Computer Systems 1
Lecture 20

CMOS and VLSI

How Logic Gates Work

Dr John T. O'Donnell
School of Computing Science
University of Glasgow

Topics

- Doped silicon
- CMOS technology
- Logic gates
- Integrated circuits
- Dynamic memory

Doped Silicon

Physics Background

Atoms

Electrons have negative charge

Nucleus contains protons with positive charge

- Like charges repel
- Unlike charges attract

Chemistry Background

- Electron Shells
  - The electrons around a nucleus are grouped into shells. Each shell has an ideal number of electrons that will fit into it, and the atom "likes" to have the outer shell filled
- Covalent Chemical Bonds
  - Two atoms with partially filled outer shells can "share" some electrons; they bond together

Insulators and Conductors

- If an atom has just enough electrons to fill the outer shell exactly, then electricity (moving electrons) can't go through it. It's an insulator.
- If an atom has only a few electrons in a big shell far from the nucleus, it's easy to get those electrons moving. It's a conductor (silver, copper, gold, …)

Semiconductors

- Silicon is a semiconductor, halfway between an insulator and a conductor.
- Its outer shell wants to have 8 electrons, but the Silicon atom provides only 4.
- The atom forms covalent bonds with its neighbors, ending with filled shells.
- This is a silicon crystal, which is a good insulator

Doped Silicon

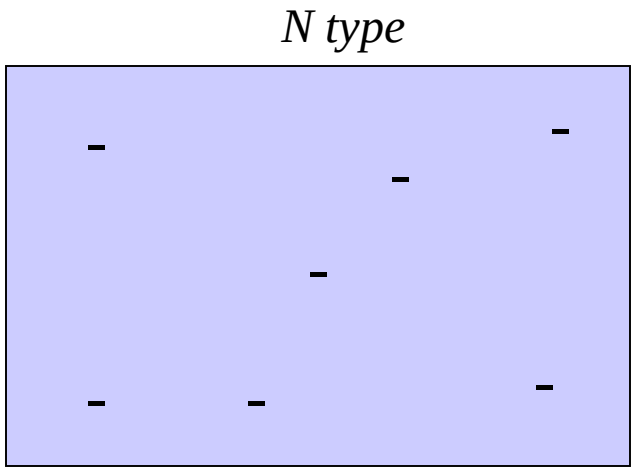
We could take a silicon crystal, but replace a few of the silicon atoms with Boron or Phosphorous.

- This would give us either one more or one fewer electrons in the outer shell
- And that means that an electrical field can more easily get some of those electrons moving, since the outer shell is not completely stable

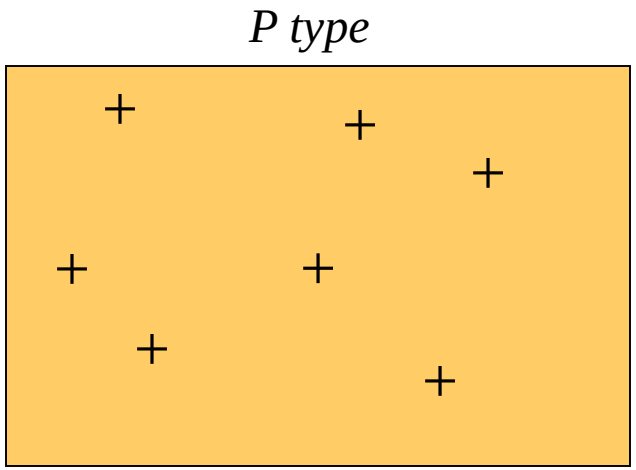
N and P Type Silicon

- Pure silicon is an insulator, but it can be doped, turning it into:
  - N type silicon, a semiconductor with negative charge carriers (called free electrons)
  - P type silicon, a semiconductor with positive charge carriers (called holes). *These are spots where an electron would like to be (would lower the energy) but isn't.*

Negative Charge Carriers

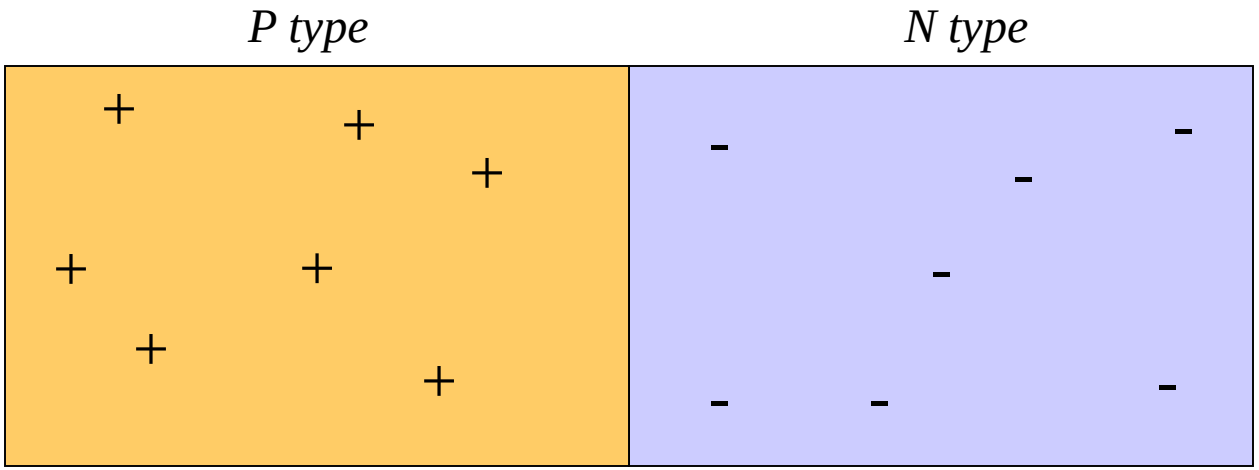


Positive Charge Carriers



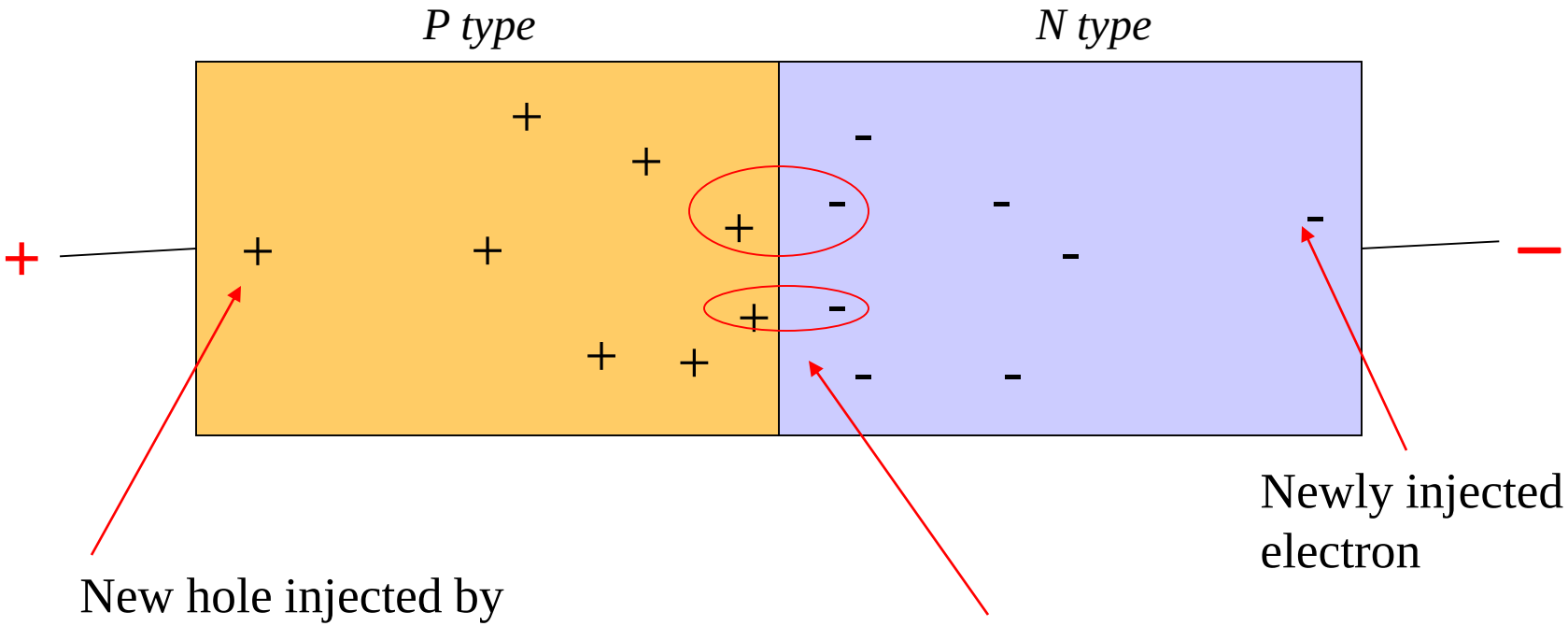
In reality, the only movable charge carriers are negative electrons. However, they behave as if there are holes that can move, and the holes act as if they have a positive charge.

Junctions



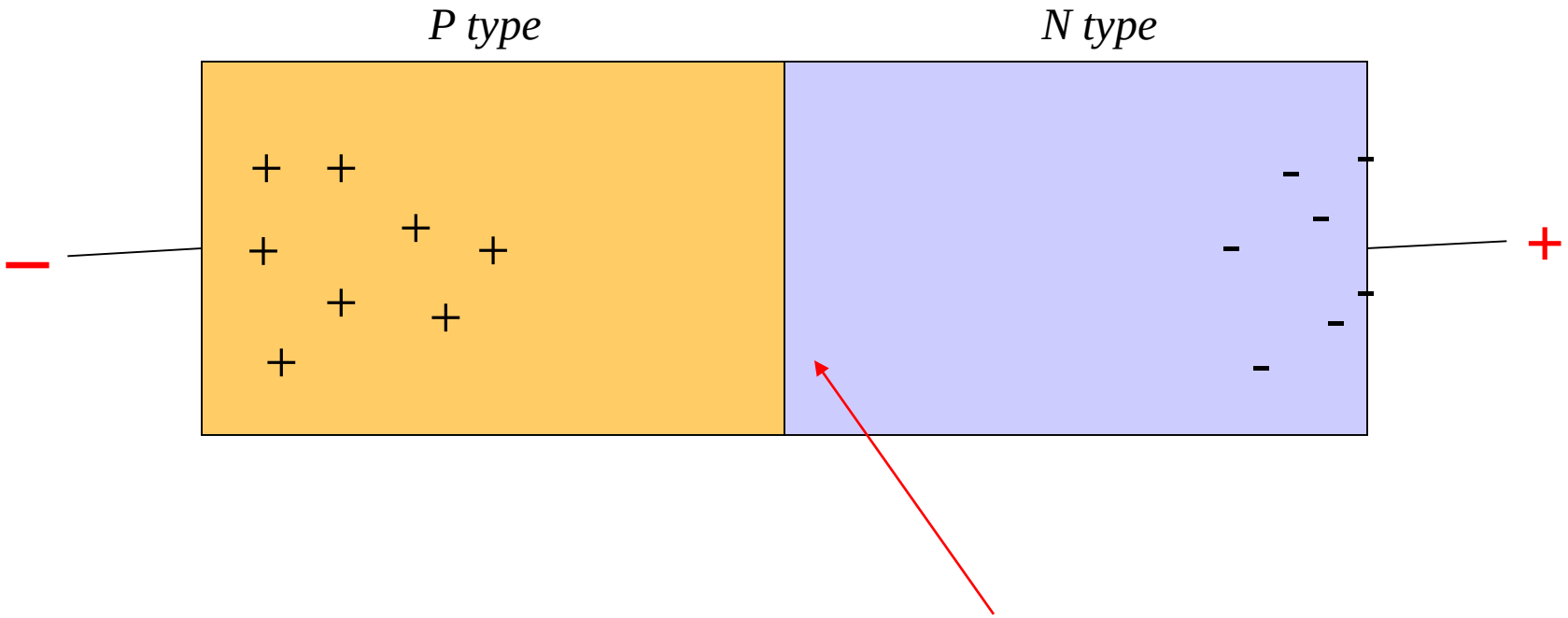
a PN junction is formed when a P-type region is fabricated adjacent to an N-type region

Forward Biased Junction



Current flows steadily!

Reverse Biased Junction



No current flows across the junction!

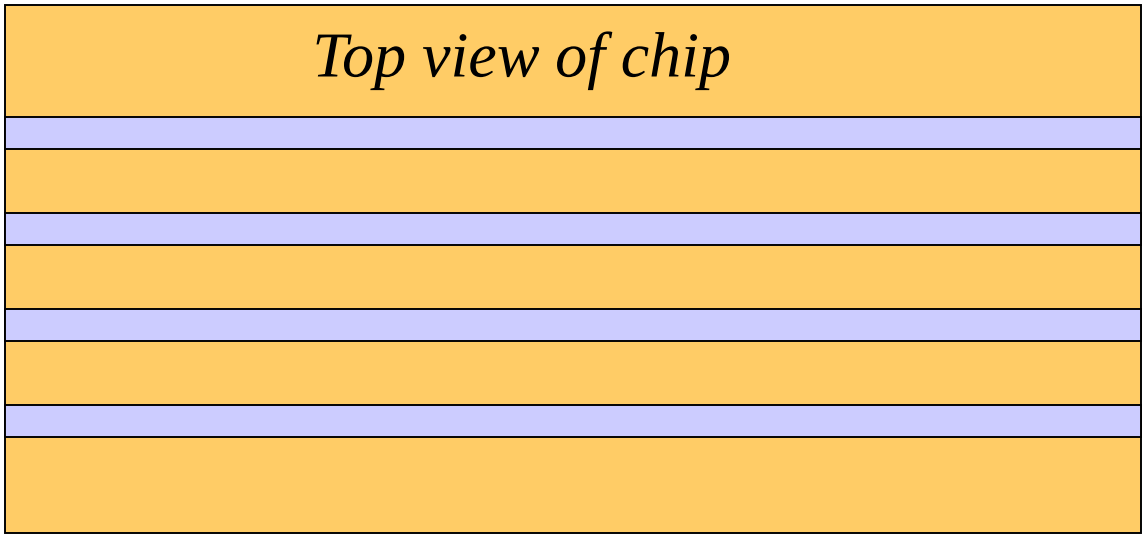
There are no free charge carriers near the junction, so it becomes an insulating crystal

The Junction Diode

A diode allows current to flow across it one direction but not the other

A PN junction does exactly that!

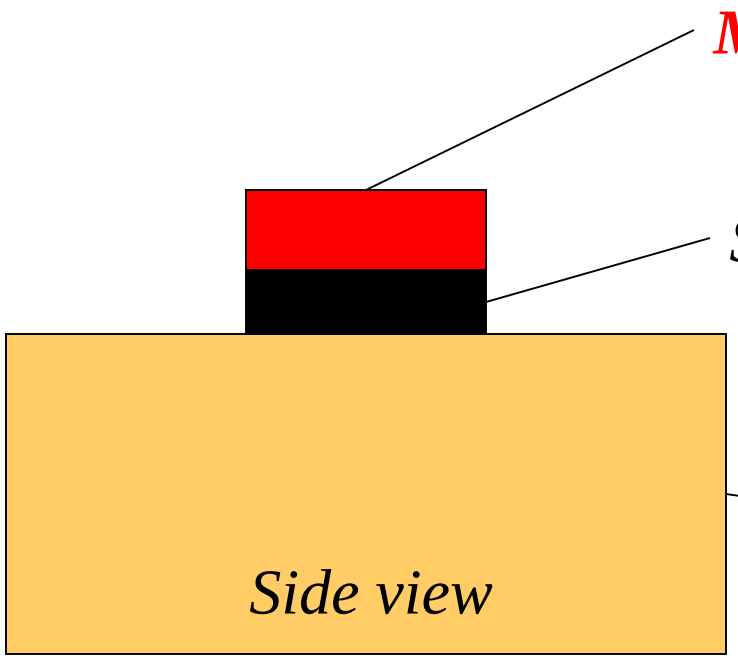
N-Type Wires



Each of the four N-type wires carries a signal safely; a short circuit is impossible because it would have to cross a reverse biased junction

CMOS Technology

Metal—Oxide—Semiconductor



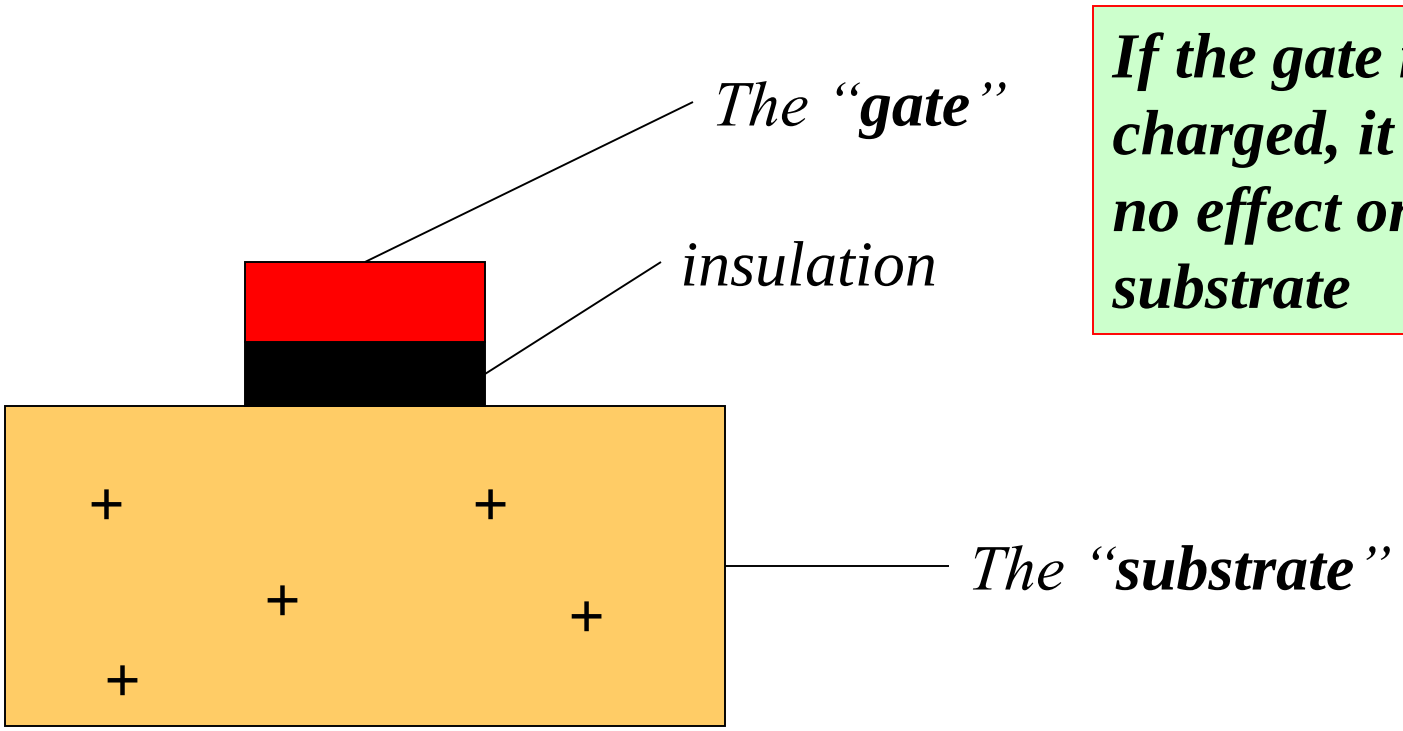
Metal: a layer of aluminum

Silicon dioxide—glass—an insulator

Semiconductor—doped silicon

This is called MOS technology

Neutral Gate

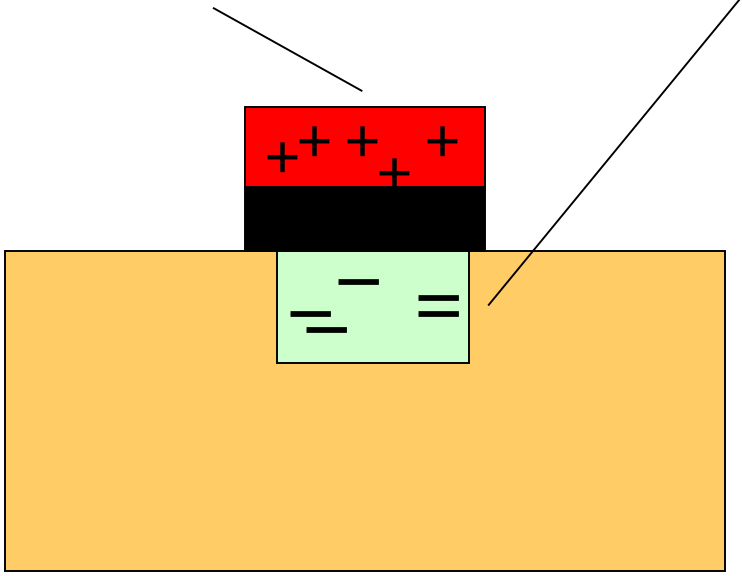


If the gate is not charged, it has no effect on the substrate

Positive Gate, P-type Substrate

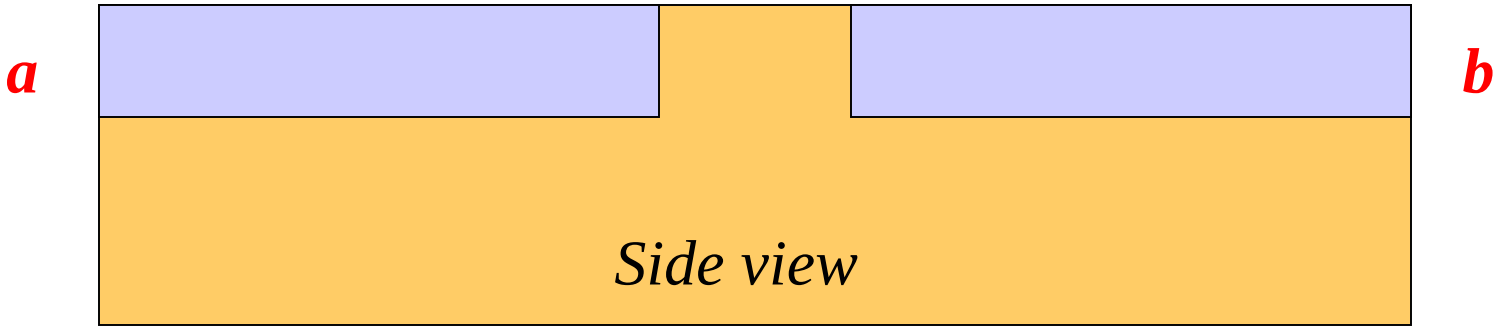
Positive power supply

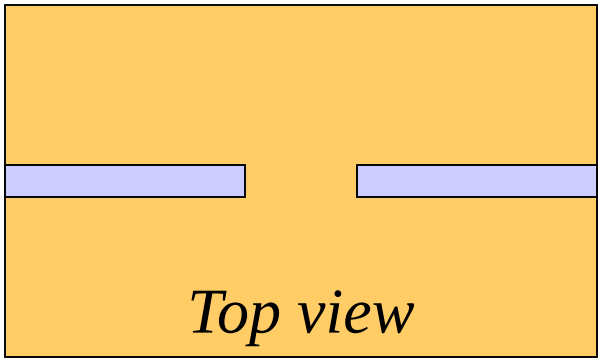
Free electrons in the substrate are attracted to the region under the gate, where they are stuck because of the insulator



If the gate is positively charged, it temporarily transforms part of the P-type substrate into N-type

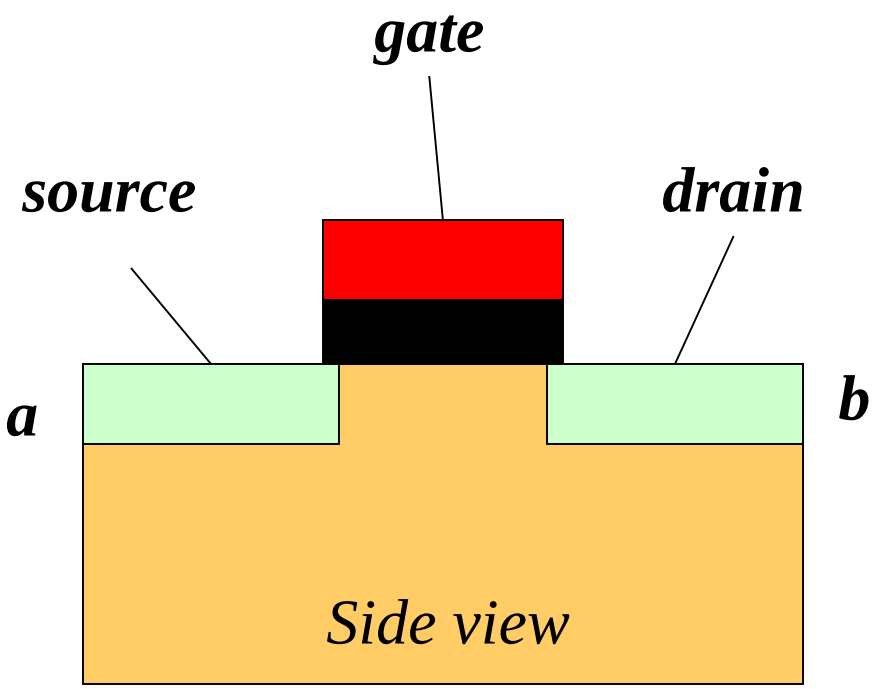
Gap in a Wire





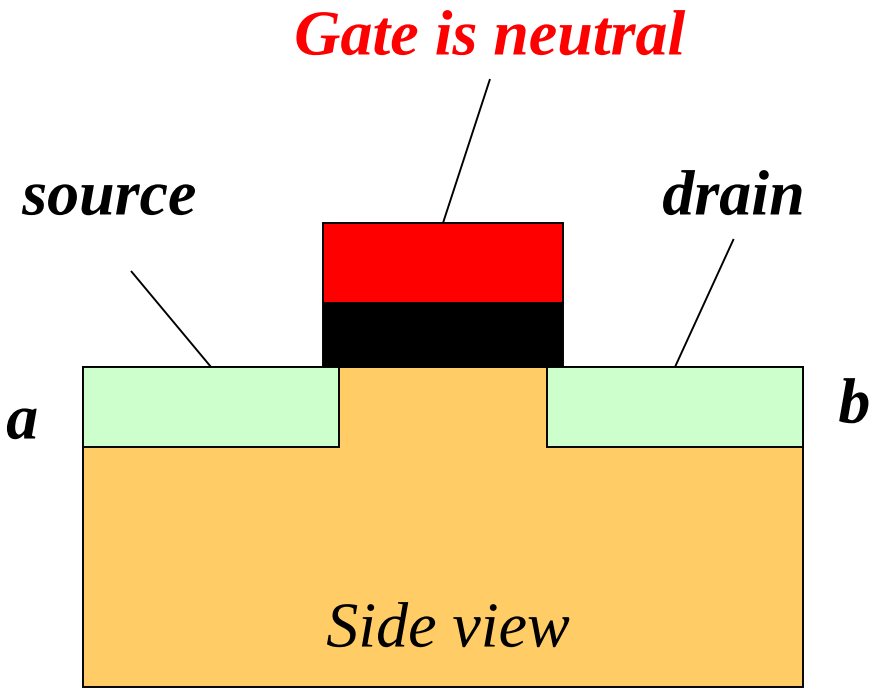
*No current can flow between **a** and **b**, in either direction, since it would have to cross a reverse biased junction*

N-Channel Pass Transistor



A pass transistor consists of a MOS capacitor built right over a gap in a wire

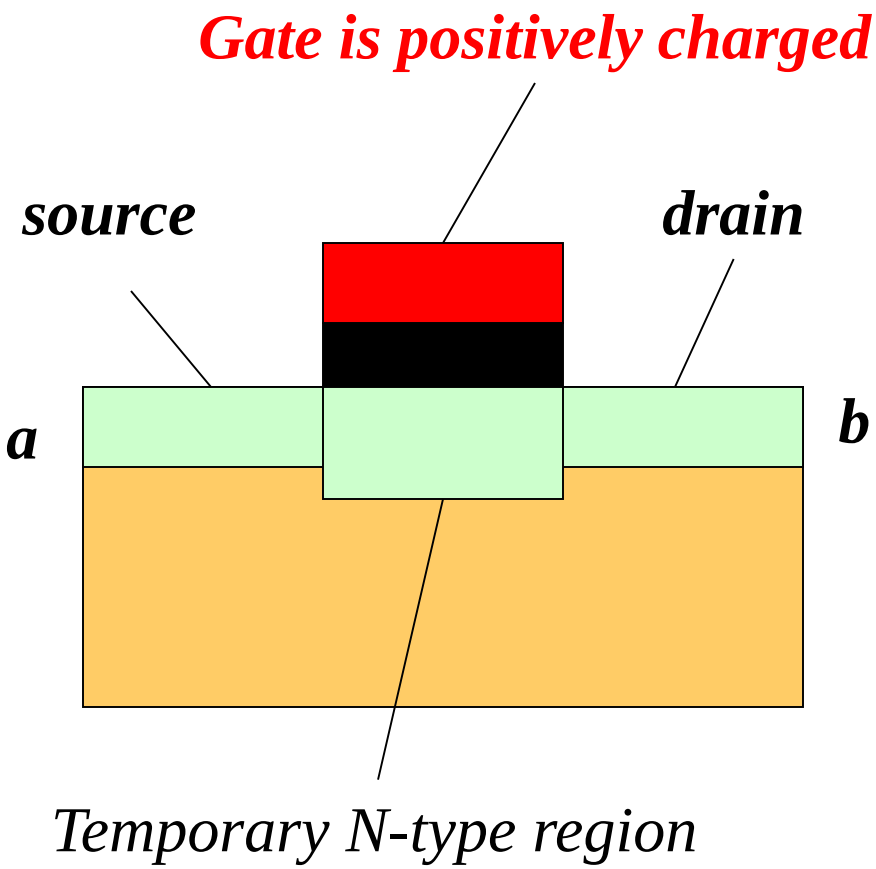
Open N-Channel



If the gate is neutral, so that the capacitor is discharged, then current cannot flow between the source and drain, and we just have a wire containing a gap

The source and drain are disconnected!

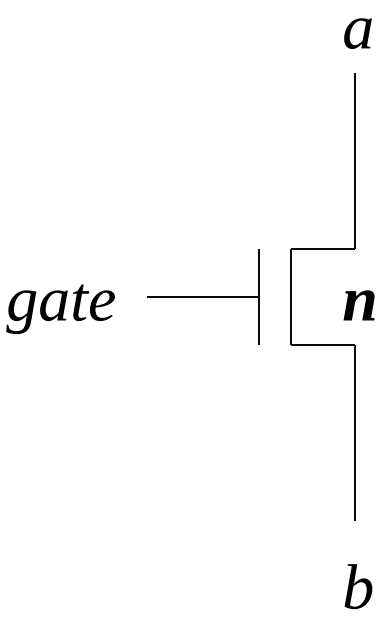
Closed N-Channel



If the gate is positivel, so the capacitor is charged, then the temporary N-type region under the gate closes the connection.

Current flows between source and drain!

A Controllable Switch



The N-channel pass transistor is a switch controlled by the gate

P-Channel Pass Transistors

The same design, with the N/P and –/+ polarity reversed, is a P-channel pass transistor

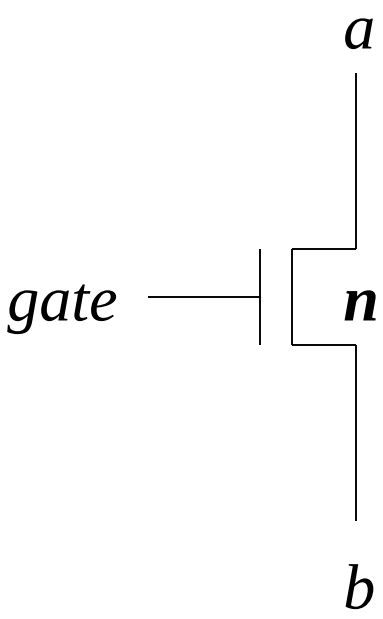
- If the gate is neutral, the wire has a gap
- If the gate is negatively charged, the source and drain are temporarily connected

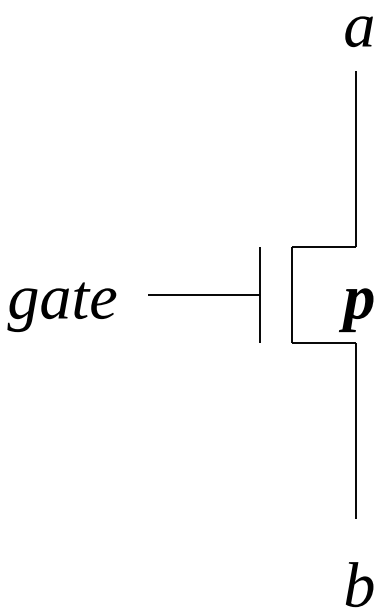
CMOS: Complementary MOS

- A circuit that contains both P-channel and N-channel devices is called CMOS
- Notice that we need both a positive voltage (to control the N-channel transistors) and a negative voltage (to control the P-channel transistors).
- CMOS is currently the dominant technology

Logic Gates

N and P Channel Transistors





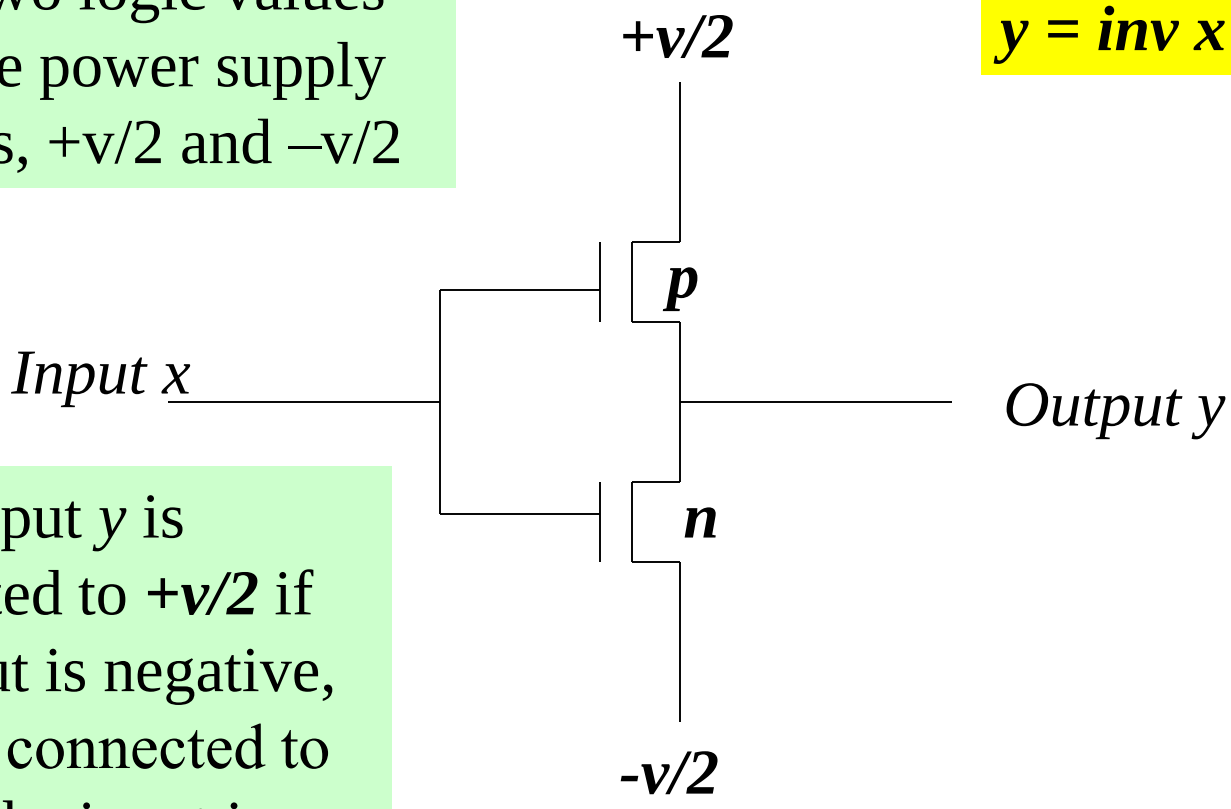
The N-channel transistor makes connection if gate is +

The P-channel transistor makes connection if gate is –

CMOS Inverter

The two logic values are the power supply values, +v/2 and –v/2

y = inv x



The output *y* is connected to ***+v/2*** if the input is negative, and it's connected to ***–v/2*** if the input is positive

Steering Logic

- Using pass transistors to control what connections are made, without always using conventional logic gates like *and2*, *or2*, *xor2*, *inv* etc.
- These logic gates (*and2* etc.) are actually *implemented using steering logic*

Synthesis of Logic Gate Circuits

- The aim: design a circuit that implements a logic function f taking some inputs x, y, …
- The method:
  - Build a network of pass transistors that connect the output to High exactly when f x y = True
  - Build another network that connects the output to Low exactly when f x y = False

Connecting “source” and “drain”

- Suppose x is on the gate of a transistor
- An N channel transistor will connect if *x is high*
- A P channel transistor will connect if *x is low*

Basic Steering Functions

- A logical conjunction (and) is implemented by transistors in series
- A logical disjunction (or) is implemented by transistors in parallel

The nor2 logic gate

A 2-input or gate, with the output inverted

a	b	Nor2 a b
0	0	1
0	1	0
1	0	0
1	1	0

The *nor2* Logic Function

z = nor2 x y

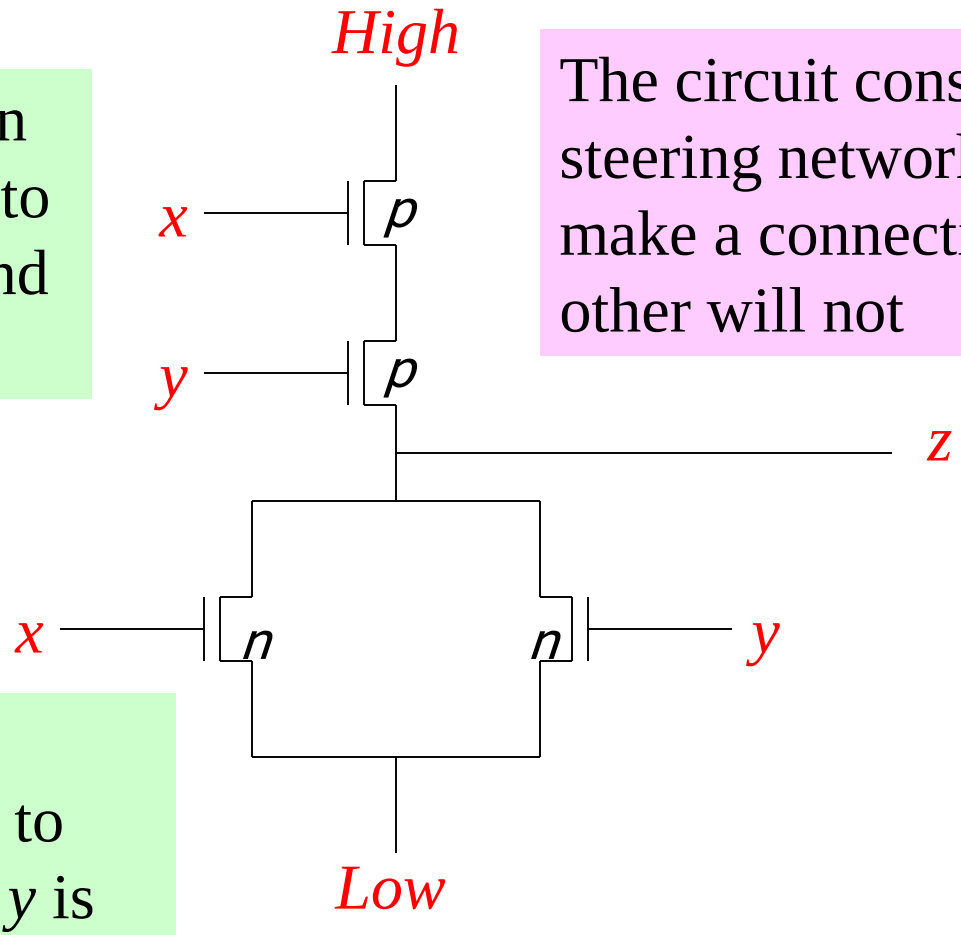
= *True* iff $\neg (x \vee y)$ = $\neg x \wedge \neg y$

= *False* iff $x \vee y$

These two expressions are used directly to construct the steering logic…

Implementation of *z = nor2 x y*

The transistors in series connect *z* to *High* if both *x* and *y* are *Low*



The circuit consists of two steering networks: one will make a connection, the other will not

The transistors in parallel connect *z* to *Low* if either *x* or *y* is *High*

Requirements of a Logic Circuit

- For all inputs, the output must be connected to either High or Low

  Otherwise, the output signal is floating – neither True nor False

- There must not be any combination of input values that cause both networks to close

  This would connect the High and Low power supplies together, resulting in a short circuit!

Integrated Circuits

Integrated Circuits

- We can fabricate many transistors on the same chip of silicon
- We can also fabricate wires connecting them directly on the chip
  - Short wires can simply be N or P type paths that are surrounded by the opposite type
  - Long wires are implemented by placing paths of aluminum on top of the surface

IC Fabrication

- A photographic process – efficient because all the devices on the chip are manufactured in parallel
- The chip is built up in stages (e.g. doping selected regions to change them from P to N type)
- Photoresist is used to mask off the portions that are to be left unchanged

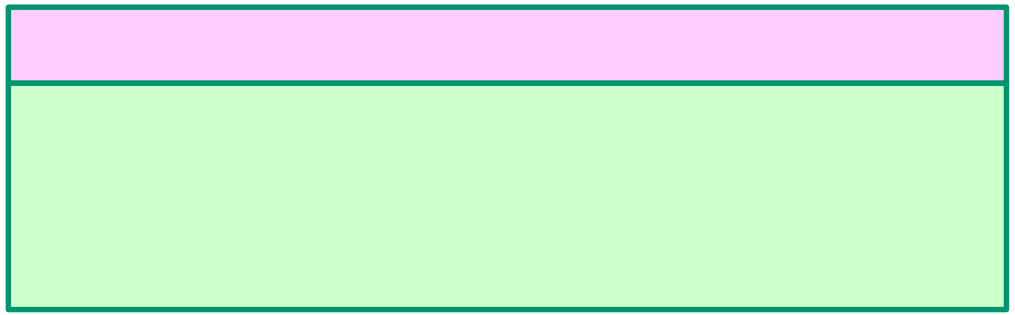
Fabricating one layer: doping

- State of the chip: the surface is P-type silicon
- Aim: change some regions (which form wires) to N-type in multiple sub-steps:
  - Photoresist, expose photoresist to patterned light, dissolve weakened photoresist, expose chip surface to doping agent, dissolve all photoresist

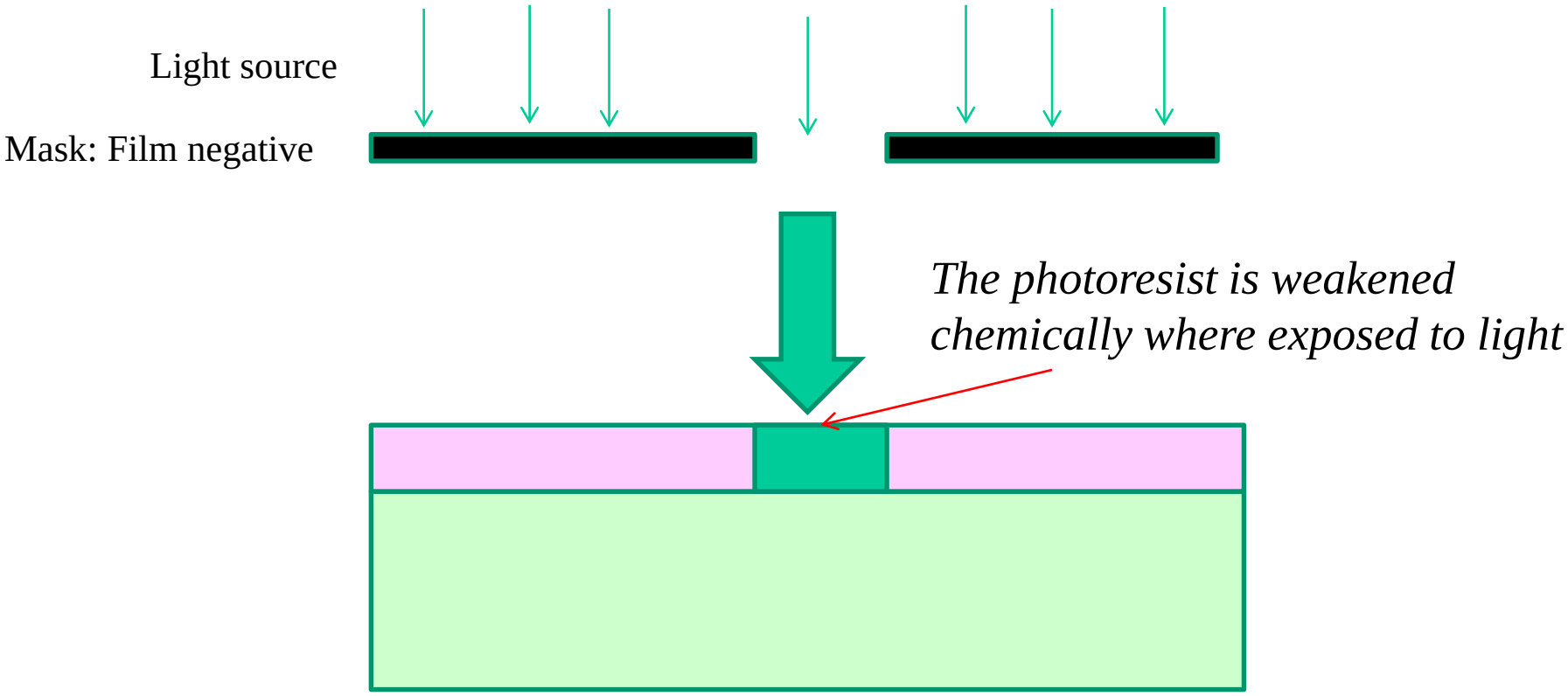
(1) Side view: P-type silicon

(2) Put on photoresist

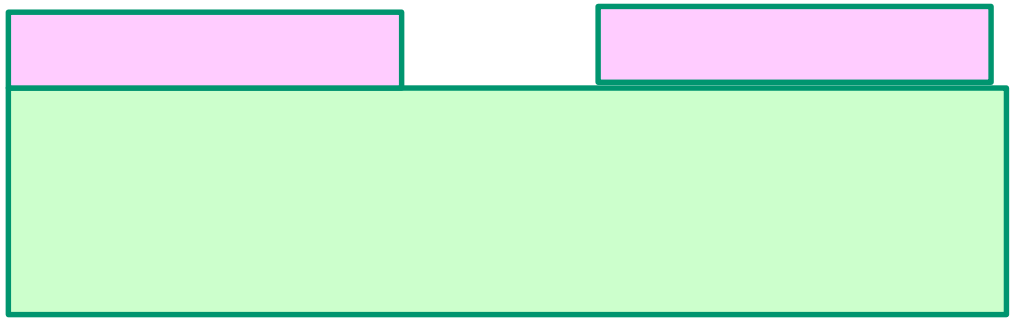
- The chip is immersed in a liquid chemical called photoresist, which coats the surface



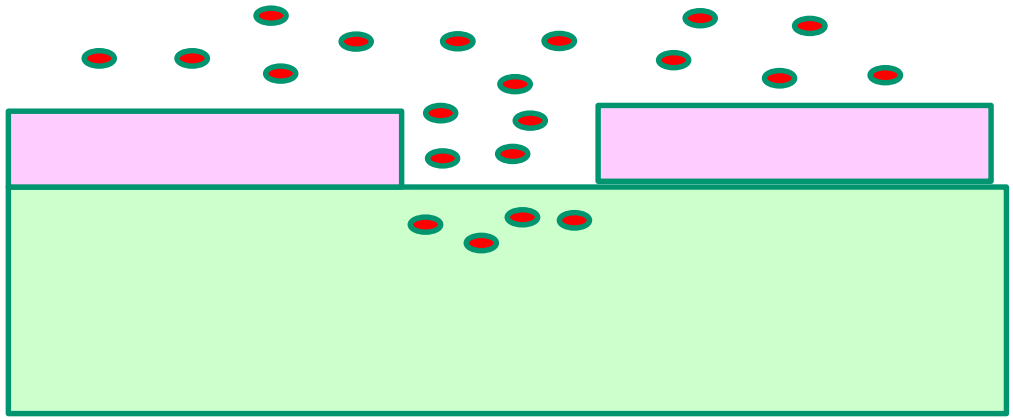
(3) Expose to light through mask (like a photographic negative)



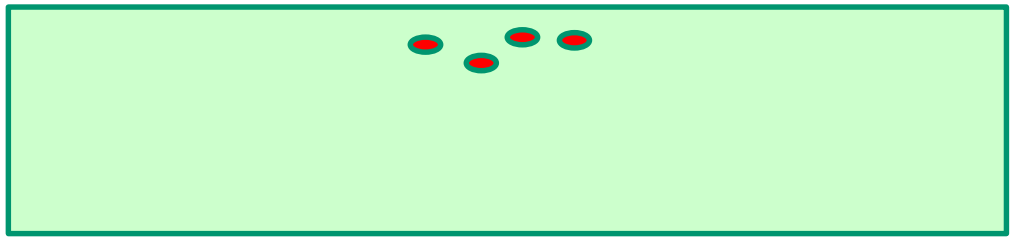
(4) Immerse chip in a moderate solvent to remove the weakened photoresist



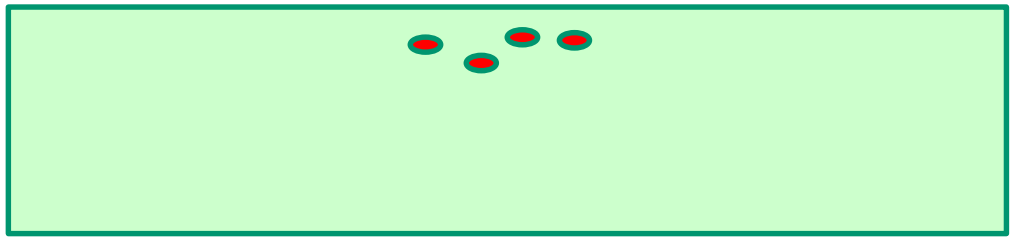
(5) Expose chip to hot gaseous N-type doping agent; some atoms go into the substrate



(6) Use a strong solvent to remove all the photoresist



(7) Result: most of the chip surface is P-type, but a selected region is now N-type



The big picture

- There are many Major Steps (change selected regions to N-type, put on insulator, put on metal layer, …)
- Each Major Step requires a sequence of small steps: photoresist, mask, …
- There may be over 100 steps
- But each step is helping to produce several billion transistors, all at once

Refinements

- The manufacturing process has continually been refined. Example:
  - Some years ago, the feature size became smaller than the wavelength of visible light!
  - Solution: switch to X-rays rather than visible light
- These refinements are the reason Moore's Law has held for so long

Dynamic Memory

Dynamic Memory

An inverter can be used to store a bit:

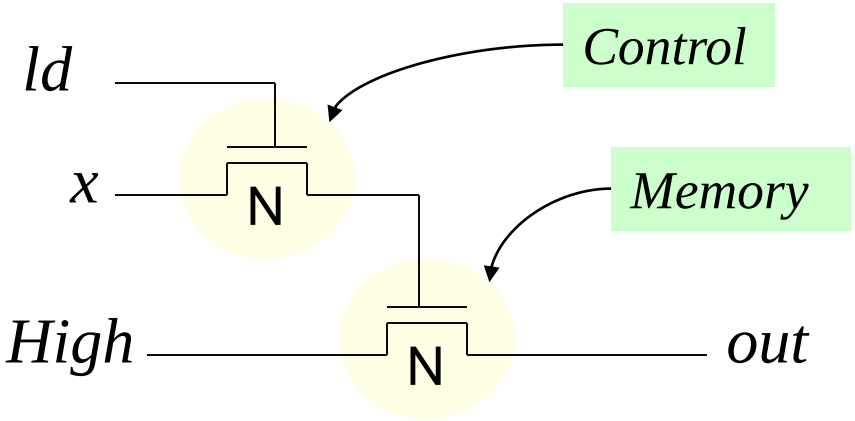
- Connect the gate to either High or Low

  This will either attract charge carriers into the channel, or it will release them allowing the channel to revert to its default type

- Then disconnect the gate

  The state of the channel will stay the same, because of the capacitor effect

Dynamic Register Bit



The stored bit is always available on *out*

When ld is High, the value of x (which must be strong) goes onto the gate of the memory bit. When ld then goes Low, this charge remains isolated (because of the capacitor effect) and the memory remains.

Refresh

- An charge that is isolated on the gate of a pass transistor will gradually dissipate.
- Eventually, the gate will not have a strong enough charge to control the channel, and the register bit become unreadable.
- To prevent this, the bit must be refreshed periodically: the gate must be reconnected to power (low or high) to restore the stored charge to its full strength.

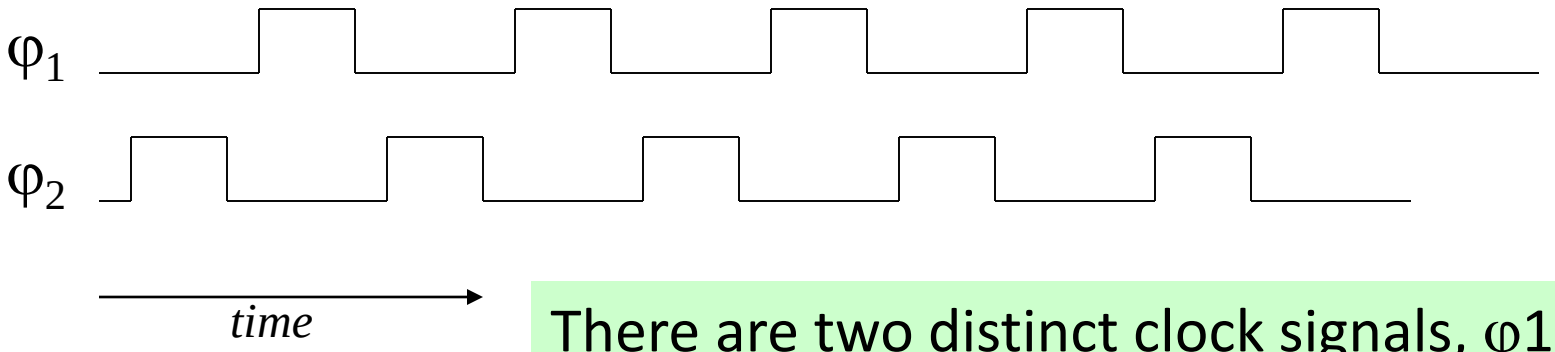
Dynamic RAM

- In DRAM chips, the memory cells are organized as a matrix: a row of columns. The address is used to activate a column, and to select a row to choose the right bit.
- In addition to serving store and fetch requests, the DRAM also does a periodic refresh operation: each bit in a column is read out and put back

Clock Signals

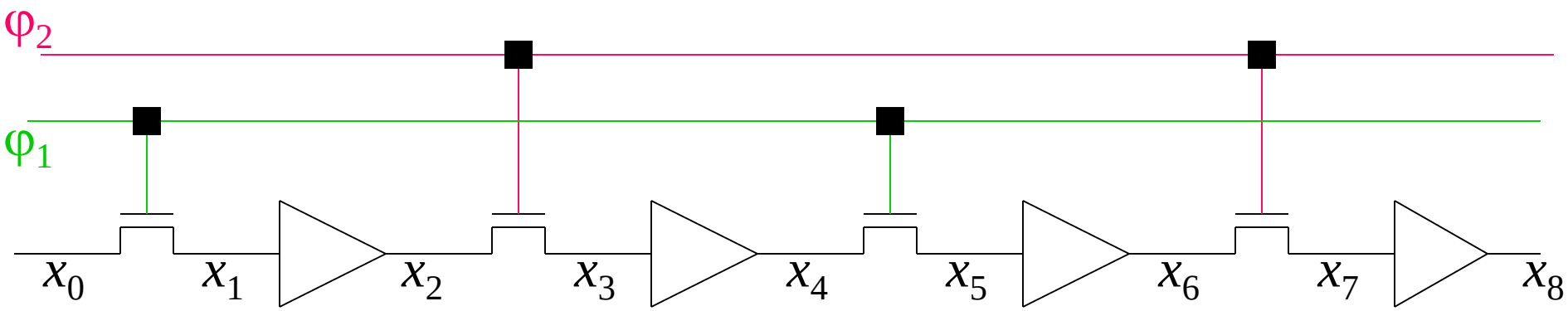
- A clock is a regular sequence of “ticks” that is broadcast globally.
- A clock value will be placed on the gates of pass transistors; during the interval that the clock is High, those transistors will be closed.

The Two-Phase Clock



There are two distinct clock signals, φ1 and φ2, which are never high simultaneously. (These can be generated from a single faster clock.)

Dynamic Shift Register



During φ_1

$x_1 = x_0, \quad x_2 = x_1$

$x_5 = x_4, \quad x_6 = x_5$

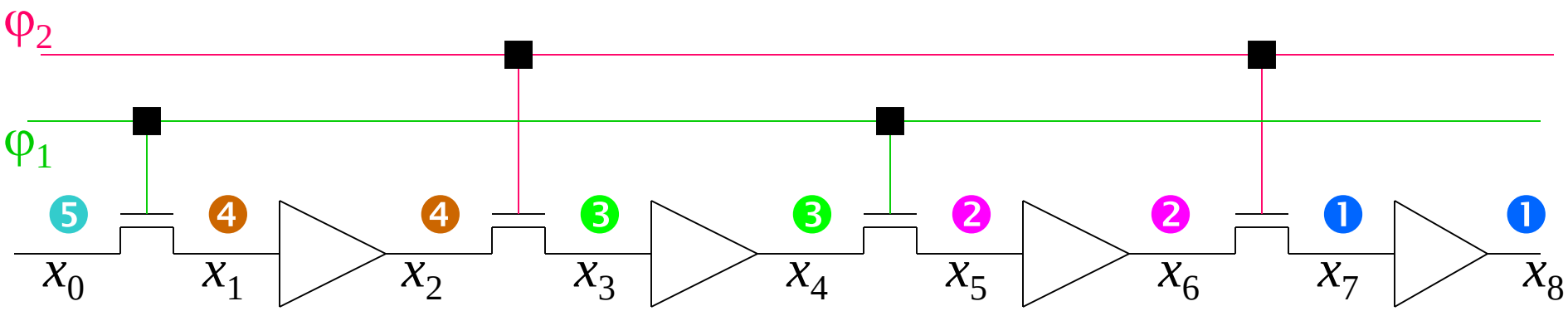
During φ_2

$x_3 = x_2, \quad x_4 = x_3$

$x_7 = x_6, \quad x_8 = x_7$

This circuit consists entirely of combinational components, with no feedback. Yet it has state, and it implements a shift register!

Initial State, with Clocks Low



During φ_1

$x_1 = x_0, \quad x_2 = x_1$

$x_5 = x_4, \quad x_6 = x_5$

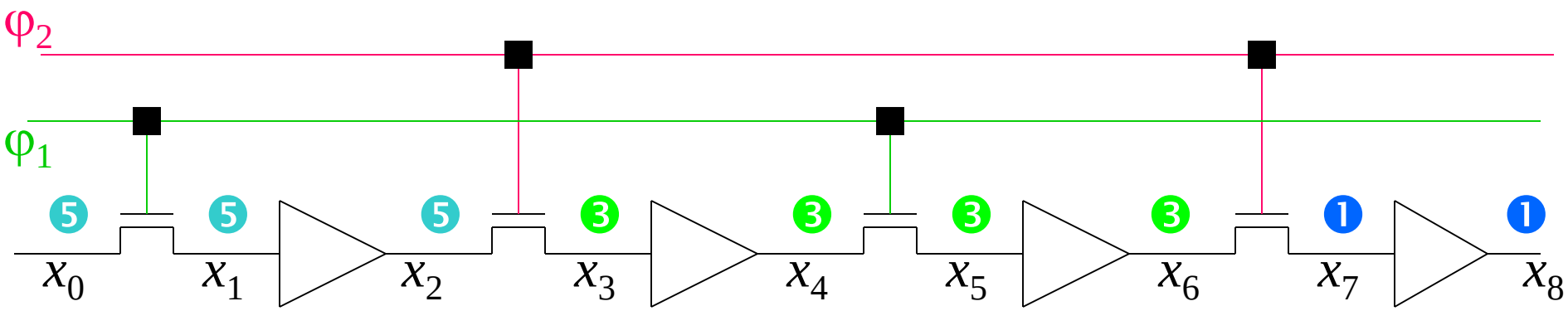
During φ_2

$x_3 = x_2, \quad x_4 = x_3$

$x_7 = x_6, \quad x_8 = x_7$

The charge on the gate x3 is isolated, so the value of x3 remains steady with value ❸ . This also causes the buffer to keep the restored signal x4 steady. Similarly, x1—x2, x5—x6, and x7—x8 remain steady.

Step 1: Clock φ_1 High



During φ_1

$x_1 = x_0, \quad x_2 = x_1$

$x_5 = x_4, \quad x_6 = x_5$

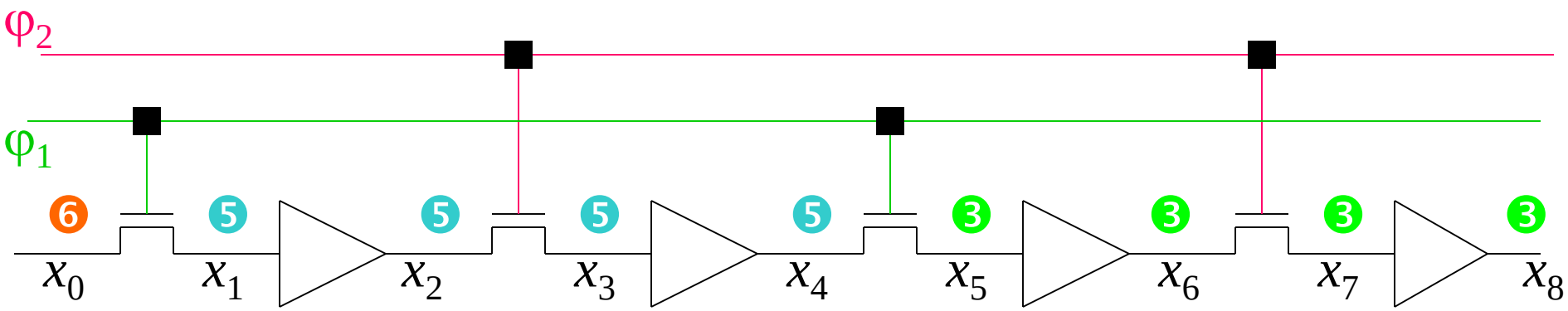
During φ_2

$x_3 = x_2, \quad x_4 = x_3$

$x_7 = x_6, \quad x_8 = x_7$

The value ❺ moves onto x1 and then x2, destroying the old value ❹ . Similarly, ❸ moves onto x5 and x6, destroying ❷, and it remains unaffected on x3—x4, since x3 is still isolated.

Step 2: Clock φ_2 High



During φ_1

$x_1 = x_0, \quad x_2 = x_1$

$x_5 = x_4, \quad x_6 = x_5$

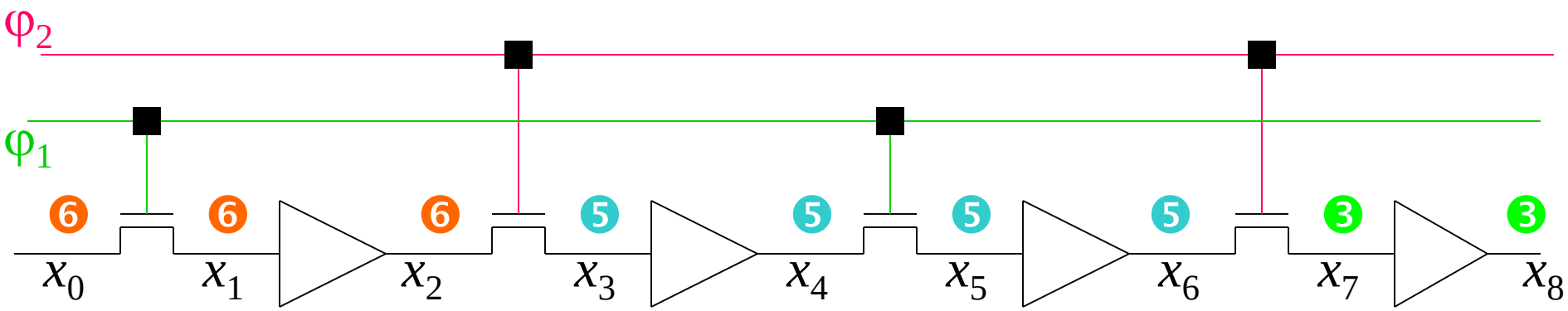
During φ_2

$x_3 = x_2, \quad x_4 = x_3$

$x_7 = x_6, \quad x_8 = x_7$

The value ❺ moves up to x3 and then x4, and ❸ moves onto x7 and x8.. A new input ❻ is arriving.

Step 3: Clock φ_1 High



During φ_1

$x_1 = x_0, \quad x_2 = x_1$

$x_5 = x_4, \quad x_6 = x_5$

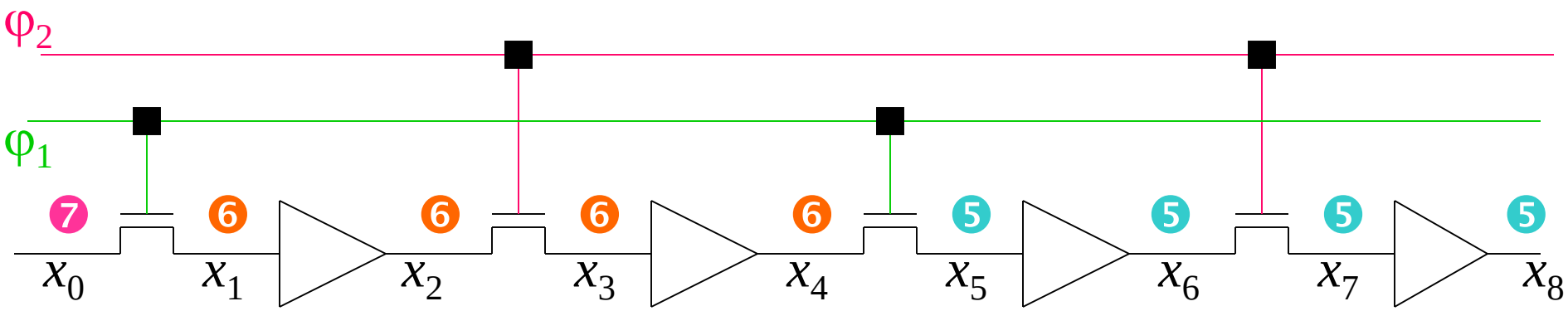
During φ_2

$x_3 = x_2, \quad x_4 = x_3$

$x_7 = x_6, \quad x_8 = x_7$

All the data values are moving along…

Step 4: Clock φ_2 High



During φ_1

$x_1 = x_0, \quad x_2 = x_1$

$x_5 = x_4, \quad x_6 = x_5$

During φ_2

$x_3 = x_2, \quad x_4 = x_3$

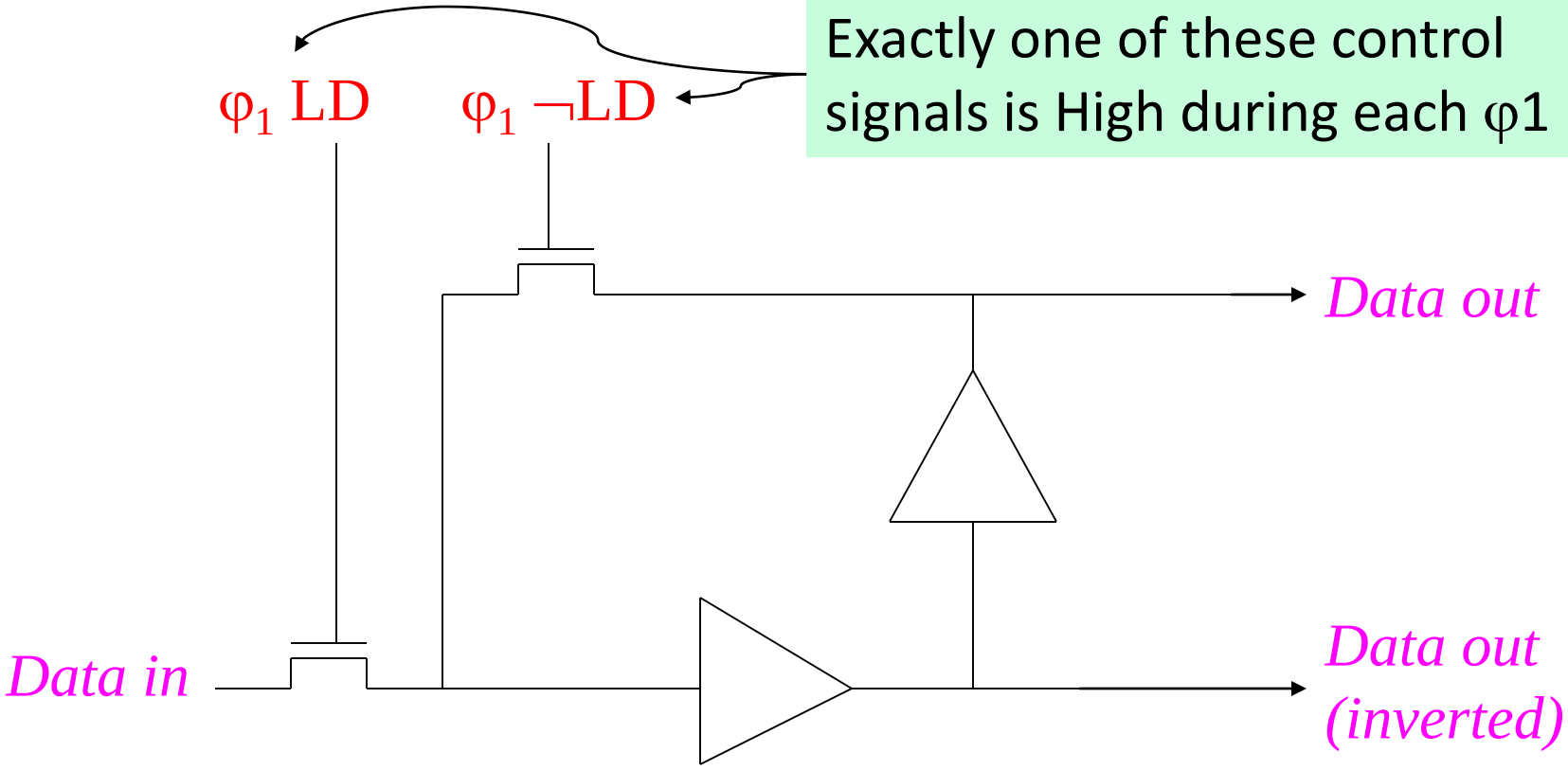
$x_7 = x_6, \quad x_8 = x_7$

All the data values are moving along… and the new input ❼ is arriving.

Pseudo-static Register Bit

- Use capacitor to store the bit, as in shift register – it is fundamentally a dynamic memory device.
- The bit value is refreshed regularly by recirculation, similarly to the dynamic shift register.

Pseudo-Static Register Cell

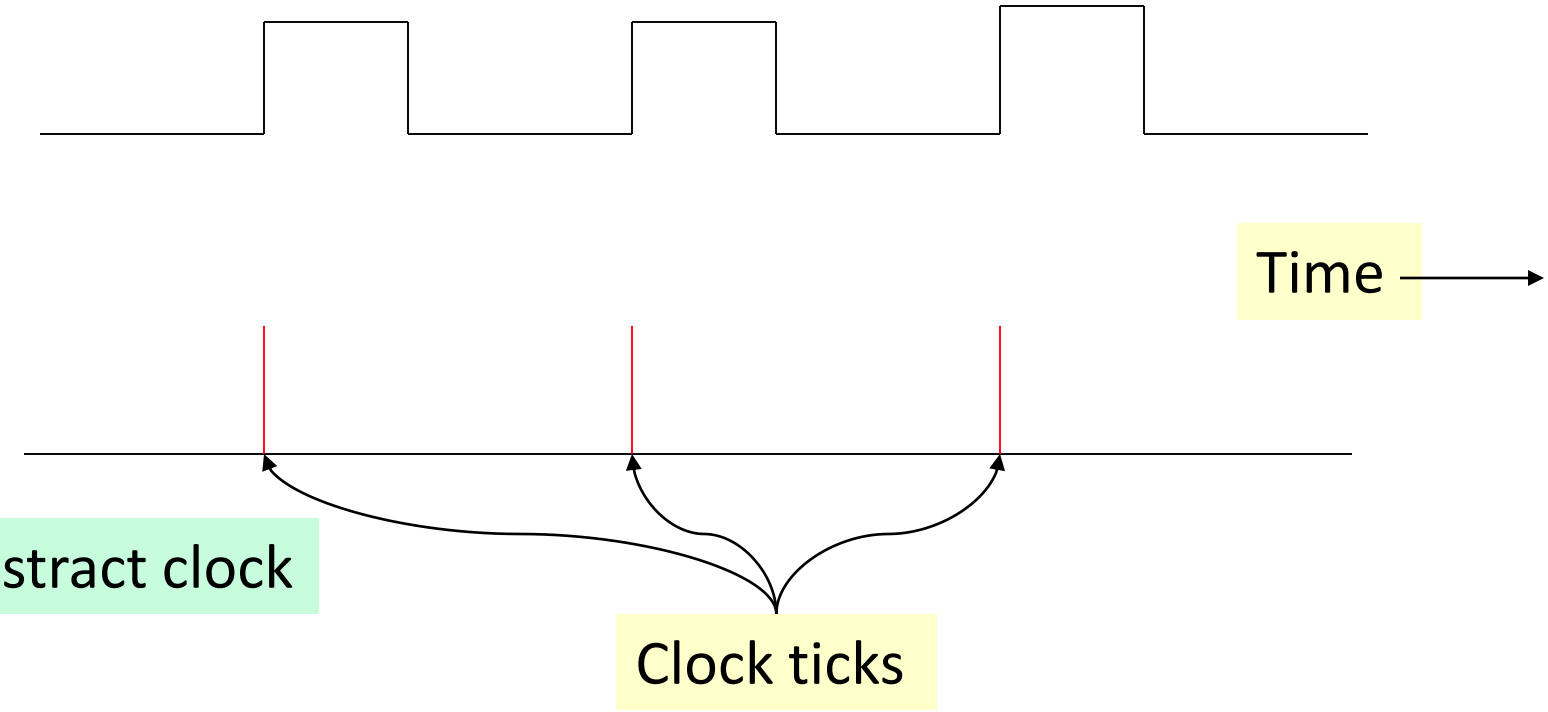


Decay, Refresh, and Feedback

- The quality of a stored bit will decay gradually
- Therefore, all electronic forms of memory require some sort of refresh (static, dynamic, pseudo-static, shift-register, …)
- This is achieved using feedback: the bit must be read out while it's still readable, and a restoring logic device must then put the bit back in full strength

Physical and Abstract Clock

Physical clock—usually multiple phases



Circuit diagram

https://xkcd.com/730/