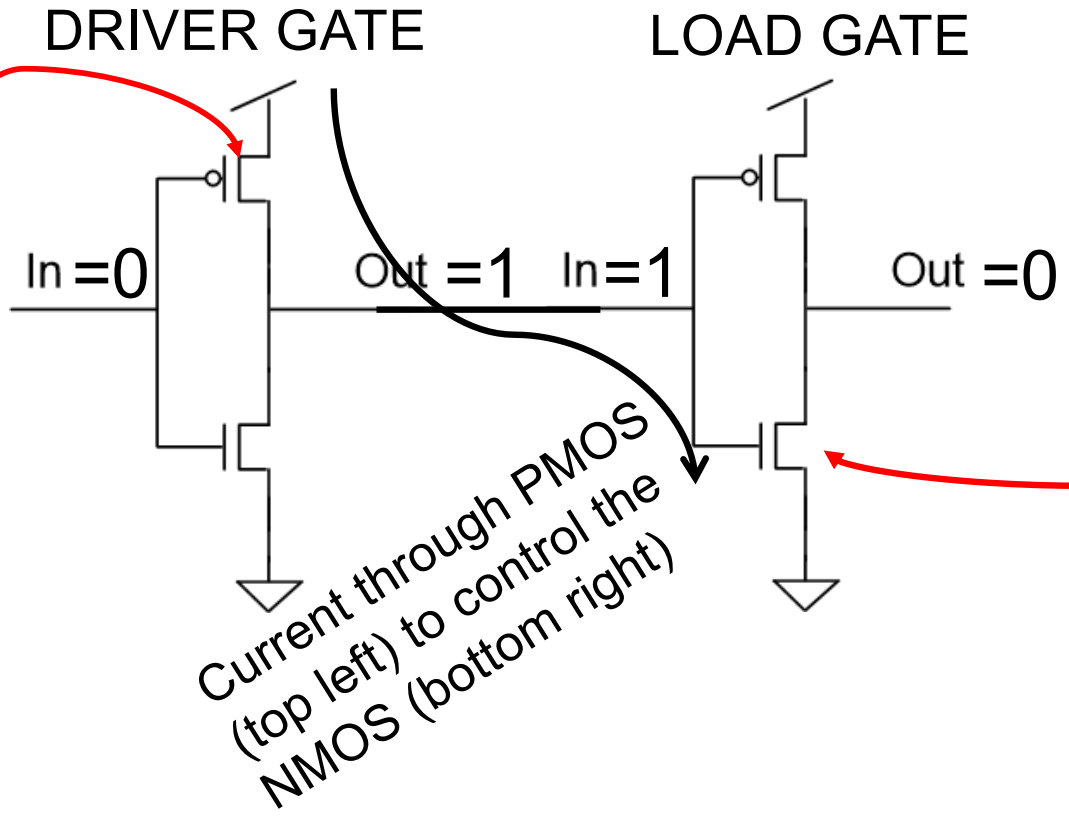
What determines the performance of NMOS/PMOS Transistors and CMOS Circuits?



DRIVER GATE LOAD GATE

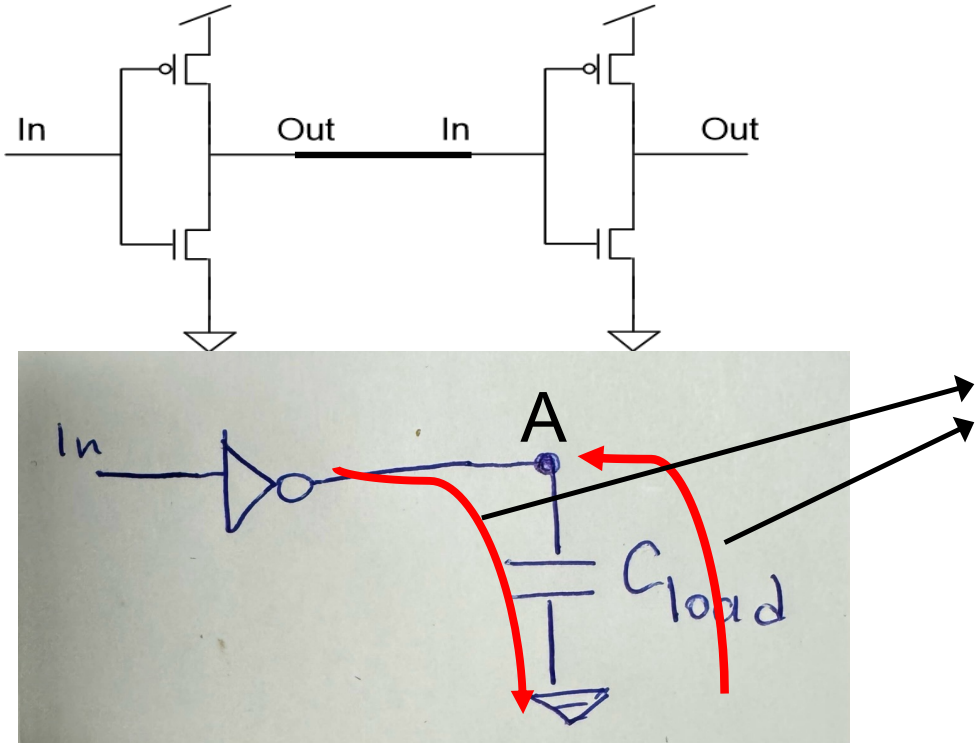


What happens when CMOS gates/transistors interact?



PMOS switch of the Driver Gate closes. Current flowing through the channel of this PMOS transistor performs 'electrical work' to create a gate voltage at the input of the Load Gate's transistors to influence the creation of a conducting channel within the **NMOS transistor in the Load Gate**

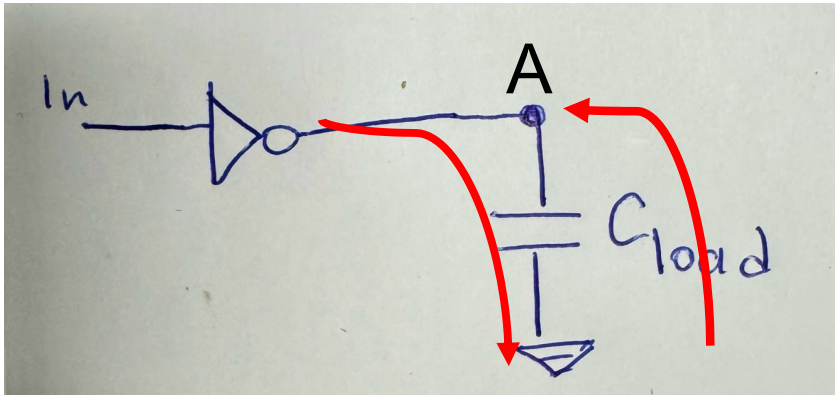
What determines the performance of NMOS/PMOS Transistors and CMOS Circuits?



Charge and discharge of the load brings the voltage at point A up to "high/1" or down to "low/0"

The 'electrical work' that needs to be performed by the Driver gate is represented as the action of charging/discharging a load capacitance (represented by the Load Inverter), which brings the voltage at the output (labeled A in picture above) of the Driver Inverter Gate to 1/0.

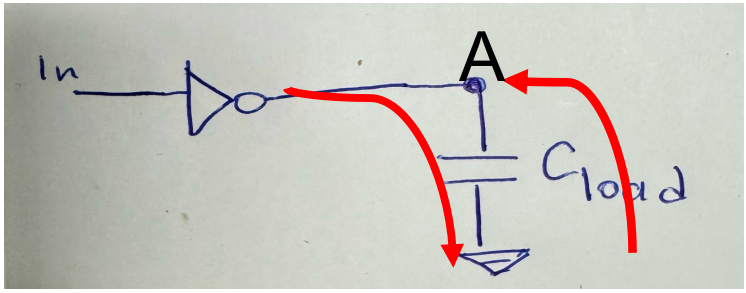
What determines the performance of NMOS/PMOS Transistors and CMOS Circuits?



Observation 1: The time it takes for point A to 'charge'/'discharge' depends on the size of the C_{load}. The bigger C_{load} is the more 'work' it takes to do this, hence the longer it will take, i.e. higher delay.

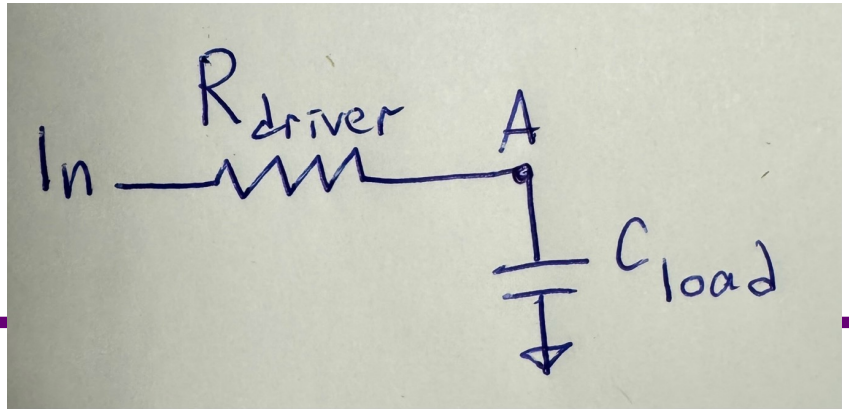
This C_{load} quantity is represented with an **intrinsic capacitance C of the transistor**(s) involved in the load side.

What determines the performance of NMOS/PMOS Transistors and CMOS Circuits?

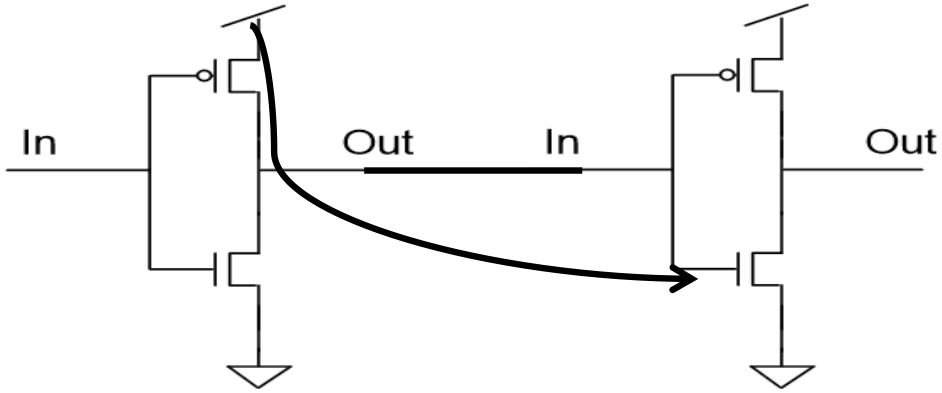


Observation 2: The time it takes for point A to 'charge'/'discharge' depends on the **strength** of the Driver. How do we define the driving strength of a transistor?

Driving strength correlates with the **intrinsic resistance R of the conducting channel** of the transistor involved on the driver side.



Putting it all together



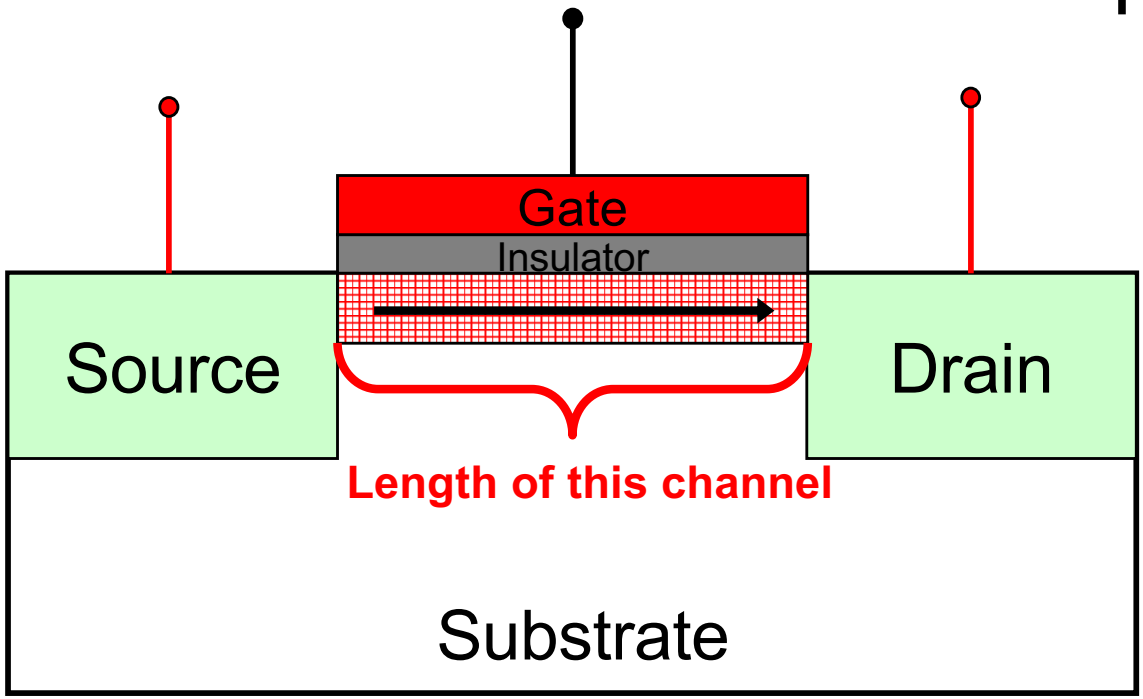
The time (circuit delay) it takes for the driver gate to change the value at its output (while connected to some load gate) depends on

how “easily” current flow can be created through the driver transistor (how big/small is its R?)

+

how “easily” this amount of current flow establishes control over the gate of the load transistor (size of the C of the load that needs to be charged/discharged)

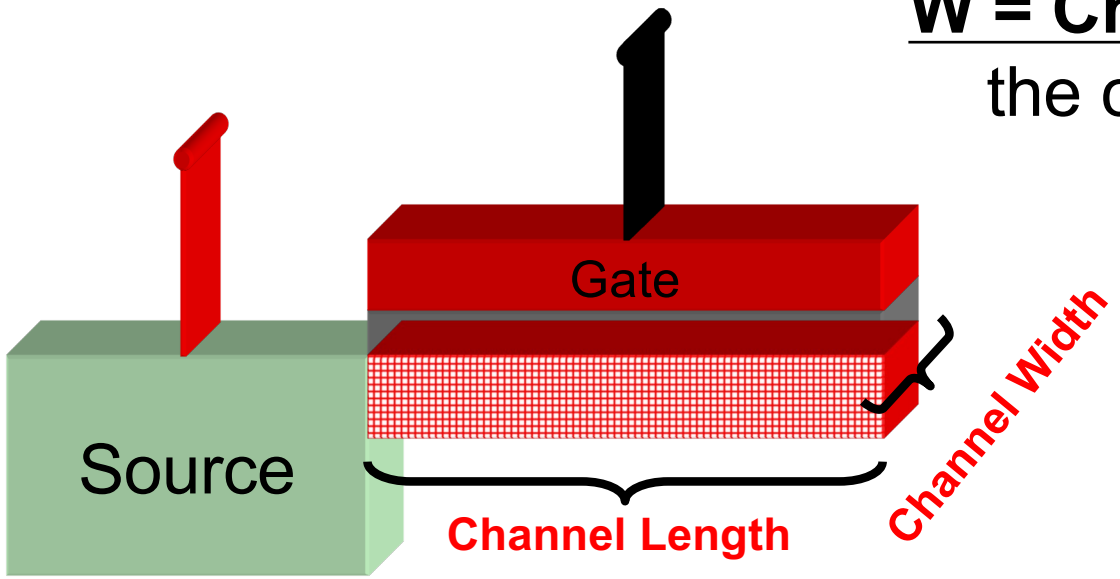
Intrinsic resistance R characterizes Drive Strength



L = Channel Length:

The longer the channel is, the higher R

Intrinsic resistance R characterizes Drive Strength



W = Channel Width: The wider the channel is, the lower R

W/L = Channel Width to Length Ratio

Largely determines the intrinsic resistance **R** of the device

Intrinsic resistance R characterizes Drive Strength

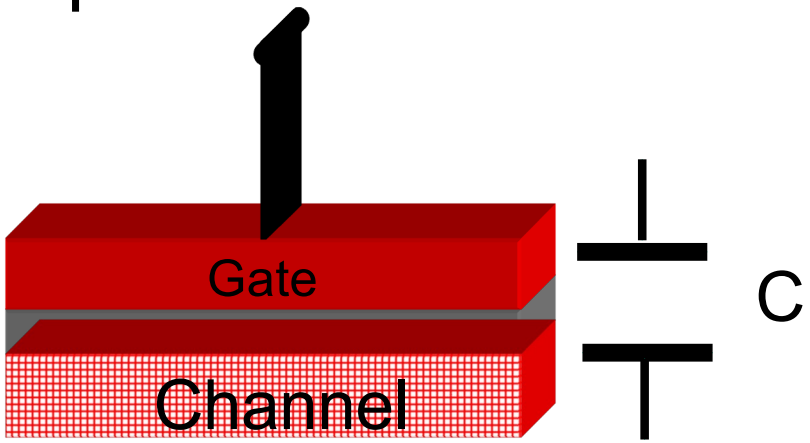
W/L = Transistor Width to Length Ratio

Largely determines the intrinsic resistance **R** of the transistor

Larger W/L indicates SMALL R – higher drive strength (circuit is likely to switch faster)

Small W/L indicates LARGE R – poorer drive strength (circuit is likely to switch slower)

Intrinsic capacitance C characterizes capacitive load burden posed by the transistor?



C is proportional to plate area **W*L**

- The Gate and the Channel act as a CAPACITOR (two plates separated by a dielectric insulation layer)
- This CAPACITANCE impacts the ability of pushing a certain amount of charge across this region

The wider the channel the higher C
The longer the channel the higher C

Intrinsic resistance R and the capacitance C characterize delay

- Just like charging capacitor
  - Q=CV=I*t
  - t=CV/I

I is related to the **R**esistance of the driver that is supplying the current I to build up the charge Q

t (time) is directly proportional to C**C**apacitance

Intrinsic resistance R and the capacitance C characterize delay

Main takeaway: the delay of signal propagation through a chain of a driver and a load is proportional to the product of R*C

$$R_{drive} * C_{load}$$

Intrinsic resistance R and the capacitance C characterize delay

Main takeaway: the delay of signal propagation through a chain of a driver and a load is proportional to the product of R*C

$$R_{drive} * C_{load}$$

This forces designers to hit the sweet spot of a tradeoff:

A good driver should have relatively wide W, yet a transistor with wide W is going to have larger C. If this transistor needs to be driven by another transistor, then it will, in turn, present itself as a “bad” load to its own driver…

Intrinsic resistance R and the capacitance C characterize delay

Main takeaway: the delay of signal propagation through a chain of a driver and a load is proportional to the product of R*C

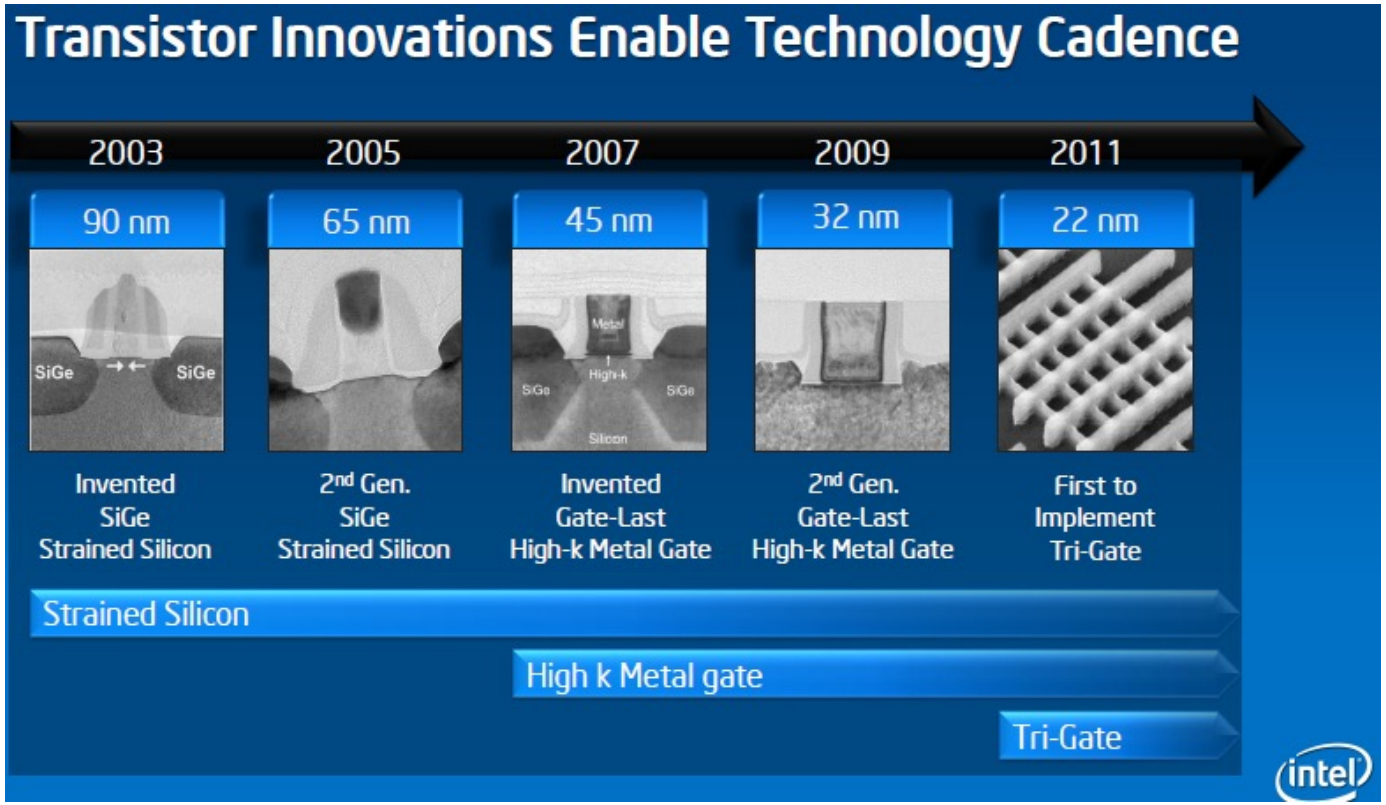
$$R_{drive} * C_{load}$$

This forces designers to hit the sweet spot of a tradeoff:

Similarly, a good load should have small (narrow) W, yet a transistor with narrow W is going to have larger R and it will turn out to be a lousy driver, when it is connected to some load downstream.

Optimal balance between drive strength and capacitive load is adjusted by determining best sizing in different connectivity topologies and speed requirements.

Technology Scaling Trend



- Process innovation drives transistor scaling
  - Strained silicon from from 90nm -> increase mobility, faster
  - High-k is High-k Metal gate from 45nm -> reduce leakage power, low power
  - Tri-gate/FinFET from 22nm -> better gate control, low power/high speed

Strained Si

- Silicon layer placed on top of a “stress” layer (SiGe)
- Causes Si atoms to slightly stretch out, move away from each other
- Strain causes the Si atoms to stretch apart by ~1%
- Benefit: improves mobility of charge carriers (electrons or holes)

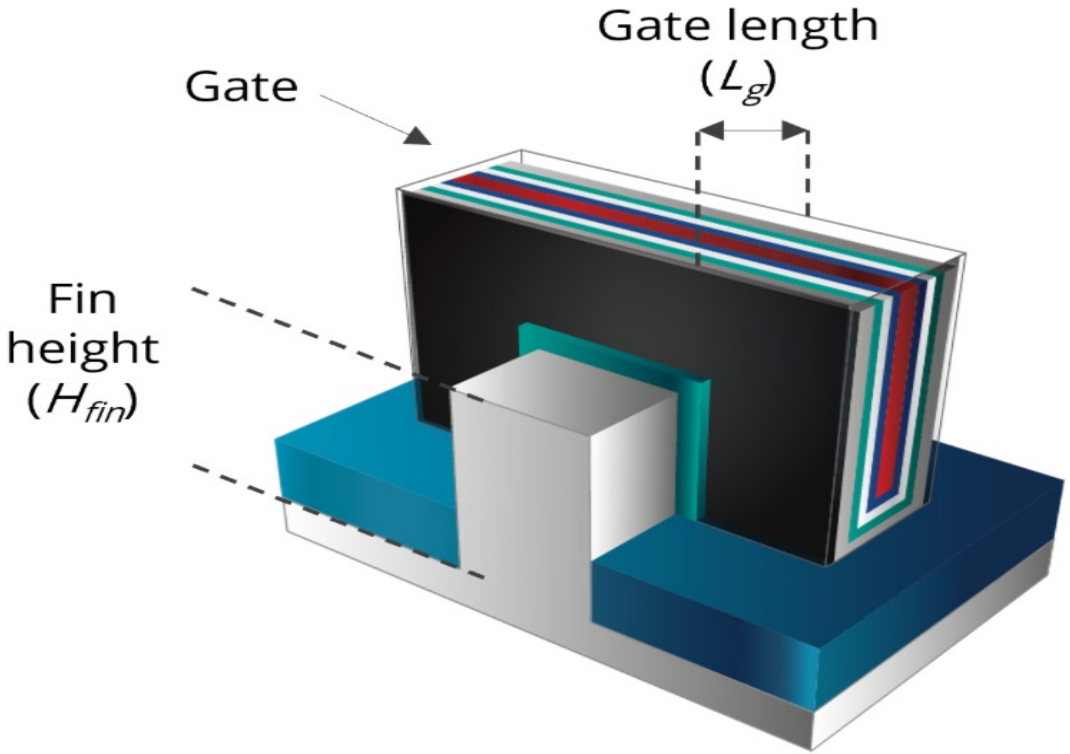
High k Gate

- High dielectric constant k
  - Improved capacity to hold charge
  - As transistors got smaller, gate became too thin to insulate effectively, too much leakage
- Replace silicon dioxide gate with a material of high dielectric constant (hafnium)

FinFET

- Thin vertical fins sticking from the planar channel surface layer
- Creates a larger total area of interaction between gate and channel
- Greater electrostatic control over the channel
- Faster and higher efficiency transistors

Design - Geometry

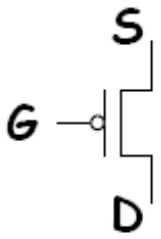


appliedmaterials.com

NMOS/PMOS as a switch

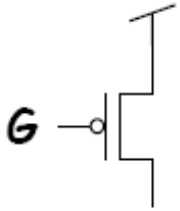
Binary logic values represented by voltages:

"High" = Supply Voltage, "Low" = Ground Voltage

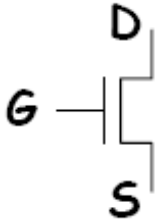


PMOS switch closes S and D when G="low"=0V

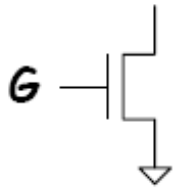
Supply Voltage = V_{DD}



PMOS switch good at Passing HIGH



NMOS switch closes D and S when G="high"=V_{DD}



NMOS switch good at Passing LOW

Ground = GND = 0V

CMOS: Inverting Logic

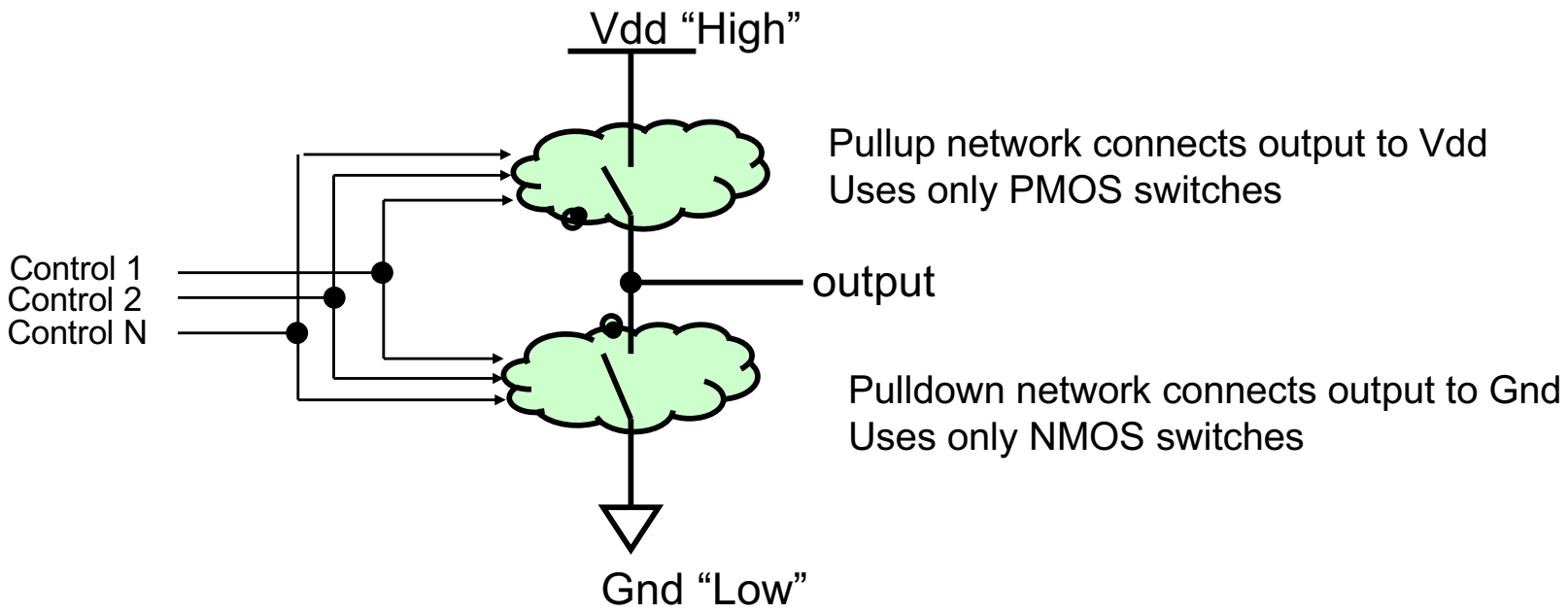
- PMOS gates are normally closed switches that are good at transmitting **only true (high)** signals
- NMOS gates are normally open switches that are good at transmitting **only false (low)** signals
- Therefore instead of directly building AND/OR gates using MOS transistors, we build NAND/NOR
- If we need AND/OR we use NAND/NOR followed by an additional inverter NOT gate

Building CMOS Logic Gate

Pulldown: realizes “0/FALSE” output, hence, should be activated when the function needs to produce “0”. So, it must be made out of NMOS transistors.

Pullup: realizes “1/TRUE” output, hence, should be activated when the function needs to produce “1”. So, it must be made out of PMOS transistors.

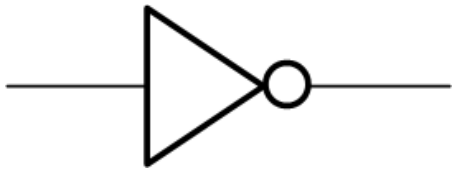
Generic CMOS Gate using a switch network



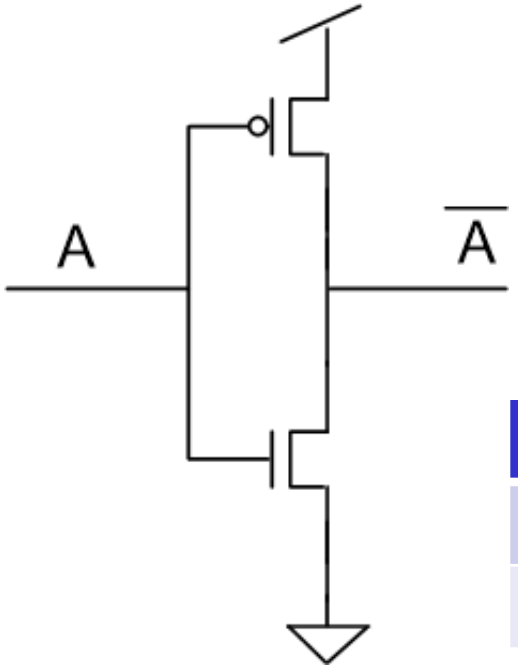
For every set of input logic values, either pullup or pulldown network makes connection to VDD or GND

- **If both connected, power rails would be shorted together**
- **If neither connected, output would float (tristate logic)**

Simple Gate



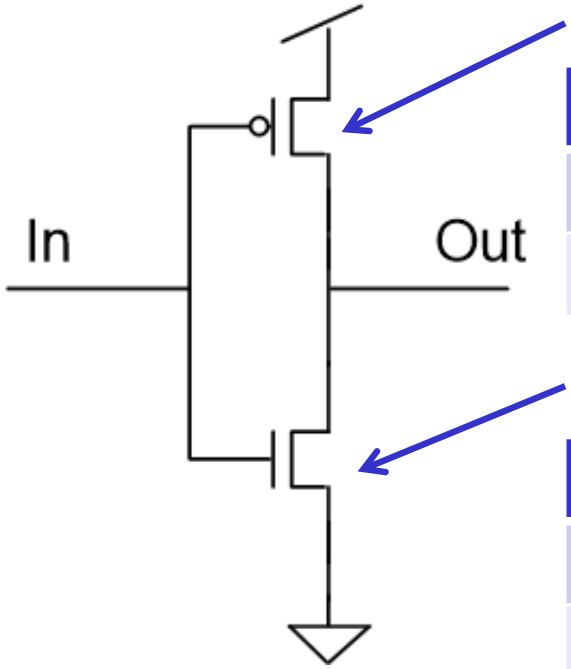
Inverter



In	Out
0	1
1	0

- When input A is high, output is low
- When input A is low, output is high

Pull-up and Pull-down Network



Pull-up network:

In	Out
0	1
1	Z (open)

Pull-down network:

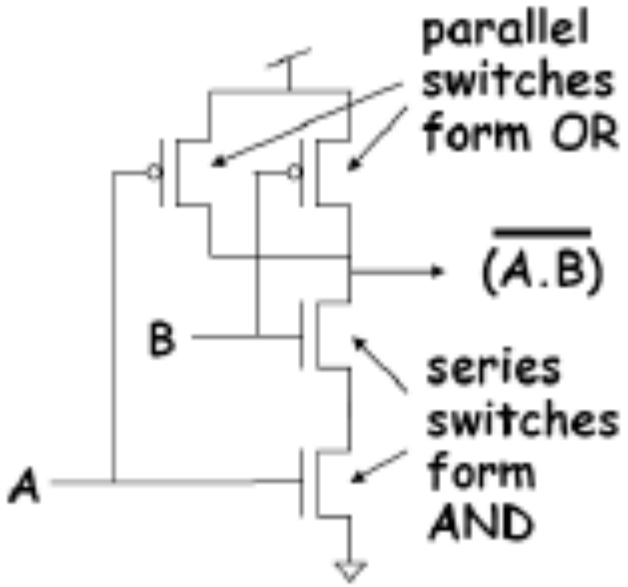
In	Out
0	Z (open)
1	0

Combined network:

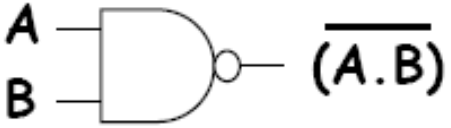
In	Out
0	1
1	0

Building a CMOS gate

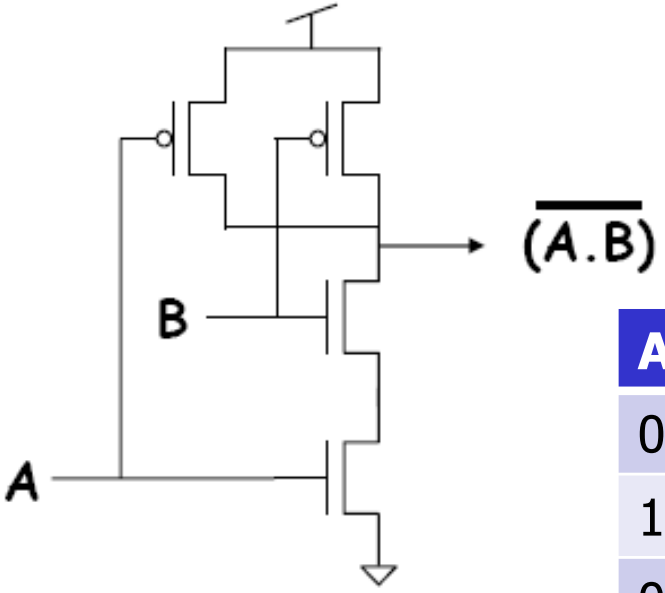
- When creating a behavior for a function F
  - Take inverted function F
  - Make network using NMOS with inputs
    - AND -in series, OR -in parall
    - Connect to ground (F=0)
- Example: NAND: $F=\overline{AB}$
- $\overline{F} = AB$



Simple Gate



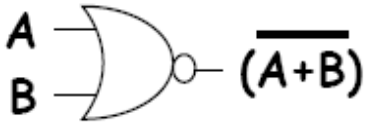
NAND Gate



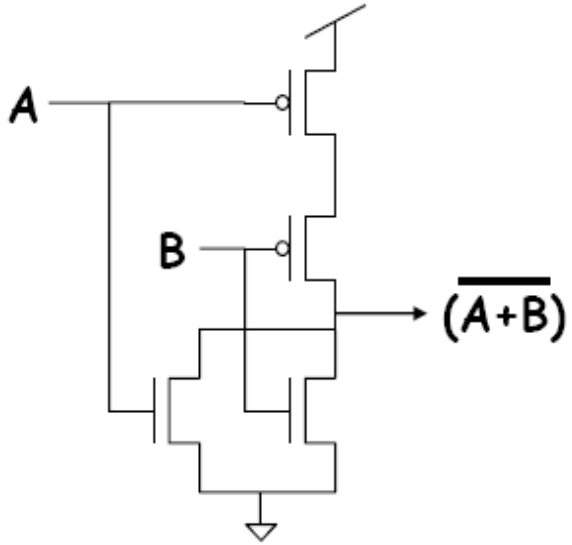
A	B	Out
0	0	1
1	0	1
0	1	1
1	1	0

- When both A and B are high, output is low
- When either A or B is low, output is high

Simple Gate



NOR Gate



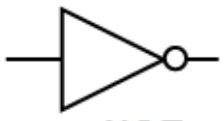
A	B	Out
0	0	1
1	0	0
0	1	0
1	1	0

- When both A and B are low, output is high
- When either A or B is high, output is low

Exercises

- How to build a NOR3 gate?
- How to build a OR3 gate?
- How to build a MUX2 gate?

Commonly Used Logic Gates



NOT

Input	Output
I	F
0	1
1	0

AND

Inputs		Output
A	B	F
0	0	0
1	0	0
0	1	0
1	1	1

NAND

Inputs		Output
A	B	F
0	0	1
1	0	1
0	1	1
1	1	0

OR

Inputs		Output
A	B	F
0	0	0
1	0	1
0	1	1
1	1	1

NOR

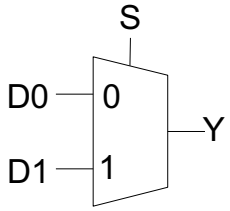
Inputs		Output
A	B	F
0	0	1
1	0	0
0	1	0
1	1	0

EXCLUSIVE OR

Inputs		Output
A	B	F
0	0	0
0	1	1
1	0	1
1	1	0

EXCLUSIVE NOR

Inputs		Output
A	B	F
0	0	1
0	1	0
1	0	0
1	1	1



MUX2 (multiplexer)

Inputs			Output
S	D1	D0	Y
0	X	0	0
0	X	1	1
1	0	X	0
1	1	X	1