

Energy and Parallel Computing

Mark Greenstreet and Ian M. Mitchell

CpSc 418 – January 26, 2018



Unless otherwise noted or cited, these slides are copyright 2017 by Mark Greenstreet & Ian M. Mitchell and are made available under the terms of the Creative Commons Attribution 4.0 International license <http://creativecommons.org/licenses/by/4.0/>

Objectives

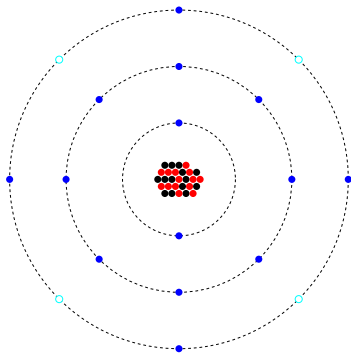
- Understand that parallel algorithms can use less energy than their sequential counterparts.
- Familiar with the technology scaling trends that lead to this.
 - ▶ Where does Moore's Law come from?
 - ▶ What is Dennard scaling (was it first proposed by Hoeneisen & Mead?)
 - ▶ What are energy-time trade-offs for real-world computers
- Aware of how this is likely to impact computing technology in the next decade or so.
 - ▶ Buying computation by the kilowatt-hour
 - ▶ What are the opportunities
 - ★ Domain specific architectures and languages.
 - ▶ Where are exponential improvements in technology happening now
 - ▶ What are energy-time trade-offs for real-world computers
- Aware of how this is likely to impact computing technology in the next decade or so.

Outline

- From silicon atoms to computers.
 - ▶ Dam transistors
 - ▶ How to make a computer
 - ▶ Classical scaling, and why it no longer applies.
- Energy performance trade-offs in real computers.
 - ▶ Going fast takes lots of energy.
 - ▶ Many slow parallel tasks can be more energy efficient than one, fast sequential task.
 - ▶ The case for dedicated co-processors.
- Guessing about the future
 - ▶ Optical technology has a bright future.
 - ▶ Dedicated co-processors means domain-specific architectures and programming models.

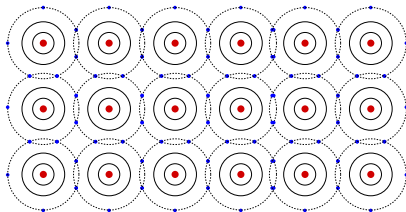
Dam Electronics

Silicon Atoms



- Silicon is atomic number 14: 14 protons ●. The most common isotope has 14 neutrons ●.
- The outermost electron shell (the valence band) has four electrons, and a capacity of eight.
- This is like carbon (atomic number 6)
- Just like carbon, silicon forms covalently bonded crystals, but silicon crystals are better than carbon crystals.
 - ▶ You can make transistors out of carbon crystals, but they aren't as fast or reliable as the ones we make out of silicon.
 - ▶ So, we discard the carbon crystals and sell them under the brand name "diamond".

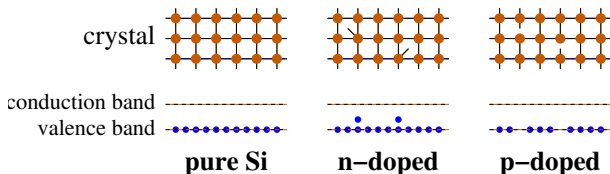
Silicon Crystals



- Silicon crystals are tetrahedral
 - ▶ but I can't draw in 3D;
 - ▶ so, we'll pretend that they make a 2D, square mesh.
 - ▶ This preserves the property that each silicon atom bonds with four neighbours.
 - ▶ And we don't need more details for this quick intro.
- Each valence electron pairs with a valence electron of a neighbouring silicon atom.
 - ▶ All of the silicon atoms have their outermost shells completed.
 - ▶ The crystal structure is very stable.
 - ▶ The electrons have little incentive to move.
 - ▶ Pure silicon is a poor conductor.

Doping

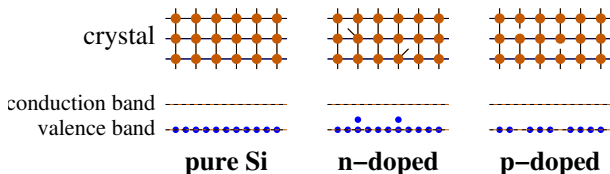
Good for semiconductors, bad for athletes.



- n-doping: add a low concentration of atoms with **5** electrons in their valence shell (e.g. phosphorus or arsenic).
 - ▶ The extra electron is in the “conduction band” and can wander around the crystal.
 - ▶ We now have an electrical conductor.
- p-doping: add a low concentration of atoms with **3** electrons in their valence shell (e.g. boron).
 - ▶ The “missing electron” in the “valence band” and can wander around the crystal.
 - ▶ The missing electrons are called **holes**.
 - ▶ We have an electrical conductor.
- p-doped silicon is a barrier to electron conduction
- Electrons are attracted to positive charges and repelled by negative ones.

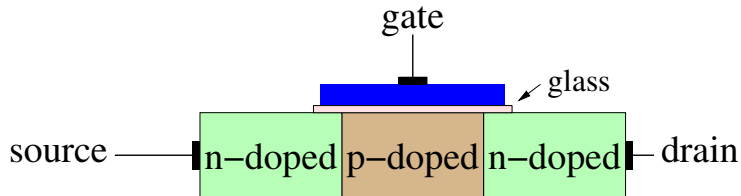
Doping

Good for semiconductors, bad for athletes.



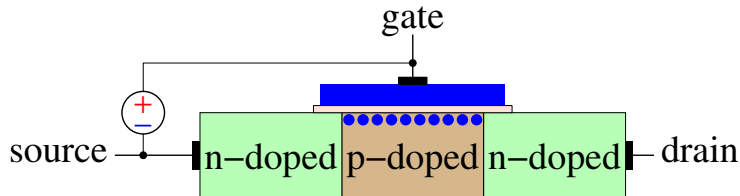
- n-doping: add a low concentration of atoms with **5** electrons in their valence shell (e.g. phosphorus or arsenic).
- p-doping: add a low concentration of atoms with **3** electrons in their valence shell (e.g. boron).
- p-doped silicon is a barrier to electron conduction
 - ▶ Likewise, n-doped silicon is a barrier to holes.
 - ▶ **We have a dam.**
- Electrons are attracted to positive charges and repelled by negative ones.
 - ▶ vice-versa for holes
 - ▶ **We have a way to raise and lower the dam.**

Transistors (i.e. MOSFETs)



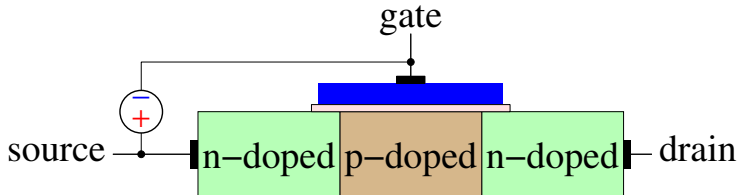
- The MOSFET starts with a $N \rightarrow P$ junction followed by a $P \rightarrow N$ junction.

Transistors (i.e. MOSFETs)



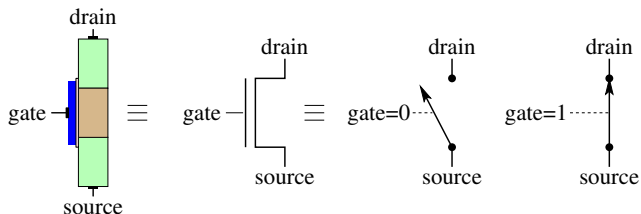
- The MOSFET starts with a $N \rightarrow P$ junction followed by a $P \rightarrow N$ junction.
- Add a “gate” electrode over the middle p-doped region.
 - ▶ The gate is separated from the rest of the transistor by a thin insulator (glass).

Transistors (i.e. MOSFETs)



- The MOSFET starts with a $N \rightarrow P$ junction followed by a $P \rightarrow N$ junction.
- Add a “gate” electrode over the middle p-doped region.
- If a positive voltage is applied to the gate:
 - ▶ Electrons that entered the p-doped region from the source are attracted towards the gate.
 - ▶ The holes in the channel are driven repelled down.
 - ▶ This allows the electrons to form a conducting bridge from the source to the drain.
 - ▶ Current can flow between the source and drain.

Transistors (i.e. MOSFETs)



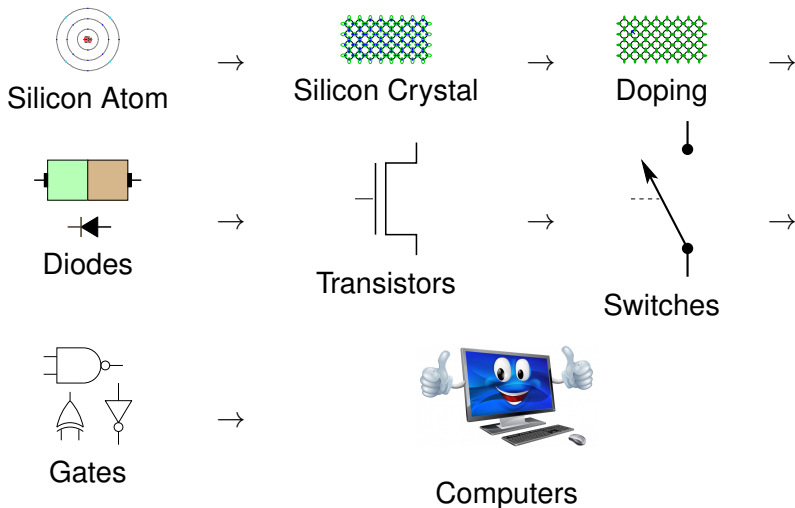
- The MOSFET starts with a N→P junction followed by a P→N junction.
- Add a “gate” electrode over the middle p-doped region.
- If a positive voltage is applied to the gate:
- If a negative voltage is applied to the gate,
 - ▶ There is no connection between the source and drain, and no current flows.
- The MOSFET is a voltage-controlled switch.

Transistors (i.e. MOSFETs)

- The MOSFET starts with a $N \rightarrow P$ junction followed by a $P \rightarrow N$ junction.
- The MOSFET is a voltage-controlled switch.
- Now that we have a switch, we can build gates.
- With gates, we can build computers.
-

Gates

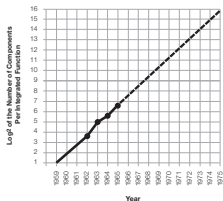
Semiconductor Summary



Computer image from

http://seniorsnoworlando.org/wp-content/uploads/2014/05/IMG_0009-1038x576.jpg

Moore's Law



- Moore's Law (original): the number of transistors on a chip will double every year from 1965 through 1975.
- Justification
 - ▶ Moore took four data points and found they could be fit reasonable well with a line on a semi-log plot. 😊
 - ▶ More seriously, Moore observed that
 - ★ Putting more transistors onto a chip allowed you do build new kinds of electronic devices.
 - ★ There would be a large market for these devices.
 - ★ The profits made from selling the chips would allow semiconductor companies to improve their manufacturing processes.
 - ★ Transistors would shrink a lot, chips would get bigger.
 - ★ Moore extrapolated until 1975 because the various technical challenges seemed solvable given plausible estimates of sales an profit.

Moore's Law – Beyond 1975

- Moore's law has enjoyed many extensions as key manufacturing issues were solved.
- The rate has gradually slowed from doubling every year to doubling every 3 or 4 years.
- Power blocked clock frequency from scaling with transistor size from roughly 2003 and beyond.
- There is a limit to scaling
 - ▶ Current products in transistors with 14nm channel length (the thickness of the “dam”).
nm = nanometer = 10^{-9} meter.
 - ▶ Chip designer working on designs with 7nm channel length.
 - ▶ Shrinking to 5nm or 3.5nm looks really difficult.
 - ▶ The spacing of silicon atoms in a silicon crystal is around 0.3nm.

Denard Scaling

What happens if we scale transistor dimensions and operating voltage by a factor λ ?

- E.g. $\lambda = 0.5$ is shrinking everything to half its previous size.
- Gate delay scales as λ .
- Clock frequency scales as $1/\lambda$.
- Energy per signal transition scales as λ^3 – this is amazing!
- Power is $\frac{\text{energy}}{\text{time}}$. Power scales as λ^2 .
- Number of devices on a chip scales as λ^{-2} .
- Power density (i.e. watts per square centimeter) is constant.
- **Conclusion:** everything gets **way** better as we shrink transistors.
 - ▶ Of course, this requires very precise manufacturing, so it took many rounds of the Moore's Law positive feedback cycle to get to where we are today.

What went wrong: The Power Wall

- To disconnect the source from the drain of the transistor, the “dam” must be above the level of the upper reservoir.
- But, the reservoirs have “waves”
 - ▶ The waves are the thermal energy of the electrons.
 - ▶ To turn off a transistor, the dam needs to be about $10\times$ higher than the average wave.
 - ▶ The dam height can be at most $\sim 40\%$ of the operating voltage.
 - ▶ This sets a lower bound for operating voltage (at room temperature) of about $0.6V$.
- Voltage hasn't scaled as predicted by classical scaling since the early 1990's.
 - ▶ Chips are faster than they should be by Denard scaling. 😊
 - ▶ They are also **way** hotter. 😞

Power is the Primary Design Concern

- In the old days, processors were designed primarily for speed.
- Now, they are designed to satisfy a power requirement.
- This impacts all forms of computing:
 - ▶ mobile devices and battery life
 - ▶ desktop devices and gaming consoles are limited by cooling
 - ▶ data centers and cloud services are limited by building cooling.
 - ★ The power bill is a major part of the operating expenses for cloud services.
 - ★ Indirectly, cloud users are buying computation by the kilowatt hour.
 - ★ Although the power bill is indirect in the billing, the financial consequences are very real.

Energy time trade-offs in real life

- The tradeoff that $E \propto T^{-2}$ from the text assumes classical scaling.
 - ▶ We can't push the operating voltage as low as assumed by such scaling laws.
 - ▶ Emperically, we get $E \propto T^{-1}$ through a combination of voltage scaling, circuit design, and architectural tradeoffs.
- Parallel computing can still be a big-win for saving energy
 - ▶ Let's say we can build processors that run $\frac{1}{10}$ the speed of a fast sequential machine. They will each use $\frac{1}{100}$ of the power.
 - ▶ If a parallel version of the computation gets perfect speed-up, we can run it on 10 slow processors in the same time as running the sequential code on one fast processor.
 - ▶ The parallel version will use $\frac{1}{10}$ of the energy.

Where does the energy go

- For a general purpose processor: instruction fetch, decode, and other control.
- For a GPU: register file accesses.
- Compared with full-custom hardware:
 - ▶ A CPU can be $1000\times$ less energy efficient.
 - ▶ A GPU can be $100\times$ less energy efficient — that's better than a CPU, but there is still plenty of room for improvement.
- The factor of $100\times$ energy waste of current architectures is begging for the next breakthrough.
 - ▶ What will that breakthrough be?

What went wrong: The Atom Wall

- Chips are now being designed where the gate length (i.e. dam thickness) is about 20 atoms.
- We need to squeeze a low concentration of dopant atoms into the channel.
- It's very hard to manufacture circuits where a few atoms makes a big difference.
 - ▶ All edges are jagged.
 - ▶ Photo-lithography (printing the circuit structures with light) is challenging because the transistors are much smaller than a wavelength of the UV light that is used.
 - ▶ Quantum mechanics becomes a big deal.
 - ▶ ...

What's next? (part 1)

- Parallel computing: how to make good use of Moore transistors without using more power.
- Optics:
 - ▶ Computer performance is often limited by chip-to-chip interconnect, e.g. the connection between a CPU and memory.
 - ▶ Glass is **much** better than copper.
 - ▶ Optical networking is standard in large data centers.
 - ▶ Optical interconnect between chips is emerging – there are clever ways to make modulate and detect light beams with silicon.
 - ▶ Wavelength-division multiplexing (WDM) is awesome – we can have hundreds of simultaneous channels on a single glass fibre by using different wavelengths of light.

What's next? (part 2)

- Higher bandwidth channels to memory
 - ▶ GPUs now use HBM and HBM2.
 - ★ This involves stacking 16 or 18 memory chips next to the GPU.
 - ★ The memory chips are connected to each other by polishing each chip down to a few tenths of a millimeter thick, etching holes in the chip, filling the holes with metal, and making connections.
 - ★ This allows $10\times$ the number of connections between the memory chips and between the memory and the GPU.
 - ▶ Cryogenic memory?
 - ★ I've read recently about a joint project between Microsoft and Rambus to look at memory that runs in liquid nitrogen.
 - ★ Silicon in liquid nitrogen has wonderful electrical properties – the waves are much smaller.
 - ★ **But**, making reliable systems has been a show-stopper because wires become extremely brittle.
 - ★ I haven't seen how Microsoft and Rambus plan to address this.
 - ▶ Nanotubes, graphene, spintronics, molecular computing, quantum computing
 - ★ Many long-shots are being explored.