Digital Logic Design

Lecture 9

Implementation technologies

Supply voltage

The lowest voltage in the system is 0V, also called **ground** or GND.

The highest voltage in the system comes from the power supply and is usually called V_{DD}

In 1970's technology, V_{DD} was generally 5V.

In the mid-1990s, transistors became too small to withstand the voltage 5V.

V_{DD} has dropped to 3.3V, 2.5V, 1.8V, 1.5V, 1.2V, or even lower to save power and avoid overloading the transistors.

V_{DD} stands for the **voltage on the drain** of a metal-oxide-semiconductor transistor, used to build most modern chips.

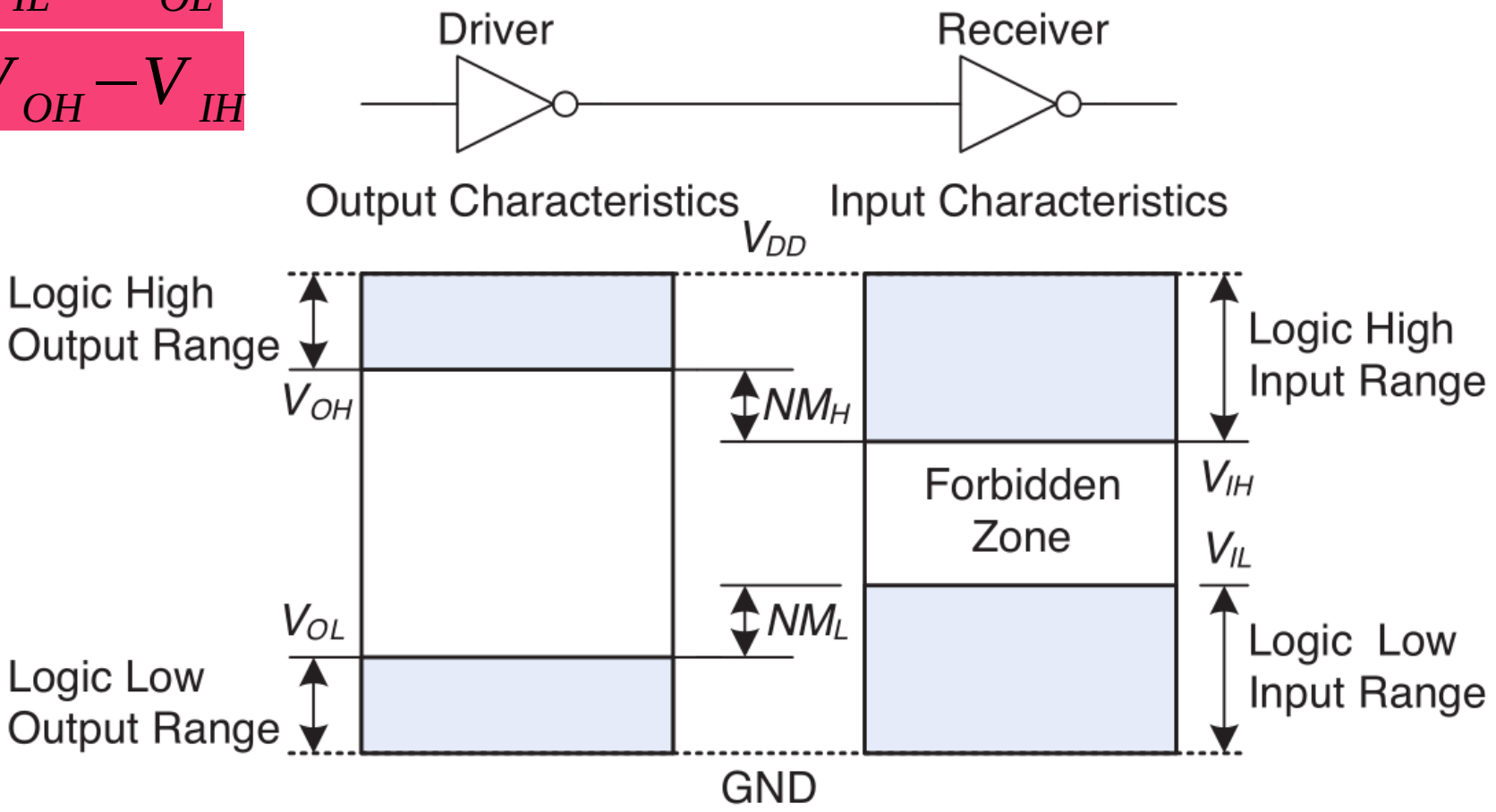
The power supply voltage is also sometimes called V_{CC}, standing for the **voltage on the collector** of a bipolar junction transistor used to build chips in an older technology.

Ground is sometimes called V_{SS} because it is the voltage on the source of a metal-oxide-semiconductor transistor.

Noise margins

$$NM_L = V_{IL} - V_{OL}$$

$$NM_H = V_{OH} - V_{IH}$$

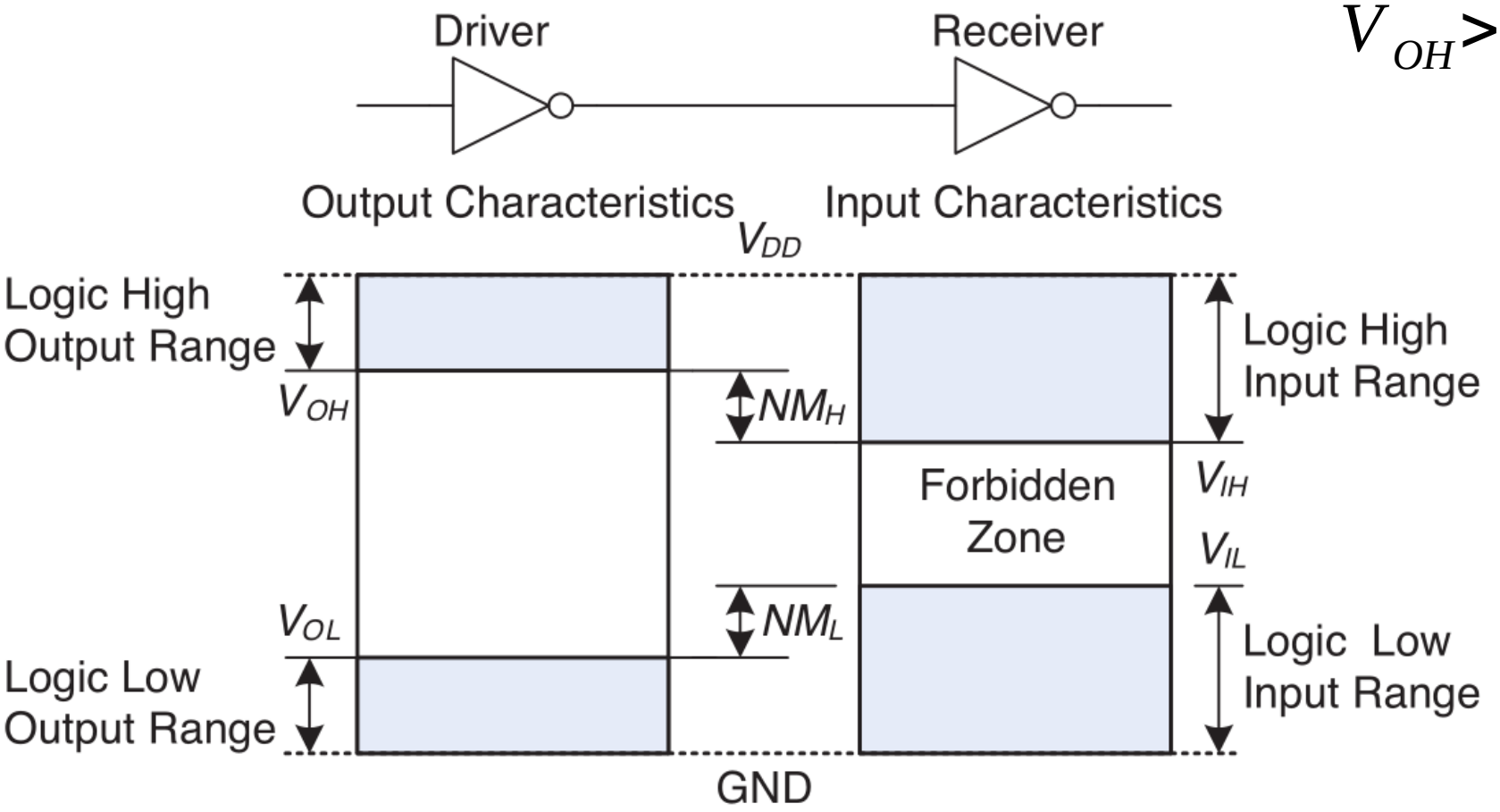


The noise margin is the amount of noise that could be added to a worst-case output such that the signal can still be interpreted as a valid input.

If the output of the driver is to be correctly interpreted at the input of the receiver, we must choose

$$V_{OL} < V_{IL}$$

$$V_{OH} > V_{IH}$$

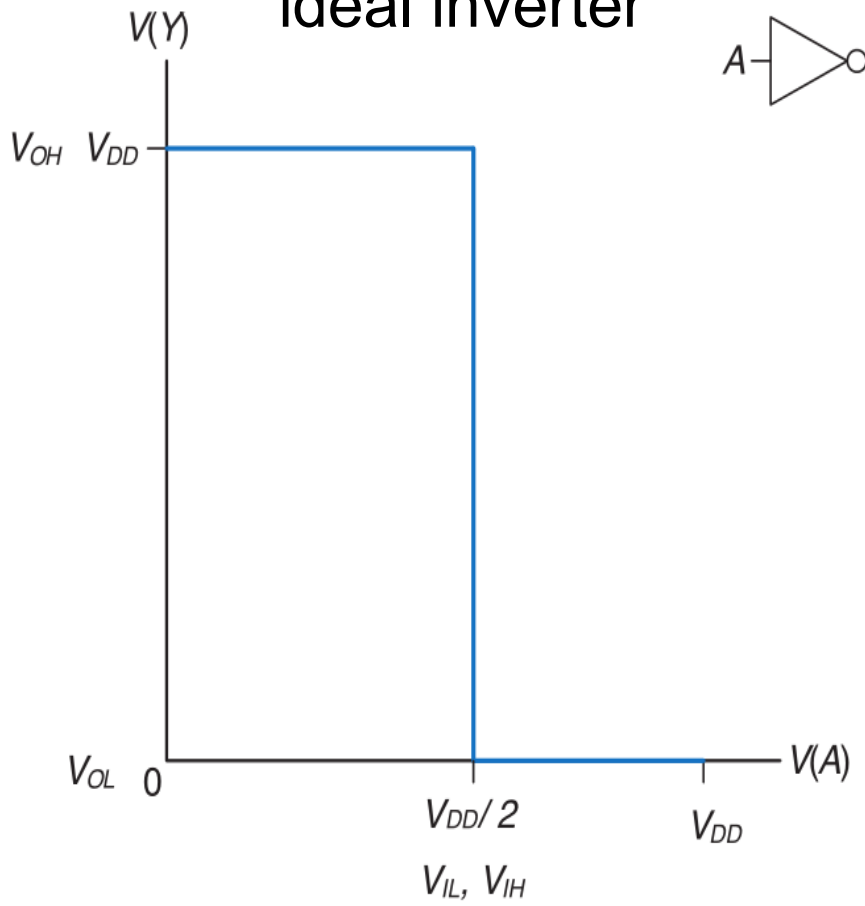


Even if the output of the driver is contaminated by some noise, the input of the receiver will still detect the correct logic level.

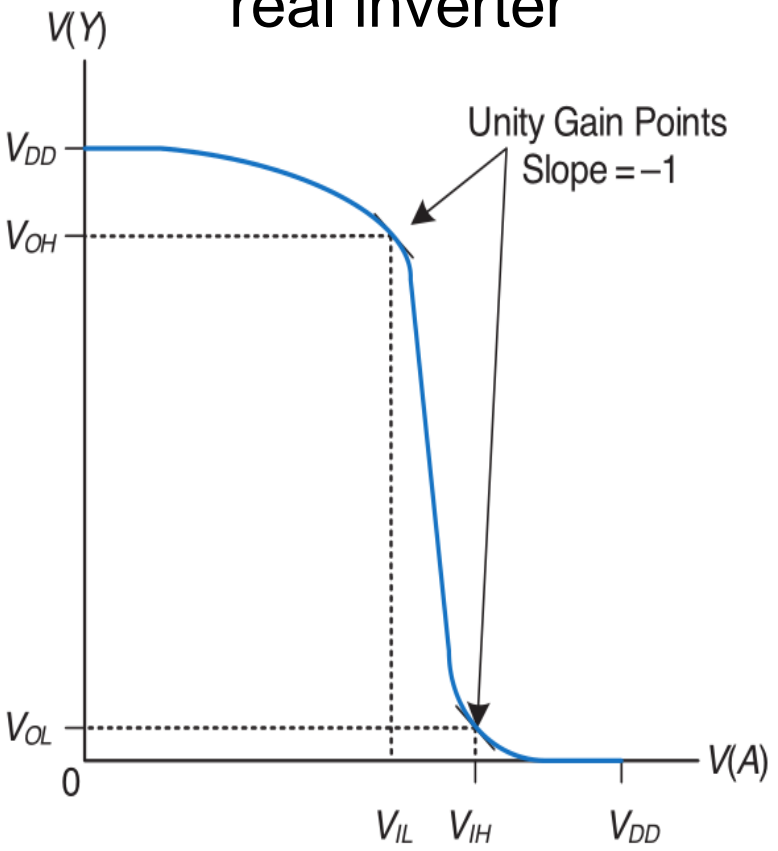
DC transfer characteristic

The **DC transfer characteristic** of a gate describe the output voltage as a function of the input voltage when the input is changed slowly enough that the output can keep up.

ideal inverter



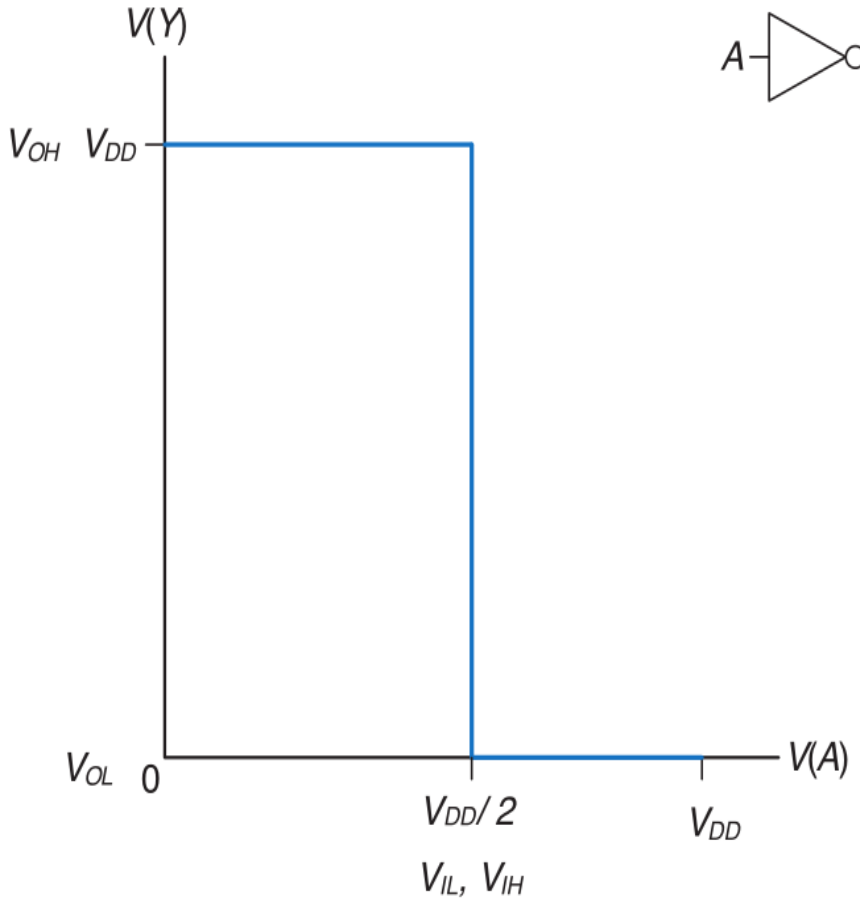
real inverter

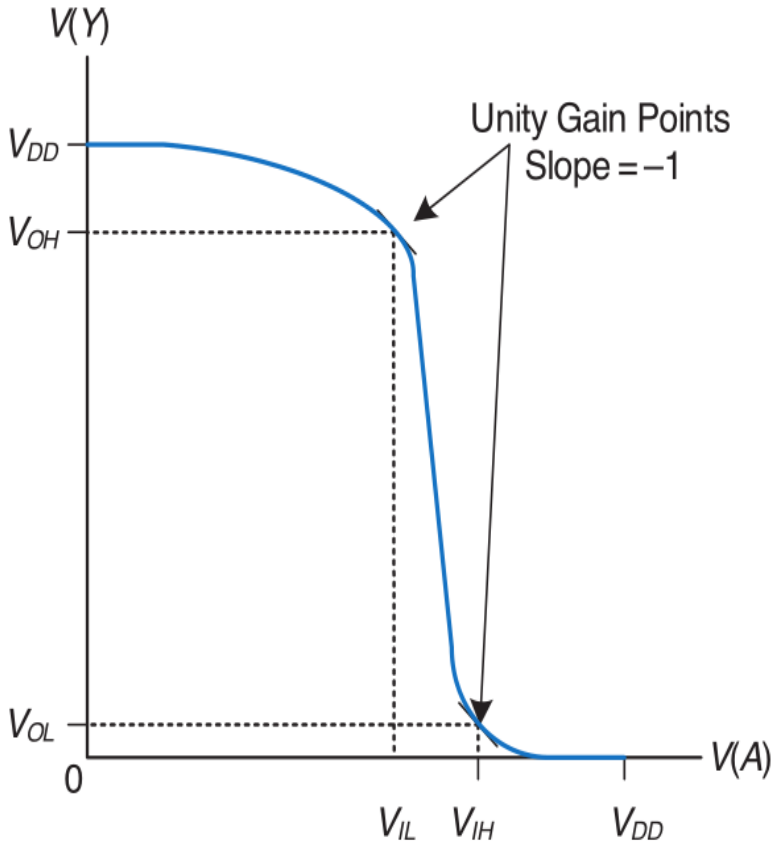


A reasonable place to choose the logic levels is where the slope of the transfer characteristic $dV(Y)/dV(A)$ is −1.

These two points are called the **unity gain points**.

Choosing logic levels at the unity gain points usually maximizes the noise margins.





Logic families

The choice of V_{DD} and logic levels is arbitrary, but all gates that communicate must have compatible logic levels.

Therefore, gates are grouped into **logic families**.

Logic Family	V_{DD}	V_{IL}	V_{IH}	V_{OL}	V_{OH}
TTL	5 (4.75–5.25)	0.8	2.0	0.4	2.4
CMOS	5 (4.5–6)	1.35	3.15	0.33	3.84
LVTTL	3.3 (3–3.6)	0.8	2.0	0.4	2.4
LVCMOS	3.3 (3–3.6)	0.9	1.8	0.36	2.7

Older chips were built using **bipolar transistors** in a technology called **Transistor-Transistor Logic** (TTL).

Current technologies used CMOS arrangement of **field-effect transistors**.

They draw less power supply or input current.

All CMOS chips are sensitive to electrostatic discharge (ESD) caused by static electricity.

Ground yourself by touching a large metal object before handling CMOS chips, lest you zap them.

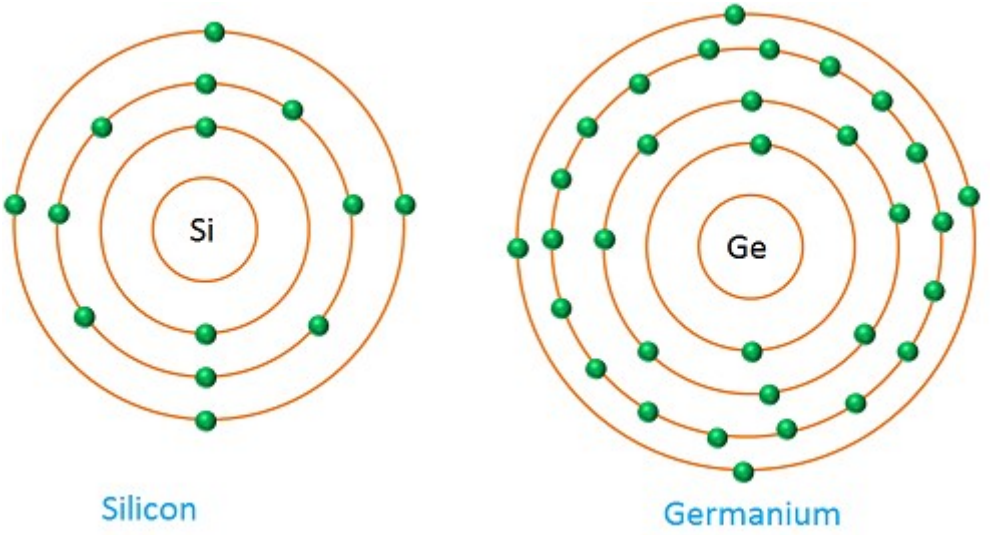
Semiconductors

Conductors conduct electric current.

Insulators don't conduct electric current.

Semiconductors can exhibit both conducting and insulating properties.

The two most commonly used semiconductor elements are **Silicon** and **Germanium**.



Silicon is the main constituent of sand and one of the most common elements on earth (silicon accounts for approximately 28% of the earth’s crust).

Doping

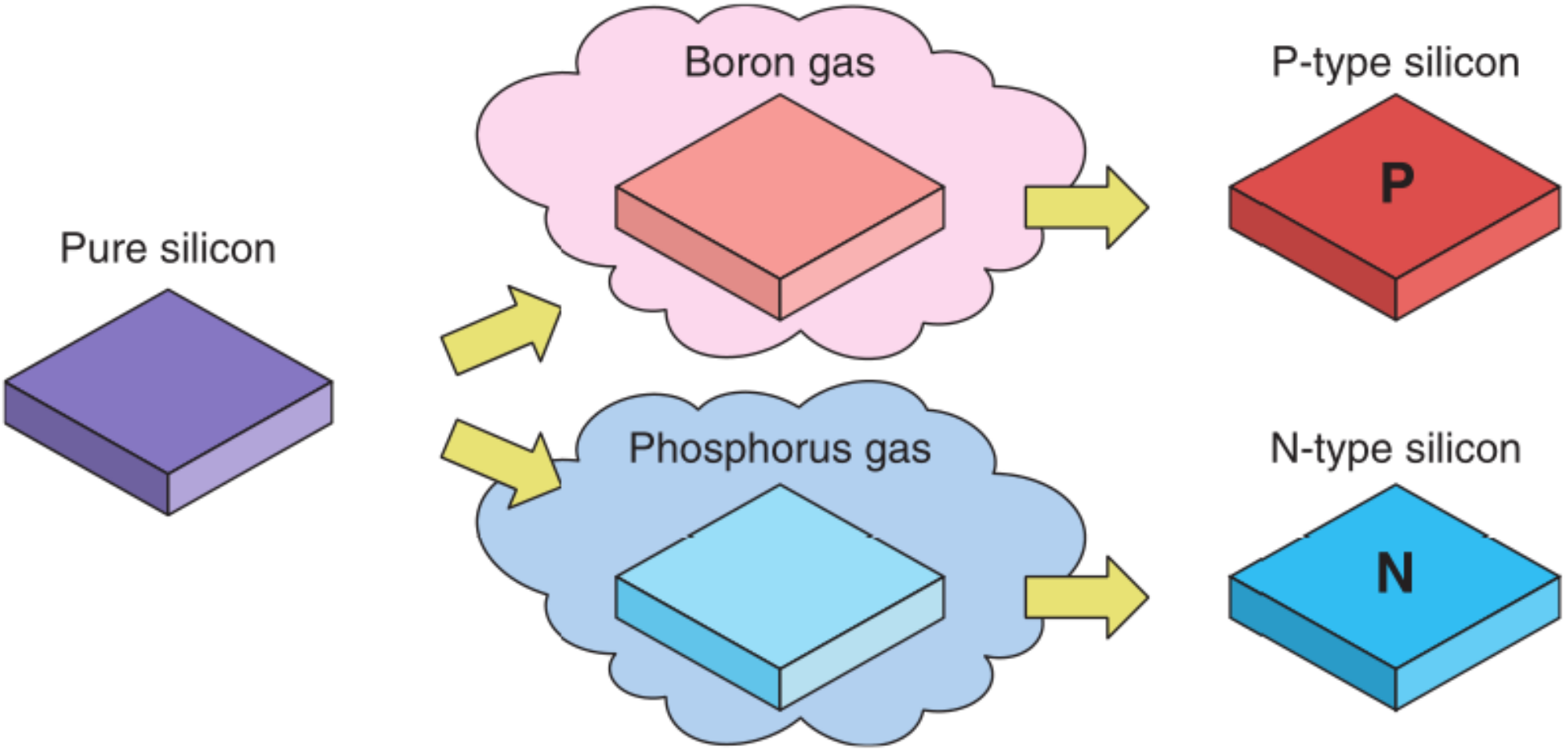
Pure crystalline silicon acts as an insulator.

Inserting certain impurities into the crystal lattice, silicon can be made to act as a conductor.

The process of inserting the impurities is known as **doping**.

The most commonly used dopants are **boron** atoms (chemical symbol: B) with three electrons in their outermost electron shells, and **phosphorus** atoms (chemical symbol: P) with five.

Silicon has four electrons in its outermost electron shell.



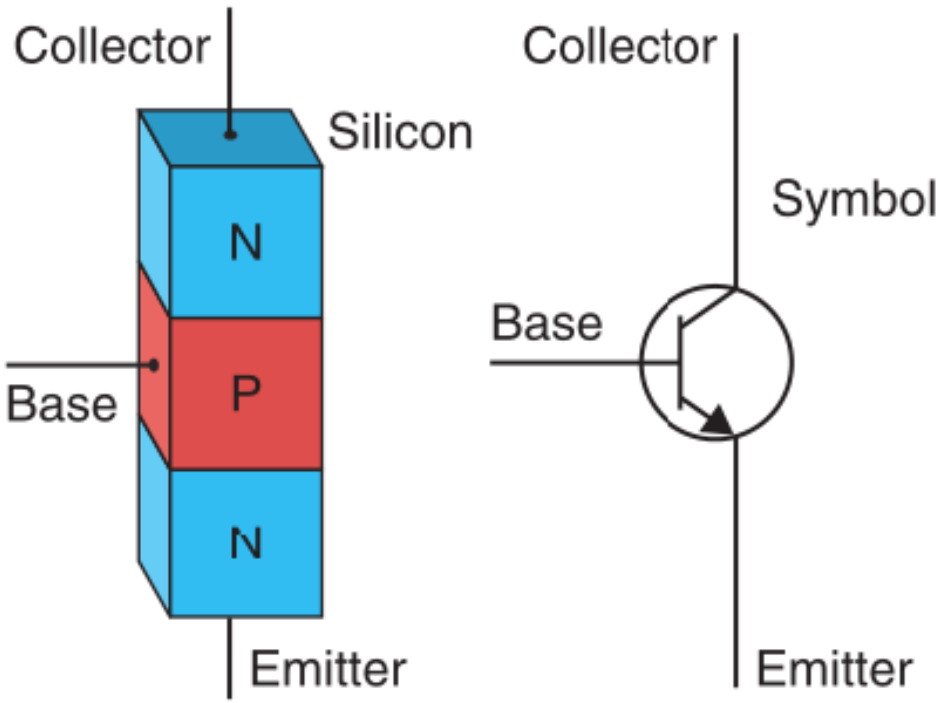
Boron-doped silicon is called **P-type silicon** and phosphorus-doped silicon is called **N-type silicon**.

Both P-type and N-type silicon are reasonably good conductors.

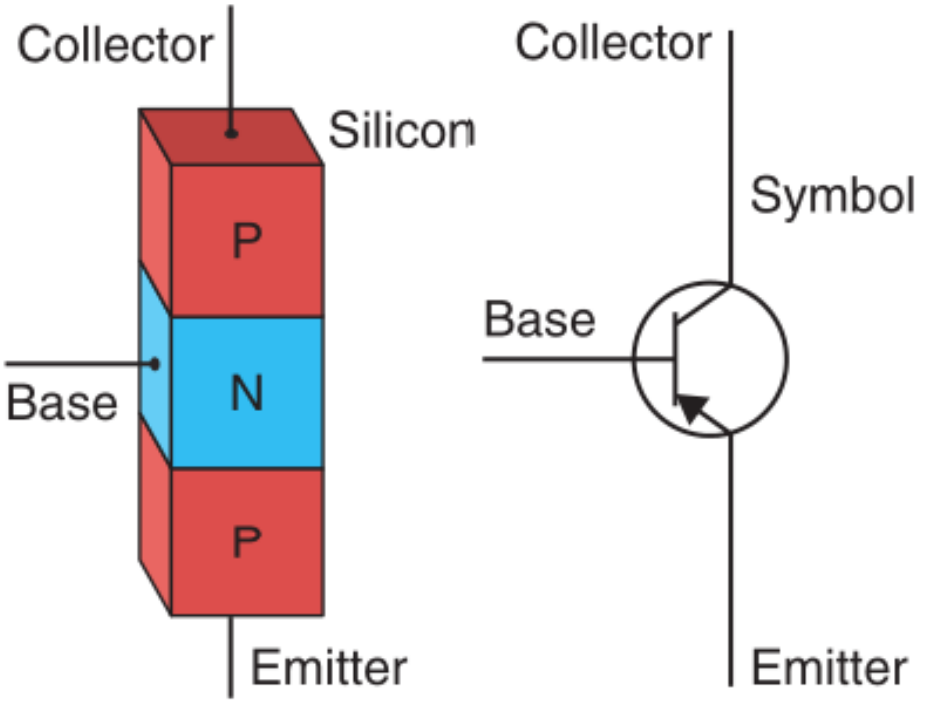
Bipolar junction transistors (BJT)

BJTs are formed from three pieces of doped silicon, called the **collector**, **base**, and **emitter**.

Two types of BJT: NPN and PNP

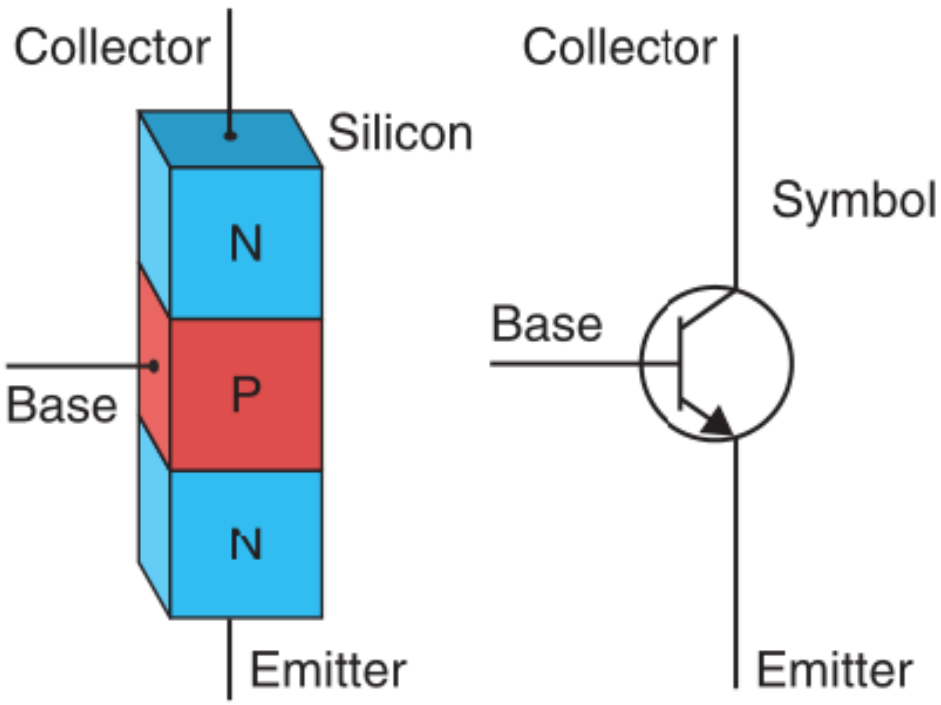


NPN bipolar junction transistor

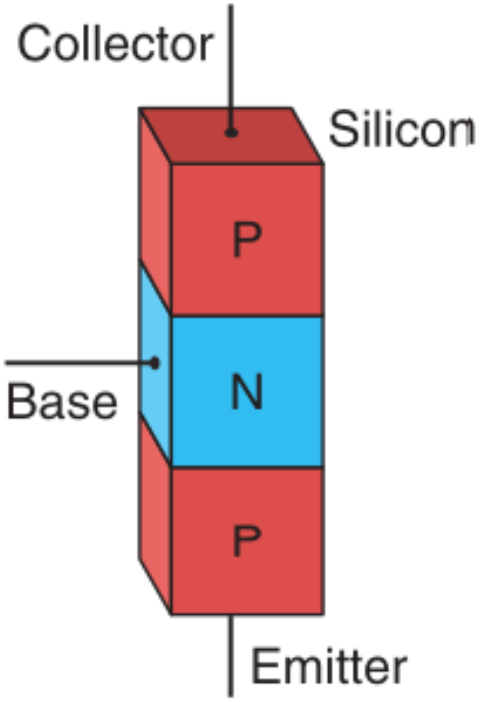


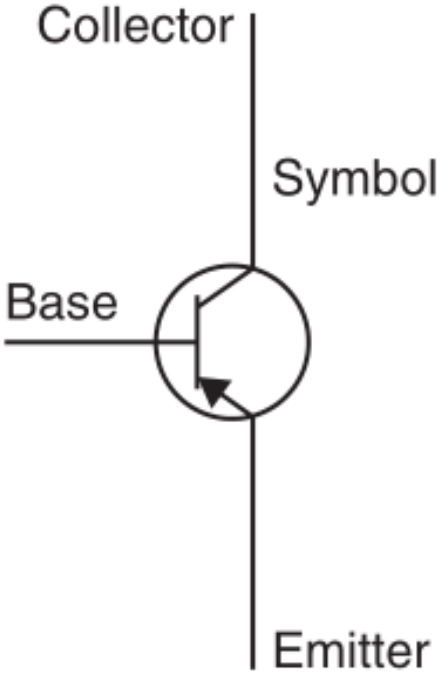
PNP bipolar junction transistor

When signals are applied to the base terminal, the transistor can be turned ON or OFF.



NPN bipolar junction transistor

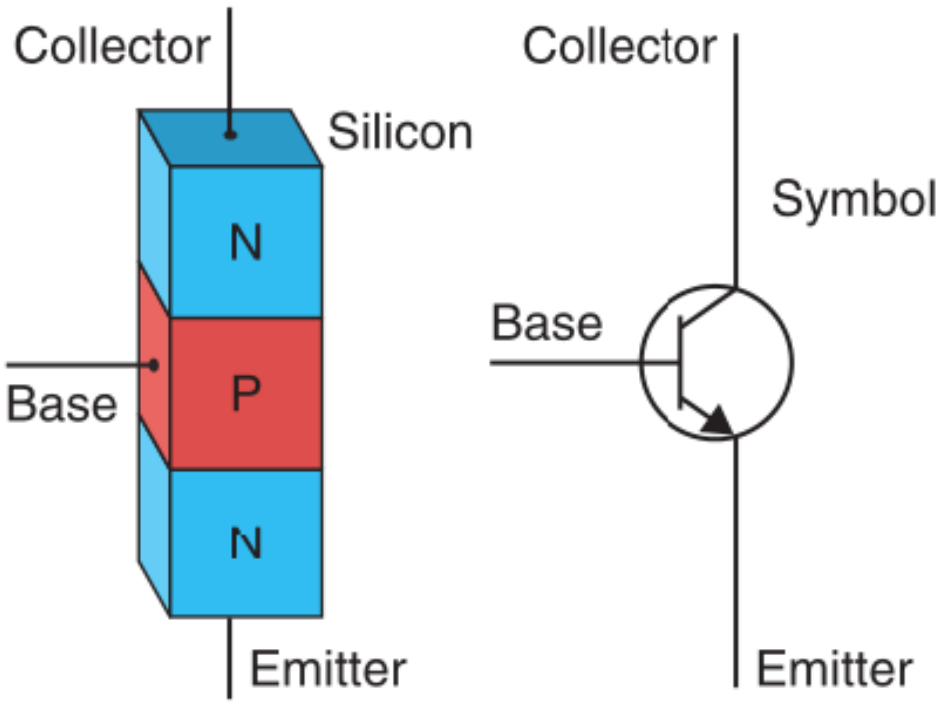




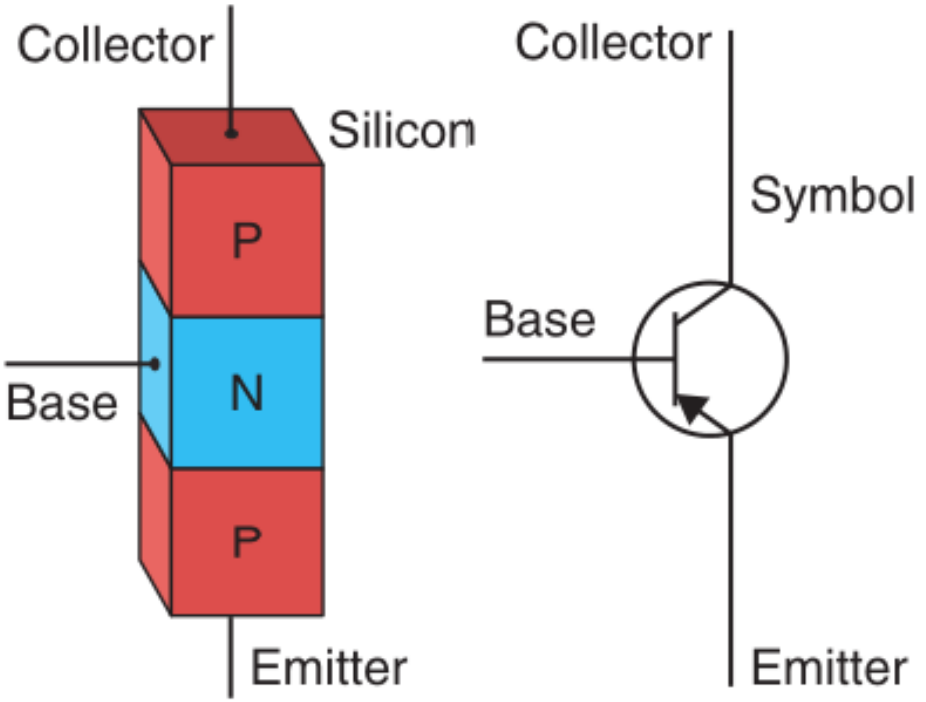
PNP bipolar junction transistor

If the transistor is turned ON, it acts like a CLOSED switch and allows current to flow between the collector and the emitter

If the transistor is turned OFF, it acts like an OPEN switch and no current flows.

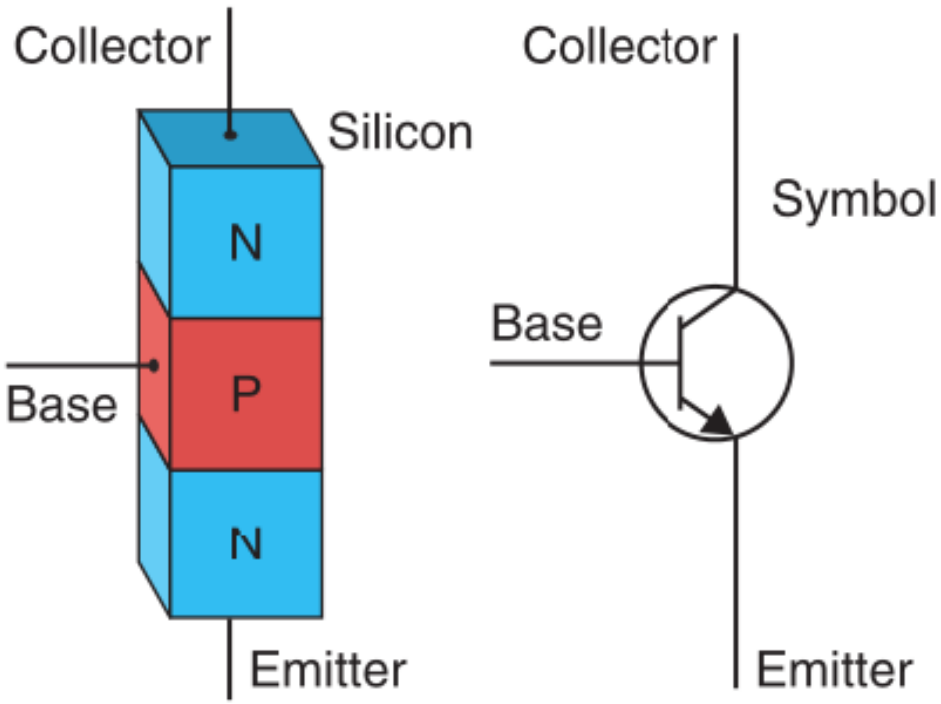


NPN bipolar junction transistor

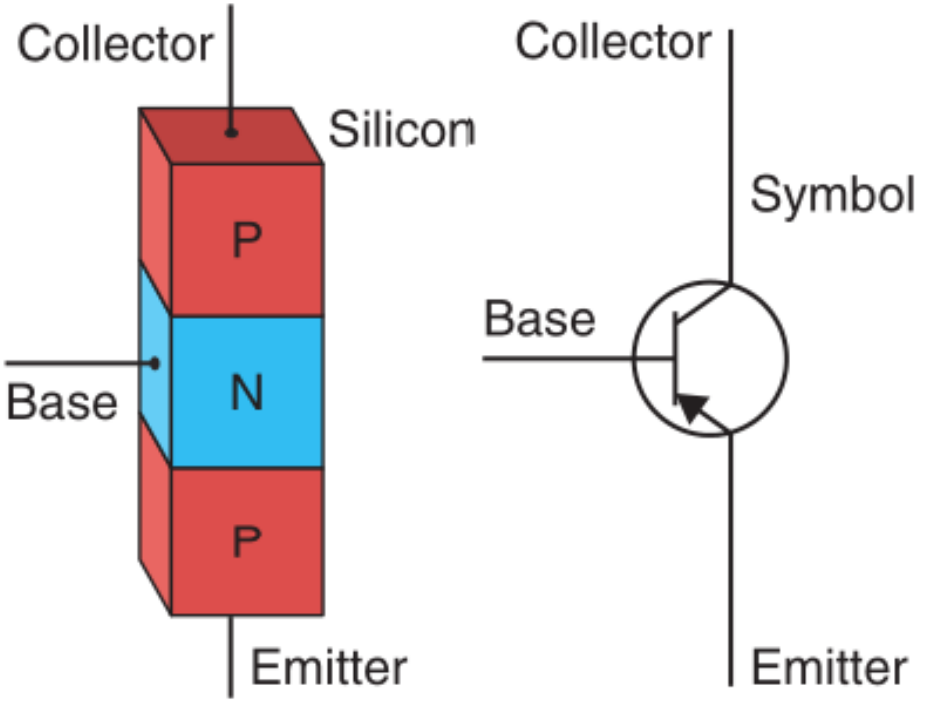


PNP bipolar junction transistor

We may think of the collector and emitter as data terminals, and the base as the control terminal.

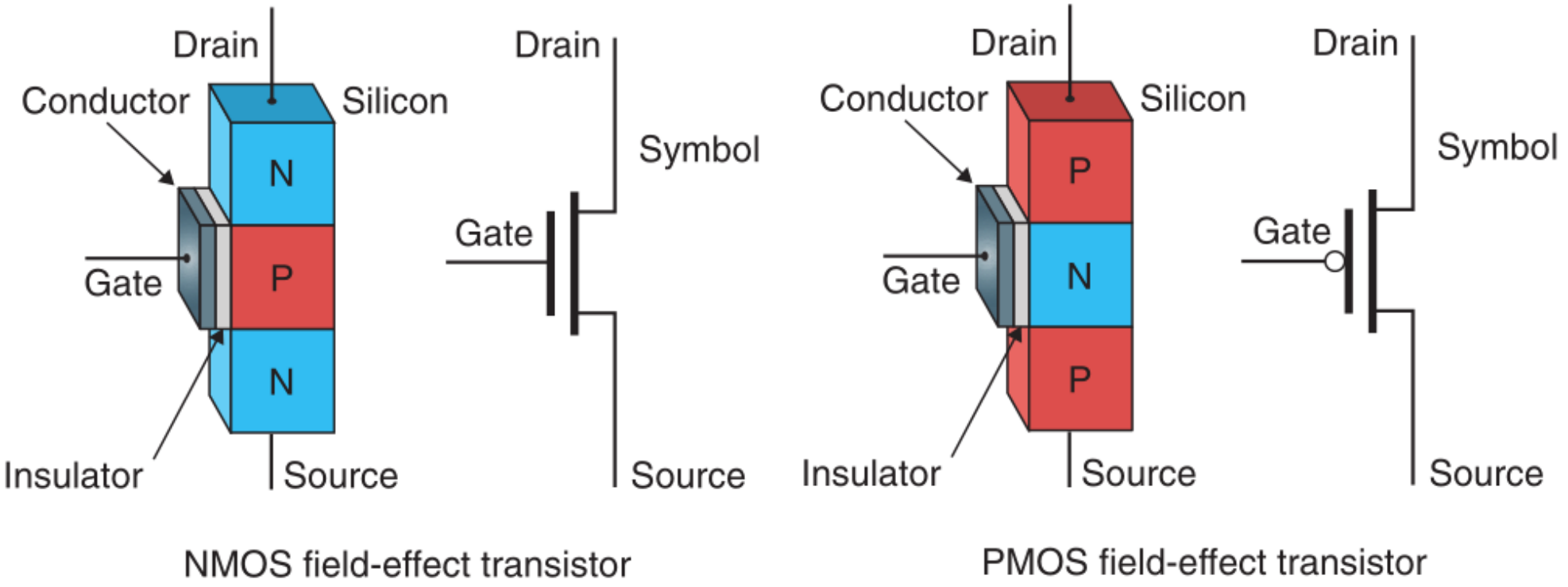


NPN bipolar junction transistor



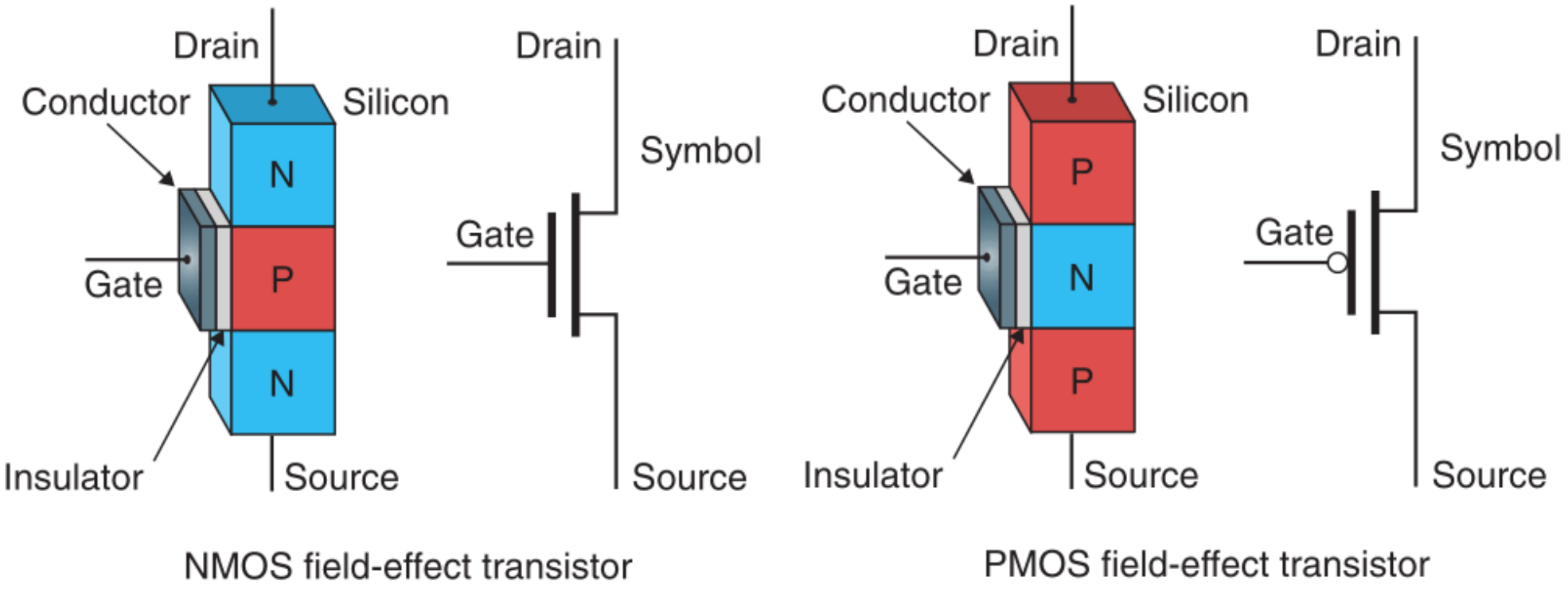
PNP bipolar junction transistor

Metal-oxide semiconductor field-effect transistors (MOSFET)



The **drain** and **source** form the data terminals and the **gate** acts as the control terminal.

Unlike bipolar devices, the control terminal is connected to a conducting plate, which is insulated from the silicon by a layer of nonconducting oxide.



NMOS field-effect transistor

PMOS field-effect transistor

When a signal is applied to the gate terminal, the plate, insulated by the oxide, creates an electromagnetic field, which turns the transistor ON or OFF – hence, the term field-effect.

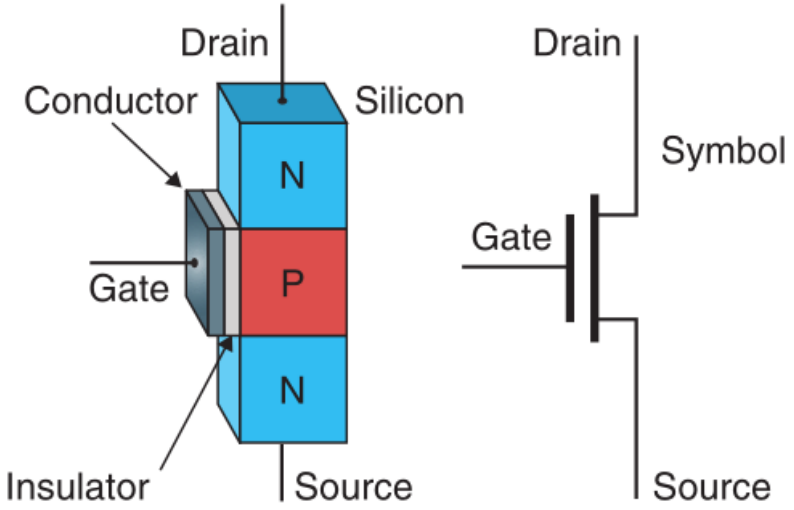
The NMOS gate input (with no circle) is active-high.

The PMOS gate input (with a circle) is active-low.

The term **channel** refers to the piece of silicon under the gate terminal; that is, the piece linking the drain and source regions.

The channel in the n-channel device is formed from P-type material, while the channel in the p-channel device is formed from N-type material.

Let's consider the n-channel device. In order to turn this ON, a positive voltage is applied to the gate.



NMOS field-effect transistor

This positive voltage attracts any negative electrons in the P-type material and causes them to accumulate beneath the oxide layer where they form a negative channel—hence, the term n-channel.

CMOS

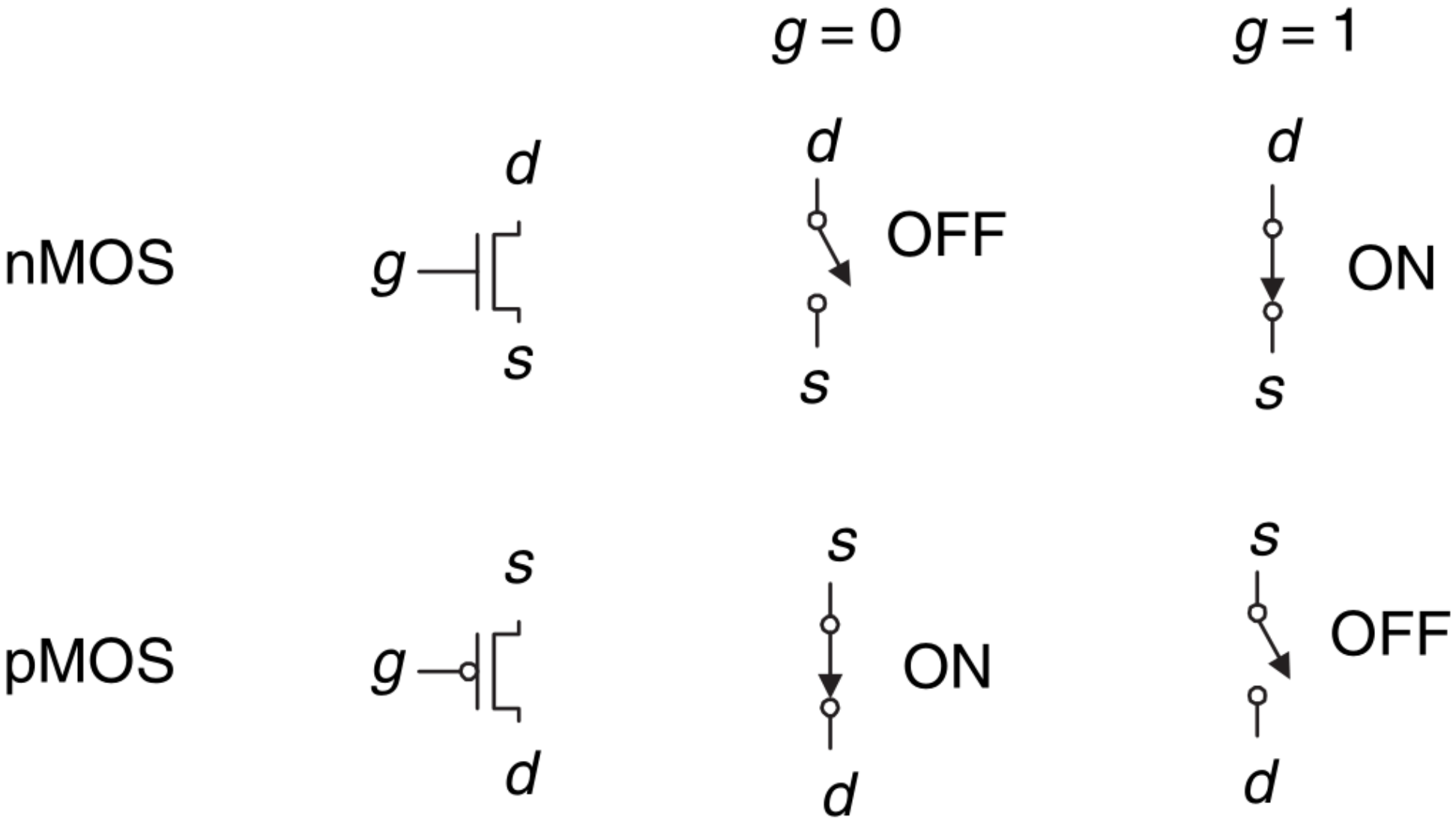
Logic gates can be created using only NMOS (n-channel) or only PMOS (p-channel) MOSFET transistors.

However, a popular implementation called **Complementary Metal-Oxide Semiconductor (CMOS)** makes use of both NMOS and PMOS transistors connected in a complementary manner.

There are other transistor technologies for creating a digital circuit.

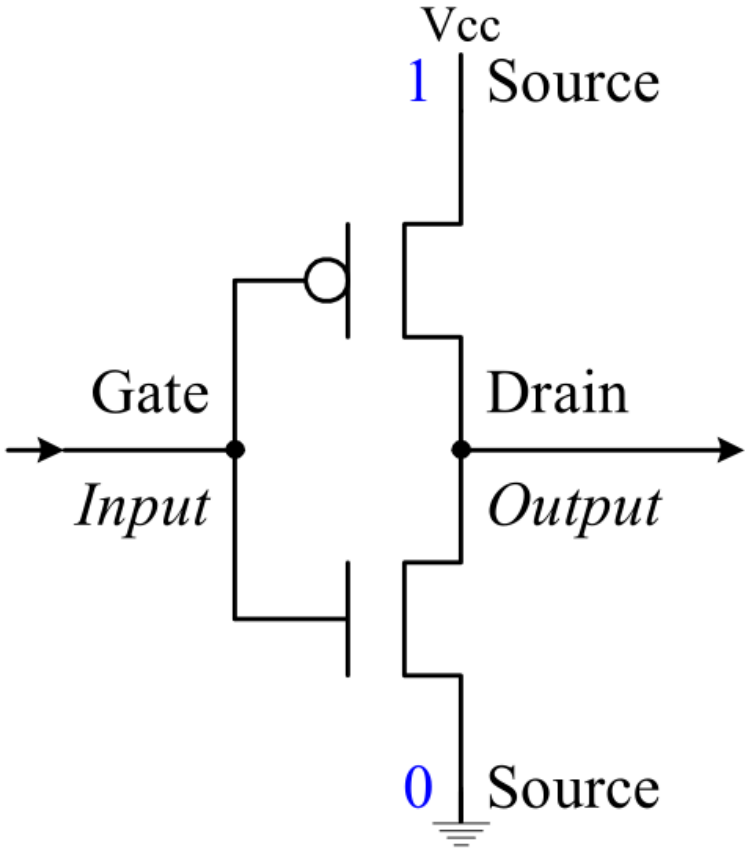
Some of these technologies are the diode-transistor logic (DTL), transistor-transistor logic (TTL), bipolar logic.

All of the low-voltage logic families use CMOS transistors



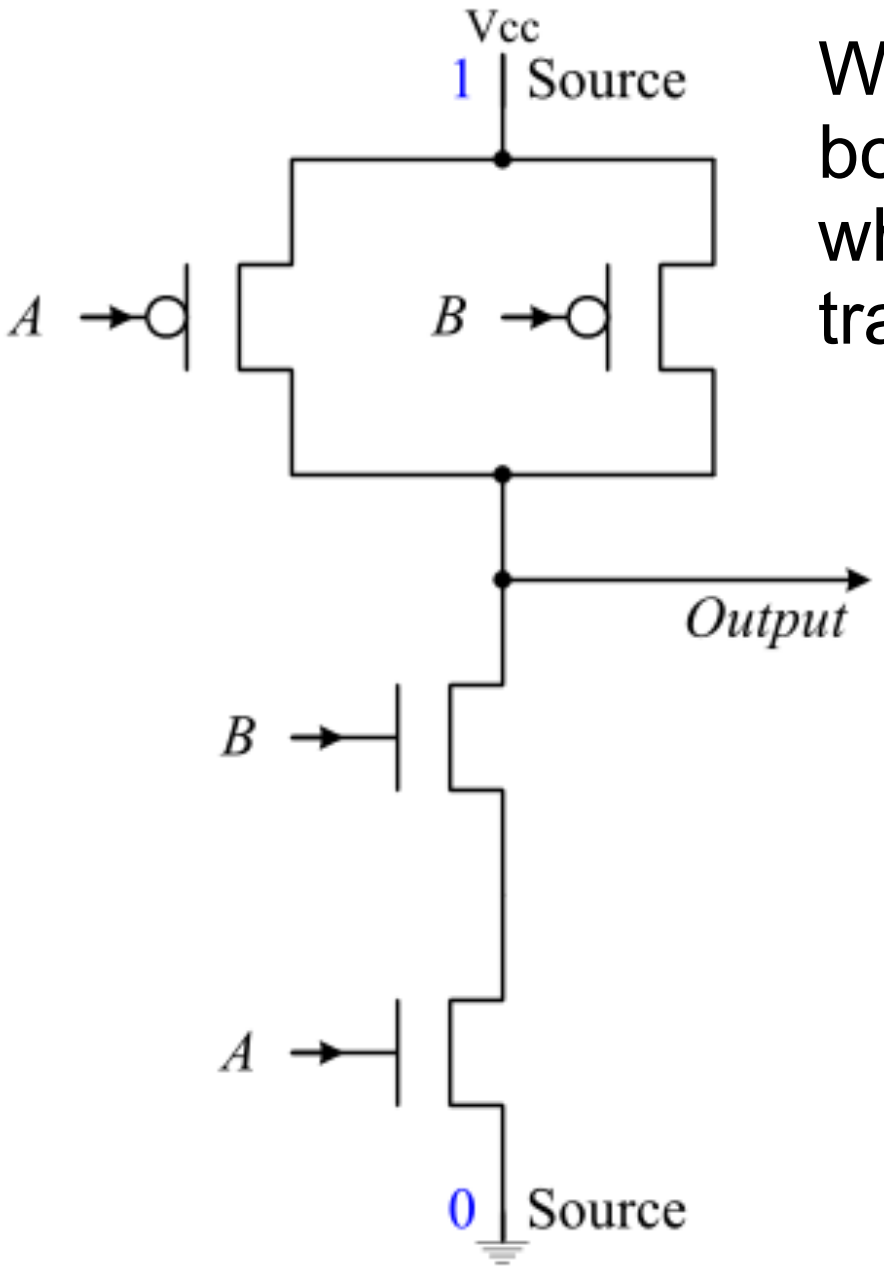
CMOS inverter

When the gate input is a 1, the bottom NMOS transistor is turned on while the top PMOS transistor is turned off. A 0 from ground will pass through the bottom NMOS transistor to the output while the top PMOS transistor will output a high-impedance Z value.



When the gate input is a 0, the bottom NMOS transistor is turned off while the top PMOS transistor is turned on. A 1 from V_{CC} will pass through the top PMOS transistor to the output while the bottom NMOS transistor will output a Z. The resulting output value is a 1.

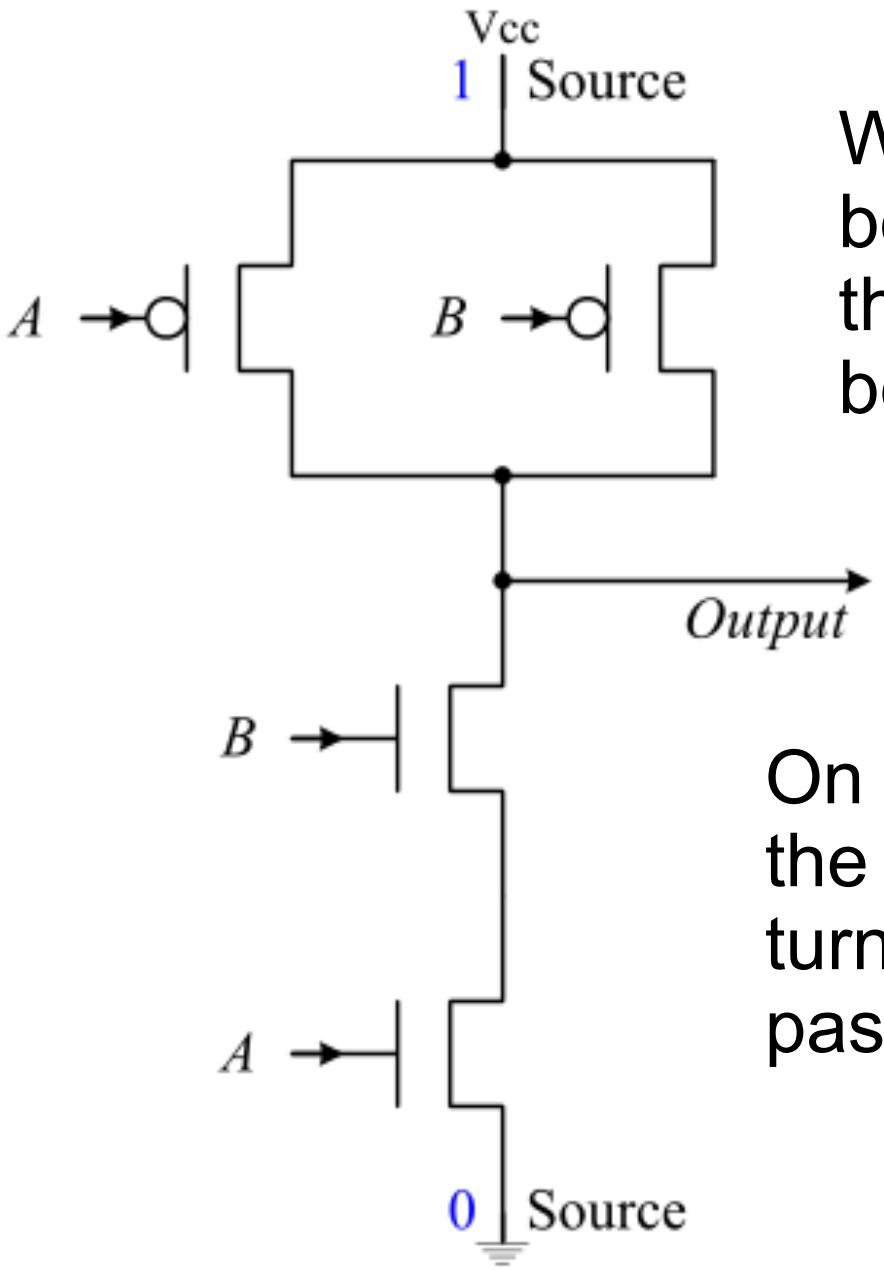
CMOS NAND circuit



When A = 1 and B = 1, the two bottom NMOS transistors are ON while the two top PMOS transistors are OFF.

A 0 from ground will pass through the two bottom NMOS transistors to the output, while the two top PMOS transistors will output a high-impedance Z value.

The resulting output value will be a 0.

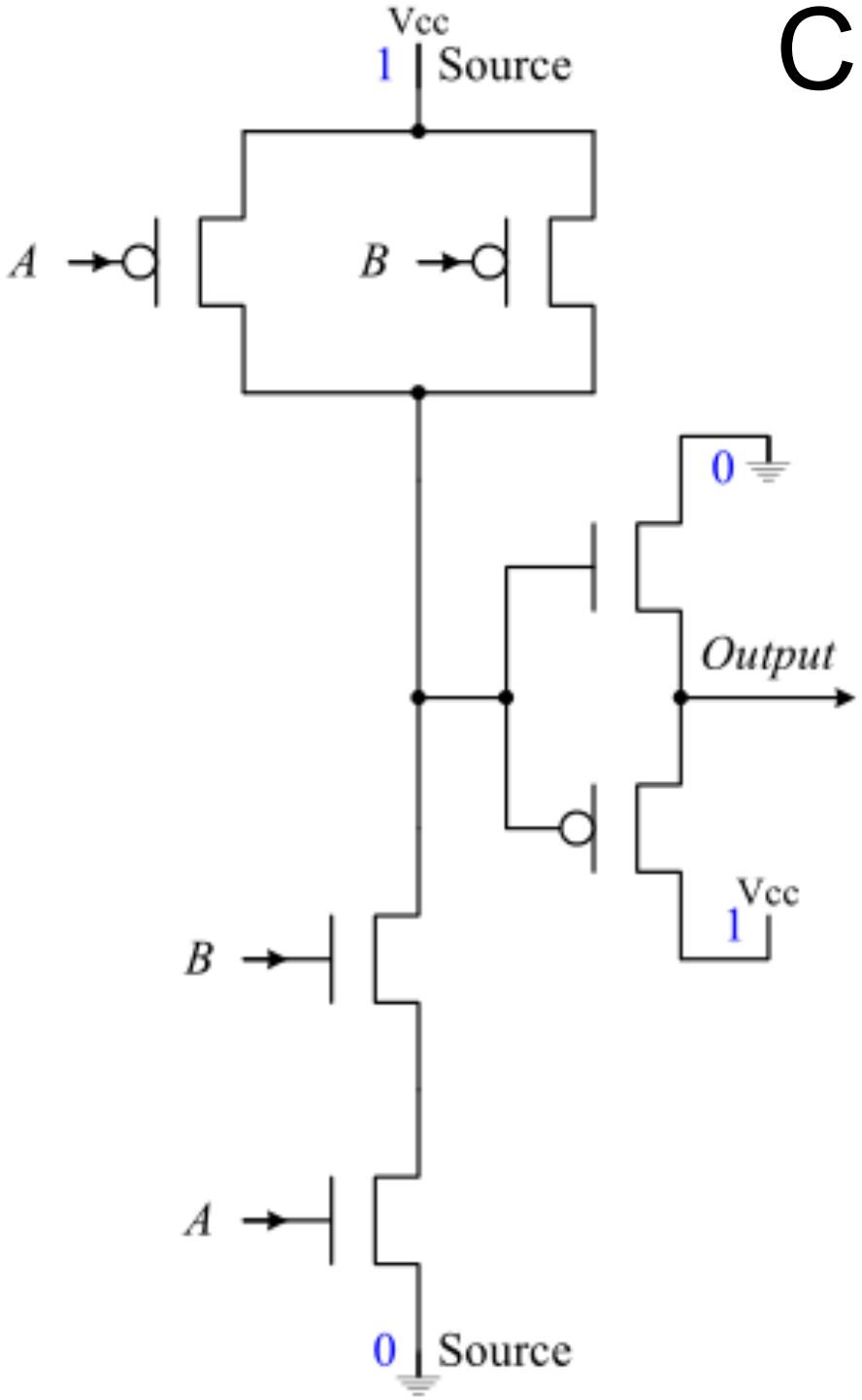


When either A = 0, or B = 0, or both equal to 0, at least one of the bottom NMOS transistors will be turned off, thus outputting a Z.

On the other hand, at least one of the top PMOS transistors will be turned on and a 1 from VCC will pass through that PMOS transistor.

The resulting output value will be a 1.

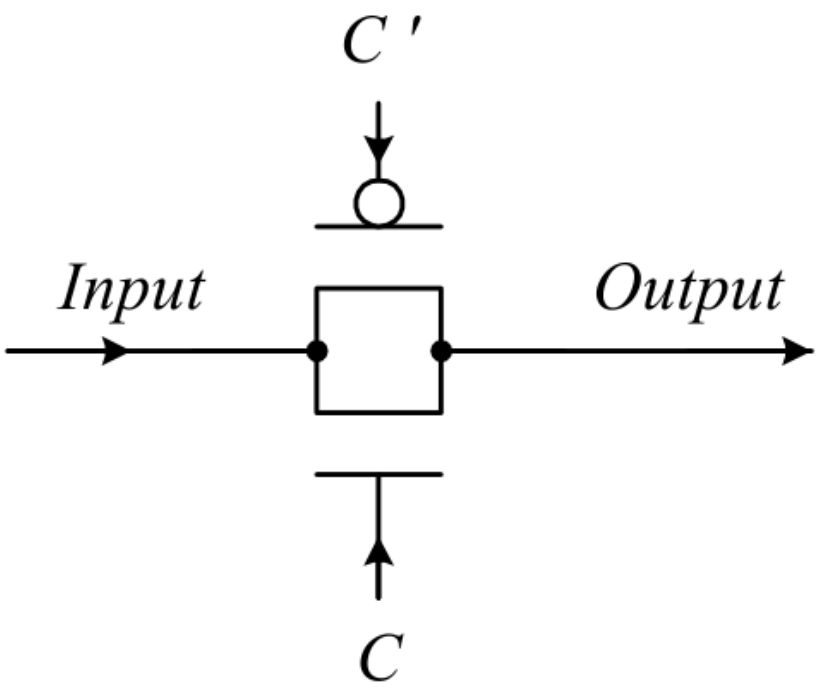
CMOS AND circuit



The AND gate circuit is the circuit for the NAND gate followed by that of the INVERTER.

Transmission gate

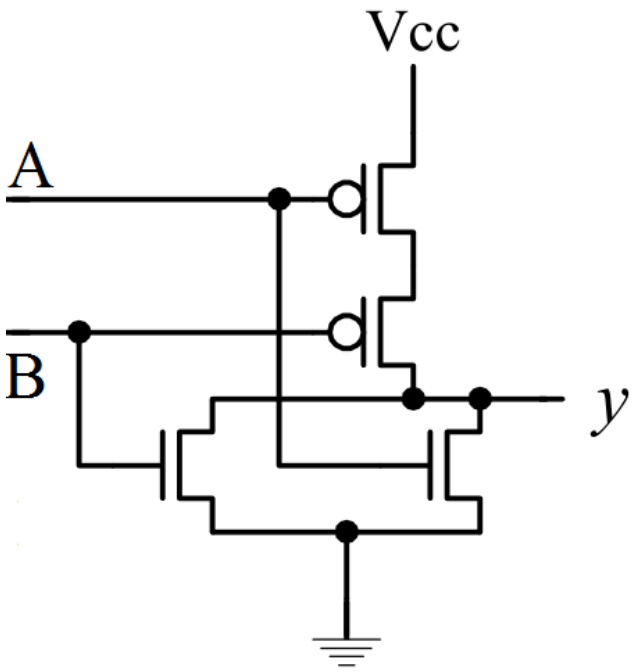
The transmission gate is such a circuit that allows both a 0 and a 1 to pass through when it is enabled. When it is disabled, it outputs the Z value.



The top PMOS transistor gate is connected to the inverted control signal C', while the bottom NMOS transistor gate is connected directly to the control signal C. Hence, both transistors are enabled when the control signal C = 1, and the circuit is disabled when C = 0.

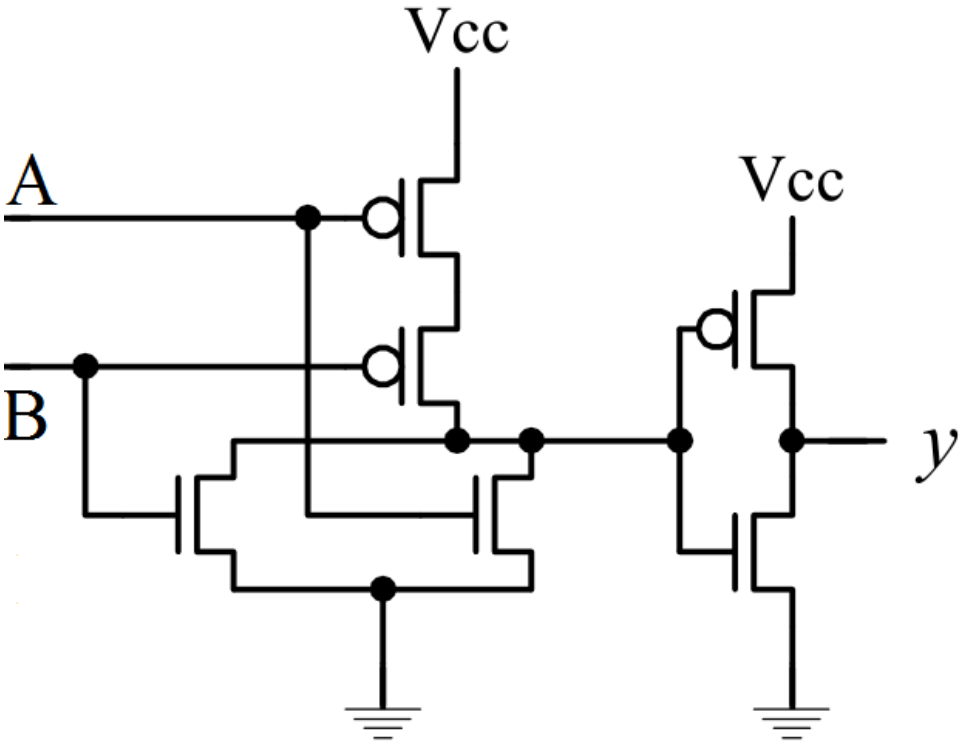
When the circuit is disabled with C = 0, both transistors will output the Z value. Thus, regardless of the input, there will be no output.

CMOS NOR and OR circuits



NOR

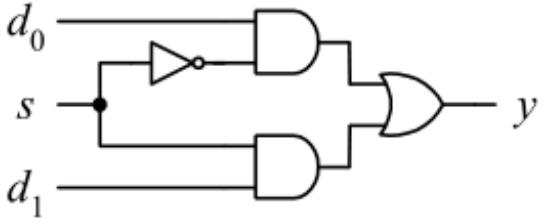
The OR gate circuit is the circuit for the NOR gate followed by that of the INVERTER.

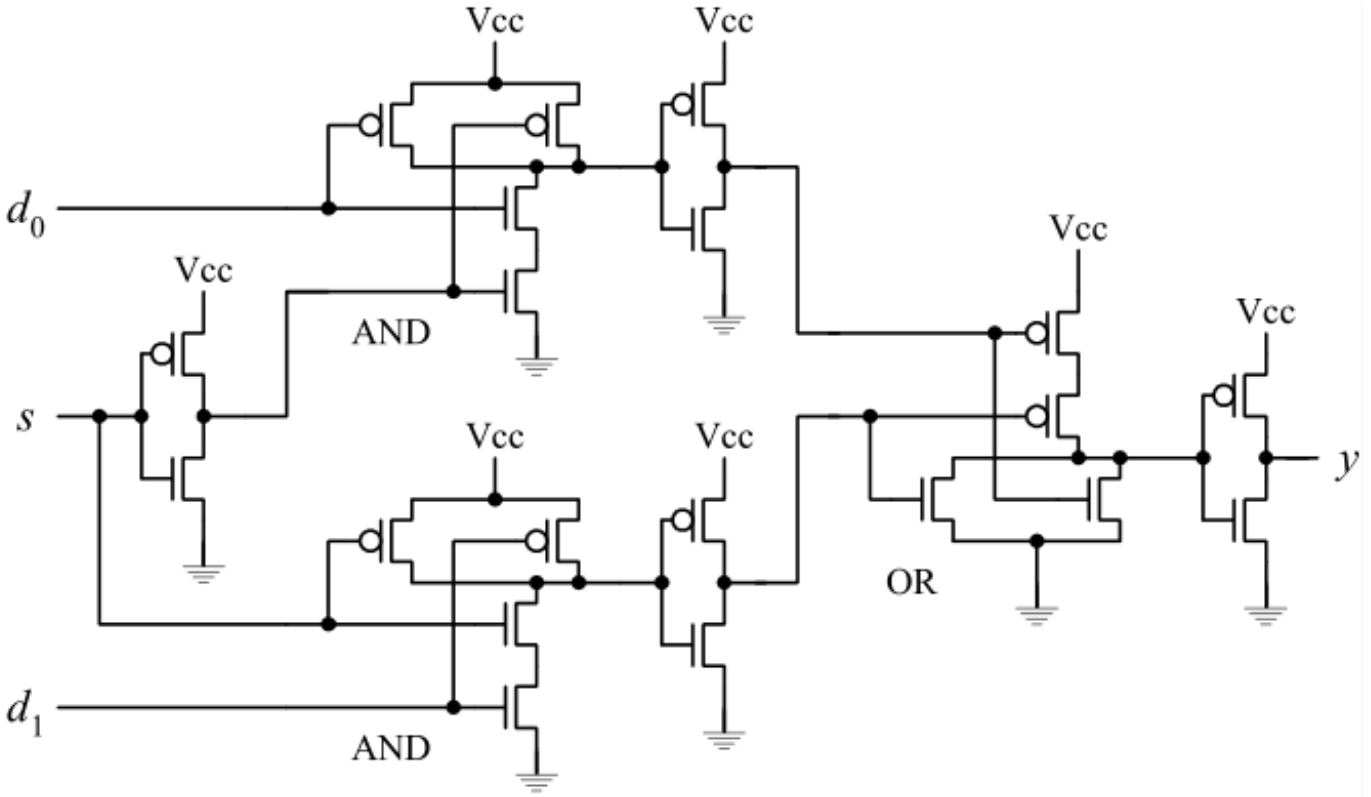


OR

2-input multiplexer CMOS Circuit

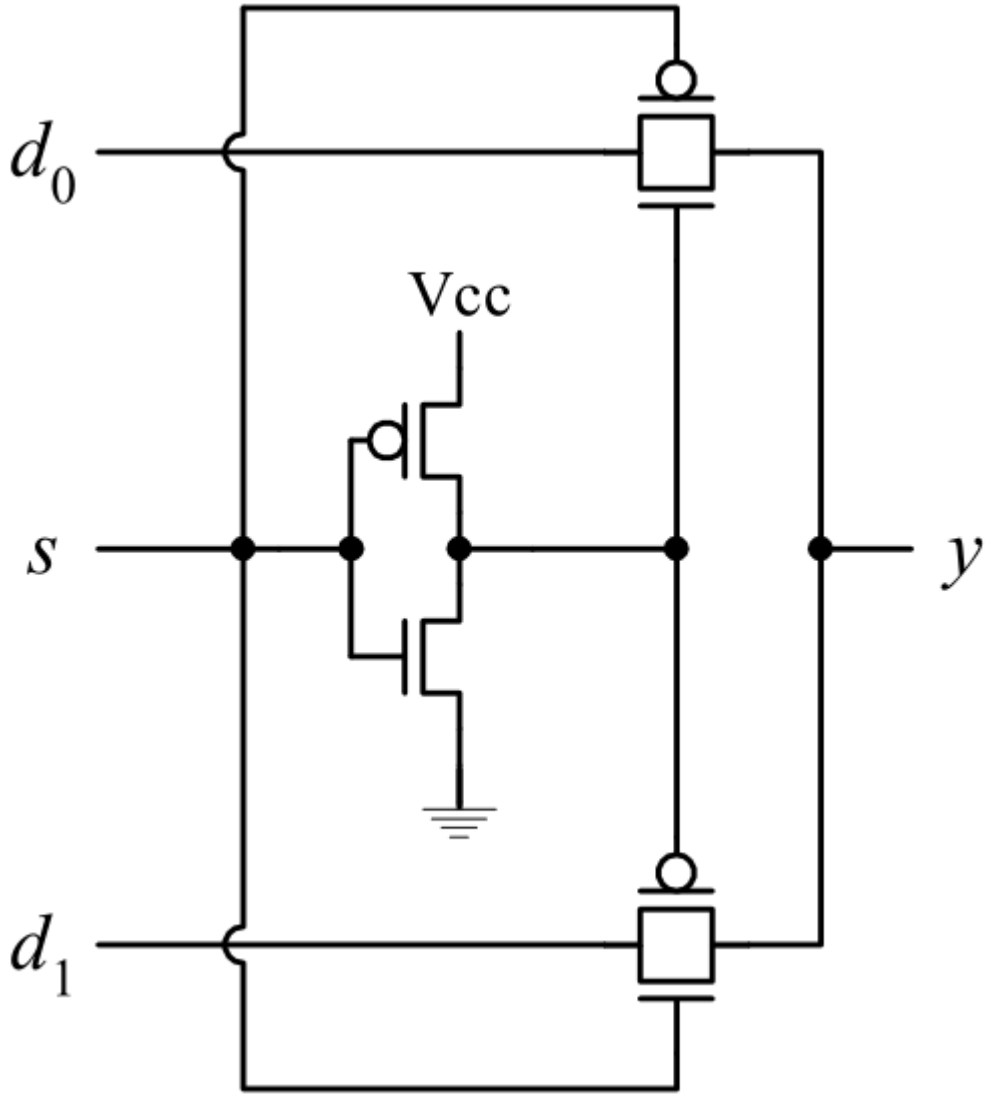
CMOS circuits for larger components can be derived by replacing each gate in the circuit with the corresponding CMOS circuit for that gate. Since we know the CMOS circuit for the three basic gates (AND, OR, and NOT) this is a simple “copy and paste” operation.



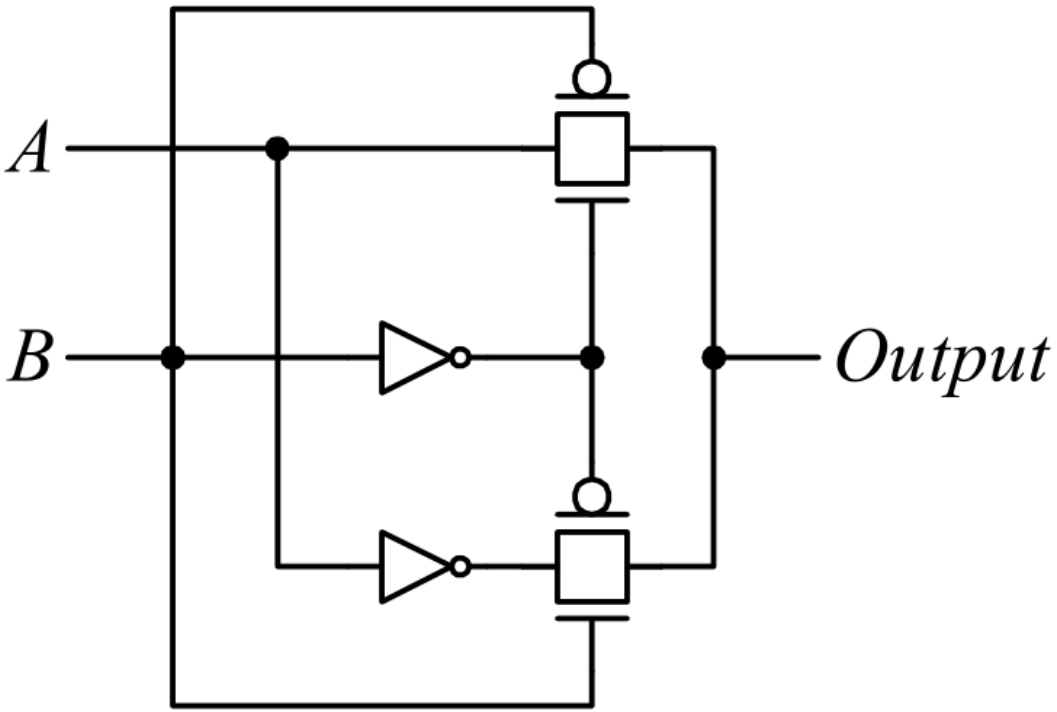


2-input multiplexer CMOS Circuit

Transistor level circuit using transmission gates:



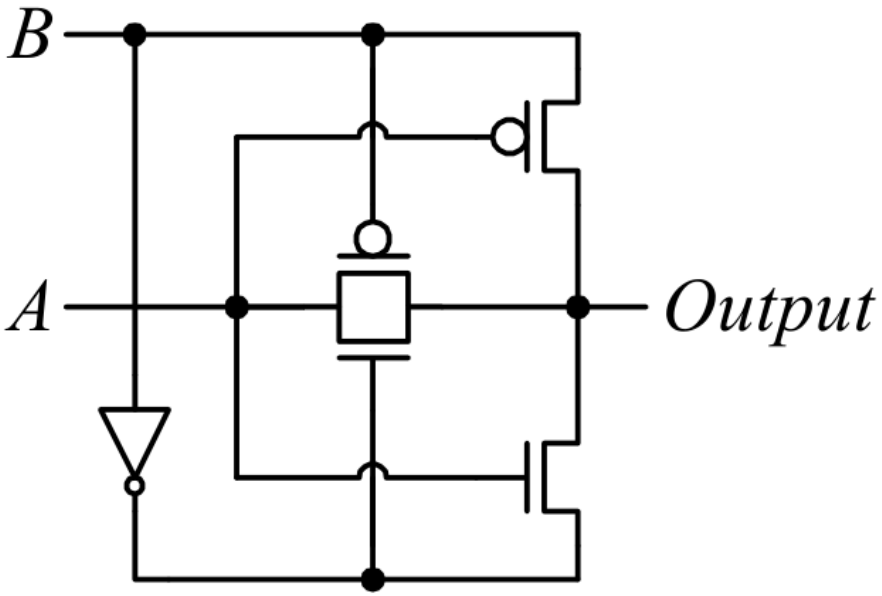
CMOS XOR Gate



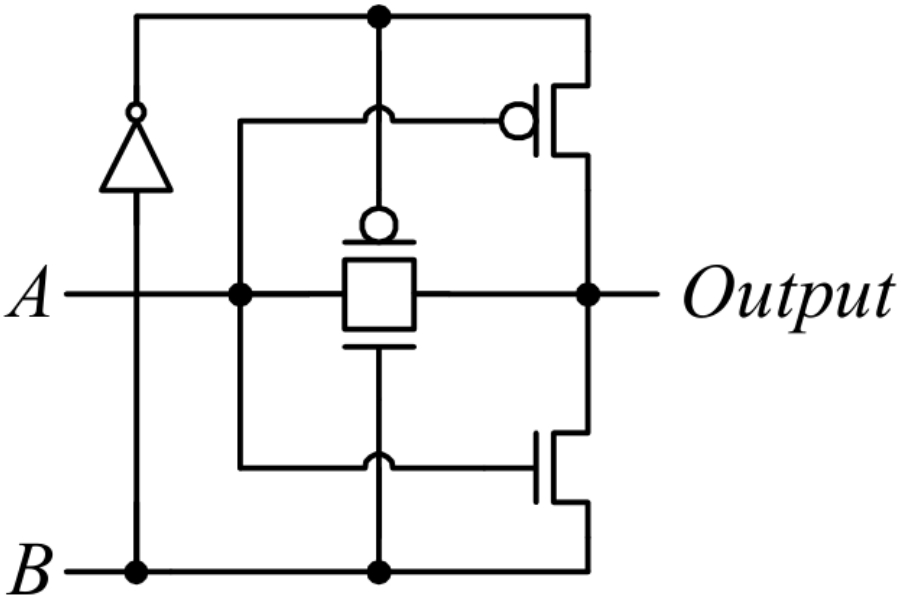
XOR

Question: How many transistors does this gate use?

Better CMOS XOR nad XNOR gates



XOR



XNOR

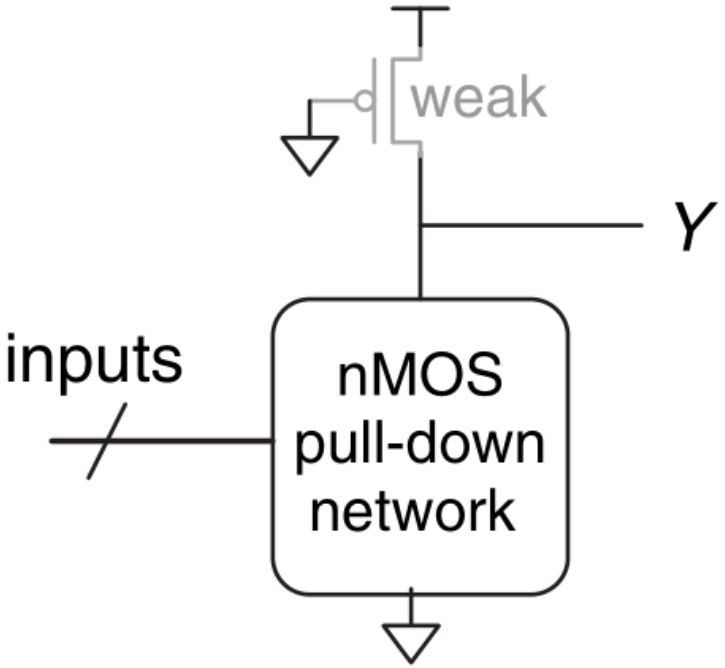
Question: How many transistors do these gate use?

Transistors in series are slower than transistors in parallel, just as resistors in series have more resistance than resistors in parallel.

pMOS transistors are slower than nMOS transistors because holes cannot move around the silicon lattice as fast as electrons.

Therefore the parallel nMOS transistors are fast and the series pMOS transistors are slow, especially when many are in series.

Pseudo-nMOS Logic



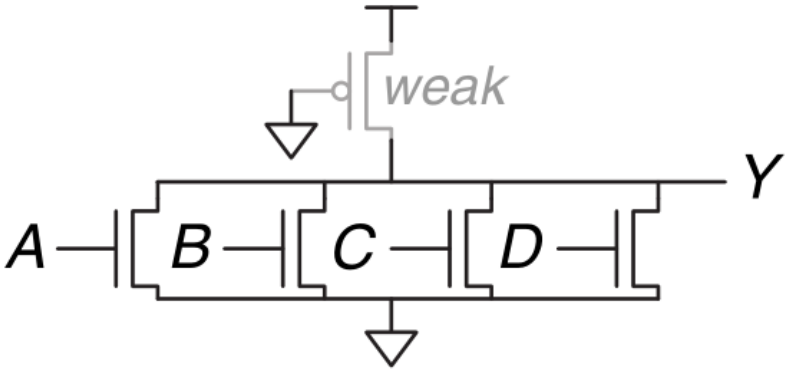
Pseudo-nMOS logic replaces the slow stack of pMOS transistors with a single weak pMOS transistor that is always ON

This pMOS transistor is called a **weak pull-up**.

The physical dimensions of the pMOS transistor are selected so that the pMOS transistor will pull the output Y HIGH weakly—that is, only if none of the nMOS transistors are ON.

If any nMOS transistor is ON, it overpowers the weak pull-up and pulls Y down close enough to GND to produce a logic 0.

Pseudo-nMOS four-input NOR gate



The advantage of pseudo-nMOS logic is that it can be used to build fast NOR gates with many inputs.

Pseudo-nMOS gates are useful for certain memory and logic arrays.

The disadvantage is that a short circuit exists between V_{DD} and GND when the output is LOW; the weak pMOS and nMOS transistors are both ON.

The short circuit draws continuous power, so pseudo-nMOS logic must be used sparingly.

Power consumption

Power consumption is the amount of energy used per unit time.

Dynamic power is the power used to charge capacitance as signals change between 0 and 1.

Static power is the power used even when signals do not change and the system is idle.

Logic gates and the wires that connect them have **capacitance**.

The energy drawn from the power supply to charge a capacitance C to voltage V_{DD} is

$$CV_{DD}^2$$

If the voltage on the capacitor switches at frequency f it charges the capacitor $\frac{f}{2}$ times and discharges it $\frac{f}{2}$ times per second.

Discharging does not draw energy from the power supply, so the **dynamic power consumption** is

$$P_{dynamic} = \frac{1}{2} C V_{DD}^2 f$$

Electrical systems draw some current even when they are idle.

When transistors are OFF, they leak a small amount of current.

Some circuits, such as the pseudo-nMOS gate have a path from V_{DD} to GND through which current flows continuously.

The total static current, I_{DD} is also called the leakage current or the quiescent supply current flowing between V_{DD} and GND.

The **static power consumption** is proportional to this static current:

$$P_{static} = V_{DD} I_{DD}$$

Memory types

Random access memory (RAM) is volatile.

It loses its when the power is turned off.

Read only memory (ROM) is nonvolatile.

It retains its data, even without a power source.

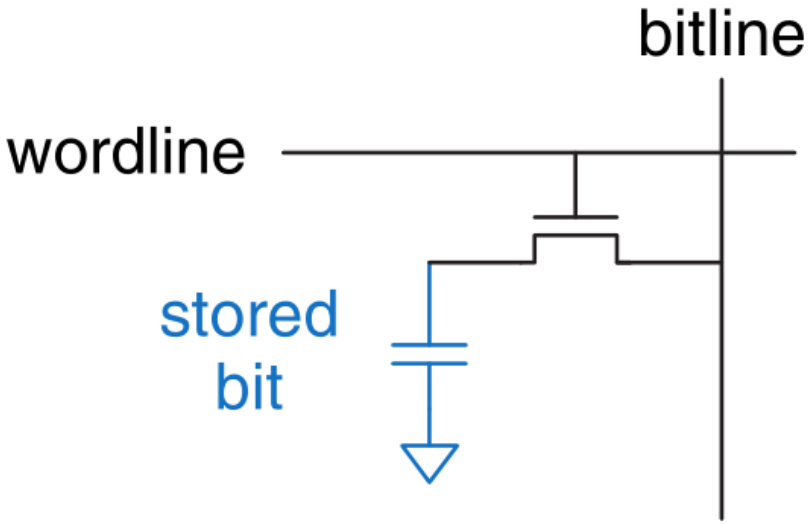
RAM and ROM received their names for historical reasons that are no longer very meaningful.

ROMs are randomly accessed too.

Most modern ROMs can be written as well as read.

The important distinction to remember is that RAMs are volatile and ROMs are nonvolatile.

Dynamic Random Access Memory (DRAM)



DRAM stores a bit as the presence or absence of charge on a capacitor.

The nMOS transistor either connects or disconnects the capacitor from the bitline.

When the wordline is asserted, the nMOS transistor turns ON, and the stored bit value transfers to or from the bitline.

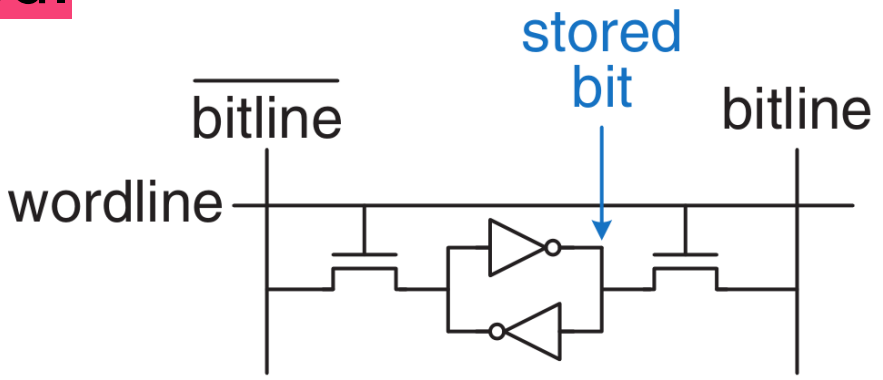
The capacitor node is dynamic because it is not actively driven HIGH or LOW by a transistor tied to V_{DD} or GND.

Reading destroys the bit value stored on the capacitor, so the data word must be restored (rewritten) after each read.

Even when DRAM is not read, the contents must be refreshed (read and rewritten) every few milliseconds, because the charge on the capacitor gradually leaks away.

Static Random Access Memory (SRAM)

SRAM is static because stored bits do not need to be refreshed.



The data bit is stored on cross-coupled inverters.

When the wordline is asserted, both nMOS transistors turn on, and data values are transferred to or from the bitlines.

Unlike DRAM, if noise degrades the value of the stored bit, the cross-coupled inverters restore the value.

Memory Type	Transistors per Bit Cell	Latency
flip-flop	~20	fast
SRAM	6	medium
DRAM	1	slow

The data bit stored in a flip-flop is available immediately at its output.

DRAM latency is longer than that of SRAM because its bitline is not actively driven by a transistor.

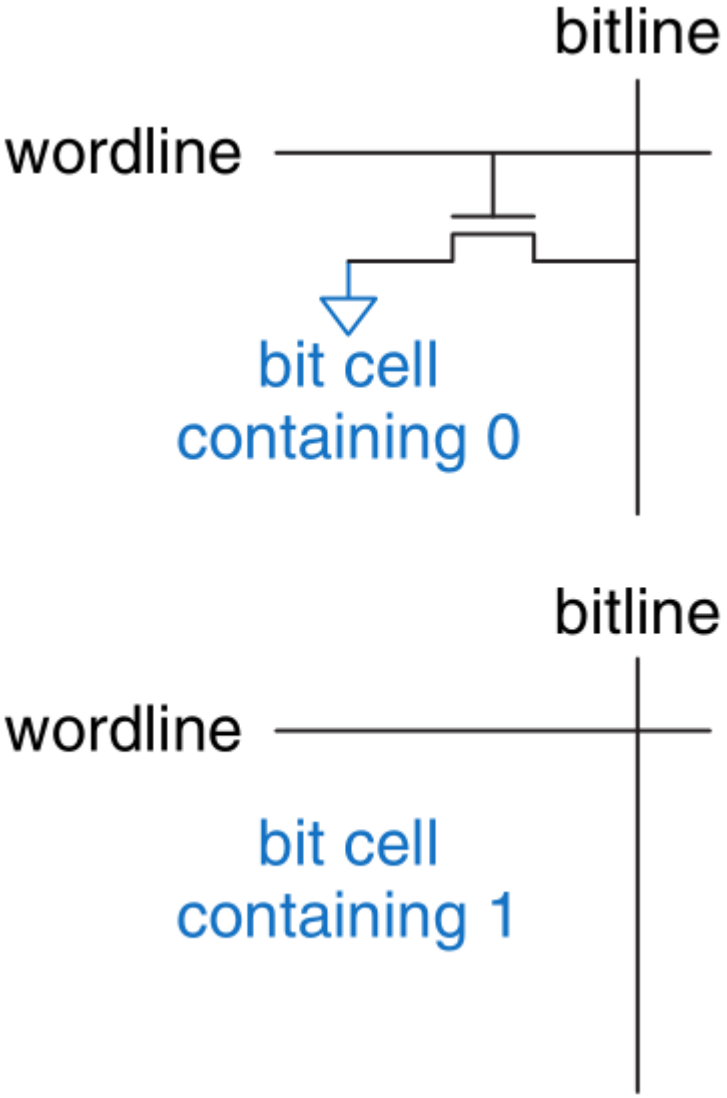
DRAM must wait for charge to move (relatively) slowly from the capacitor to the bitline.

DRAM also has lower throughput than SRAM, because it must refresh data periodically and after a read.

Synchronous DRAM (SDRAM) uses a clock to pipeline memory accesses.

Double data rate (DDR) SDRAM, called simply DDR, uses both the rising and falling edges of the clock to access data, thus doubling the throughput for a given clock speed.

Read only memory (ROM)



ROM stores a bit as the presence or absence of a transistor.

To read the cell, the bitline is weakly pulled HIGH.

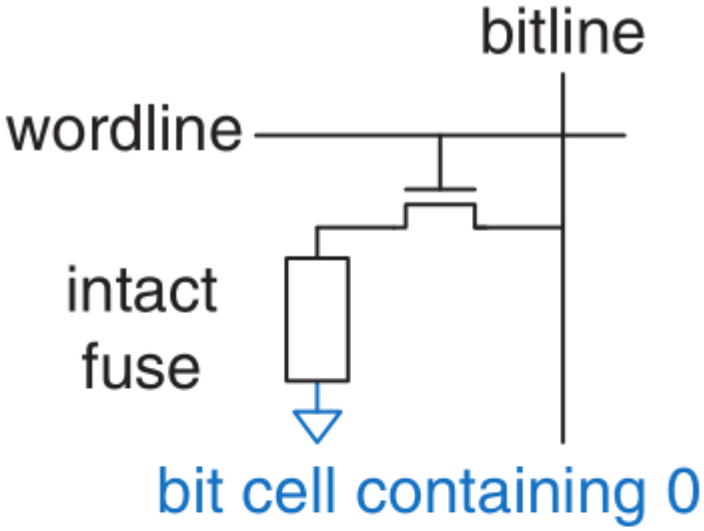
Then the wordline is turned ON.

If the transistor is present, it pulls the bitline LOW.

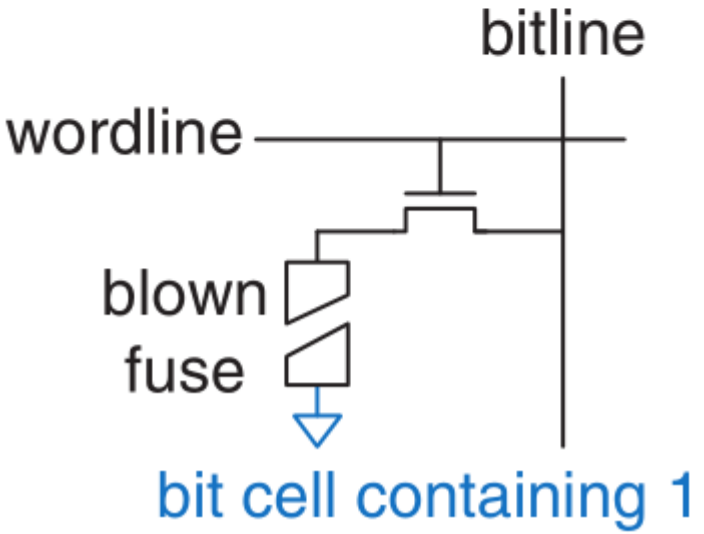
If it is absent, the bitline remains HIGH.

The ROM bit cell is a combinational circuit and has no state to “forget” if power is turned off.

The contents of the ROM bit cell are specified during manufacturing by the presence or absence of a transistor in each bit cell.



A programmable ROM (PROM) places a transistor in every bit cell but provides a way to connect or disconnect the transistor to ground.



Reprogrammable ROMs provide a reversible mechanism for connecting or disconnecting the transistor to GND.

Examples: EEPROM, Flash.

Erasable PROMs (EPROM) replace the nMOS transistor and fuse with a floating-gate transistor.

Examples:

Electrically erasable PROMs (EEPROM), Flash

EEPROM bit cells are individually erasable.

Flash memory erases larger blocks of bits and is cheaper because fewer erasing circuits are needed.