Unit 22

Semiconductor Devices

This unit online, with its interactives:
science-playground.com/s/12f22

Look at the charger of your phone. Inside that small white box, 230 V of alternating current from the wall becomes a steady 5 V that always pushes the same way, and the phone counts the charge going in, switches the screen's pixels and plays music, all with the same kind of part: a sandwich of two slightly different kinds of silicon. One sandwich lets current through one way only. Billions of them, switching on and off, can do arithmetic.

This unit is about the simplest of these devices. A **p–n junction diode** passes current one way and blocks it the other; a **Zener diode** holds a voltage steady; a few diodes together turn a.c. into d.c.; and diodes and transistors make **logic gates**, the switches that decide 1 or 0 in every computer.



Figure 22.1 *John Bardeen, William Shockley (seated) and Walter Brattain at Bell Laboratories in 1948, a few months after the first transistor worked.*

Photo: AT&T; photographer: Jack St. (last part of name not stamped well enough to read), New York, New York., Public domain

A STORY FROM SCIENCE

Shockley, the transistor and Silicon Valley

William Bradford Shockley (1910–1989), an American physicist, led the team at Bell Laboratories in New Jersey that set out after the Second World War to replace the hot, fragile vacuum tubes in telephone exchanges with something solid. In December 1947 **John Bardeen** and **Walter Brattain** made the first working **transistor**, a slab of germanium with two gold contacts pressed onto it; a few weeks later Shockley worked out a better design, the junction transistor, built from p–n junctions like the ones in this unit. The three shared the **Nobel Prize in Physics in 1956** (the full story).

In 1956 Shockley went to California to start his own company making transistors from silicon, near Stanford University. He was a brilliant physicist but a suspicious, overbearing manager, and within a year eight of his best young engineers walked out together. They started Fairchild Semiconductor, and people who left Fairchild in turn started Intel and dozens of other companies. The valley of orchards south of San Francisco became **Silicon Valley**. Shockley spent his later years promoting discredited ideas about race and intelligence, and died estranged from most of his colleagues.

22.1 Formation of p–n junction

Have you ever wondered what happens when two kinds of silicon meet?

A crystal of silicon can be **doped** to make it one of two kinds (Grade 11, Unit 25). In **p-type** silicon (doped with a trivalent element such as boron), the current is carried mostly by **holes**, missing electrons that behave like positive charges: holes are its **majority carriers**. In **n-type** silicon (doped with a pentavalent element such as phosphorus), the majority carriers are free **electrons**. Both kinds are electrically neutral: each free electron has left behind a positive donor ion, each hole a negative acceptor ion.

When a p-type and an n-type region are made in one single crystal, side by side, the boundary between them is a **p–n junction**. The whole device is called a **junction diode**, or simply a junction. (It can't be made by pressing two separate pieces together: the surfaces would never touch closely enough. The two regions are made by doping different parts of one crystal.)

Diffusion. Just after the junction forms, there are very many holes on the p side and very few on the n side, and very many free electrons on the n side and very few on the

p side. Like ink spreading in water, carriers spread from where there are many to where there are few: holes **diffuse** from the p side into the n side, and electrons from the n side into the p side. When an electron meets a hole, it falls into it and the two vanish: they **recombine**.

The depletion layer. So a thin region on each side of the junction is left with almost no free carriers. It is called the **depletion layer** (it has been depleted, emptied, of carriers). It is very thin, between about 0.01 μm and 1 μm, the thinner the more heavily the silicon is doped.

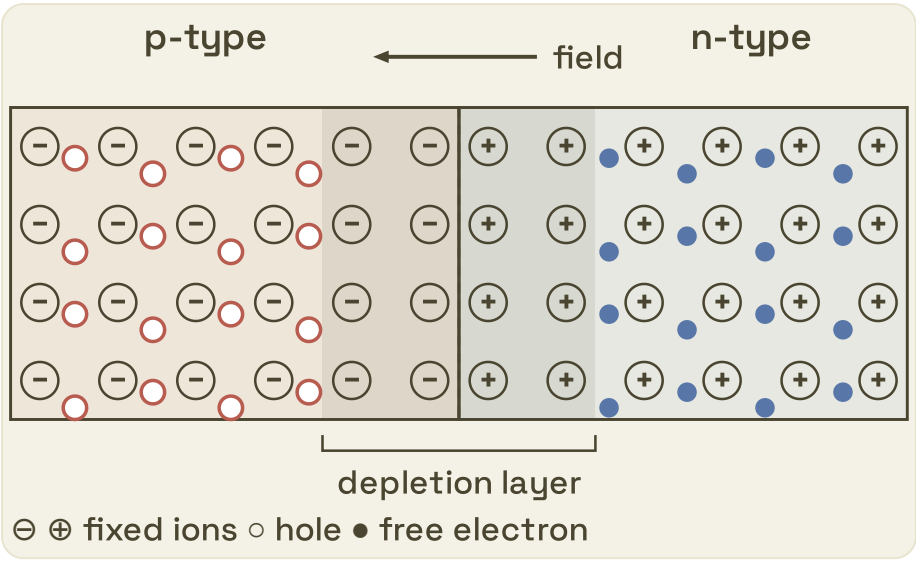


Figure 22.2 *A p–n junction. Electrons from the n side and holes from the p side have crossed and recombined, leaving the depletion layer: the fixed negative acceptor ions on the p side and positive donor ions on the n side, with no free carriers between them. Their electric field, from the n side to the p side, makes the potential barrier.*

The potential barrier. The depletion layer isn't neutral. The n side of it has lost electrons and is left with its fixed positive donor ions; the p side has lost holes and is left with its fixed negative acceptor ions. So the p side of the junction becomes slightly **negative** and the n side slightly **positive**. These charges make an **electric field** across the depletion layer, pointing from the n side to the p side. It pushes electrons back towards the n side and holes back towards the p side, so it opposes any more diffusion. Soon the two balance and diffusion stops.

The potential difference across the depletion layer that this field makes is the **potential barrier** (or barrier potential). It works like a small built-in voltage that a carrier must climb to cross. It is about **0.15–0.3 V for germanium** and about **0.7 V for sili-**

con, the exact value depending on how heavily the semiconductor is doped (and on its temperature).

CAREFUL

The barrier is not a cell

You can't measure the barrier potential by putting a voltmeter across a diode, and you can't get a current from it. Where the meter's wires touch the p and n silicon, new junctions form whose voltages cancel the barrier exactly. A diode lying on the bench is in balance: no current, no energy to give.

Check 1

In the depletion layer of a p–n junction:

a there are extra free electrons on the p side

b there are fixed negative ions on the p side and fixed positive ions on the n side, but almost no free carriers

c there are fixed positive ions on the p side and fixed negative ions on the n side

d holes and electrons are both present in large numbers

Check 2

The electric field across the depletion layer points:

a from the n side to the p side, so it opposes further diffusion of holes and electrons

b from the p side to the n side, so it helps diffusion

c along the junction, parallel to it

d nowhere: the depletion layer is neutral

22.2 Biasing of p–n junction

What happens when a battery is joined across the junction?

Applying a potential difference across a p–n junction to make it work is called **biasing** it. There are two ways round:

i. forward biasing

ii. reverse biasing

1. Forward bias

When the **p side is connected to the positive** terminal of a battery and the **n side to the negative** terminal, the junction is **forward biased**.

The battery's positive terminal pushes the holes in the p side towards the junction, and its negative terminal pushes the electrons in the n side towards it. The battery's voltage works against the barrier: the **depletion layer becomes thinner** and the barrier lower. Once the applied voltage is more than the barrier potential (0.7 V for silicon), the majority carriers, holes from the p side and electrons from the n side, cross the junction in large numbers. The current they carry is the **forward current**, and it grows quickly as the forward voltage is raised. A forward-biased diode has a **low resistance**.

2. Reverse bias

When the **p side is connected to the negative** terminal and the **n side to the positive** terminal, the junction is **reverse biased**.

Now the battery pulls the holes away from the junction towards its negative terminal, and the electrons away towards its positive terminal. The **depletion layer becomes wider** and the barrier higher, adding to the battery's voltage. The majority carriers can't cross, and the diode's **resistance is very high**. A very small current still flows, carried by the few **minority carriers** (the few electrons on the p side and holes on the n side, made by heat), which the barrier's field sweeps across. It is the **reverse current**, typically a few microamperes or less.

Notice that the barrier **opposes the forward current** (the majority carriers must climb it) but **helps the reverse current** (its field carries the minority carriers across).

The symbol of a diode

A junction diode is drawn as a **triangle against a bar** (Figure 22.3). The triangle is the p side (the **anode**) and the bar the n side (the **cathode**). The triangle points the way **conventional current** (and holes) flows when the diode is forward biased; the electrons flow the other way. On a real diode, a band painted round one end marks the cathode.

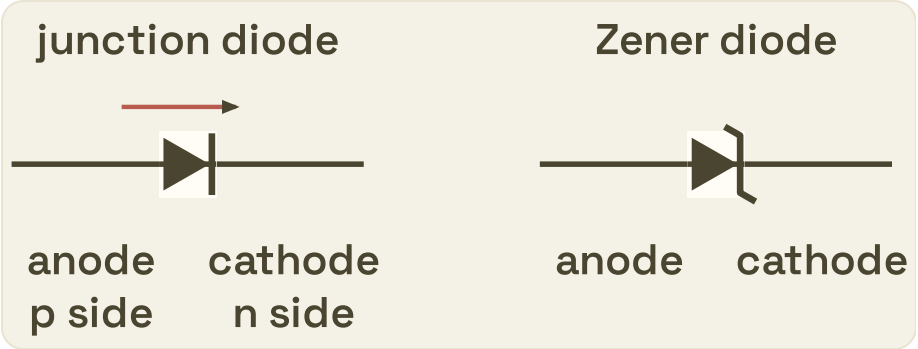


Figure 22.3 *The symbols of a junction diode (left) and a Zener diode (right). The triangle is the p side, the anode; the bar the n side, the cathode, marked on a real diode by a painted band. Forward biased, conventional current flows the way the triangle points. A Zener diode's bar has bent ends, like a flattened Z.*

Check 3

Which way round must a silicon diode be connected for it to conduct?

a cathode (the banded end) to the positive terminal

b anode (the p side) to the positive terminal, with more than about 0.7 V across it

c either way, as long as the voltage is above 0.7 V

d anode to the positive terminal, with any voltage at all

22.3 Characteristics of a junction diode

How does the current through a diode change as the voltage across it is raised?

The graph of the **current through a junction diode against the potential difference across it** is called its **characteristic** (or I–V characteristic). It is plotted in two parts, forward and reverse, because the currents are so different: milliamperes forward, microamperes reverse.

Forward characteristics

The **forward characteristic** is the graph of the current against the forward bias voltage. The circuit is in Figure 22.4: a battery feeds a rheostat used as a potential divider, so the voltage across the diode can be raised from zero a little at a time; a voltmeter reads the voltage across the diode, and a **milliammeter** in series the current.

As the forward voltage is raised from zero, almost no current flows at first, because the carriers can't yet get over the potential barrier. When the voltage reaches about the barrier, the **knee voltage** V_k (about **0.7 V for silicon**, **0.3 V for germanium** at

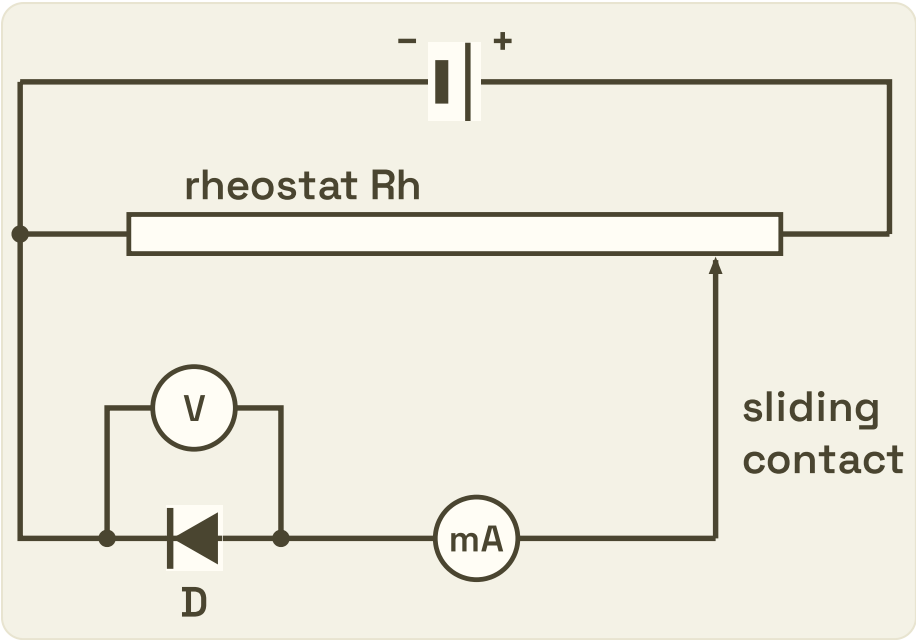


Figure 22.4 *Measuring a diode's forward characteristic. The rheostat Rh, connected across the battery as a potential divider, sets the voltage V across the diode D, read by the voltmeter; the milliammeter reads the current. The p side (anode) of the diode is towards the positive terminal.*

room temperature), the current begins to rise, and beyond the knee it rises **very steeply** (in fact exponentially: each extra 0.06 V multiplies it by about ten), so that a few tenths of a volt more give hundreds of milliamperes. In this part the diode's resistance is low. If the forward voltage is pushed further, the current becomes so large that it heats the junction, and the diode is **burnt out**: that is why a diode in a circuit always has a resistance in series to limit its current.

Reverse characteristics

The **reverse characteristic** is the graph of the current against the reverse bias voltage. The circuit is the same with the diode turned round (p side to the negative terminal), and with a **microammeter** in place of the milliammeter, since the reverse current is so small; the voltmeter needs a much larger range (Figure 22.6). The diode's resistance now is very high, about 10^5 Ω or more.

As the reverse voltage is raised, the reverse current, carried by minority carriers, is small and hardly changes: there are only so many minority carriers, and all of them are already being swept across. This nearly constant current is the **reverse saturation current**. But at a certain large reverse voltage the current suddenly **rises**

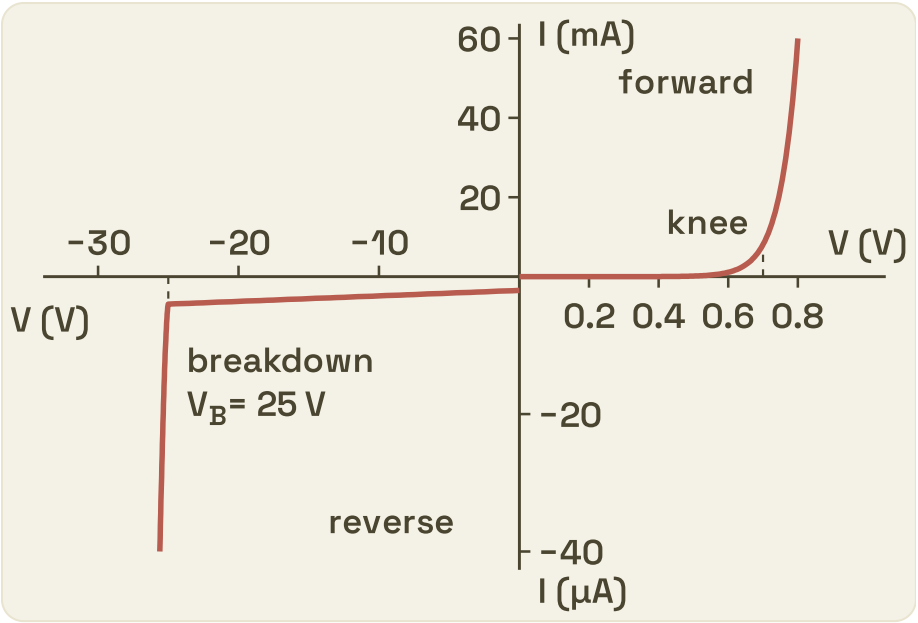


Figure 22.5 *The I–V characteristic of a silicon diode. Forward (right): almost no current until the knee voltage, about 0.7 V, then a steep rise, in milliamperes. Reverse (left, a different scale): a tiny current of a few microamperes, nearly constant, until the breakdown voltage V_B, where the current suddenly shoots up.*

sharply. This is **breakdown**, and the voltage at which it happens is the **reverse breakdown voltage** V_B (about 25 V in Figure 22.5; tens to hundreds of volts for ordinary diodes).

What happens at breakdown? In the very strong field of the wide depletion layer, the few minority carriers crossing it are accelerated so hard that they knock electrons out of the covalent bonds of the silicon atoms they hit. Those freed electrons are accelerated in turn and knock out more: the number of carriers grows like an avalanche, and so does the current (**avalanche breakdown**). In a heavily doped junction, with a very thin depletion layer, the field is so strong that it pulls electrons straight out of their bonds (**Zener breakdown**). Unless something limits the current, the large current at breakdown heats the junction and destroys an ordinary diode.

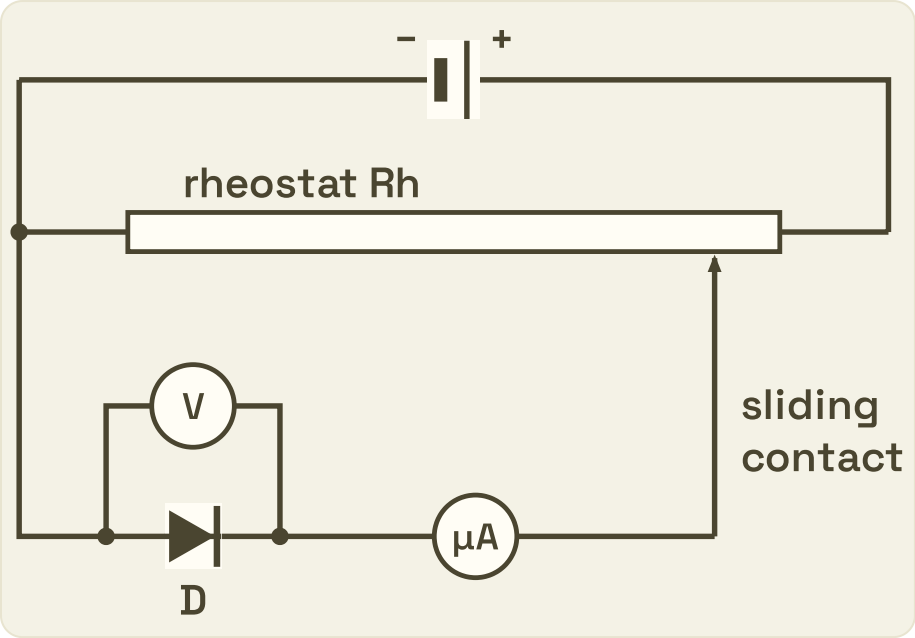


Figure 22.6 *Measuring the reverse characteristic: the diode is turned round, so its n side is towards the positive terminal, and a microammeter (μA) reads the tiny current.*

CAREFUL

Two scales on one graph

In a drawn characteristic, the forward current is in **milliamperes** and the forward voltage in tenths of a volt, while the reverse current is in **microamperes** and the reverse voltage in tens of volts. Drawn to a single scale, the reverse current would be invisible and the forward rise a vertical line at zero. When you draw it, mark both scales.

Check 4

A silicon diode is reverse biased at 5 V. Raising the reverse voltage to 10 V (still well below breakdown):

a doubles the reverse current, as Ohm's law says

b leaves the reverse current almost unchanged

c makes the diode conduct strongly

d reverses the direction of the current

22.4 Zener diode

Can the breakdown that ruins an ordinary diode be put to good use?

A **Zener diode** is a specially made diode, heavily doped, designed to **work in the breakdown region**. Forward biased, it conducts like any diode. Reverse biased, it conducts too, once the reverse voltage exceeds a definite value, its **Zener breakdown voltage** V_Z, and above that the voltage across it stays **almost constant** however much the current changes. Zener diodes are made with breakdown voltages from about 2 V to a few hundred volts (3.3 V, 5.1 V and 12 V are common).

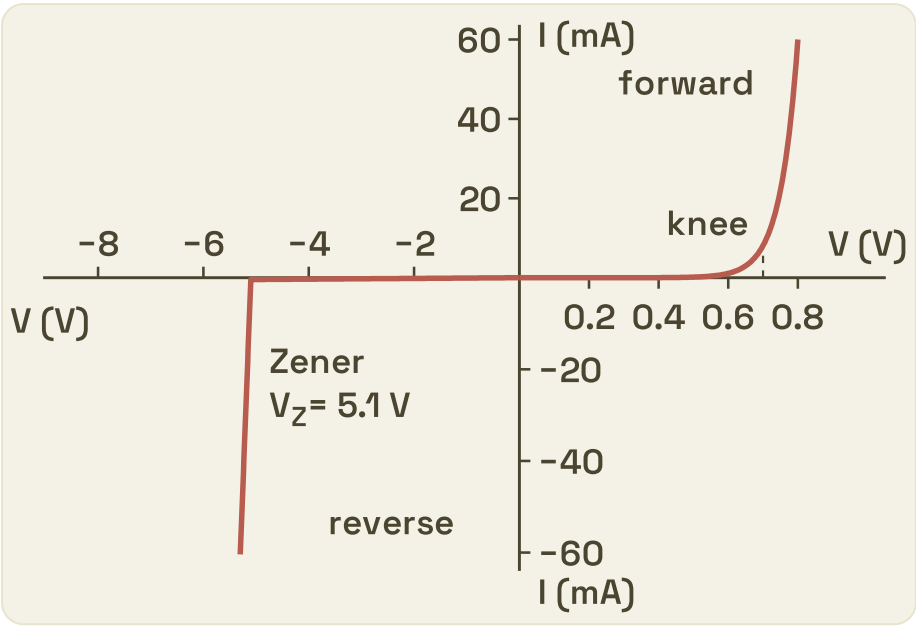


Figure 22.7 *The characteristic of a Zener diode with $V_Z = 5.1$ V. Forward, it is like an ordinary diode. Reverse, it carries almost nothing until 5.1 V, then the current rises steeply while the voltage across it hardly changes: along the near-vertical line the voltage is held at V_Z.*

This is what a Zener diode is used for: keeping a voltage steady, **voltage regulation**. An electronic circuit that turns a varying d.c. input into a steady d.c. output is called a **voltage regulator**. In the simplest one (Figure 22.8), the Zener diode is connected **reverse biased** across the load, with a resistor R in series with the supply. If the input rises, more current flows through R and the Zener, the extra voltage drops across R, and the voltage across the Zener and the load stays at V_Z. If the input falls (but stays above V_Z), the Zener takes less current, and again the output stays at V_Z.

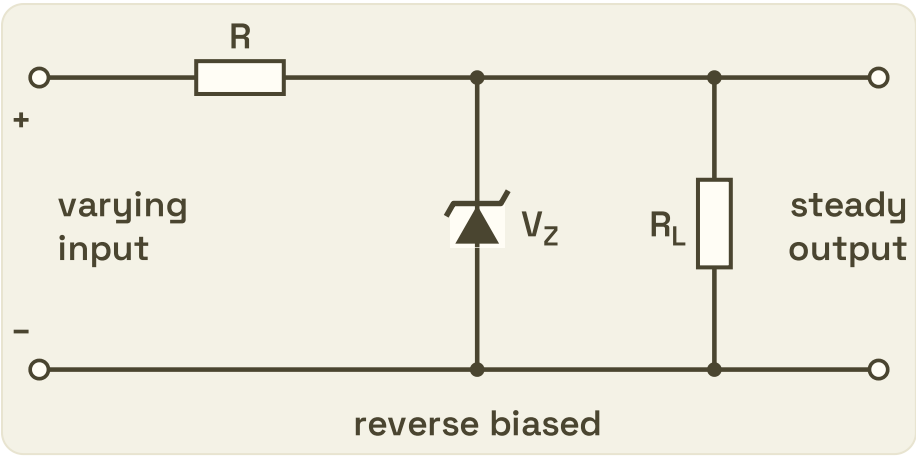


Figure 22.8 *A Zener diode as a voltage regulator. The Zener is reverse biased (its bar towards the positive side) across the load R_L. Whatever the varying input, as long as it is above V_Z, the extra voltage drops across the series resistor R, and the load gets a steady V_Z.*

The **symbol** of a Zener diode (Figure 22.3) is a diode's with the bar (the n side) bent at its ends, like a flattened letter Z.

WORKED EXAMPLE

Example 22.1: why a diode conducts one way only

A p–n diode conducts electricity when forward biased and does not conduct when reverse biased. Explain.

Answer

Forward biased: the battery pushes the majority carriers (holes in the p side, electrons in the n side) towards the junction, against the barrier. The depletion layer becomes thinner, its resistance falls, and once the voltage is above the barrier potential the majority carriers cross the junction: a large current flows.

Reverse biased: the battery pulls the majority carriers away from the junction. The depletion layer becomes wider, the barrier higher and the resistance very high. The majority carriers can't cross; only a tiny current of minority carriers flows. So the diode practically does not conduct.

Applications of diodes

Diodes are used:

a.

WORKED EXAMPLE

Example 22.2: is n-type silicon charged?

An n-type semiconductor has a large number of free electrons. Is it electrically neutral?

Answer

Yes. Each free electron comes from a **donor atom** (such as phosphorus) that was itself neutral. When the donor gives up its fifth electron, it becomes a **positive ion** fixed in the crystal. So for every extra free electron there is one fixed positive charge, and the crystal as a whole stays neutral.

as **rectifiers**, to turn a.c. into d.c. (Section 22.5);

b. as **switches** in logic circuits in computers (Section 22.7);

c. as **Zener diodes**, in voltage regulation;

and in many other ways: as **light-emitting diodes** (LEDs), which give out light when forward biased, as **photodiodes** and **solar cells**, which make a current when light falls on the junction, and to protect circuits from a battery put in the wrong way round.

DID YOU KNOW?

Light from a junction

In a forward-biased junction, every electron that falls into a hole gives up its energy. In silicon that energy becomes heat, but in some semiconductors, such as gallium arsenide and gallium nitride, it comes out as a photon of light whose color is set by the band gap. That is an LED. Red LEDs came in 1962; efficient blue ones took thirty more years and won Isamu Akasaki, Hiroshi Amano and Shuji Nakamura the 2014 Nobel Prize in Physics, because blue plus a yellow phosphor gives the white LED bulbs now in nearly every home in Nepal.

22.5 Junction diode as rectifier

How does a phone charger turn the a.c. from the wall into d.c.?

An electronic device that converts a.c. power into d.c. power is called a **rectifier**, and the process is **rectification**.

Principle. A junction diode has a low resistance when forward biased and a very high resistance when reverse biased. An alternating e.m.f. reverses its direction every half cycle, so its two halves bias a diode in opposite directions: if the first half cycle biases it forward, the next biases it in reverse. A diode in an a.c. circuit therefore **conducts only during alternate half cycles**, those that bias it forward. Its current flows one way only.

Half-wave rectifier

The simplest rectifier is a single diode in series with the load resistance R_L, fed from the secondary of a transformer (Figure 22.9). During the half cycle that makes the diode's anode positive, the diode conducts and the current flows through R_L; during the other half, the diode is reverse biased and no current flows. The output is a series of half-waves, all in the same direction, with gaps between them: **half-wave rectification**. Half of the input's energy is simply not used.

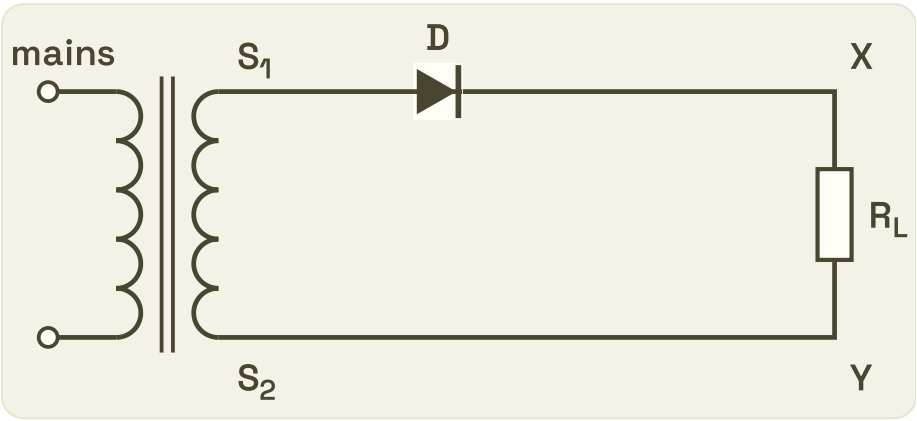


Figure 22.9 *A half-wave rectifier. The transformer steps the mains voltage down; the diode D conducts only in the half cycles that make its anode (the top of the secondary) positive, so current flows through the load R_L one way only, from X to Y.*

Full-wave rectifier with two diodes (center-tap)

A full-wave rectifier uses **both** halves of the a.c. One way needs a transformer whose secondary has a connection at its middle, a **center tap** (Figure 22.10). Its two ends, S_1 and S_2, go through diodes D_1 and D_2 to the same end of the load; the center tap goes to the other end.

During one half cycle, S_1 is positive with respect to the center tap and S_2 negative: D_1 is forward biased and conducts, D_2 is reverse biased. During the next half cycle, S_2 is positive: D_2 conducts and D_1 is off. Either way, the current flows **through the load in**

the same direction. Each diode works for half of each cycle, and the load gets a current in both halves. The drawback: the transformer needs a center tap, and each half of its secondary gives only half the voltage.

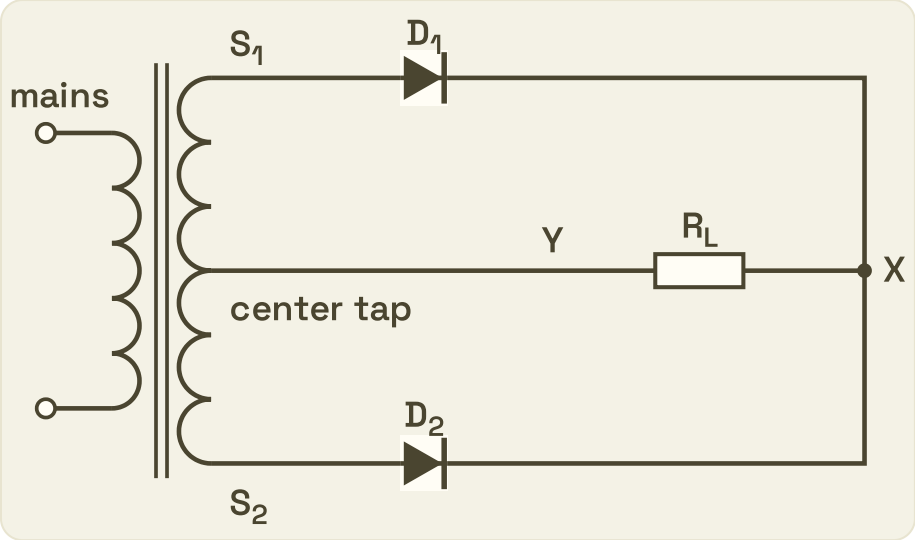


Figure 22.10 *A full-wave rectifier with two diodes and a center-tapped transformer. When S_1 is positive, D_1 conducts; when S_2 is positive, D_2 conducts. Either way the current flows through R_L from X to Y, towards the center tap.*

Full-wave bridge rectifier

The usual full-wave rectifier, in almost every charger, uses **four diodes** D_1, D_2, D_3 and D_4 arranged like a **Wheatstone bridge** (Figure 22.11). Two opposite corners, A and C, are connected to the ends S_1 and S_2 of the transformer's secondary; the other two corners, B and D, to the load resistance R_L. The transformer's primary P is connected to the a.c. mains.

When the a.c. is applied to the primary, at some instant the positive half of the cycle makes A positive and C negative. Diodes **D_1 and D_3** are forward biased and conduct, and the current flows along **S_1 → A → D_1 → B →** R_L **→ D → D_3 → C → S_2**. Meanwhile D_2 and D_4 are reverse biased and don't conduct.

During the next half cycle, A is negative and C positive. Now **D_2 and D_4** conduct, and the current flows along **S_2 → C → D_2 → B →** R_L **→ D → D_4 → A → S_1**, while D_1 and D_3 are off.

In both halves the current flows through R_L **from B to D**, in the same direction. So a d.c. output appears across R_L in both halves of the cycle: **full-wave rectification**

(Figure 22.12), with no center tap needed.

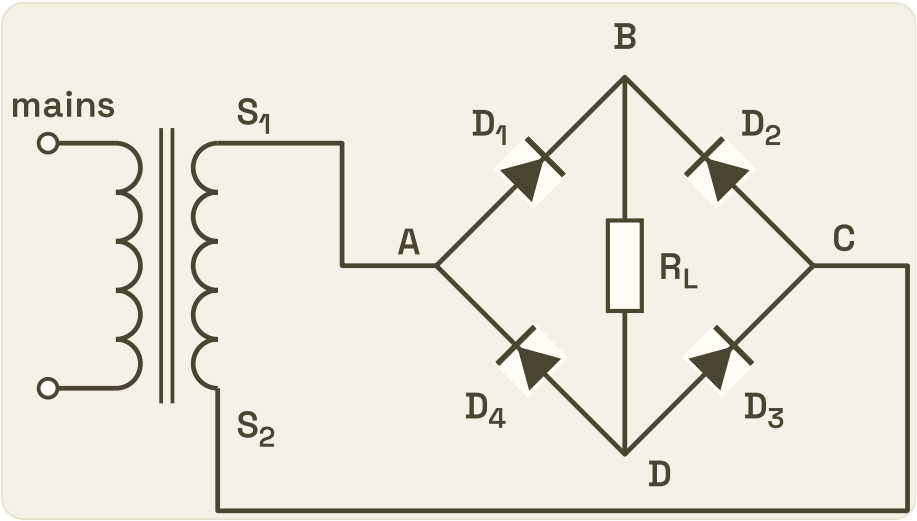


Figure 22.11 *A bridge rectifier. The transformer's secondary feeds the corners A and C of a bridge of four diodes; the load R_L joins the other two corners, B and D. When A is positive, D_1 and D_3 conduct; when C is positive, D_2 and D_4 conduct. Either way the current flows through R_L from B to D.*

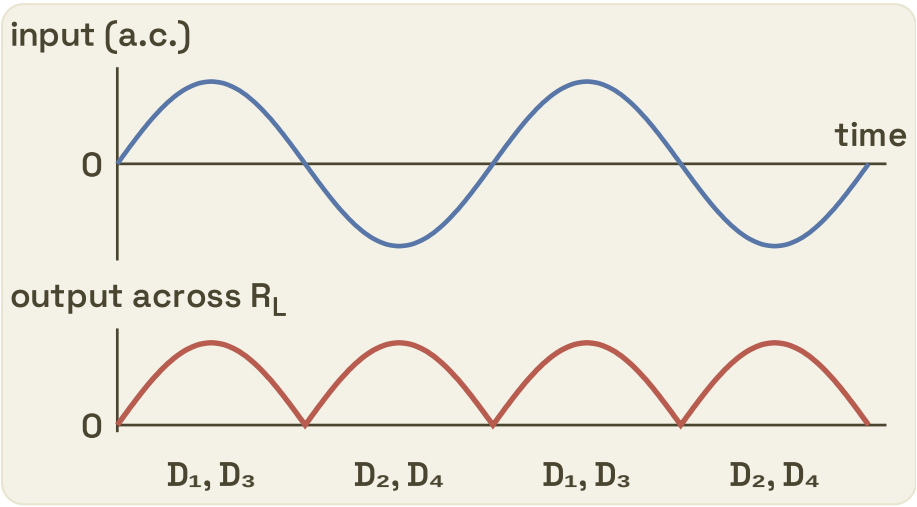


Figure 22.12 *The input and output of a bridge rectifier. The input (top) is the a.c. from the transformer. The output (bottom) is a hump in every half cycle, all positive: in the first half D_1 and D_3 conduct, in the second D_2 and D_4, and so on.*

Pulsating d.c. and smoothing

The output of a rectifier flows one way only, but it isn't steady: it rises from zero to a peak and falls back to zero 100 times a second (for 50 Hz mains, full wave). Such a current is called **pulsating d.c.** To smooth it, a large **capacitor** (a filter capacitor) is connected across the load. It charges up to the peak voltage during each hump, and between humps, when the rectifier's output falls, it keeps the voltage up by discharging slowly through the load. What remains is a nearly steady voltage with a small ripple. A voltage regulator, such as a Zener diode, then removes the ripple altogether.

WORKED EXAMPLE

Example 22.3: what the diode does

What is the role of a diode in a rectifier circuit?

Answer

A diode lets current flow in one direction only. In a rectifier it conducts in the half cycles that forward bias it and blocks the others, so the current through the load always flows the same way: the a.c. becomes d.c. In a **half-wave** rectifier the diode simply blocks the negative halves; in a **full-wave** rectifier the diodes are arranged so that the negative halves are also sent through the load in the same direction as the positive halves.

WORKED EXAMPLE

Example 22.4: pulsating d.c.

What is meant by pulsating d.c. output in rectifiers?

Answer

Pulsating d.c. is a current (or voltage) that flows in **one direction only** but keeps changing in size: it rises from zero to a peak and falls back again in every half cycle. It is what a rectifier gives before any filtering. It is not smooth like the d.c. of a battery; a filter capacitor (and a regulator) is needed to make it steady.

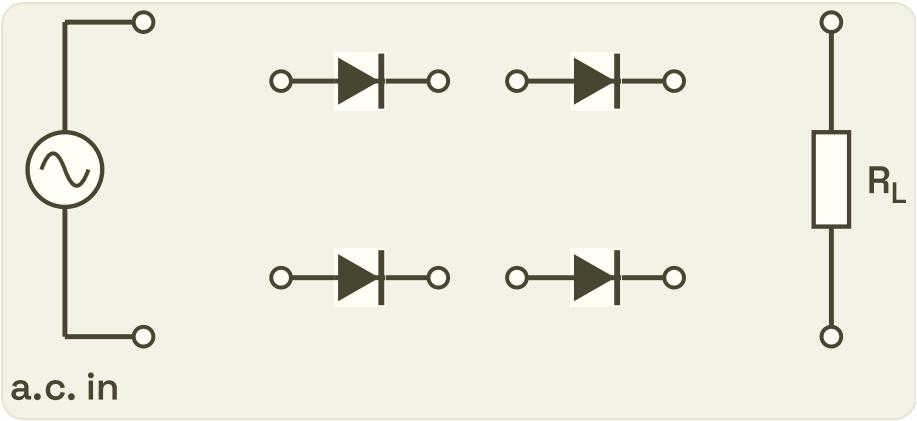


Figure 22.13 *For Test yourself 22.3 ii: an a.c. input, four diodes and a load resistor, not yet connected.*

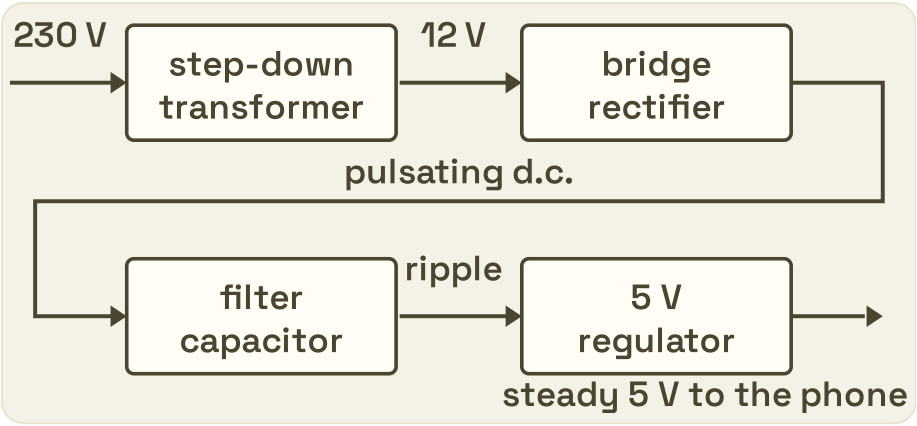


Figure 22.14 *The charger's circuit, in blocks: an enclosed step-down transformer, a bridge rectifier, a filter capacitor C, a 5 V regulator, and the load (the phone).*

22.6 Logic gate

How can electricity make a decision?

A digital circuit that either lets a signal through or stops it is called a **gate**. It lets the signal through only when certain logical conditions are satisfied, so such circuits are called **logic gates**. In a logic gate every input and output is either a high voltage (**1**, for example 5 V) or a low one (**0**, near 0 V). Simple gates can be built from **diodes** (and resistors), but a gate that turns a 1 into a 0 needs a **transistor**, and the gates inside a computer are built entirely of transistors, billions of them on one chip.

Logic gates are the building blocks of **digital systems** such as computers, mobile phones and calculators. Each gate follows a definite logical relationship between its input and output voltages.

The basic logic gates are of three types:

1. the **OR** gate
2. the **AND** gate
3. the **NOT** gate

Each is drawn with its own **symbol**, and what it does is given either by a **truth table** or by a **Boolean expression**.

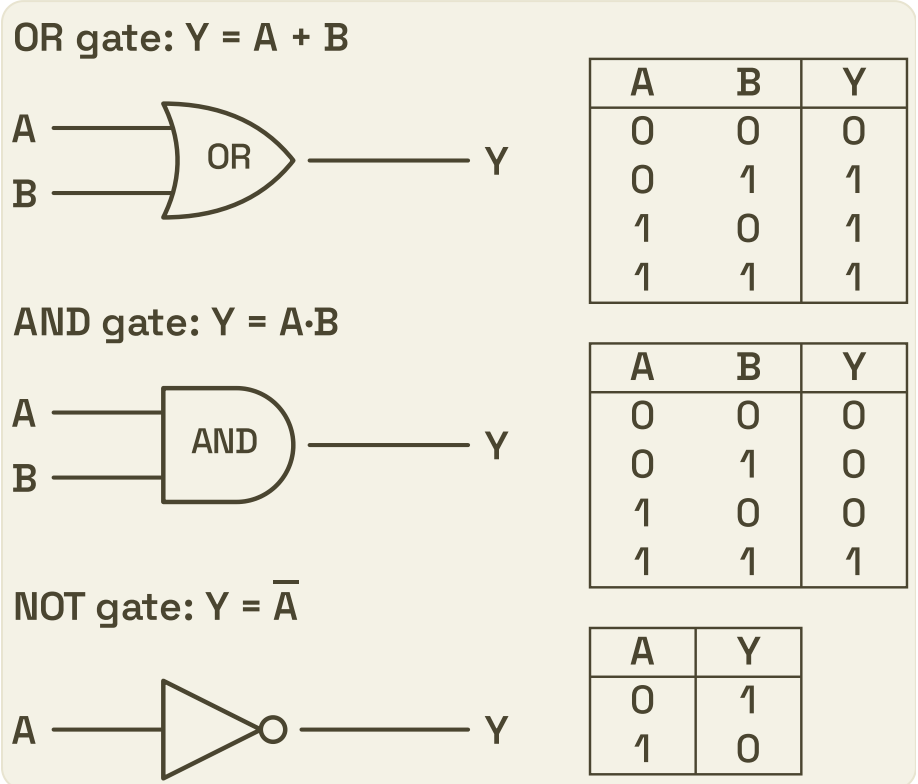


Figure 22.15 *The symbols of the three basic gates, with their truth tables. OR: $Y = A + B$, 1 if either input is 1. AND: $Y = A \cdot B$, 1 only if both are. NOT: $Y = \bar{A}$, the opposite of its input.*

22.7 Truth table and Boolean expression

Truth table

A **truth table** is a table showing **all the possible combinations of the inputs** of a logic gate (or circuit) and the **output** for each. It is also called a table of combinations. With two inputs there are $2 \times 2 = 4$ combinations; with three, 8; with n inputs, 2^n. For example, the OR function:

A	B	A or B
0	0	0
0	1	1
1	0	1
1	1	1

Boolean algebra

George Boole, an English mathematician (1815–1864), invented an algebra of logical statements, which can only be **true** or **false**. Its variables are called **Boolean variables**: true is written **1** and false **0**. These 1 and 0 are not numbers to be added in the ordinary way; they are names for the two values. In electronic circuits, 1 and 0 stand for the two states of a circuit: on and off, a high voltage and a low one.

In Boolean algebra:

- the **addition sign (+)** means **OR**: A + B is read “A or B”;
- the **multiplication sign (· or ×)** means **AND**: A·B is read “A and B”;
- a **bar** over a variable means **NOT**: $\bar{A}$ is read “not A”.

So in Boolean algebra $1 + 1 = 1$ (true or true is true), which is why the symbols must not be confused with arithmetic.

A STORY FROM SCIENCE

Boole's algebra and Shannon's switches

George Boole was the son of a shoemaker in Lincoln, England, who taught himself Latin, Greek and mathematics and became a professor in Ireland. In 1854 he published *The Laws of Thought*, which turned reasoning into algebra with only two values. For eighty years it was a curiosity of logic. Then in 1937 **Claude Shannon**, a 21-year-old student in the United States, showed in his master's thesis that circuits of switches and relays obey exactly Boole's algebra: switches in series are AND, switches in parallel are OR. So any logical rule can be built as a circuit, and a circuit can be simplified by algebra before it is built. It has been called the most important master's thesis of the century; every chip designed since rests on it.

The OR gate (Y = A + B)

An **OR gate** is a logic gate whose output is **1 if any one or more of its inputs is 1**.

Figure 22.16 shows a two-input OR gate made of two ideal diodes D_1 and D_2: A and B are the inputs, Y the output, and R the load resistor between Y and earth (0 V). Each input is either 0 V (0) or, say, 5 V (1).

- **Case I:** $A = 0$, $B = 0$. Neither diode has a voltage to drive it, so no current flows through R, and the output is 0 V: $\mathbf{Y = 0}$.
- **Case II:** $A = 1$, $B = 0$. D_1 is forward biased and conducts, D_2 doesn't. Current flows from A through D_1 and R to earth, and Y is at the input's voltage: $\mathbf{Y = 1}$.
- **Case III:** $A = 0$, $B = 1$. D_2 conducts and D_1 doesn't: $\mathbf{Y = 1}$.
- **Case IV:** $A = 1$, $B = 1$. Both diodes conduct: $\mathbf{Y = 1}$.

Summarized as a truth table:

A	B	Y = A + B
0	0	0
0	1	1
1	0	1
1	1	1

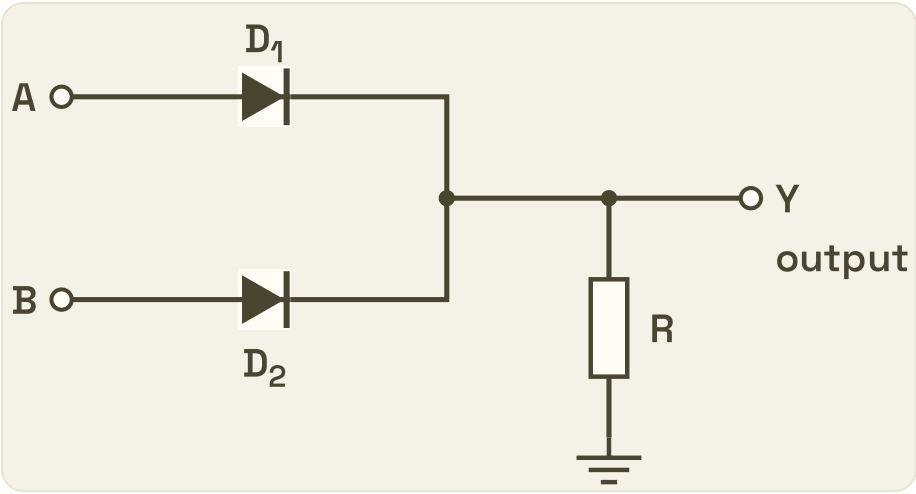


Figure 22.16 *A two-input OR gate made of two diodes. Each input, A or B, is joined through its own diode, pointing towards Y, to the output Y; the resistor R joins Y to earth.*

The AND gate (Y = A · B)

In an **AND gate**, the output is **1 only when every input is 1**; a single 0 at any input makes it 0.

A two-input AND gate can also be made of two ideal diodes, turned the other way round (Figure 22.17): the output Y is joined through the resistor R to the + terminal of a 1 V (or 5 V) supply, and each diode's anode is at Y, its cathode at an input. Each input is either 0 V (0) or the supply's voltage (1).

- **Case I:** A = 0, B = 0. Both diodes are forward biased: current flows from the supply through R and both diodes into the inputs at 0 V. Y is held at 0 V: **Y = 0**.
- **Case II:** A = 1, B = 0. D_2 conducts and holds Y at B's 0 V: **Y = 0**.
- **Case III:** A = 0, B = 1. D_1 conducts and holds Y at 0 V: **Y = 0**.
- **Case IV:** A = 1, B = 1. Both inputs are at the supply's voltage, so neither diode is forward biased. No current flows through R, so there is no voltage drop across it, and Y is at the supply's voltage: **Y = 1**.

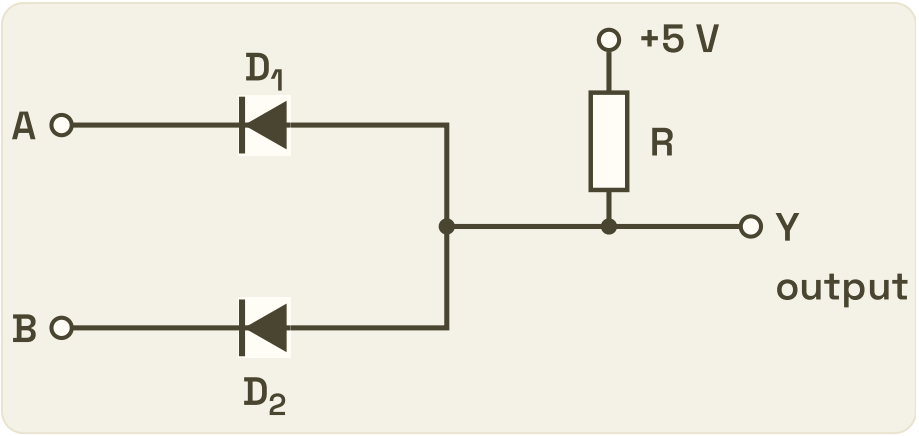


Figure 22.17 *A two-input AND gate made of two diodes. Each diode points from the output Y to an input, A or B; the resistor R joins Y to the + terminal of the supply.*

A	B	Y = A · B
0	0	0
0	1	0
1	0	0
1	1	1

Check 5

In the diode OR gate, A = 1 and B = 0. Which diodes conduct?

a only D_1, the one joined to A
b only D_2
c both
d neither

The NOT gate ($Y = \bar{A}$)

A **NOT gate** has one input and one output, and it **inverts** its input: if the input is 1 the output is 0, and if the input is 0 the output is 1. It is also called an **inverter**.

A NOT gate can't be made of diodes and resistors alone: diodes can only pass on a voltage, never turn a low one into a high one. It needs a **transistor** used as a switch (Figure 22.18). The input A is joined through a resistor R_1 to the transistor's base; its collector is joined through R_2 to the supply +V and is the output Y; its emitter is earthed.

- When **A = 0**, no current flows into the base, so the transistor is **off** (like an open switch). No current flows through R_2, so there is no drop across it, and Y is at +V: **Y**

= 1.

- When **A = 1**, a current flows into the base and the transistor is fully **on** (like a closed switch from collector to emitter). The output is joined almost directly to earth: **Y = 0**.

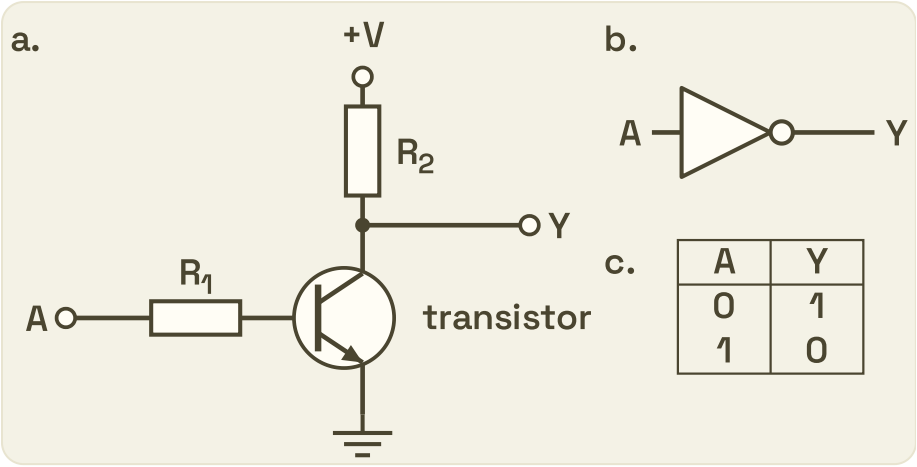


Figure 22.18 *A NOT gate from a transistor: a. the circuit, b. its symbol, c. its truth table. Input 0: the transistor is off and Y is pulled up to +V through R_2 (1). Input 1: the transistor is on and Y is joined to earth (0).*

The NOR gate (Y = $\overline{A + B}$)

An OR gate followed by a NOT gate is a **NOR gate** (NOT-OR). The NOT gate inverts the OR gate's output (Figure 22.19), so its output is 1 only when **both inputs are 0**. Its symbol is the OR symbol with a small circle (the sign of NOT) at its output.

- **Case I:** A = 0, B = 0: A + B = 0, so $Y = \overline{A + B} = 1$.
- **Case II:** A = 0, B = 1: A + B = 1, so Y = 0.
- **Case III:** A = 1, B = 0: A + B = 1, so Y = 0.
- **Case IV:** A = 1, B = 1: A + B = 1, so Y = 0.

A	B	Y = $\overline{A + B}$
0	0	1
0	1	0
1	0	0
1	1	0

So a NOR gate gives a high output (1) **only when both inputs are low** (0).

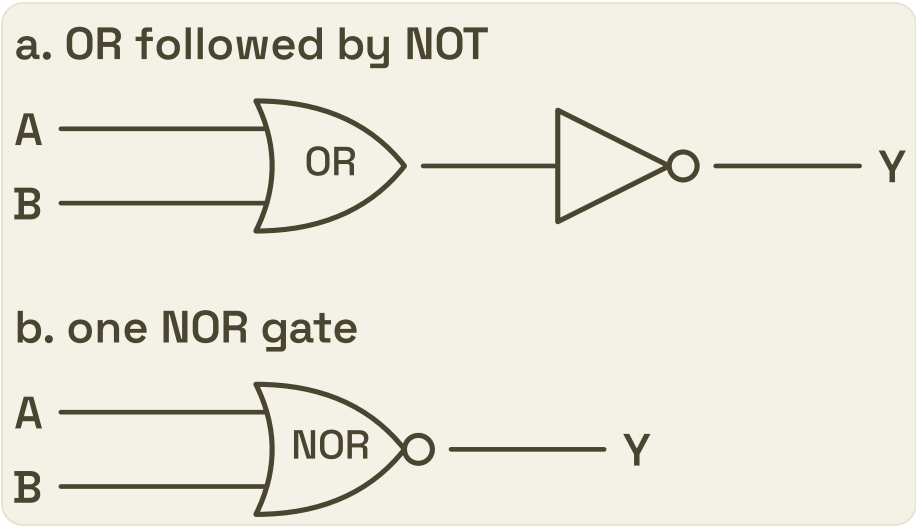


Figure 22.19 *a. An OR gate followed by a NOT gate. b. The same thing as one gate: the NOR gate, drawn as an OR with a small circle at its output.*

The NAND gate ($Y = \overline{A \cdot B}$)

An AND gate followed by a NOT gate is a **NAND gate** (NOT-AND): a transistor NOT circuit placed after the diode AND circuit. Its symbol is the AND symbol with a small circle at its output (Figure 22.20).

- **Case I:** A = 0, B = 0: A·B = 0, so $Y = \overline{A \cdot B} = 1$.
- **Case II:** A = 0, B = 1: A·B = 0, so Y = 1.
- **Case III:** A = 1, B = 0: A·B = 0, so Y = 1.
- **Case IV:** A = 1, B = 1: A·B = 1, so Y = 0.

A	B	Y = $\overline{A \cdot B}$
0	0	1
0	1	1
1	0	1
1	1	0

So a NAND gate gives a high output (1) when **either or both inputs are low** (0), and 0 only when both are 1.

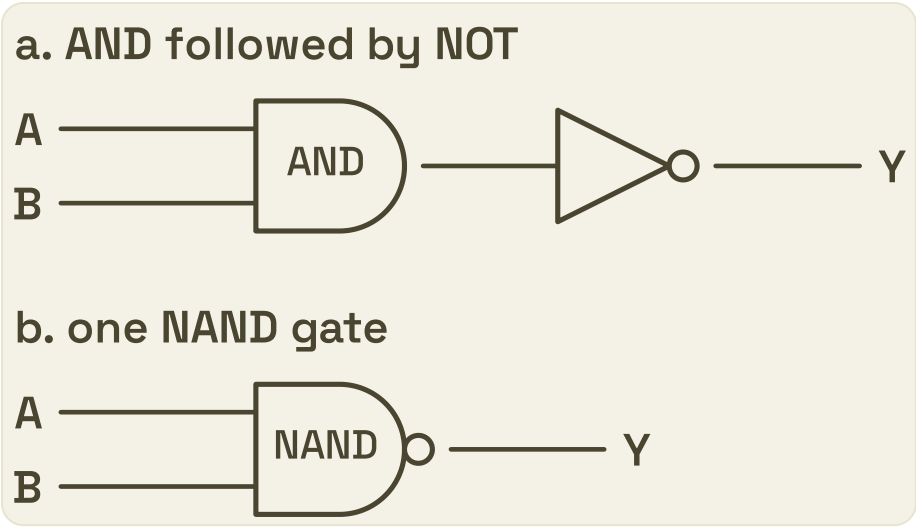


Figure 22.20 *An AND gate followed by a NOT gate is the same as one NAND gate, drawn as an AND with a small circle at its output.*

Check 6

The output of a two-input gate is 0 only when both inputs are 1. The gate is:

a AND **b** NOR
c NAND **d** OR

NAND gate as a universal gate

The NAND gate is called a **universal gate**: NAND gates alone, used over and over, can make every other gate, and so any logic circuit at all.

- **NOT from NAND:** join both inputs of one NAND gate together. With both inputs equal to A, $Y = \overline{A \cdot A} = \bar{A}$.
- **AND from NAND:** a NAND gate followed by a NAND used as a NOT: $Y = \overline{\overline{A \cdot B}} = A \cdot B$. Two gates.
- **OR from NAND:** invert each input with a NAND used as a NOT, then feed both into a NAND: $Y = \overline{\bar{A} \cdot \bar{B}} = A + B$ (by De Morgan's law). Three gates.

In the same way, the **NOR gate is a universal gate** too:

- **NOT from NOR:** both inputs joined: $\overline{A + A} = \bar{A}$.
- **OR from NOR:** a NOR followed by a NOR used as a NOT: $\overline{\overline{A + B}} = A + B$.

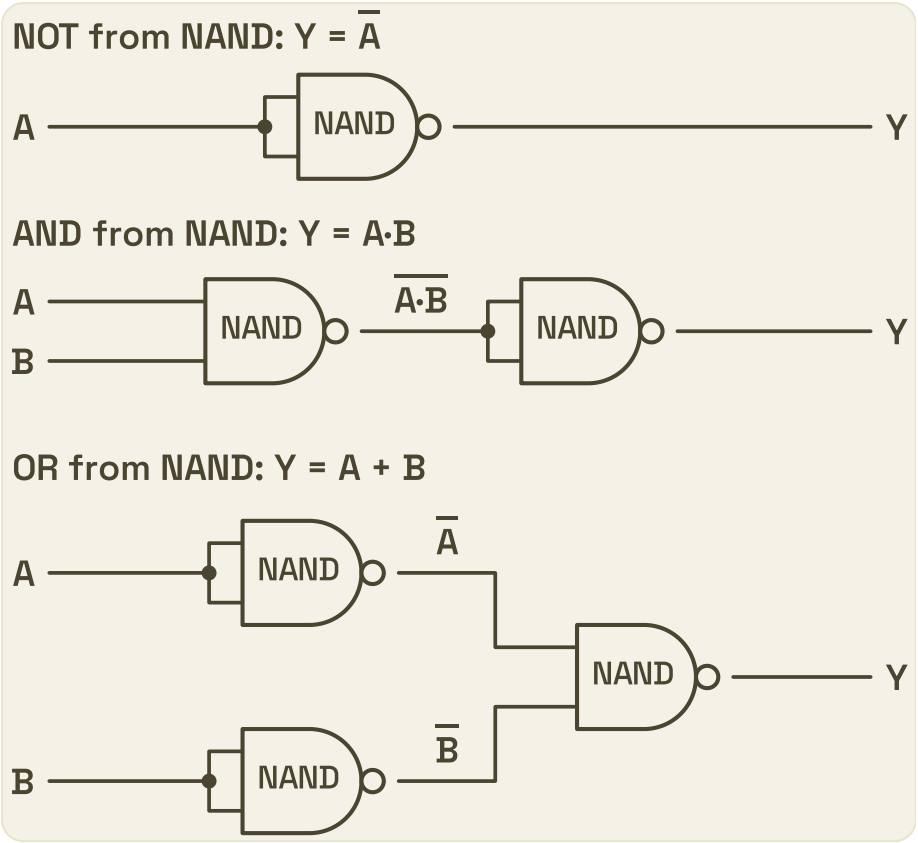


Figure 22.21 *NOT, AND and OR made of NAND gates only. A NAND with its two inputs joined is a NOT gate; a NAND followed by such a NOT is an AND; two of these NOTs, giving $\bar{A}$ and $\bar{B}$, feeding a NAND make an OR.*

- **AND from NOR:** invert each input with a NOR, then feed both into a NOR: $\overline{\bar{A} + \bar{B}} = A \cdot B$.

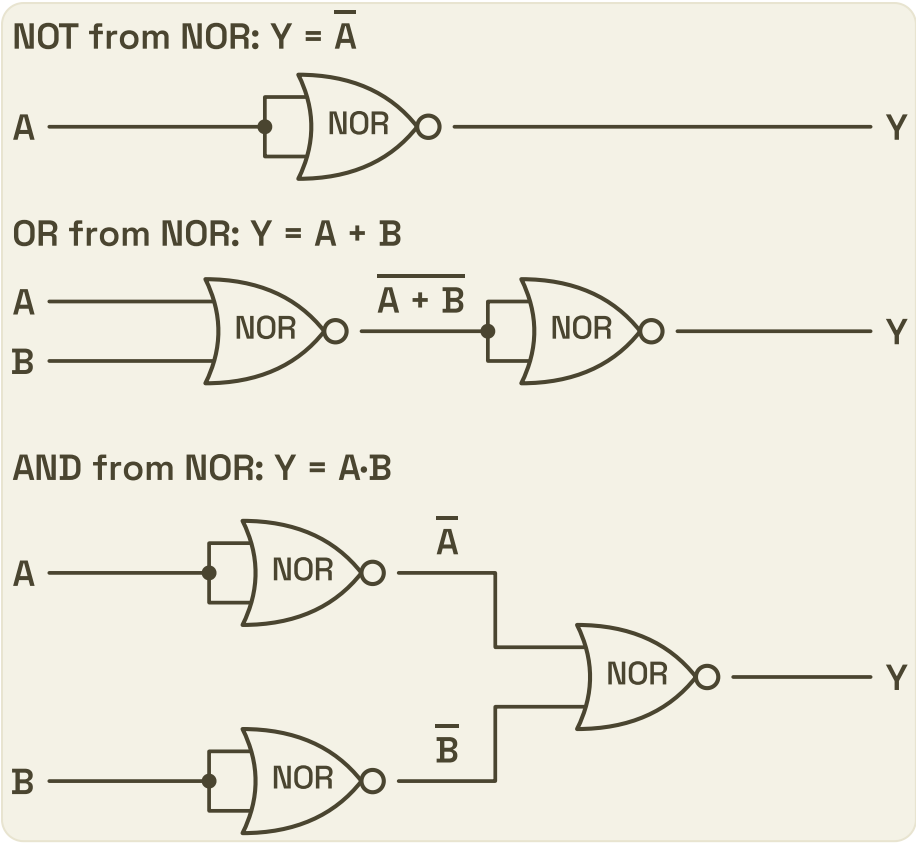


Figure 22.22 *NOT, OR and AND made of NOR gates only: a NOR with its inputs joined is a NOT gate, a NOR followed by such a NOT is an OR, and $\bar{A}$ and $\bar{B}$ into a NOR make an AND.*

DID YOU KNOW?

Why chips are built of NAND

A NAND gate is the simplest gate to make in the CMOS transistors of modern chips (four transistors), and an AND gate is literally a NAND gate followed by an inverter. So engineers think of NAND as the natural brick. In the USB drive in your pocket, the memory itself is called **NAND flash**, because its cells are wired in strings like the transistors of a NAND gate.

WORKED EXAMPLE

Example 22.5: NAND as a building block

Why are NAND gates called the digital building blocks?

Answer

Because NAND gates alone, used repeatedly, can produce all three basic gates: NOT (a NAND with its inputs joined), AND (a NAND followed by a NAND-NOT) and OR (two NAND-NOTs feeding a NAND). Since every logic circuit can be built of OR, AND and NOT gates, it can be built of NAND gates only. That is why the NAND gate is called a **universal gate**, a digital building block.

WORKED EXAMPLE

Example 22.6: an OR gate feeding a NOR gate

The output of an OR gate is connected to both inputs of a NOR gate. Draw the logic circuit of this combination of gates and write its truth table.

Answer

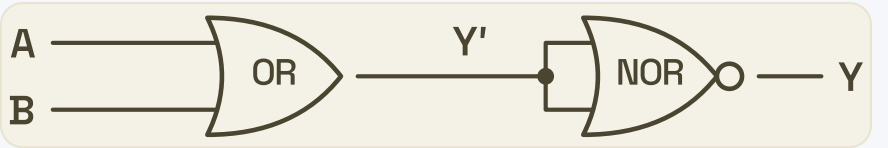


Figure 22.23 *The OR gate's output $Y' = A + B$ goes to both inputs of the NOR gate.*

The OR gate gives $Y' = A + B$. The NOR gate, with both inputs equal to Y′, gives

$$Y = \overline{Y' + Y'} = \overline{Y'} = \overline{A + B}$$

So the combination works as a single **NOR gate**:

A	B	Y' = A + B	Y
0	0	0	1
0	1	1	0
1	0	1	0
1	1	1	0

WORKED EXAMPLE

Example 22.7: the output of a circuit

Find the output of the logic circuit below for every combination of inputs.

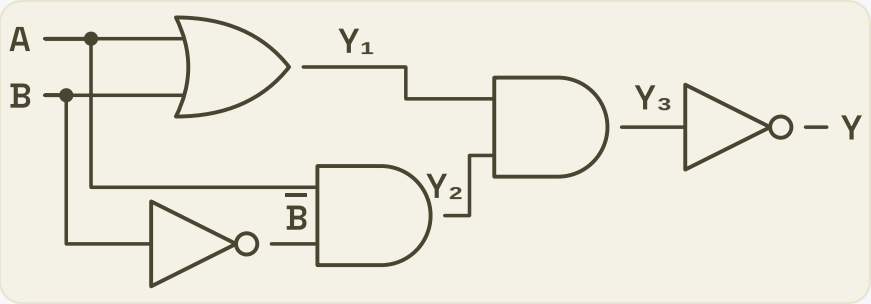


Figure 22.24 *The circuit for Example 22.7. A and B go to an OR gate; B also goes to a NOT gate, whose output and A go to an AND gate; the OR's and the AND's outputs go to a second AND gate, followed by a NOT gate.*

Answer

Name each wire and work along the circuit, column by column:

- the OR gate: $Y_1 = A + B$;
- the NOT gate: $\bar{B}$;
- the first AND gate: $Y_2 = A \cdot \bar{B}$;
- the second AND gate: $Y_3 = Y_1 \cdot Y_2$;
- the last NOT gate: the output $Y = \overline{Y_3}$.

A	B	$\bar{B}$	$Y_1 = A + B$	$Y_2 = A \cdot \bar{B}$	$Y_3 = Y_1 \cdot Y_2$	$Y = \overline{Y_3}$
0	0	1	0	0	0	1
0	1	0	1	0	0	1
1	0	1	1	1	1	0
1	1	0	1	0	0	1

The output is 0 only when A = 1 and B = 0. (Since $A \cdot \bar{B}$ is already 1 only when A + B is, $Y_3 = A\bar{B}$, so $Y = \overline{A\bar{B}}$.)

IN SHORT

Key ideas of Unit 22

- A p–n junction is p-type and n-type regions in one crystal. Carriers diffuse across and recombine, leaving a depletion layer (0.01–1 μm) of fixed ions with a potential barrier: about 0.7 V for silicon, 0.15–0.3 V for germanium.
- Forward bias (p to +): the layer thins, the resistance is low, and above the knee voltage a large forward current of majority carriers flows. Reverse bias (p to −): the layer widens, the resistance is very high, and only a tiny reverse current of minority carriers flows, until breakdown.
- The characteristic: forward, nothing until the knee, then a steep rise (mA); reverse, a few μA, nearly constant, up to the breakdown voltage, where the current shoots up.
- A Zener diode works in reverse breakdown: the voltage across it stays at V_Z, so it regulates voltage.
- Rectifiers turn a.c. into d.c.: half-wave (one diode), full-wave with a center tap (two diodes), full-wave bridge (four). The output is pulsating d.c.; a capacitor smooths it.
- Logic gates: OR (Y = A + B), AND (Y = A·B), NOT ($Y = \bar{A}$), NOR ($\overline{A + B}$), NAND ($\overline{A \cdot B}$). OR and AND can be made of diodes; NOT needs a transistor.
- NAND and NOR are universal gates: each alone can make NOT, AND and OR, and so any logic circuit.

Exercise

Choose the correct option

Check 7

1. The barrier potential difference is the potential difference across:

a the terminals of a cell
b the depletion layer of a p–n junction
c the plates of a capacitor
d the ends of a conductor

Check 8

2. The current through a p–n junction is negligible when the junction is:

a forward biased
b reverse biased
c positive biased
d negative biased

Check 9

3. A reverse-biased p–n junction has:

a a very narrow depletion layer
b almost no current
c a very low resistance
d a large current

Check 10

4. Which device is used to convert a.c. into d.c.?

a a transformer
b a rectifier
c a diode
d a triode

Check 11

5. A Zener diode can be used as a:

a rectifier
b voltage regulator
c oscillator
d voltage divider

Check 12

6. All logic gates can be made by repeated use of:

a OR gates
b AND gates
c NOT gates
d NAND gates

Check 13

7. What is the minimum number of NAND gates needed to make a NOT gate?

a 1
b 2
c 3
d 0

Check 14

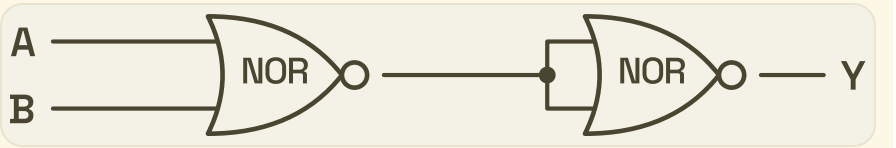


Figure 22.29 *For question 8: A and B into a NOR gate, whose output goes to both inputs of a second NOR gate.*

8. Which gate does the logic circuit above give?

a an OR gate
b an AND gate
c a NAND gate
d a NOT gate

Check 15

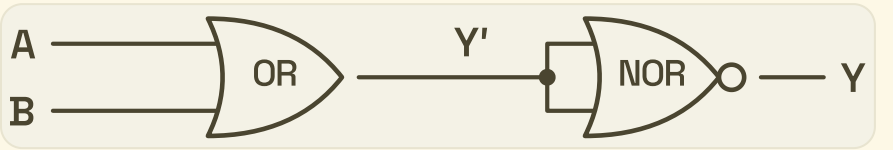


Figure 22.30 *For question 9: A and B into an OR gate, whose output $Y' = A + B$ goes to both inputs of a NOR gate.*

9. Which gate does the logic circuit above give?

a an OR gate
b an AND gate
c a NAND gate
d a NOR gate

Check 16

10. Boolean algebra is essentially based on:

a logic
b truth
c numbers
d symbols

Answer these questions

11.

a. Explain the formation of the potential barrier and the depletion layer in a p–n junction.

b. Explain the forward and reverse characteristics of a p–n junction diode.

12. a. What is a junction diode? Explain a full-wave rectifier using two diodes.
    b. What will happen if one of the diodes is damaged?
    c. The output of a full-wave rectifier is pulsating d.c. Which device can be used to get a constant output voltage? Draw its symbol and its characteristic.

13. a. What are logic gates? Give the truth table of a two-input NAND gate, with its symbol.
    b. Explain why the NAND gate is called a universal gate.

14. a. Draw a circuit diagram to realize an AND gate from NOR gates, with its truth table.
    b. Which gate will you get if you connect a NOT gate at the output of the circuit of part a?
    c. Show, using a truth table, that the circuit given below acts as an OR gate.

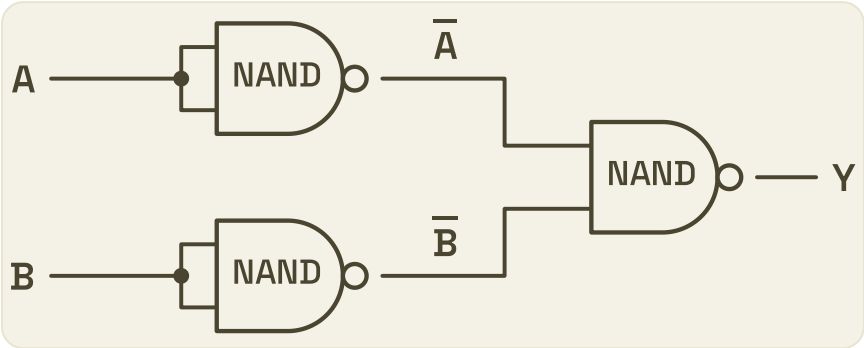


Figure 22.31 *For question 14 c: A and B each go to both inputs of a NAND gate; the two outputs go to a third NAND gate.*

A companion to your textbook, not a replacement for it: study from the book your school uses, and where the two differ, follow your teacher. Found a mistake? Please tell us.