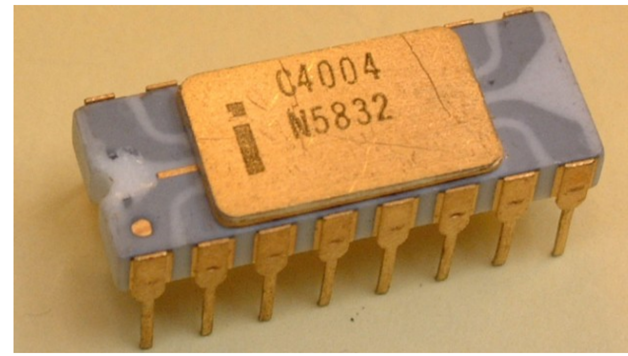
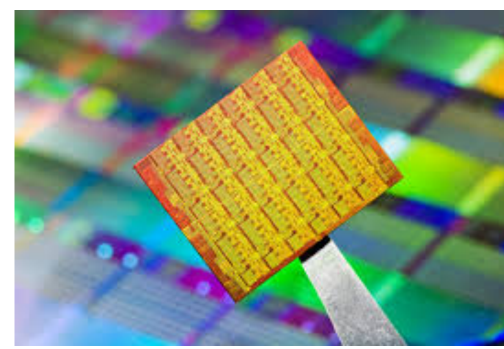
The Future of Computing
Towards the Post Moore's Law Era

Stefano Markidis
KTH Royal Institute of Technology

Processor Evolution

- In 1971 a small company, called Intel, released the 4004, its first microprocessor:
  - The 12 mm^2 chip contained 2,300 transistors (switches representing 0 and 1). The gap between the transistors was 10,000 nm (about as big as a red cell)
- Today, the latest Intel chip is Skylake:
  - The size is ten times the 4004 but at the space of 14 nm (invisible to light miscroscope)
- **Is there a law that relates the 4004 and the Skylake processor?**





Moore’s Law - 1965

- “The complexity for minimum component costs has increased at a rate of roughly a factor of two per year. Certainly over the short term this rate can be expected to continue, if not to increase.”

Electronics Magazine, *Cramming more components onto integrated circuits*, 1965

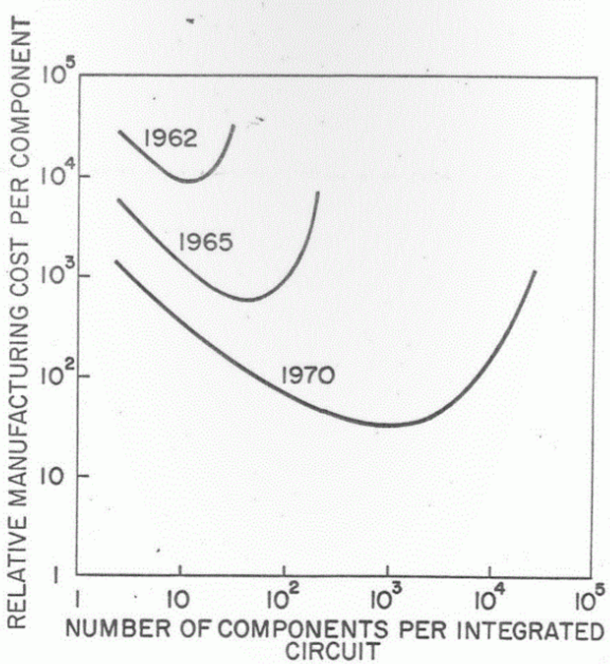
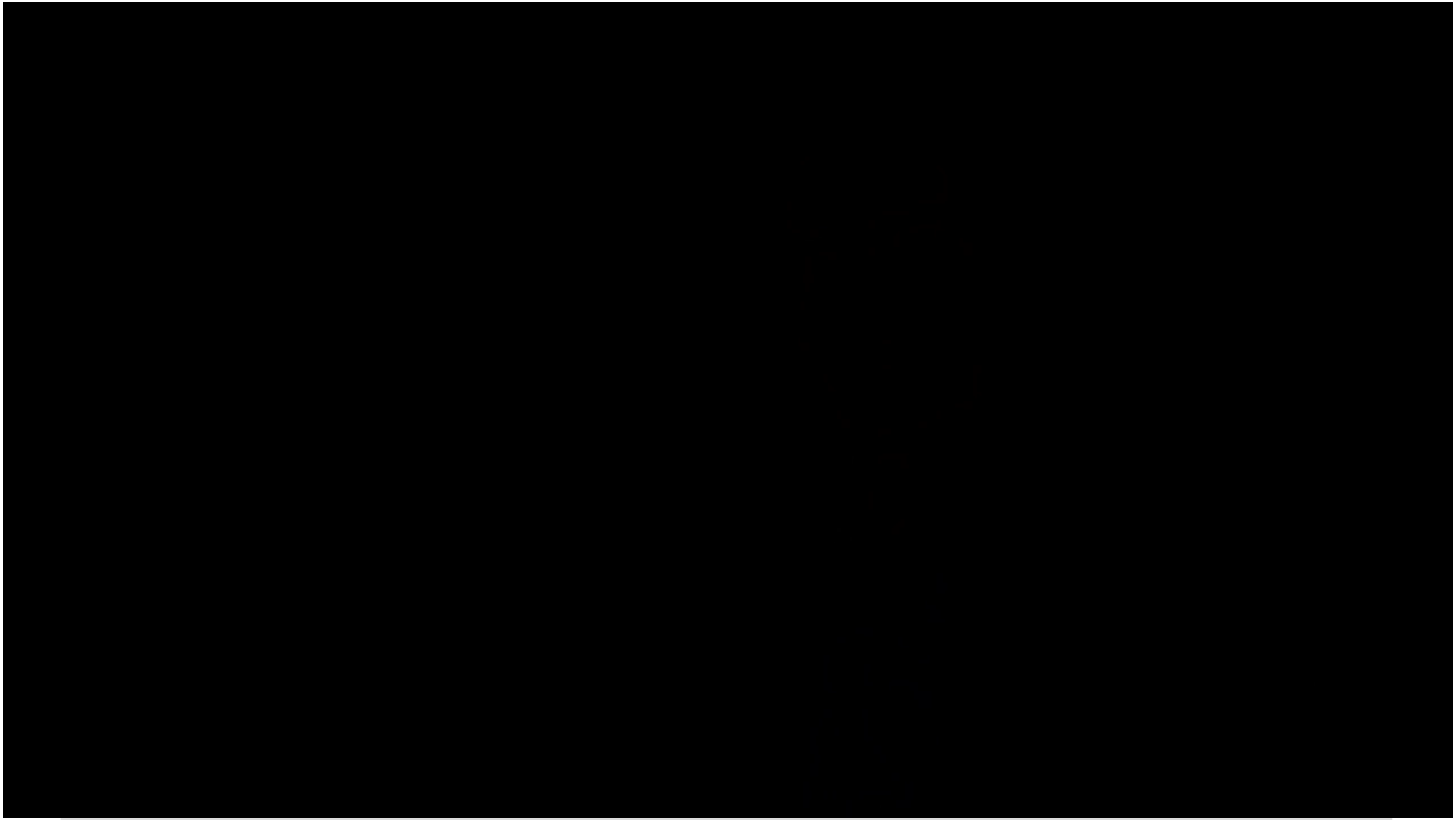


Fig. 1 Estimated relative cost per component vs complexity for a typical integrated function for three different times.

Moore’s Law in Action

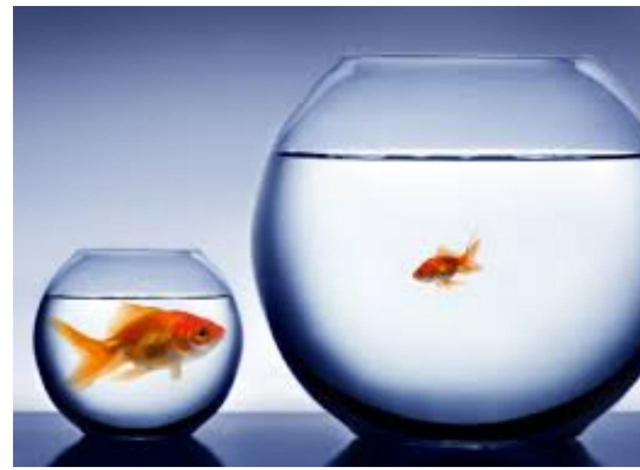


https://www.youtube.com/watch?v=T6UlUA8jU48

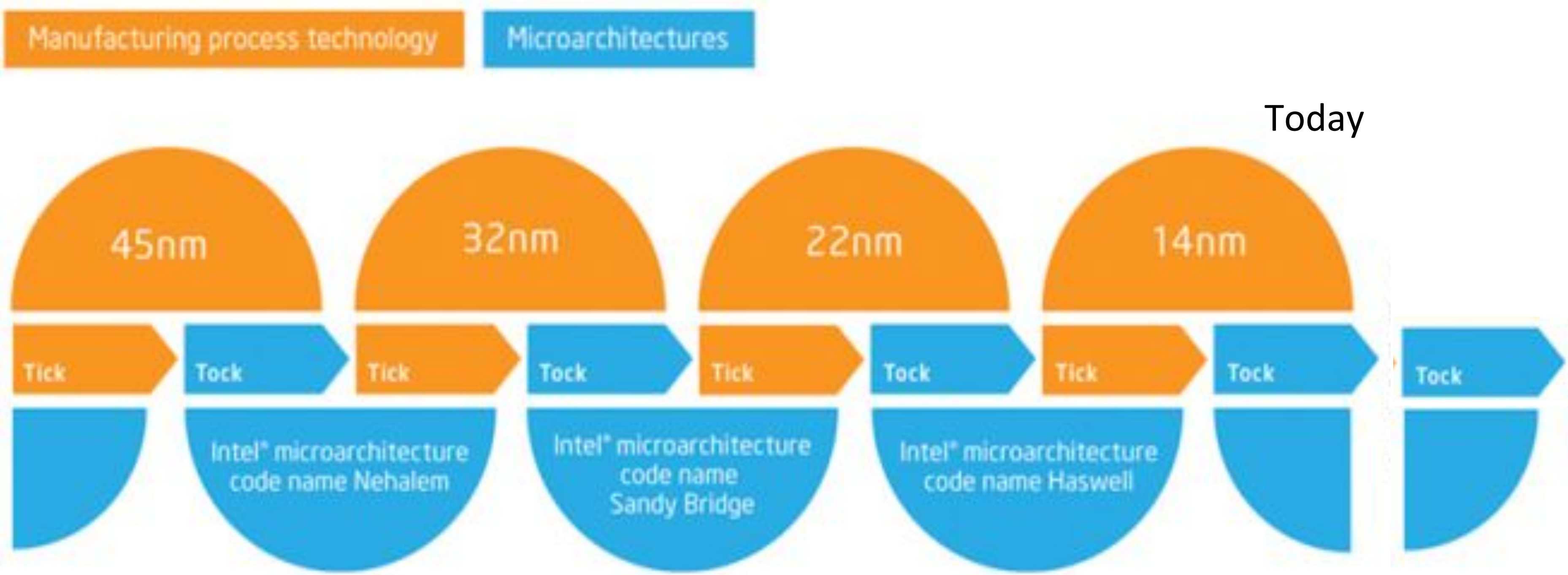
How does Moore's Law Work?

Transistors have the quality of **getting better as they get smaller**

- A smaller transistor can be turned on and off **with less power** at **a greater speed** than a larger one.
- This means that you could use more and faster transistor **without using more power** or **generating more waste heat**
- Thus the chips could get denser as well as better



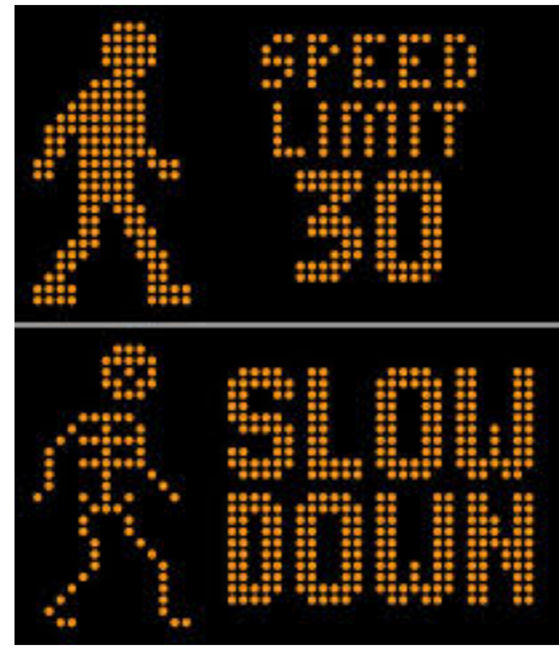
First Cracks in Moore's Law



- Intel used to release chips on a **tick-tock schedule** = every the other chip would use a **new manufacturing process**.
- Since 2006, Intel moved from 65→ 45 → 32 → 22 → 14 in this fashion.
- Intel released 14nm Broadwell (tic) and Skylake (toc) chips. The upcoming Kaby Lake processor will be still a 14nm processor, so we are having a tic toc toc schedule.

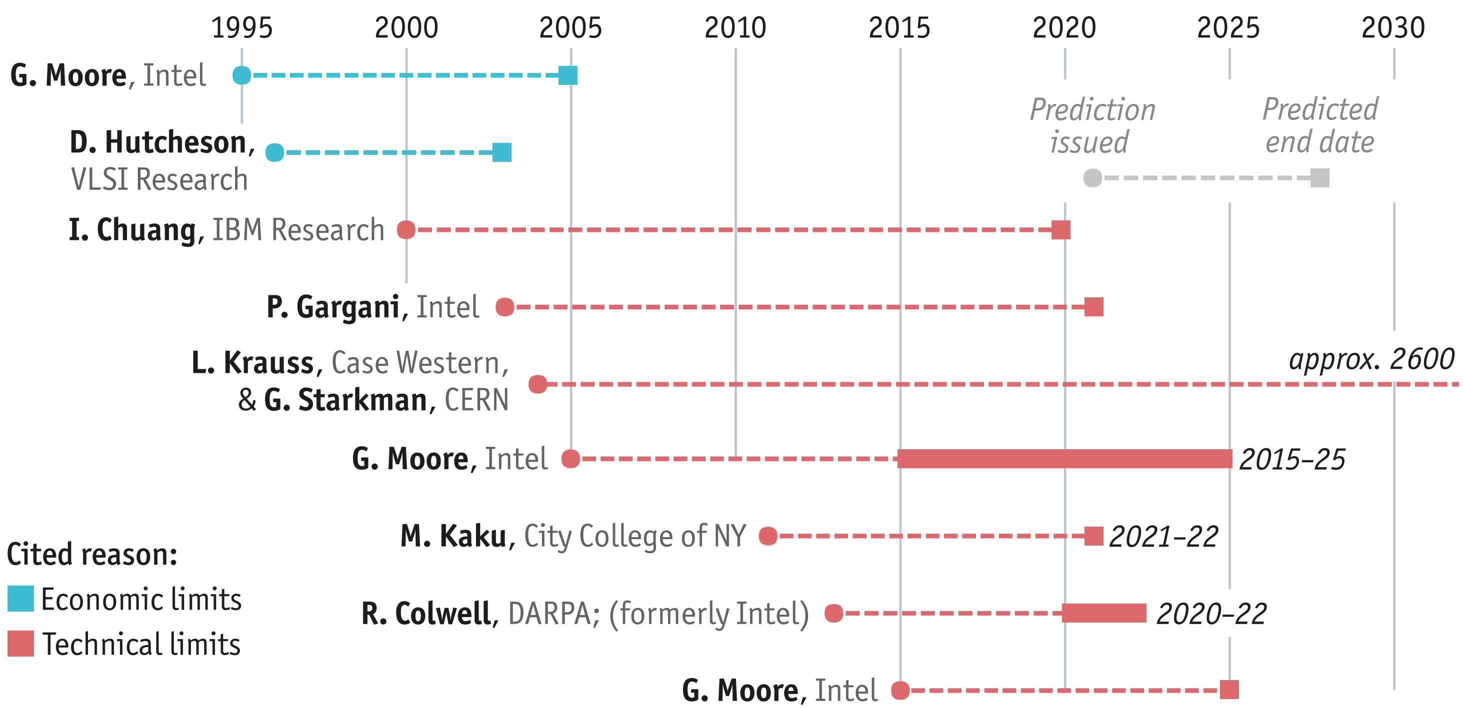
Why Moore's Law is Slowing down?

- For sometimes now, **making transistor smaller does not make them more energy-efficient.** Thus the operating speed of high-end chips has been on a plateau since mid-2000s.
- While the benefits of making things smaller have been decreasing, **the cost has been rising**.



The Law of Moore's Law

"The number of people predicting the death of Moore's law doubles every two years." – *Peter Lee VP MS research*



Sources: Intel; press reports; *The Economist*

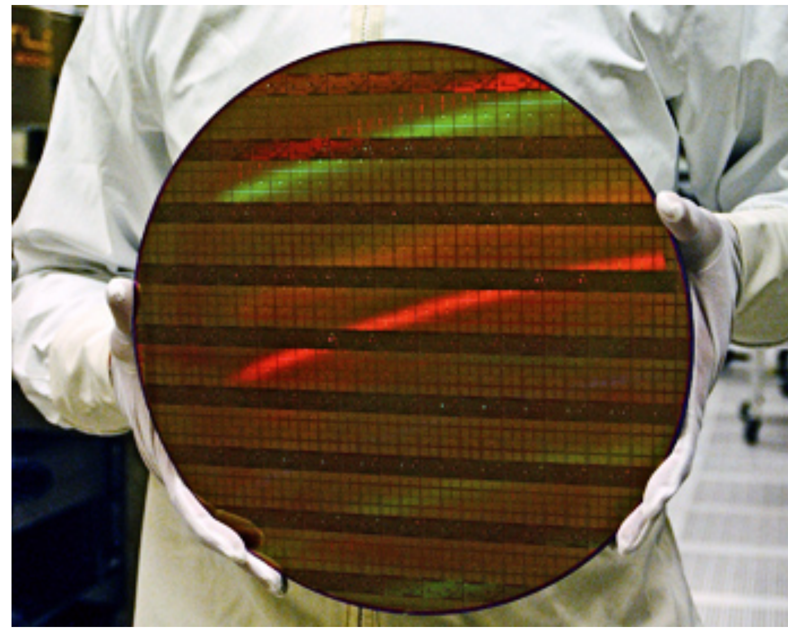
Physical Limits to Moore's Law

- The components are approaching a fundamental limit of smallness: the atom.
  - A Skylake transistor is around 100 atoms across.
  - For smaller transistor you need trickier designs and extra materials.
- And as chips get harder to make, fabs (semiconductor fabrication plant) get even more expensive
  - to make 5nm chips would cost 1/3 of Intel's current annual revenue.

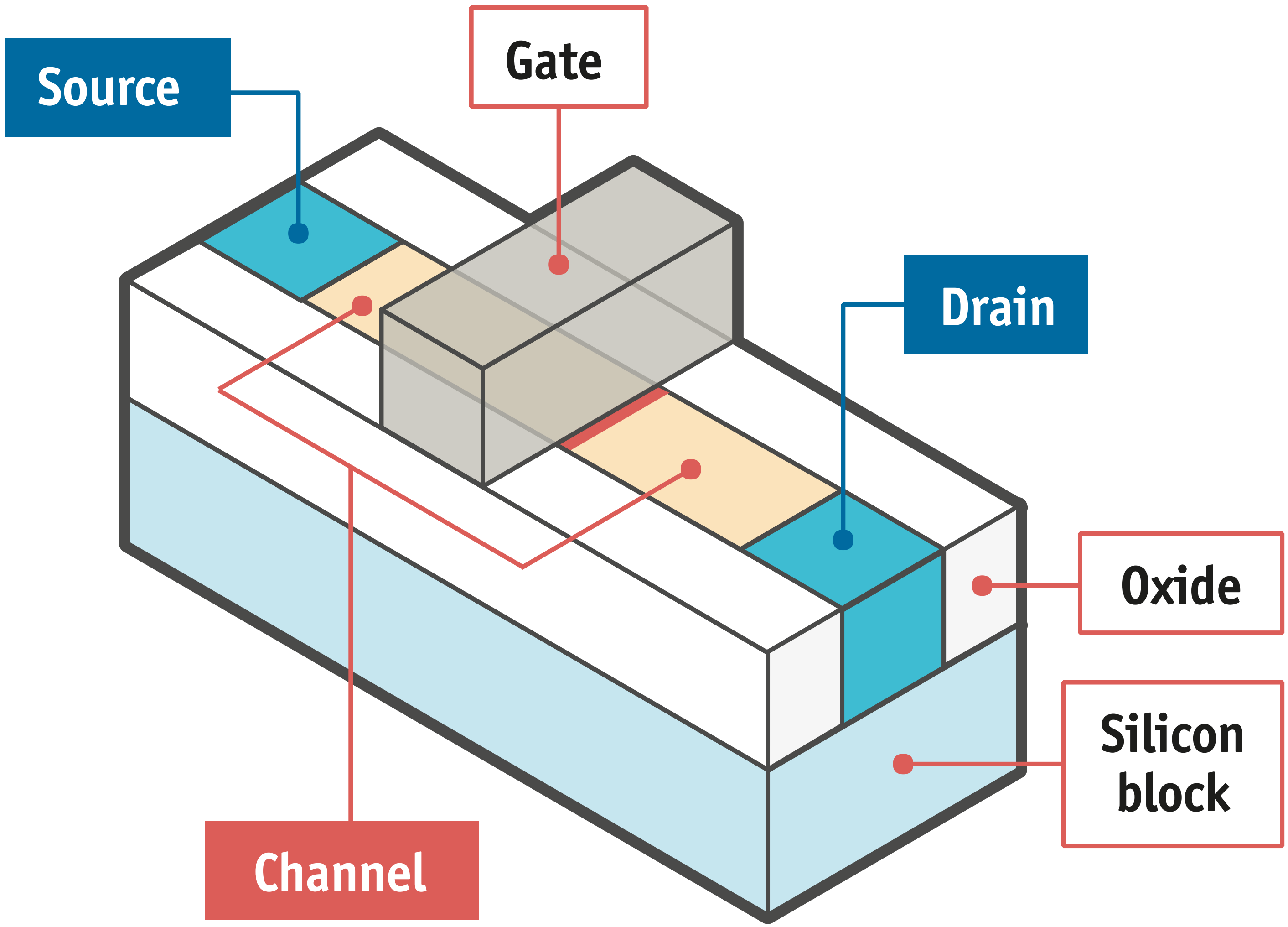
The Manufacturing Wall

Manufacturers will be able to produce chips on the 10-nanometer manufacturing process, expected to arrive in **2018**, and maybe one manufacturing process after that, **but that's it**!

From an economic stand point, Moore's law is over.

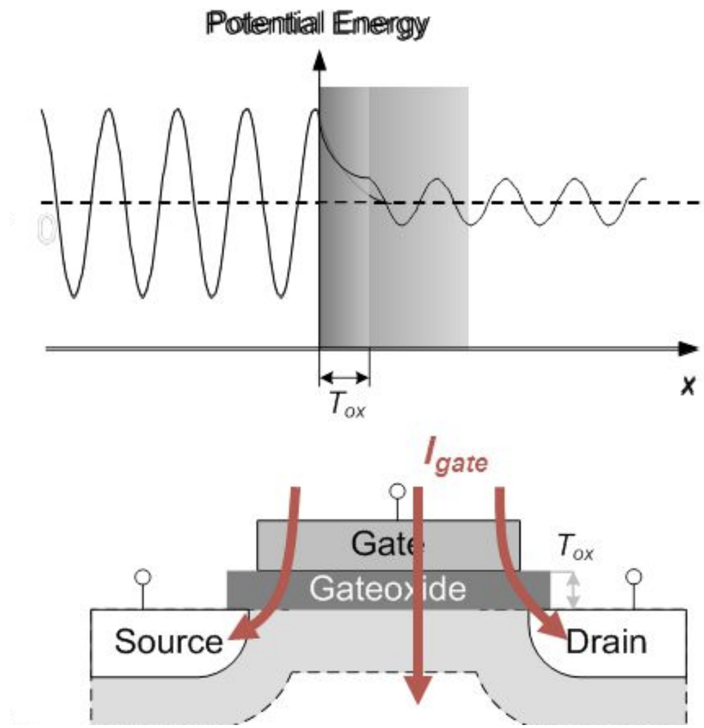


Standard Transistor

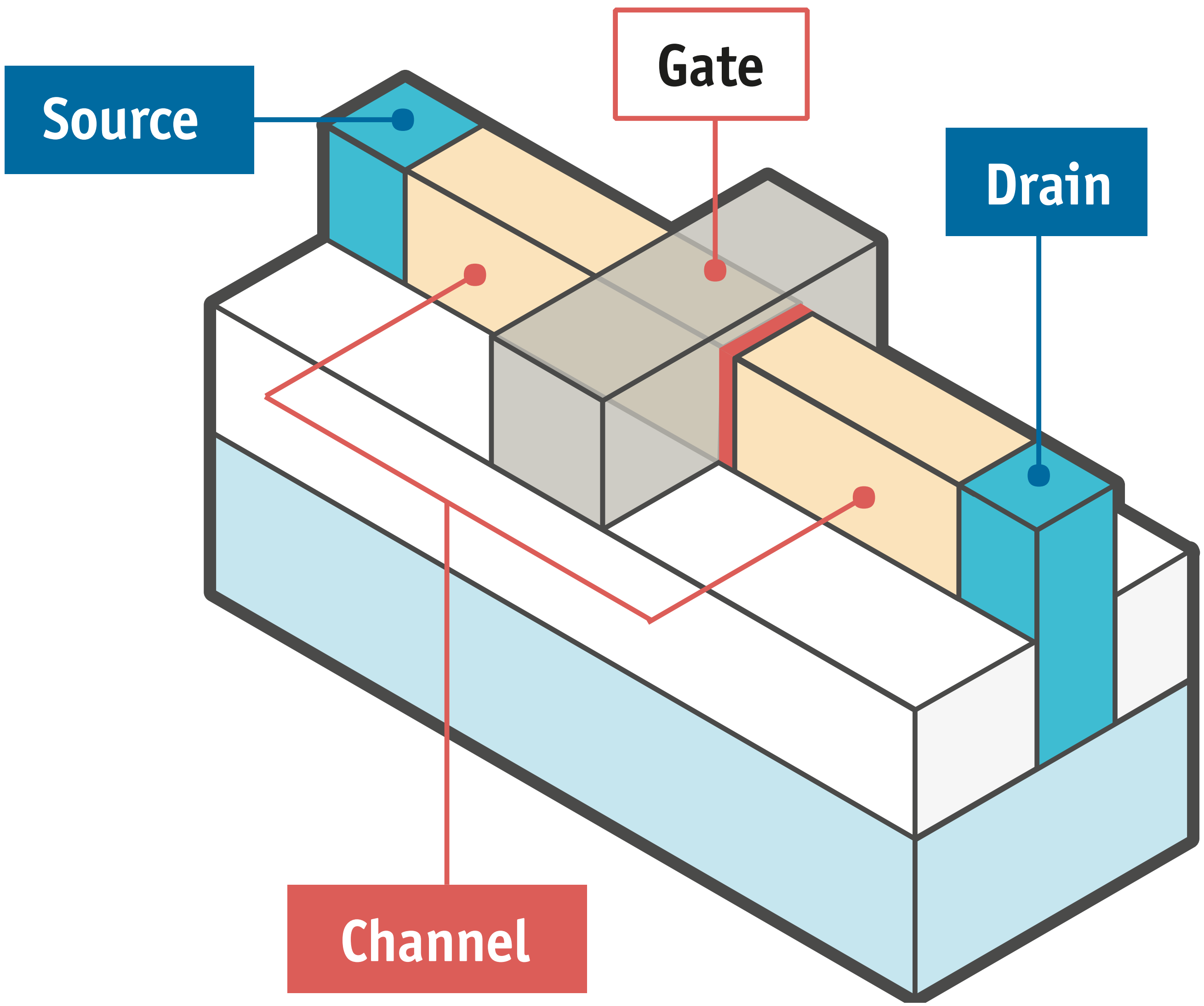


Problem no. 1 – Leakage Current

- The sources and drain are very close together (≈20nm).
- The **channel can leak** with a **residual current** flowing even when the device is off, wasting power and generating heat.
- 2 broad changes are needed to address the **leakage current** problem:
  - **Design of the transistor (topology)**
  - **Find replacement for silicon**

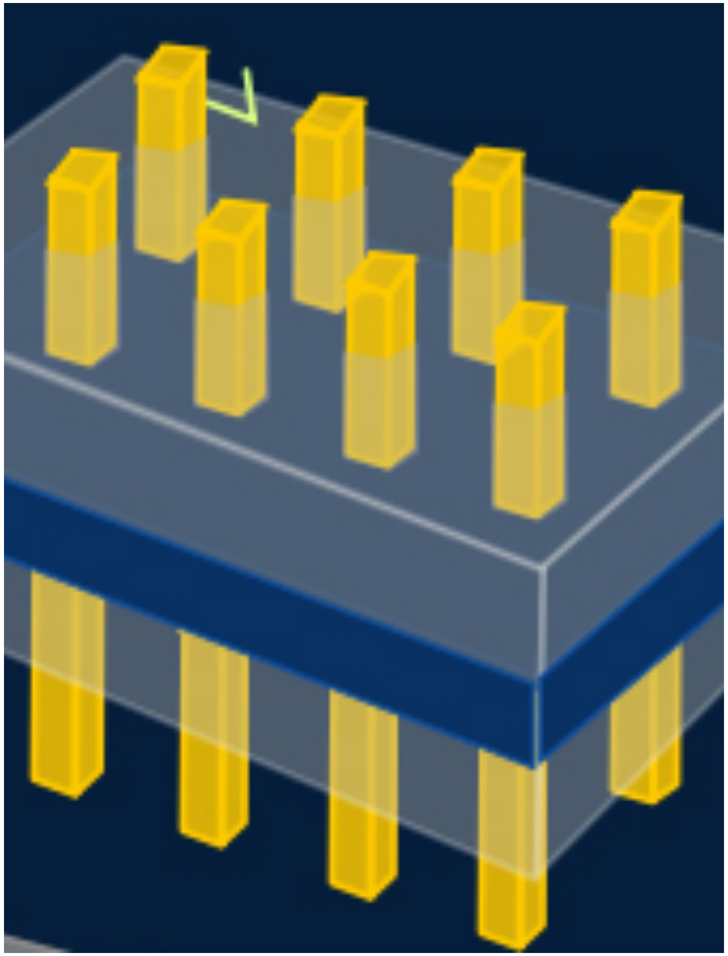


Going 3D - finFET



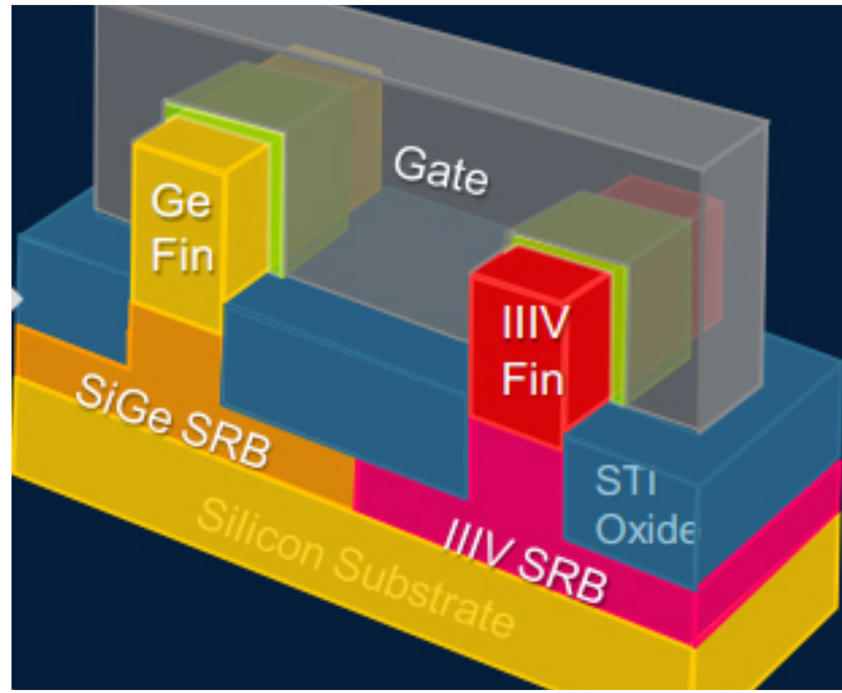
Next Logical Step – “Gate-all-around” Transistor

- The channel is surrounded by its gate on all four sides.
- These transistors are expected by early **2020s** and they will allow to build chips with features **5nm** apart



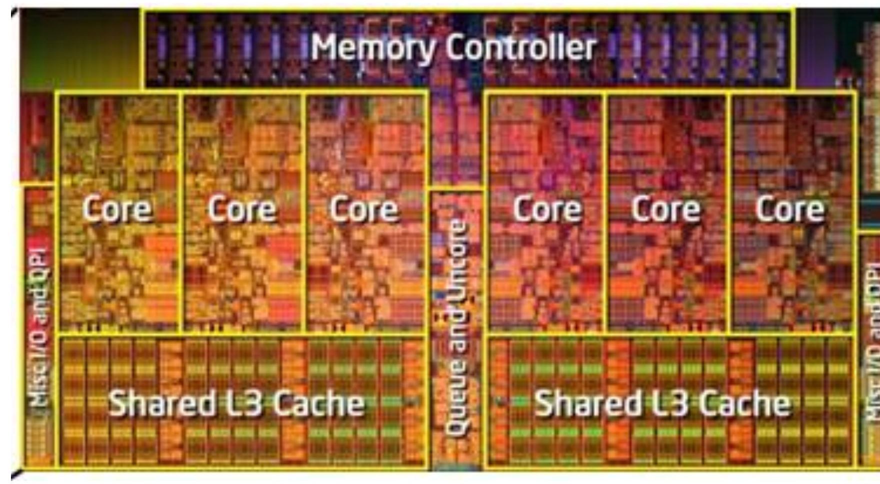
Materials Matter! III-V FinFET

- Chipmakers are experimenting with materials **beyond silicon**.
- The goal is to have materials with **better conductivity** than silicon. This means lower power usage and transistors can switch on and off faster.
- Silicon-Germanium alloy (SiGe) for channels. Also alloys of Indium, Gallium and Arsenide (III-V materials) used.



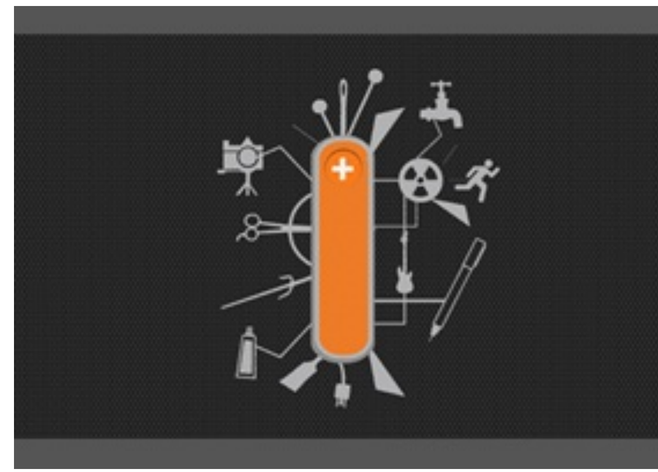
Keep Moore's Law Alive: Duplicate Circuits

- Increasing the speed clock, power consumption increases at the cubic power. Since the middle of past decade **clock speed barely increases**.
- Chipmakers responded to this by **duplicating chip existing circuitry → multicore chips**.
- The basic idea is that to have several slower chips might give better results than relying on a single speedy one.



Keep Moore's Law Alive: Specialize!

- The most widely used chips, such as Intel and those based on ARM's Cortex design (found in smart-phones) are **generalists**, which makes them **flexible**.
- They can do a bit of everything but **excel at nothing**.
- Tweaking hardware to make it better at dealing with **specific mathematical tasks** can provide 100 to 1,000x performance.
- When Moore's law was strong, little incentive to customize processing, **but now trade-off is changing**!

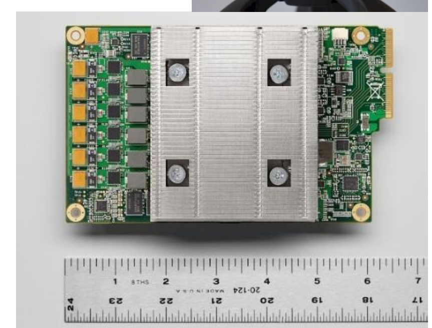


Some Examples of Specialized Processors

- **Nvidia and AMD** are the best-known examples of **graphics chips** to improve the visuals of video games. Now they are used in HPC.

- **Movidius's Myriad 2** chip is a special purpose chip for computer vision (applications from robotics to self-driving cars to augmented reality)

- **Google TPU** is a custom chip for deep learning and AI applications, specifically designed to run TensorFlow.



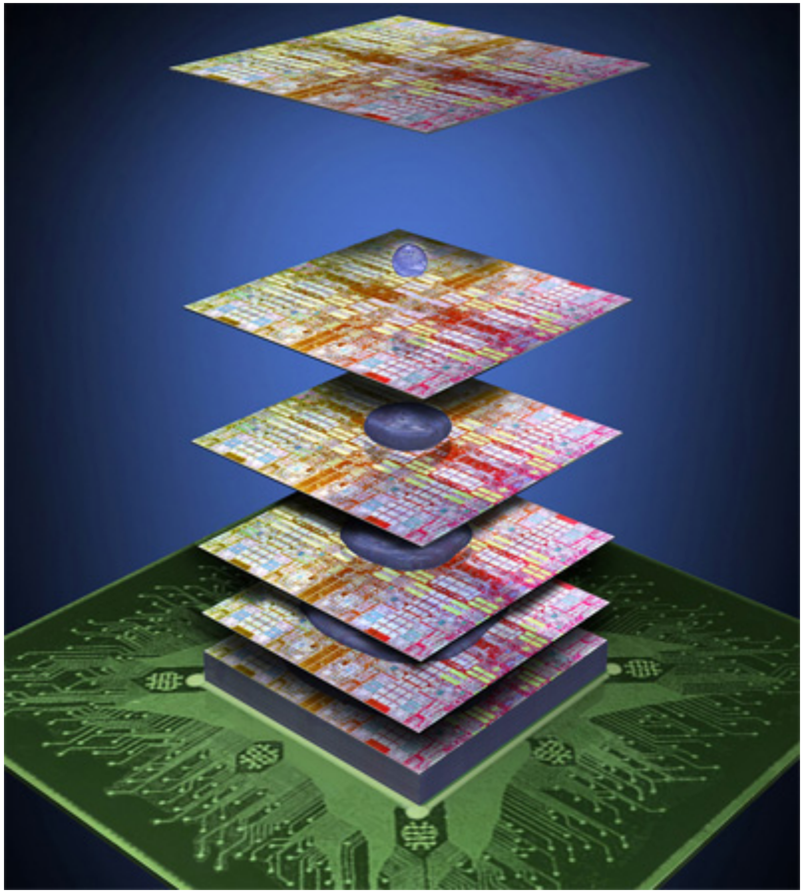
- **Anton** is a special purpose system for molecular-dynamics simulations.

Chips for Data Centers

- Best target for specialized logic might be data centers, large computing warehouses that power the servers running internet.
- Because of the sheer of volume of information they process, data centers will always be able to find a use for **a chip that can only do one thing but do it very well, i.e. MS BING uses FPGA**
- Data access becomes bottleneck.

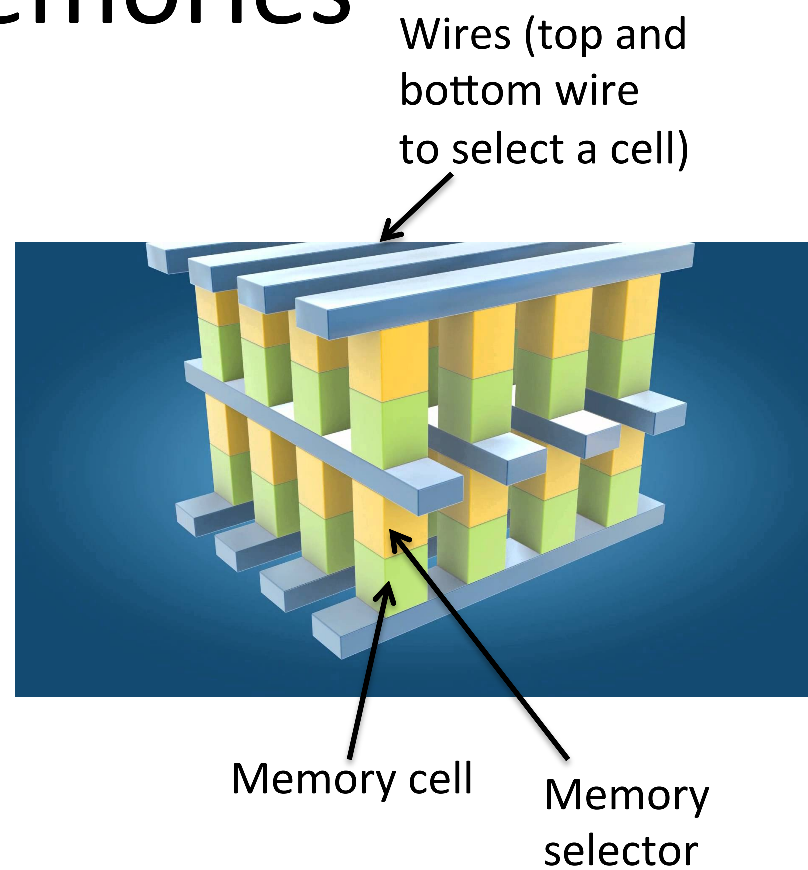
3D Chip Stacking

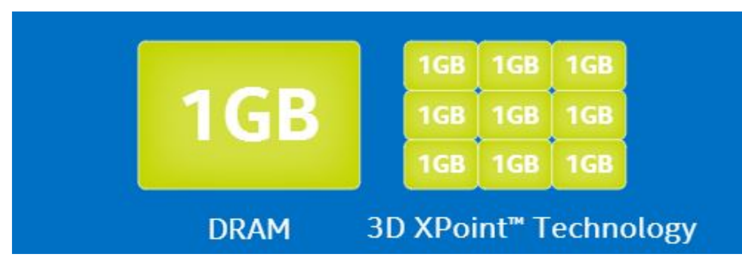
Modern chips are essentially flat but a number of companies are now working on stacking chips on top of each other



3D Stack Memories

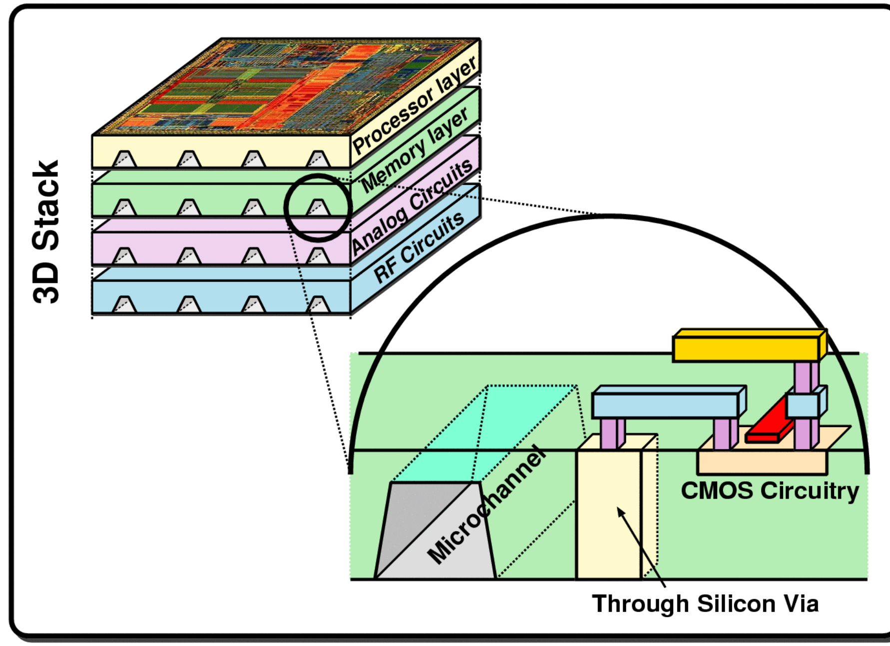
- Samsung already sells storage systems made from vertical stacked flash memories
- Last year, Intel and Micron announced a new memory technology called 3D Xpoint that also uses stacking.
- Advantages:
  – Non-volatile (Persistent)
  – 8x-10x higher density than DRAM





Putting Together Memory and CPU

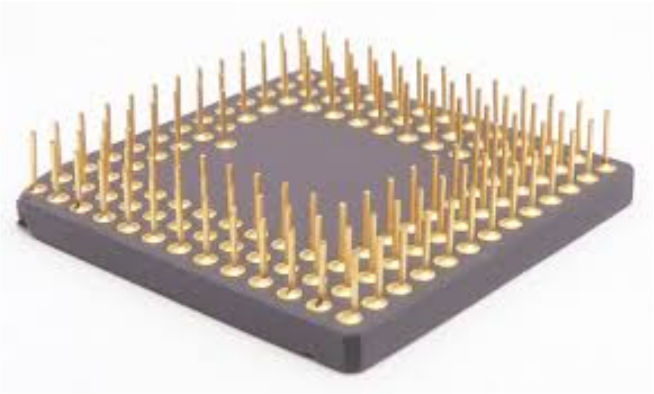
- IBM is working on chip stacks in which slices of **memory are sandwiched between slices of processing units**.
- A traditional computer's main memory is housed *cm* away from the processor
  - Moving the memory inside the chip cuts those distances from *cm* to *μm*, which allows to **move data quicker** and with **lower power consumption.**



Challenges in 3D Stacking

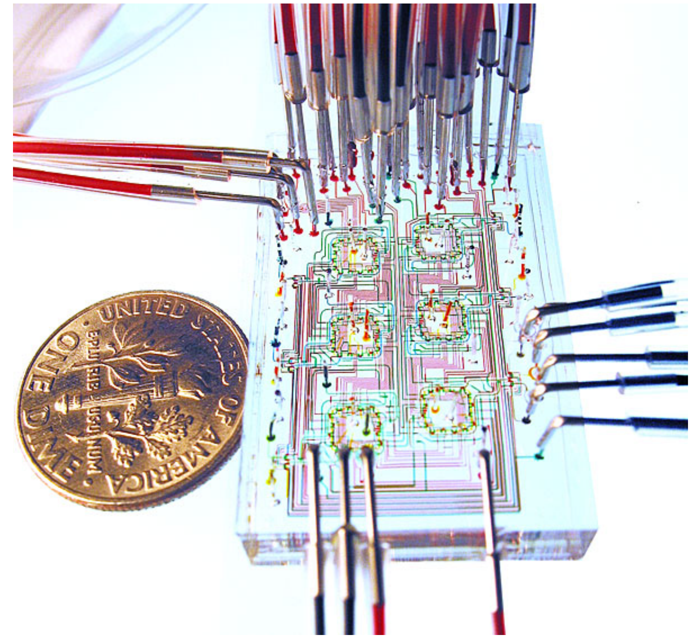
- **Heat**: as more layers are added, the volume of the chips, where heat is generated, grows faster than the surface where the heat is removed.

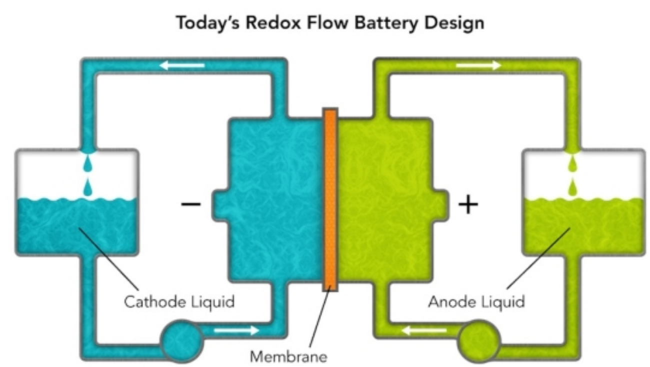
- **Getting electricity in the chip**. 80% of the pins in a chip are reserved to feed electricity (the rest is for I/O). In 3D this constraint multiplies, as the number of pins must serve a much more complicated chip



Micro-fluid and Flow Batteries

- Both heat and feeding electricity problems can be solved by fitting 3D chips with a micro internal plumbing
  - **Microfluid channels** can carry cooling liquid into the heart of the chip
  - The liquid can **deliver energy** as well: power is provided by two liquids that meet on either side of a membrane, producing electricity.



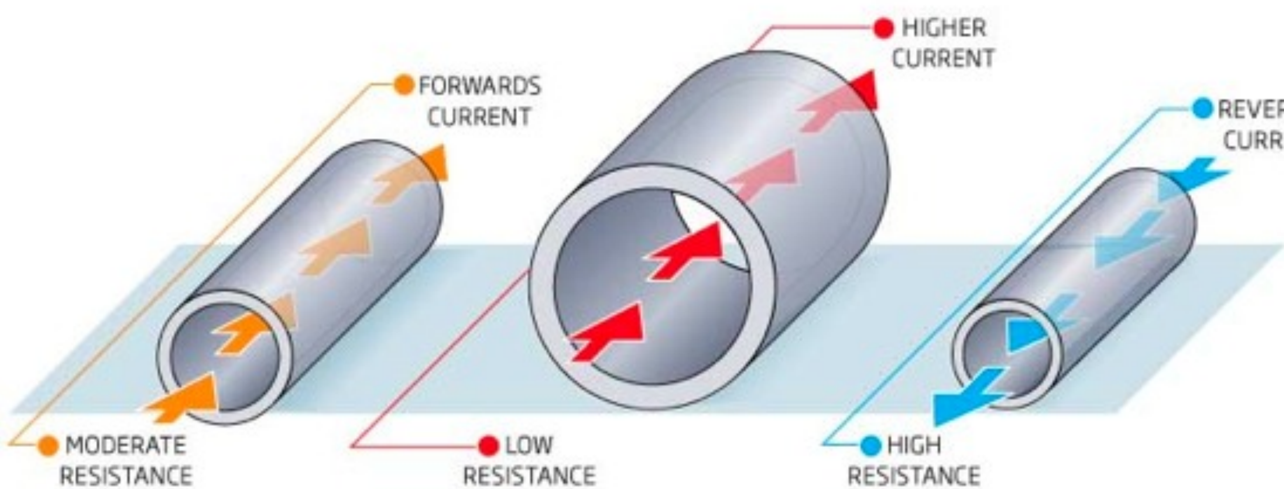


Beyond Moore's Law – Neuromorphic Computing

- Neuromorphic computing was developed by Carver Mead in the late 80s to use Very Large Scale Integration (VLSI) systems consisting of electronic **analogical** circuits to mimic neuro-biological systems, i.e. our brain.
- The HW implementation of neuromorphic processors can be realized with memristors.

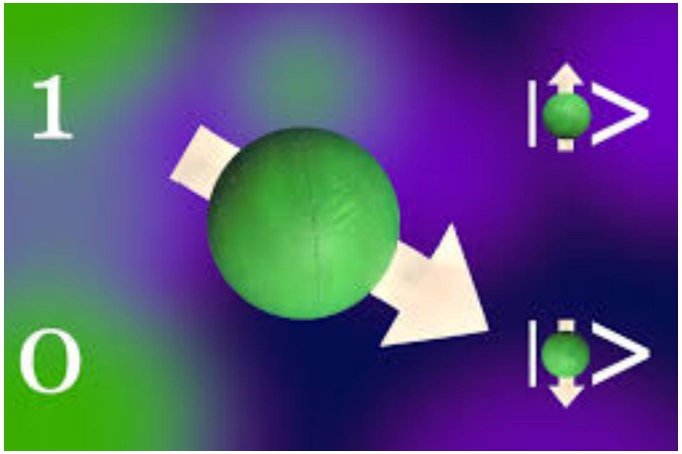
Memristor – a New Electronic Component

- Memristor is an electronic component that increases the resistance as electricity passes through it in one way and reduces the resistance when electricity goes in the opposite way
- When no current flows, the memristor remembers its last resistance value (It can store data).
- Its existence was predicted in 1971 and realized in 2008 by HP

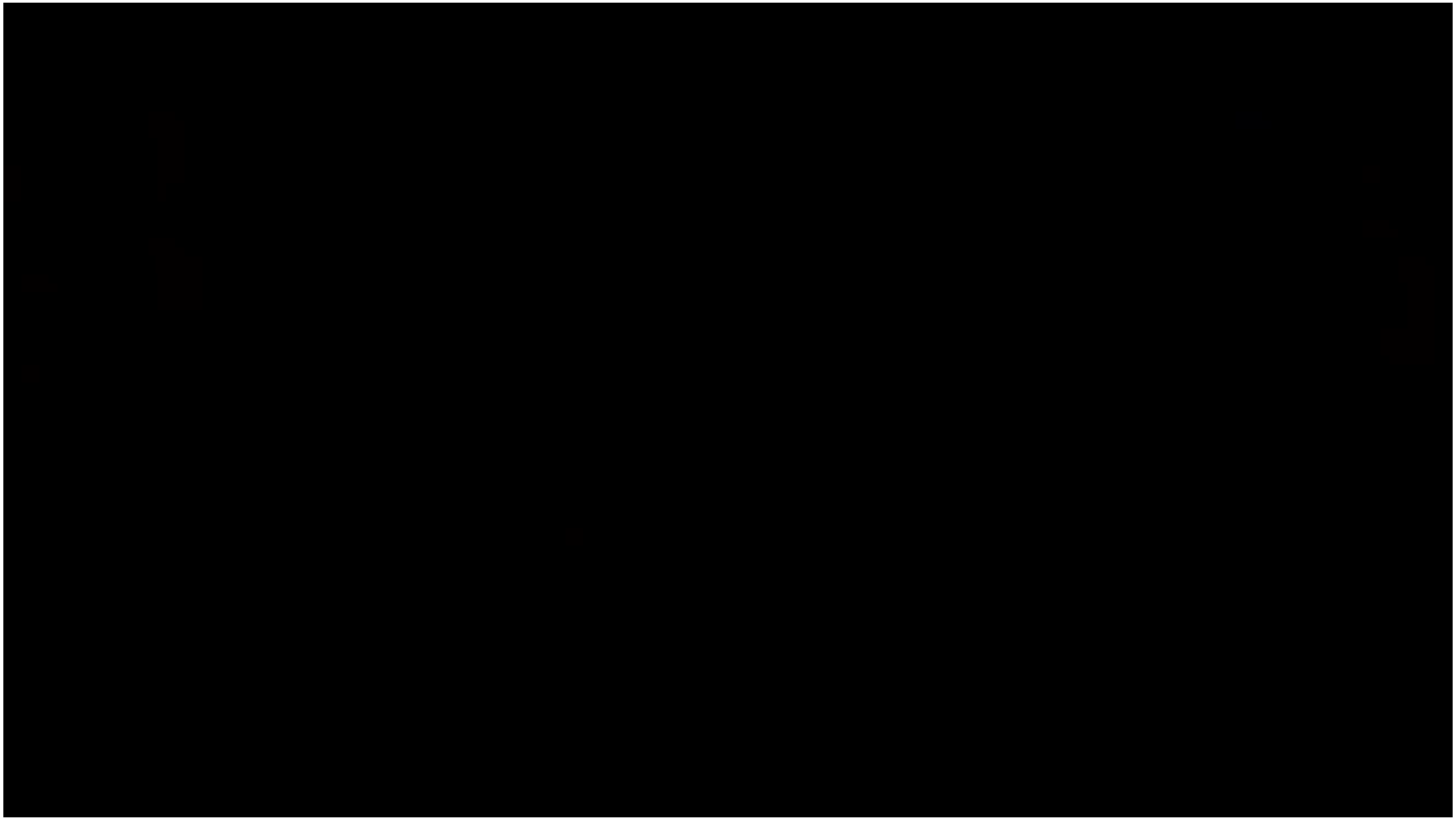


Beyond Moore's Law - Quantum Computing

- QC is a computing model that uses QM phenomena, such as **superposition**, **entanglement** and **quantum tunneling** to perform operations on data.
- It could offer a speed advantage for some mathematical problems: sorting of unordered list, Fourier transform, integer factorization, QM simulations.
- In many cases, we still don't know whether a given quantum algorithm will be faster than the best-known classical one.



Quantum Computing

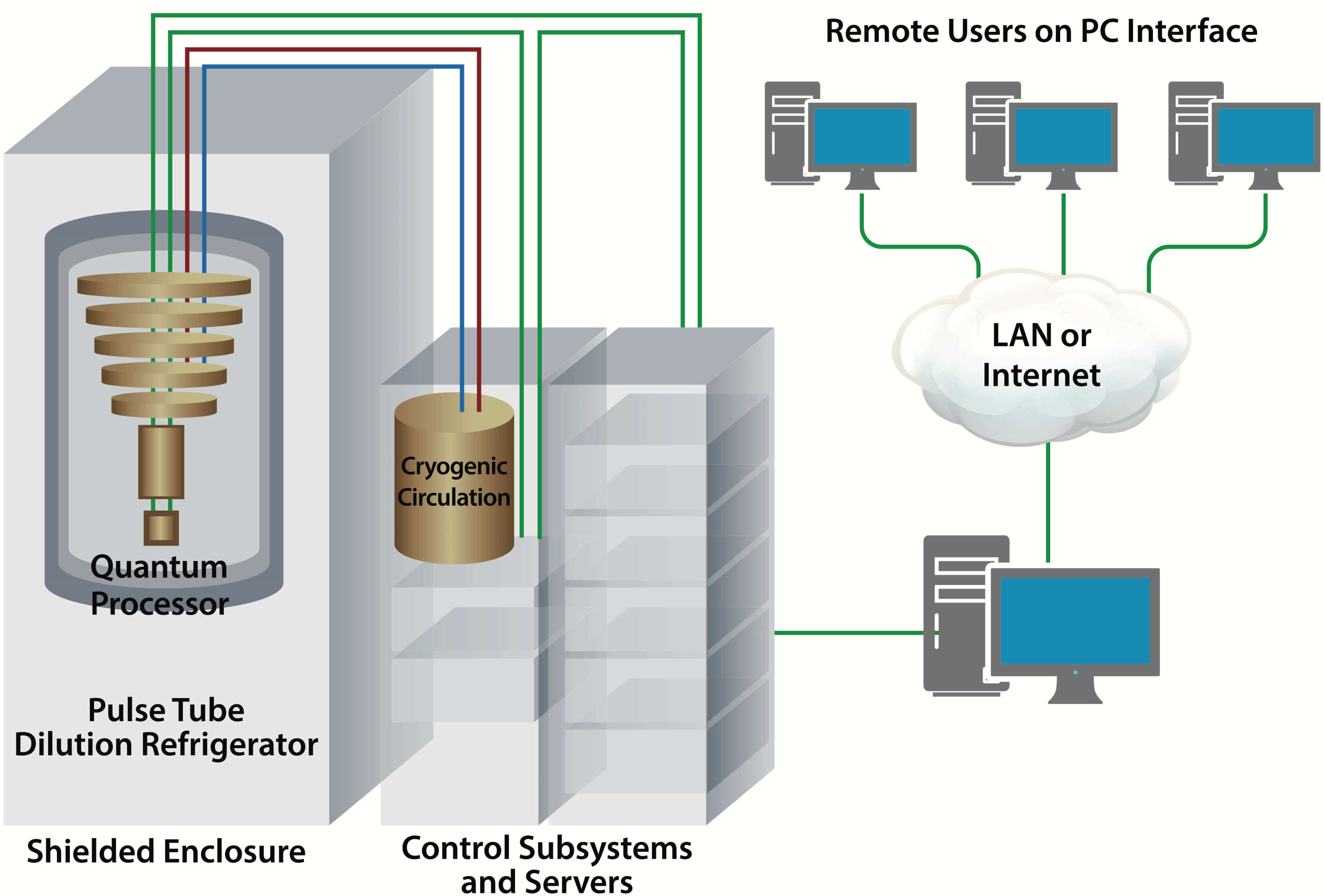


https://www.youtube.com/watch?v=4ZBLSjF56S8

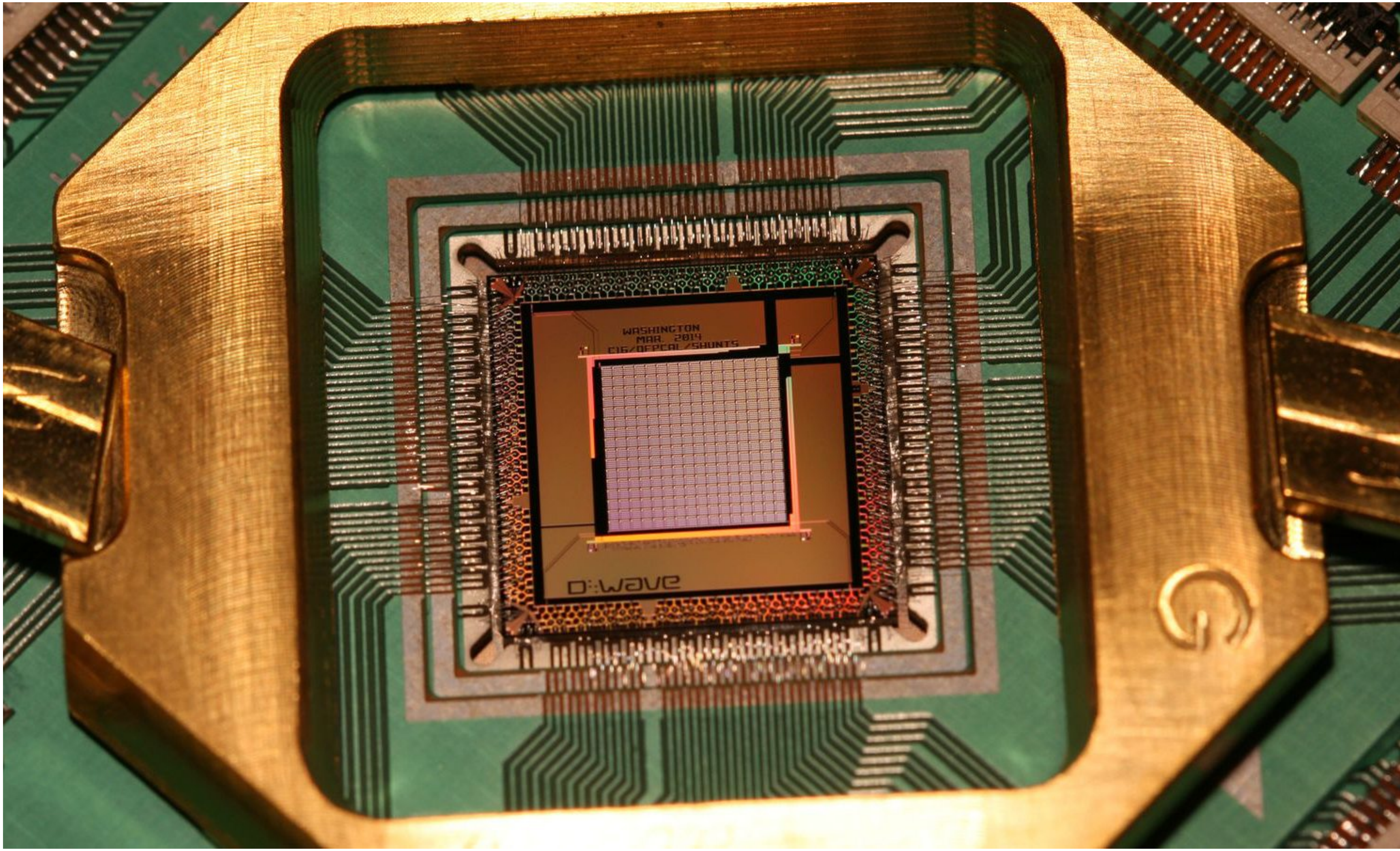
D-Wave 2X Machine

- It is one of the world's first **commercially** available quantum computers.
- Limited to one mathematical problem**: finding the lowest value** of **complicated functions** by quantum annealing
- Google, NASA, LANL and Lockheed Martin own a D-WAVE 2X.

D-Wave 2X Machine

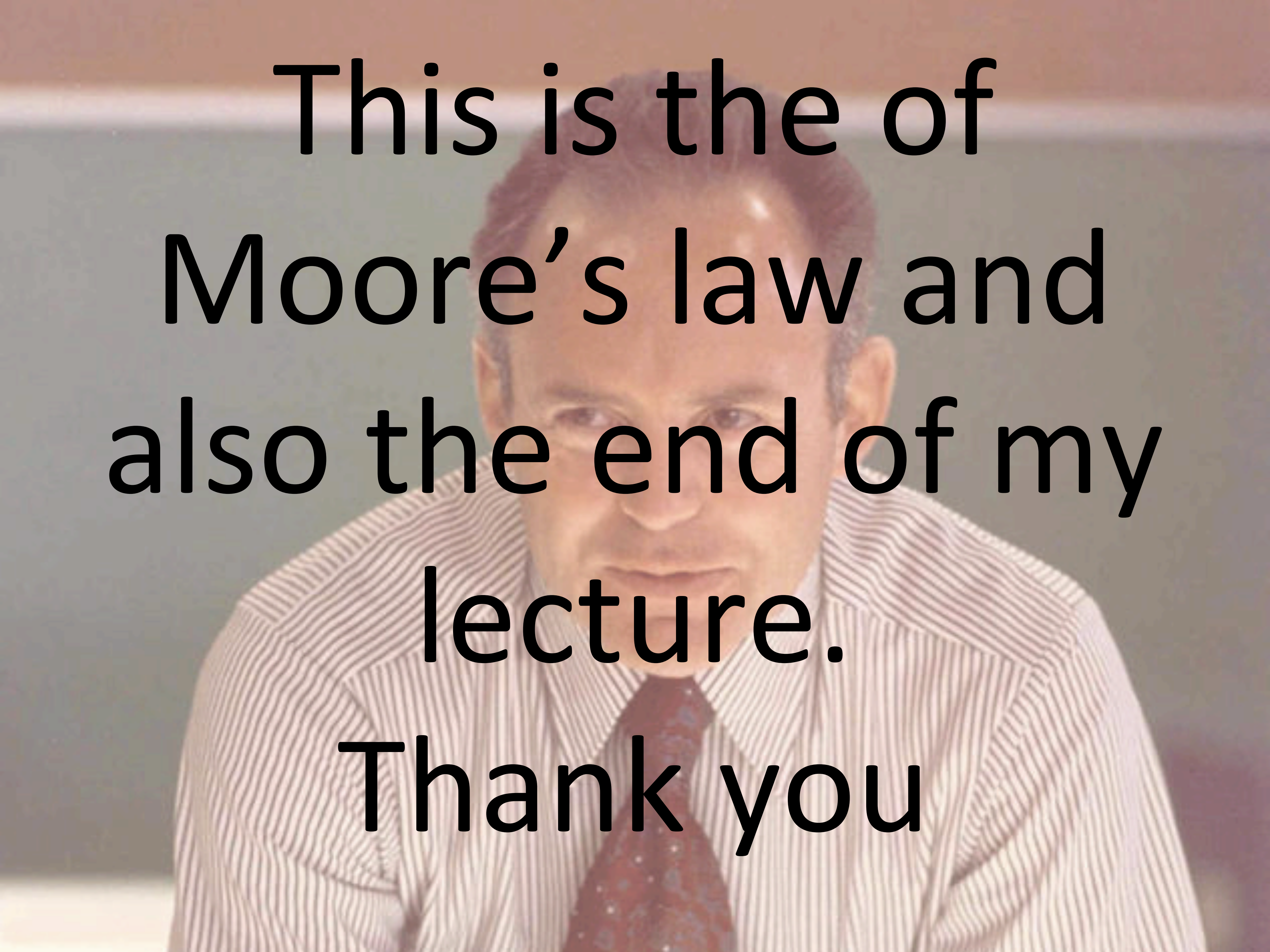


D-Wave 2X Processor



The end of Moore's law

- **Moore's law made our life simple**: computers got better in a **predictable way** and at a **predictable rate**.
- As Moore's law slows down, we are forced to make tough choices between three key metrics: **performance**, **power** and **cost**.
- Progress will become **less predictable**, but will make our life more fun and creative.



This is the of Moore’s law and also the end of my lecture.

Thank you