



CS305: Computer Architecture

Trends in Computer Architecture

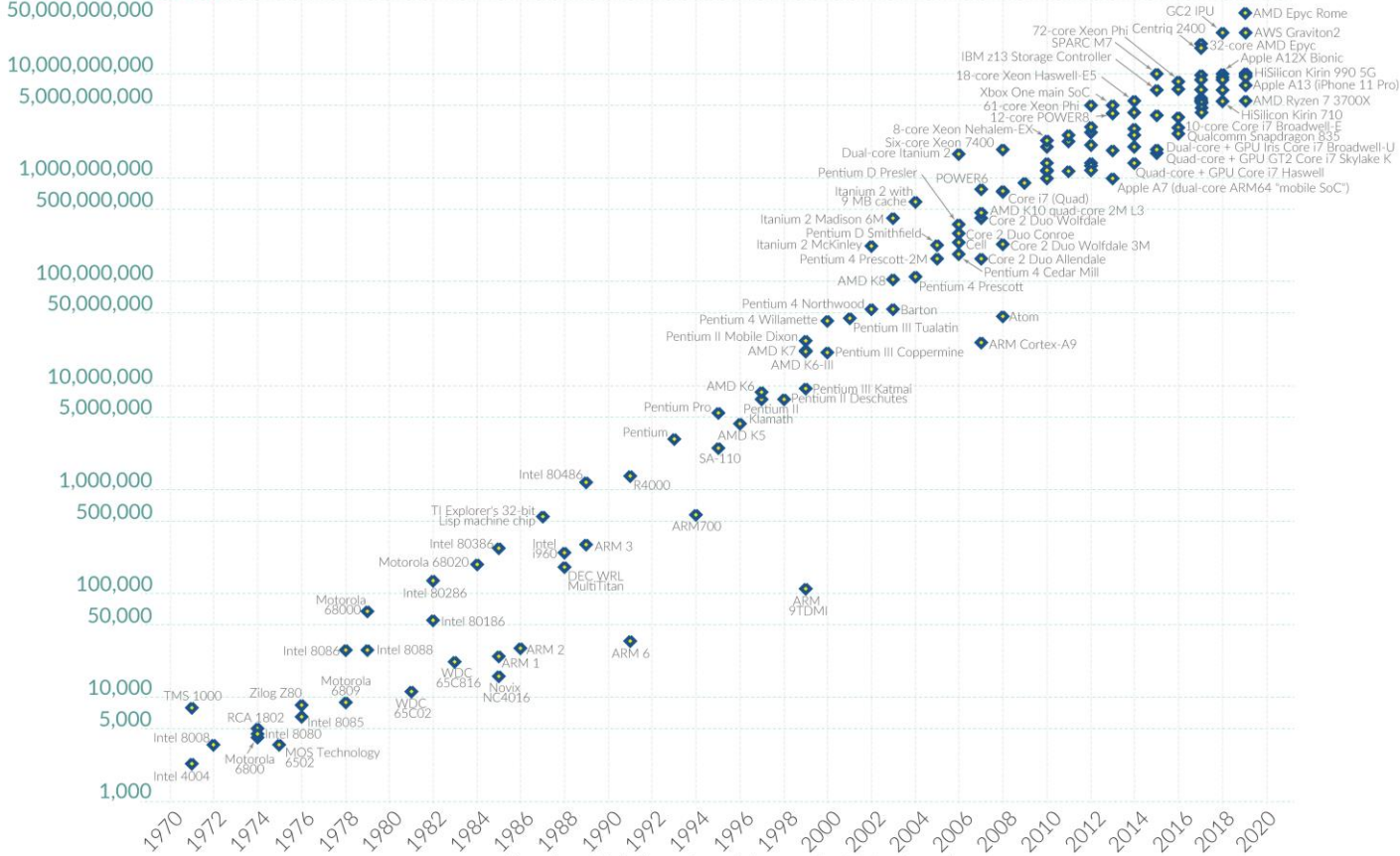
<https://www.cse.iitb.ac.in/~biswa/courses/CS305/main.html>

<https://www.cse.iitb.ac.in/~biswa/>

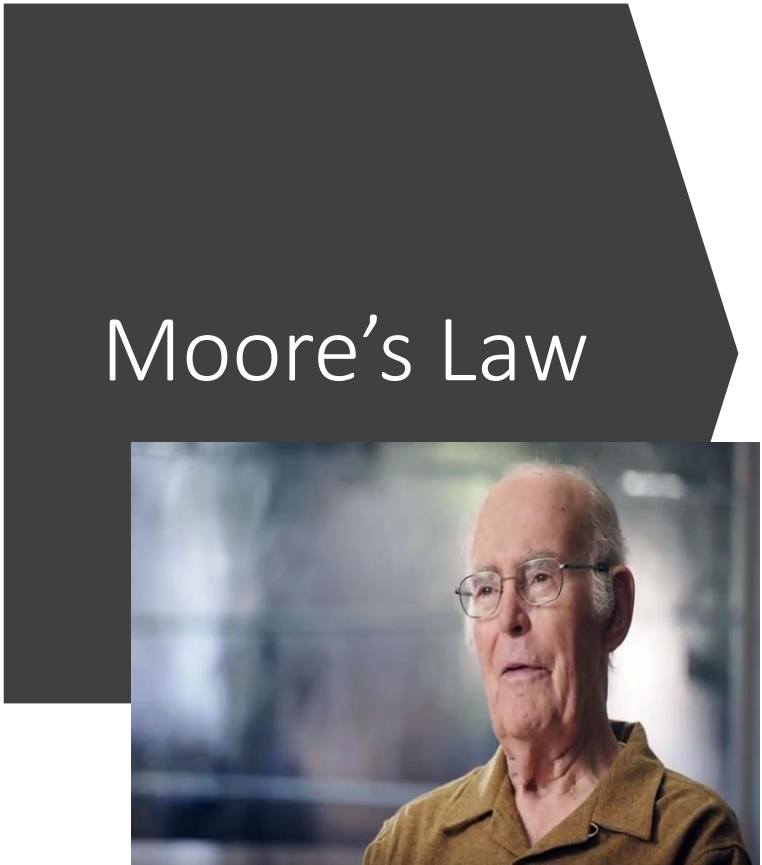
Moore's Law: The number of transistors on microchips doubles every two years

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

Transistor count



Data source: Wikipedia (wikipedia.org/wiki/Transistor_count) OurWorldinData.org – Research and data to make progress against the world's largest problems. Licensed under CC-BY by the authors Hannah Ritchie and Max Roser.



Cache/core size doubling 😊

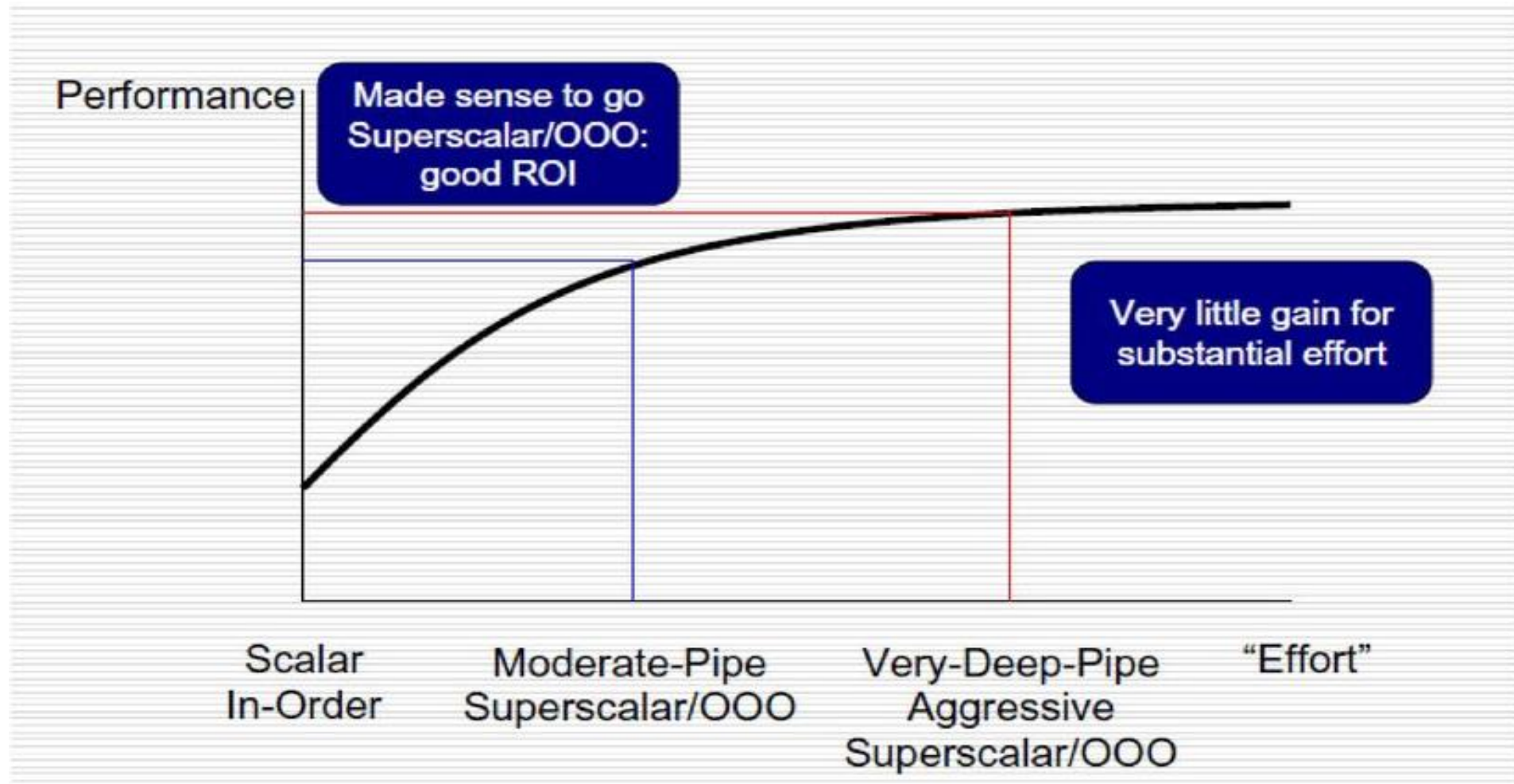
Dennard Scaling

as transistors get smaller ->

their power density stays constant

the power use stays in proportion with area

ILP Wall



The Power Wall

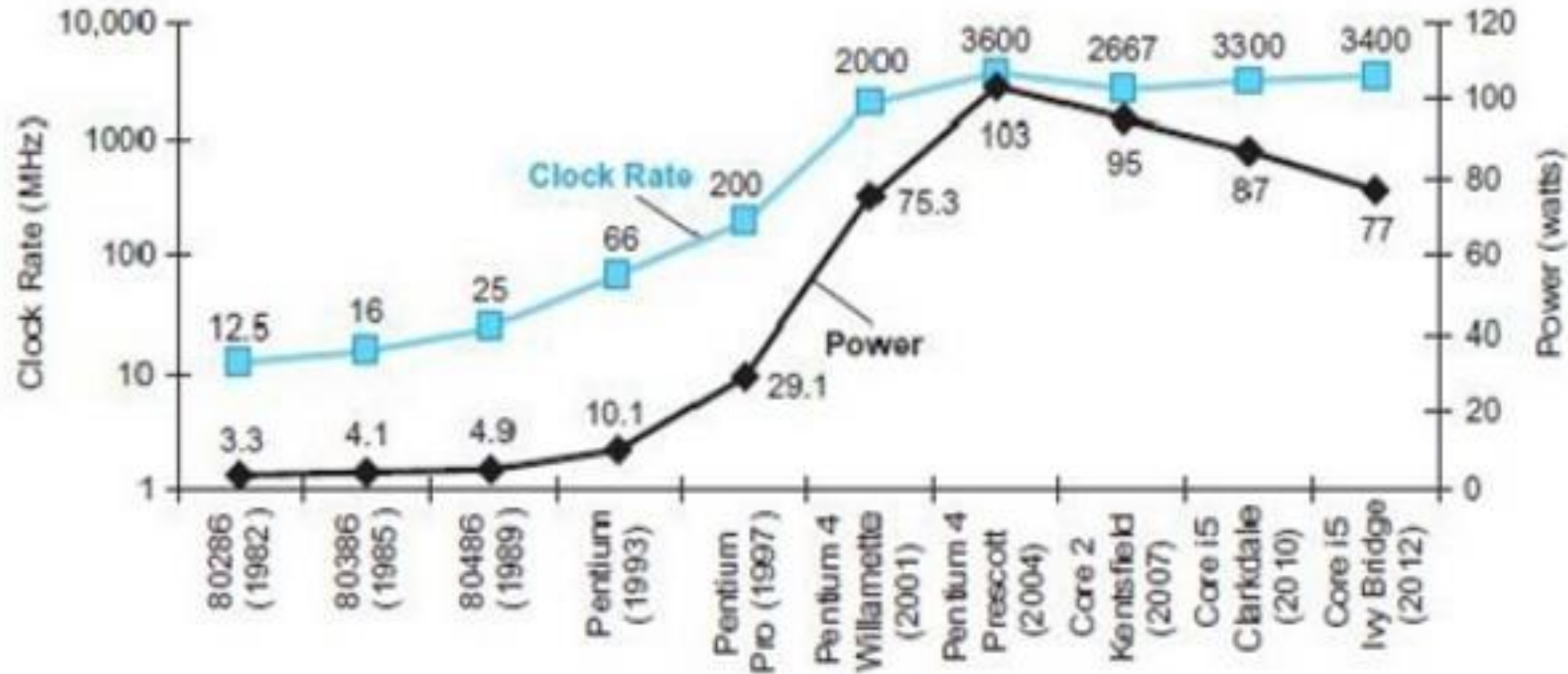
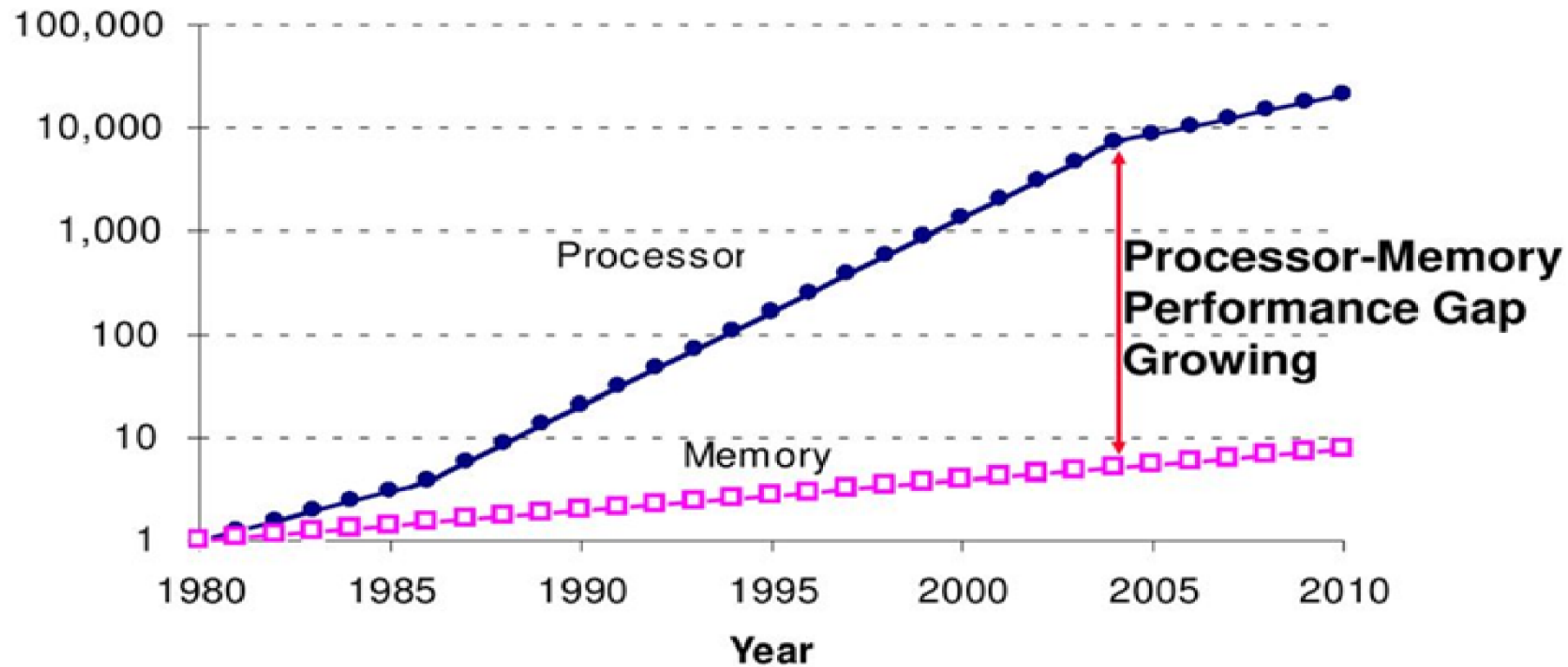
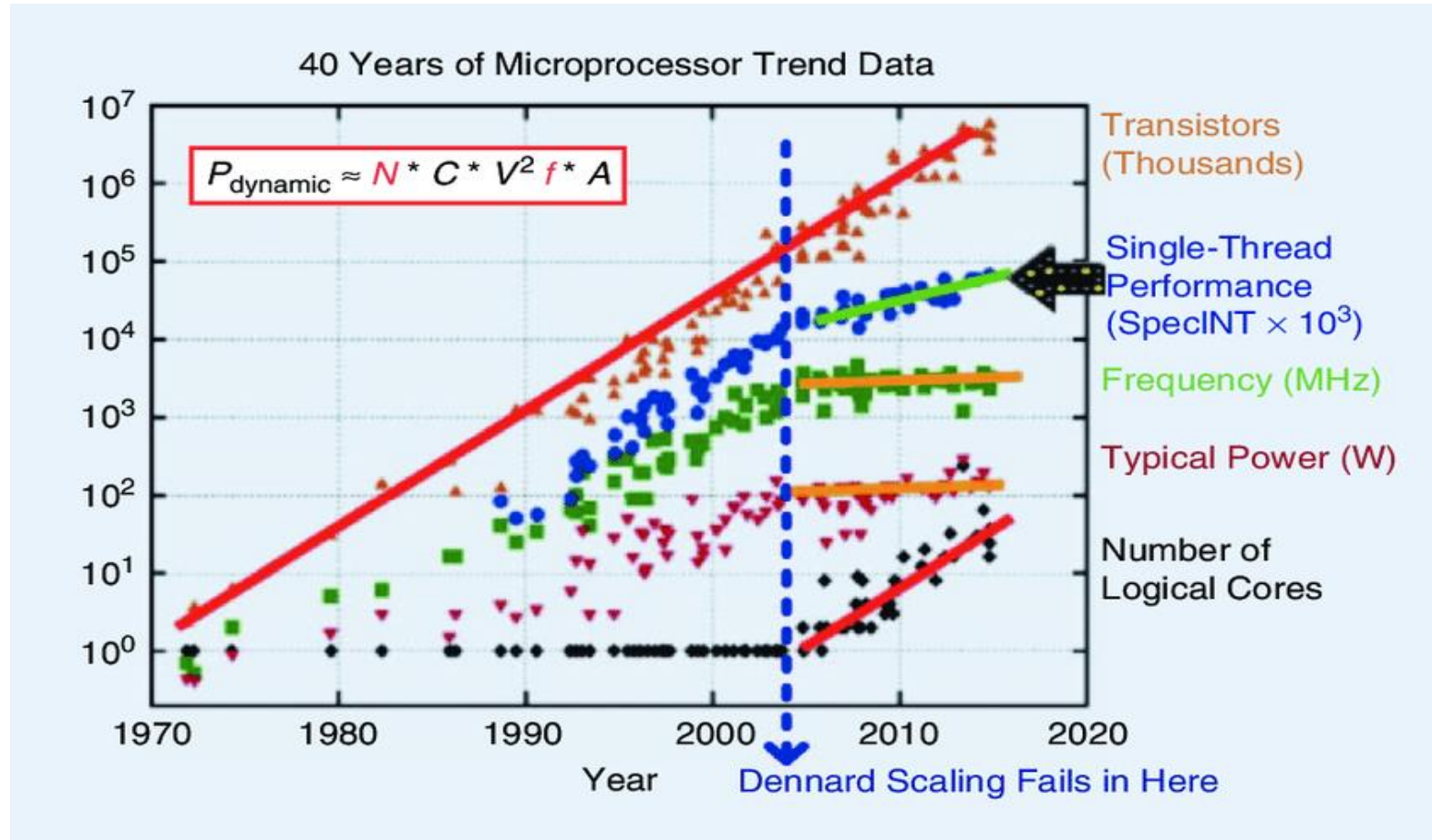


Fig.1.3 Clock rate and power for Intel x86 microprocessors

Memory Wall




All in one



Dark Silicon

Before 2006, transistor scaling (Moore's Law) has mostly been followed by voltage scaling (Dennard scaling).

Around 2006, Dennard scaling failed such that it cannot follow Moore's Law.

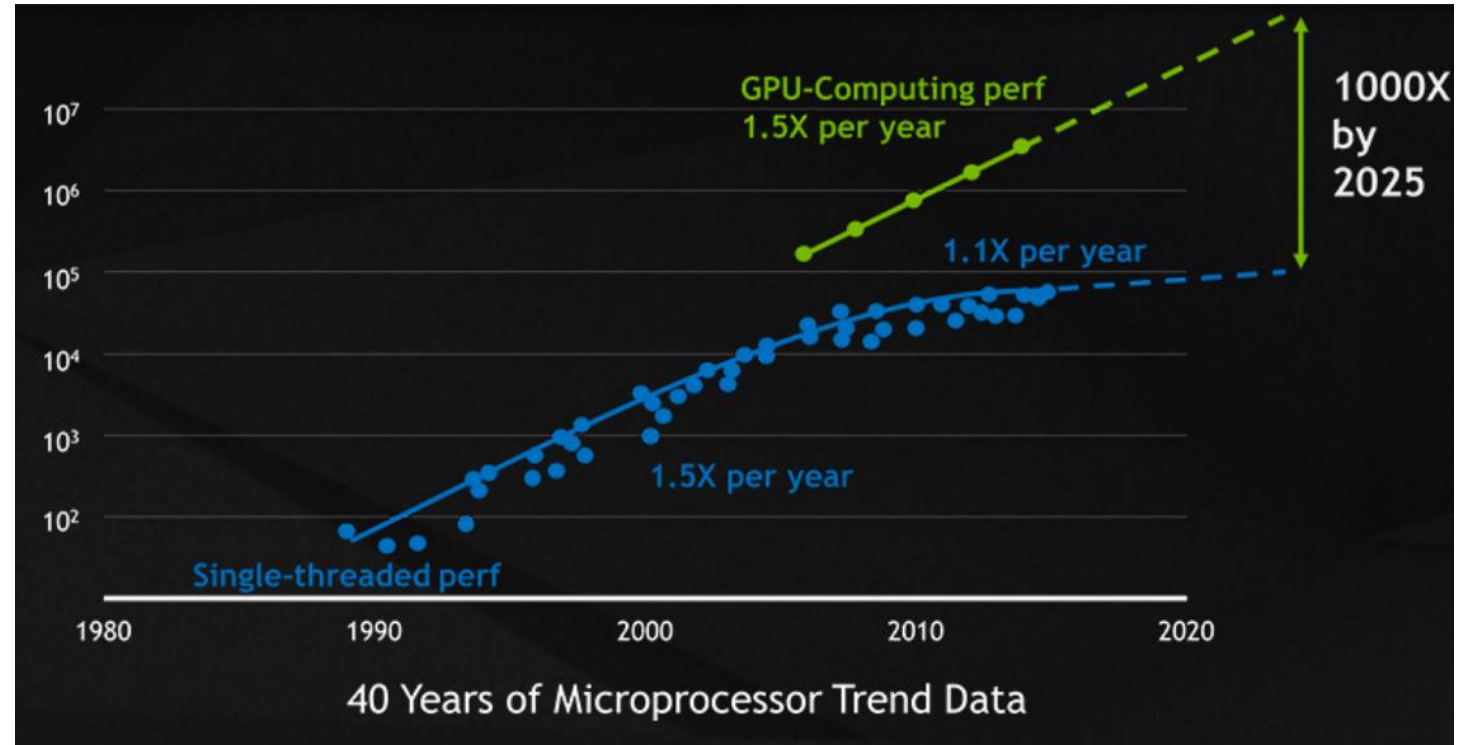
The extra transistors brought by Moore's Law can no longer be powered on because it would violate the thermal design power (TDP) constraint  These unpowered/unused transistors are "dark silicon".



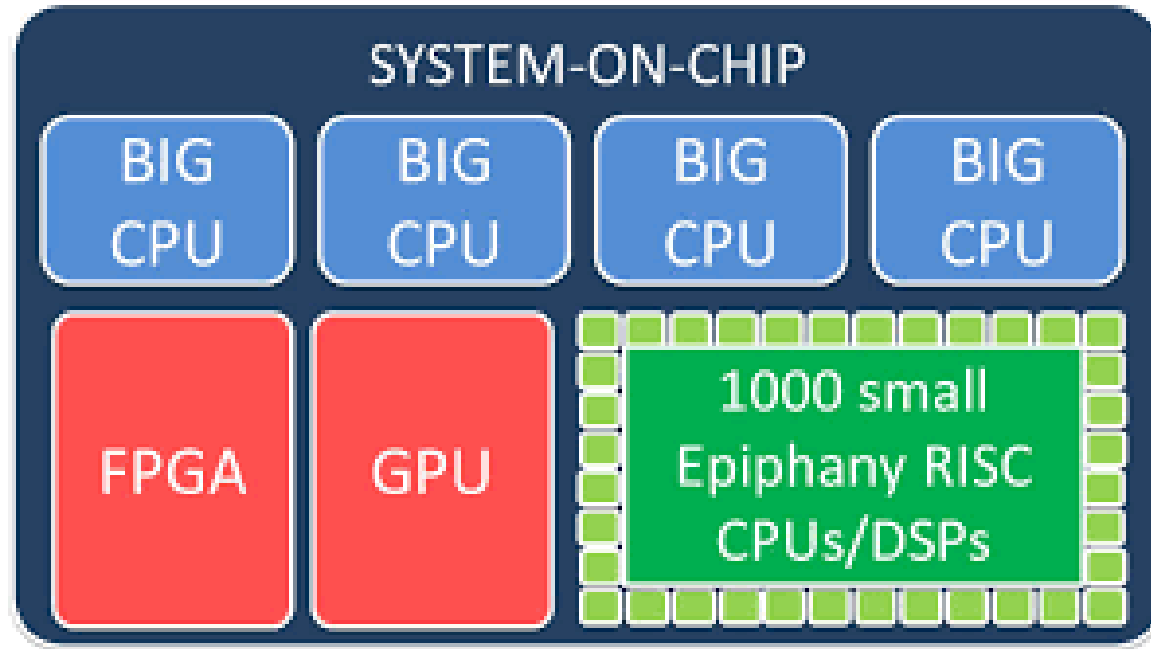
2021: 50 Years of Microprocessors

GPUs (World of Teraflops) to GPGPUs

- SIMD (single instruction Multiple Data) model



Heterogenous Systems



Google's TPU (Tensor Processing Unit)

<https://spectrum.ieee.org/the-accelerator-wall-a-new-problem-for-a-post-moores-law-world>

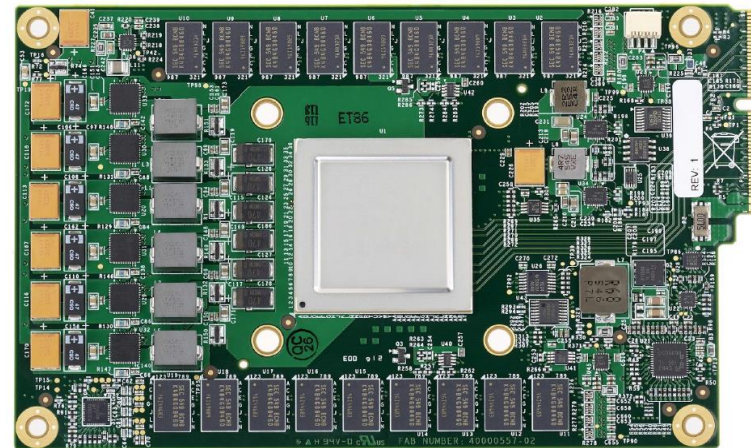


Figure 3. TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

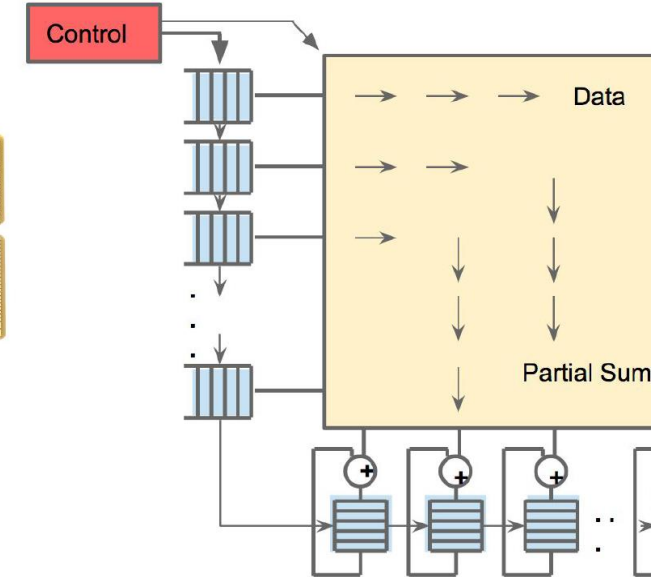
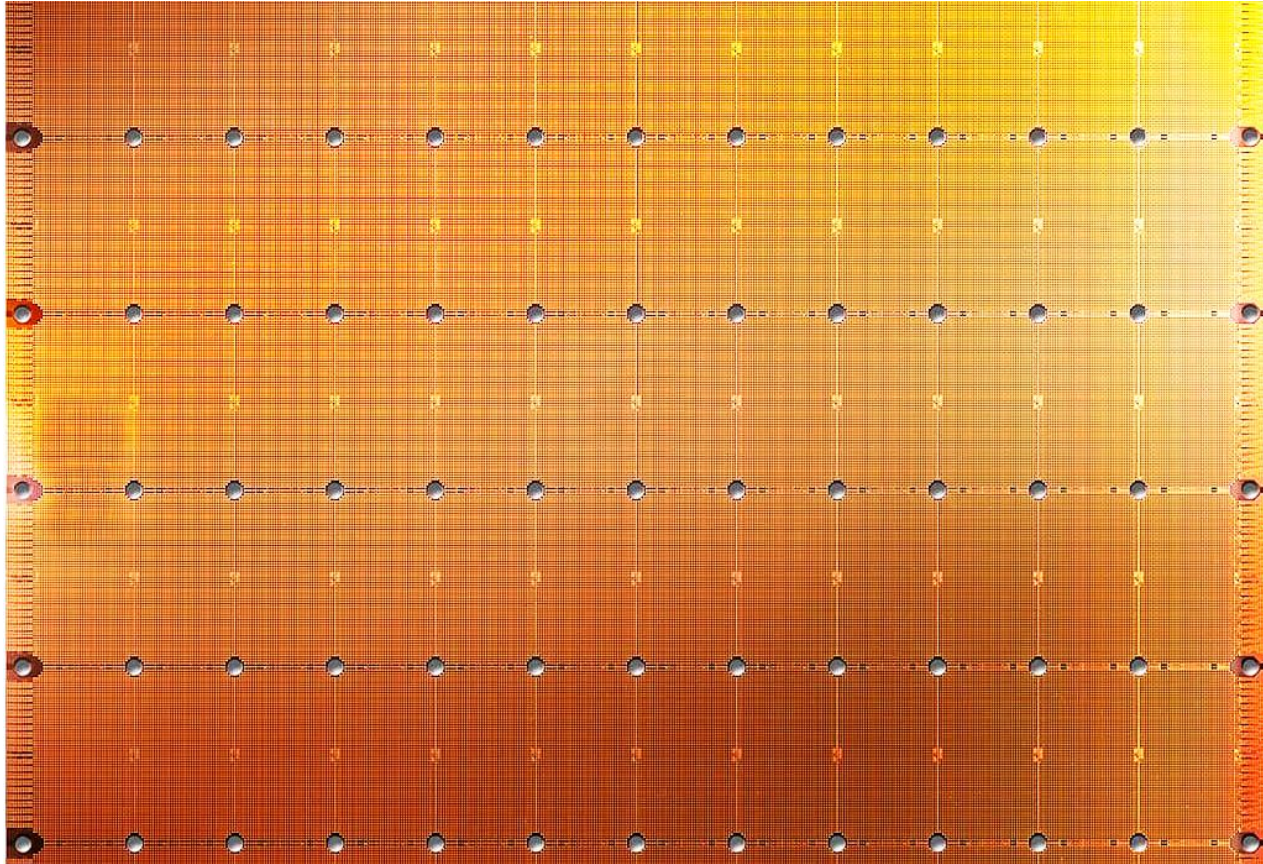


Figure 4. Systolic data flow of the Matrix Multiply U has the illusion that each 256B input is read at once, a update one location of each of 256 accumulator RAM

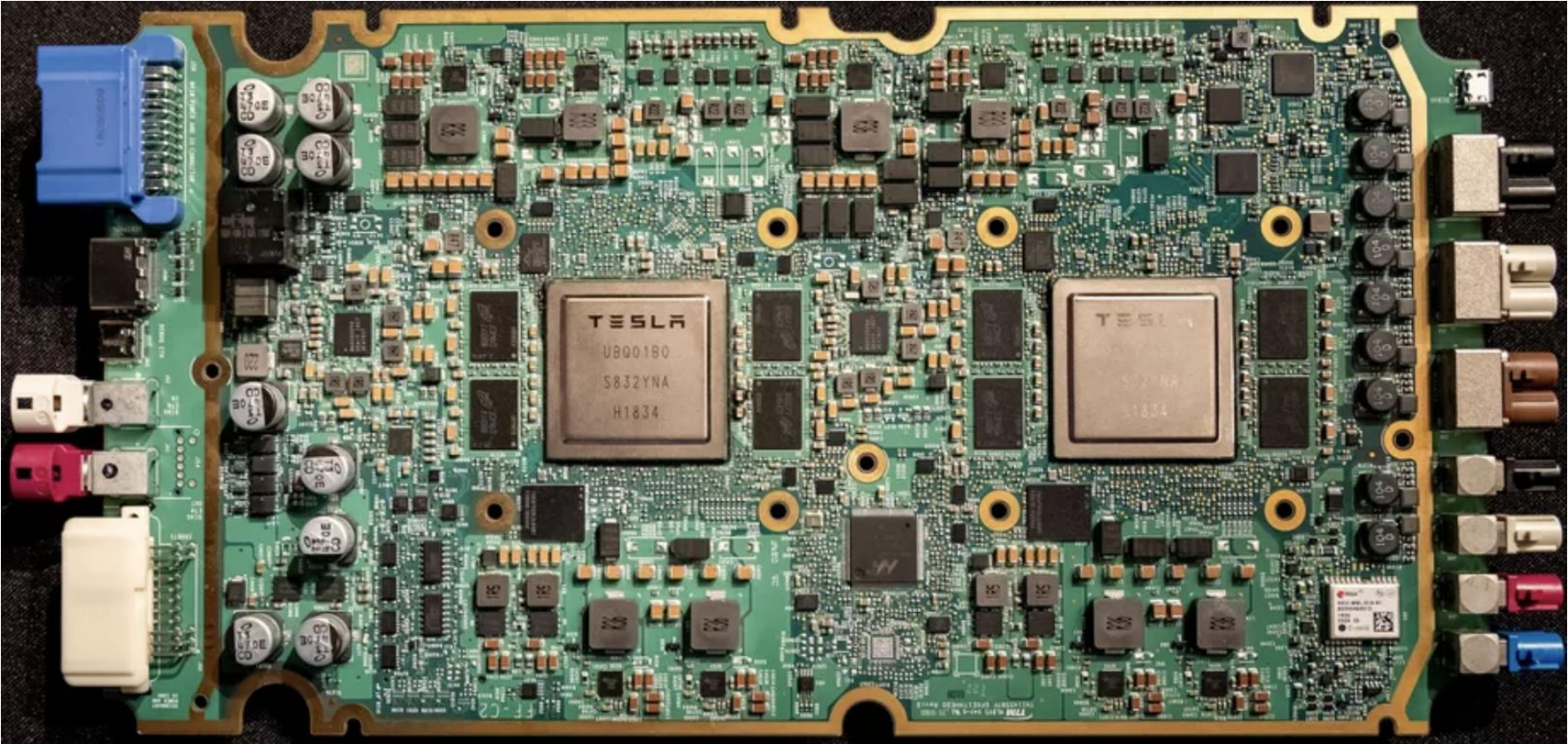


- *The largest ML accelerator chip*
- *400,000 cores*

<https://www.anandtech.com/show/14758/hot-chips-31-live-blogs-cerebras-wafer-scale-deep-learning>

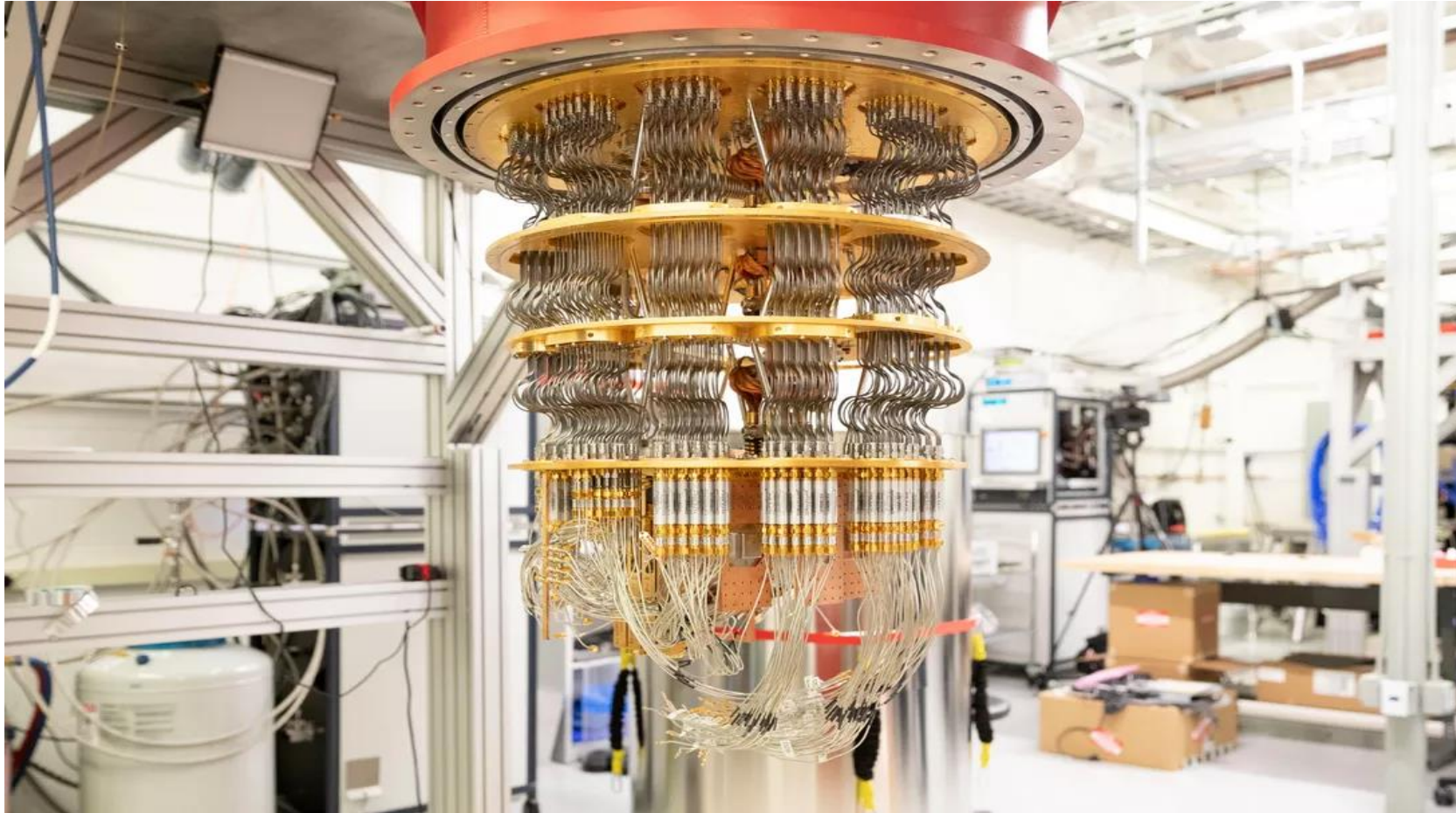
Cerebras's Wafer Scale Engine

Tesla Self Driving Car



Computer Architecture

Google's Sycamore Quantum Computer



*Finished a task in
200 seconds that
would take a CPU
10,000 years 😊*

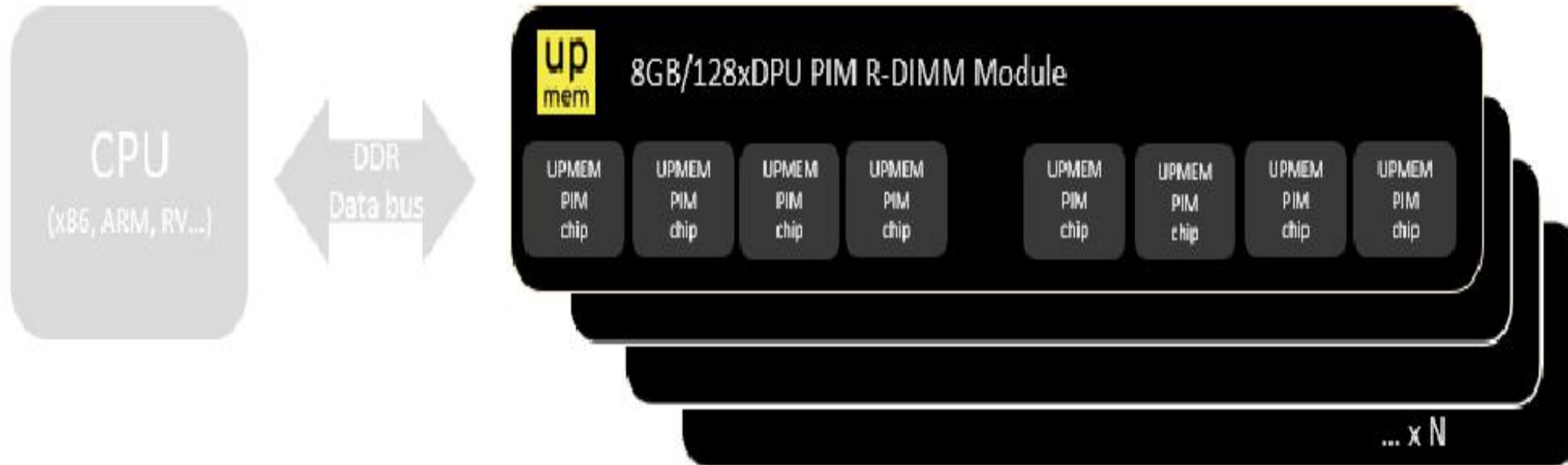
Google's datacenter



The Supercomputer



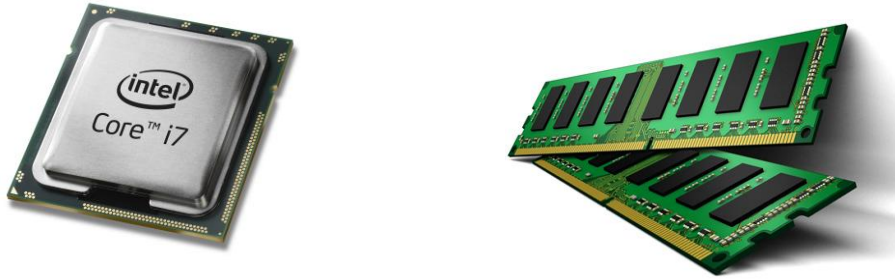
Processing in Memory



Intel Optane: Non-volatile memory



What did we cover in CS305?



Key Takeaways: Moore's law -> ISA abstraction ->
Common case fast -> parallelism, pipelining, prediction, locality

Thanks, Thanks, and Thanks

All the TAs for all the hard work. Appreciate it.

All the students: online semester, COVID-19

Hope you have learnt the 10K feet view if not the 10K/1K feet view.