# Lecture-1 (Logistics and Introduction) CS422-Spring 2020

Biswa@CSE-IITK

# Instructor

Biswa ~~(Biswabandan, Sir, Prof., Dr., Er., *-Biswa~~)

Contact: KD 203, email with [CS422] as the subject, biswap@cse.iitk.ac.in
Office Hours: Friday, 12 noon

Teaching and Research Interests:
Computer Architecture, Arch-OS interface, and micro-architecture Security

# Helper Threads

Vishal


Neelu

# Course Logistics

- https://www.cse.iitk.ac.in/users/biswap/CS422-2020/evaluation.html

| Rank | IPC (public traces) | Speedup | Contestant(s) |
|---|---|---|---|
| 1 | **3.881** | +39.6% | Chirag Sakhuja, Anjana Subramanian, Pawanbalakri Joshi, Akanksha Jain and Calvin Lin (The University of Texas at Austin) |
| 2 | **3.812** | +37.2% | Arpit Gupta, Parv Mor, Hritvik Taneja and Biswabandan Panda (Indian Institute of Technology Kanpur) |
| 3 | **3.773** | +35.8% | André Seznec (IRISA/INRIA) |
| 4 | **3.491** | +25.6% | Nayan Deshmukh, Snehil Verma, Prakhar Agrawal, Biswabandan Panda, Mainak Chaudhuri (Indian Institute of Technology Kanpur) |
| 5 | **3.056** | +10.0% | Kenichi Koizumi, Kei Hiraki and Mary Inaba (The University of Tokyo, Japan) |

Instruction Prefetching @ISCA 2020 ☺

# Expectations

No open-screens (no nomophobics): No open smart-phones (phones) & laptops/tablets. Keep your phones in silent mode

Open-screens will affect (distract) you, your friends, and me

Ask questions & participate in in-class discussions (worth bonus points)

Understand, implement, and analyze ideas (Hard work and honesty)

Slides will not contain everything. So attend lectures.

# Expectations

| | |
|---|---|
| Timing | Classes start at 12 PM, not 12.10/15 PM |
| Cheating | In any form will lead to zero points. Grade will be capped down (one level). To prevent capping down, you have to build architectural tools. |
| Dropping CS422 | Not allowed after Jan 15th 2020. Drop the course before that. Why? It will affect your group points. |

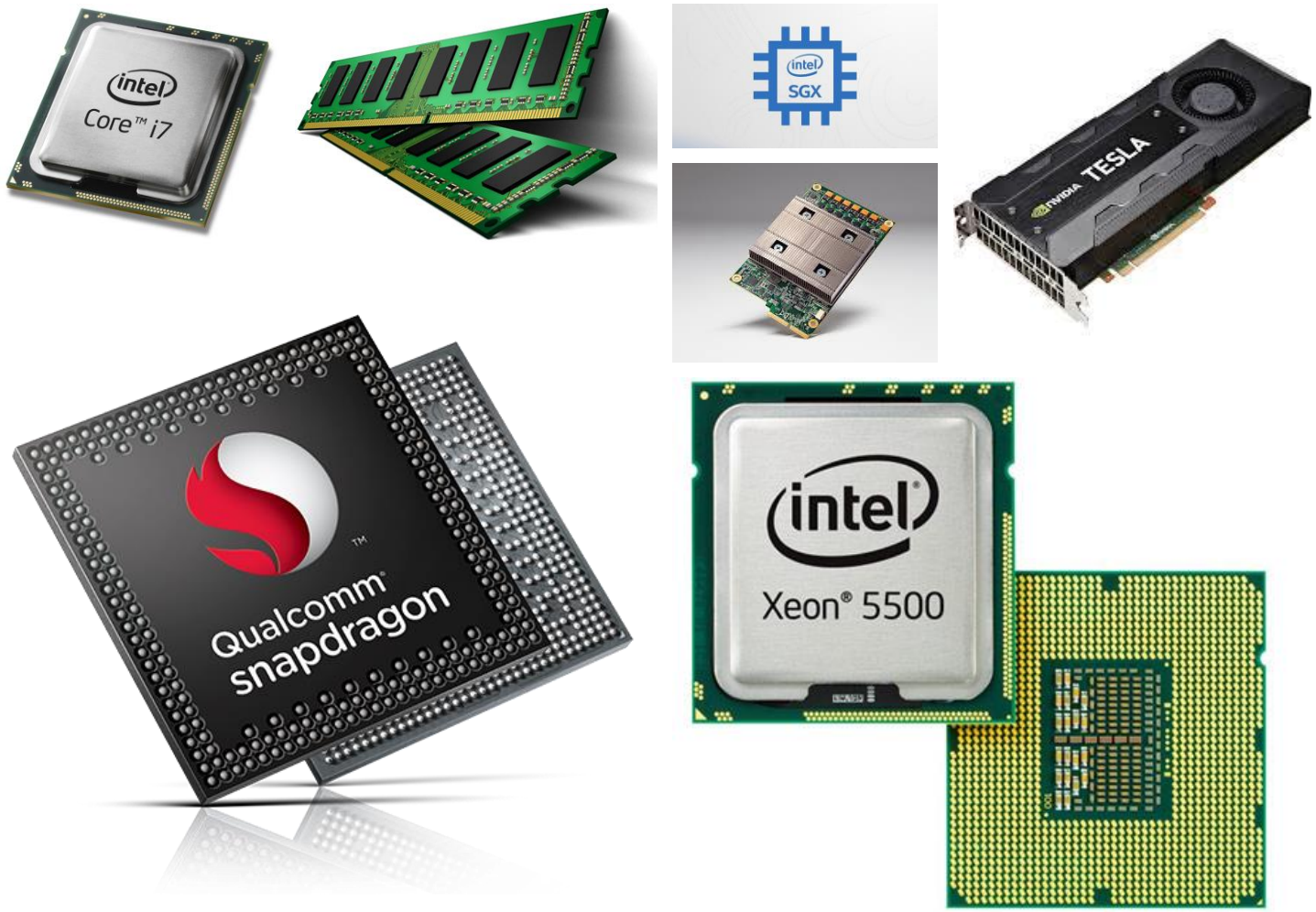
# Expectations

Ditch your excuses.

Participate in class/Piazza regularly.
Do not fear about your doubts. Just communicate.
We (you, T.A., and me) will try our best to address it.

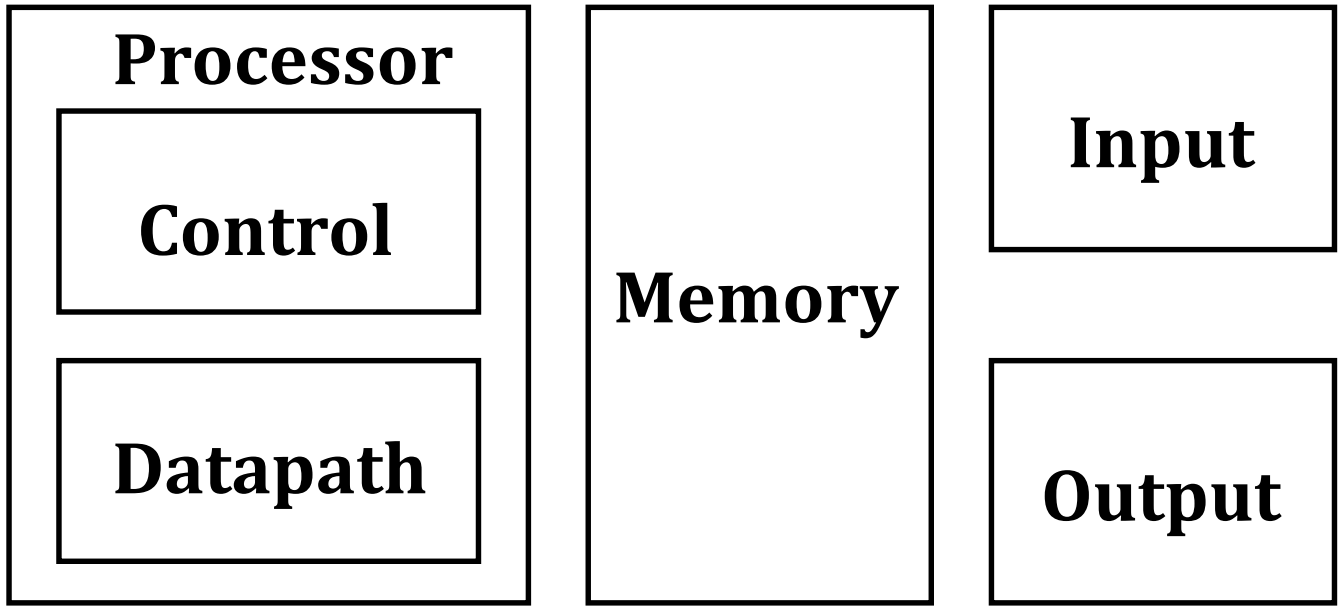
Just shout if you do not like something about me or about the course. However, be on the right side and then shout.

# Course


intel
Core™ i7
intel
SGX
TESLA
Qualcomm
snapdragon
intel
Xeon® 5500

# CS422

Since 1946 all computers have had 5 components

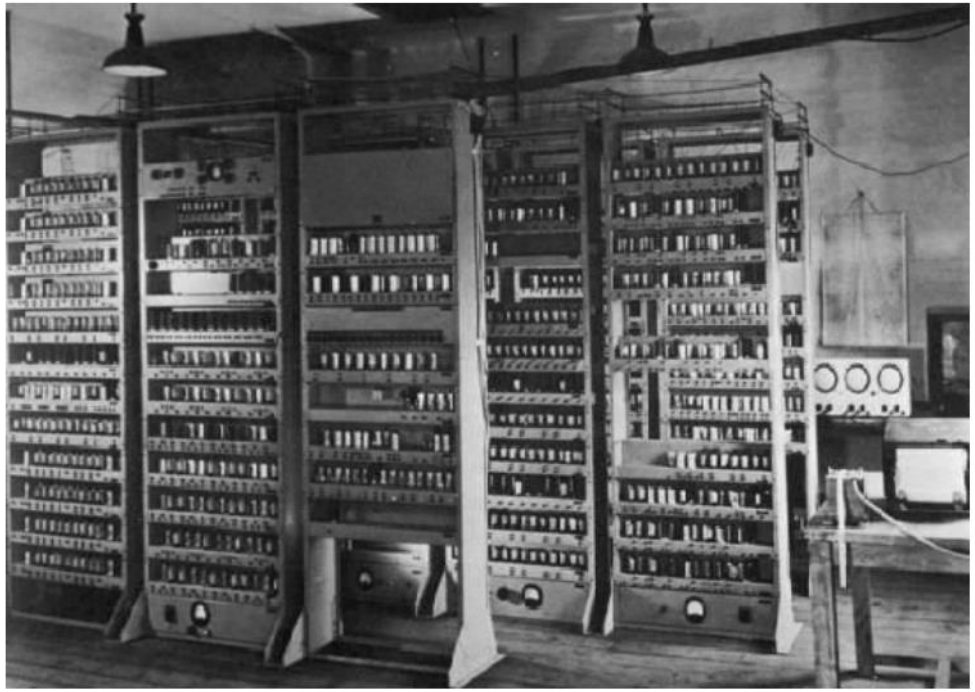


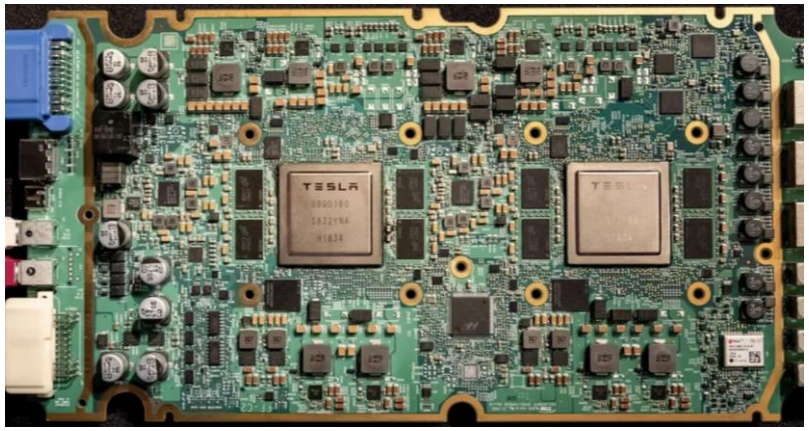

So What ?

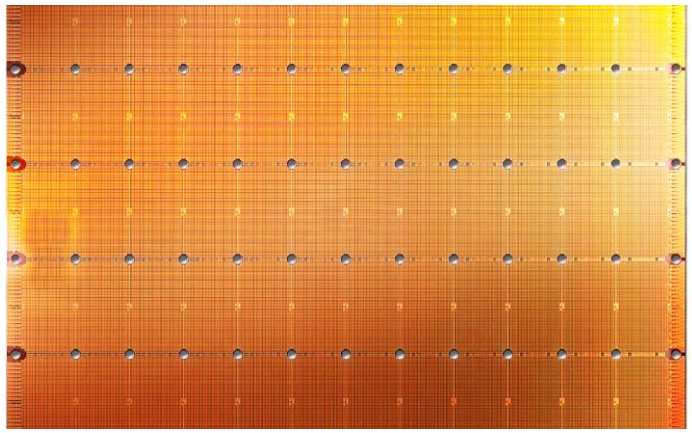
# Again

- 1950s to 1960s: Computer Arithmetic
- 1970s to mid 1980s: Instruction Set Design, especially ISA appropriate for compilers
- 1990s: Design of CPU, memory system, I/O system, Multiprocessors, Networks
- 2020s: Self adapting systems? Self organizing structures? DNA Systems/Quantum Computing?

# Then and Now

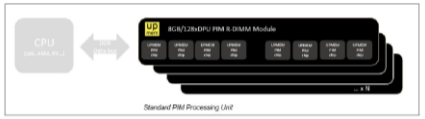




## UPMEM PIM-DRAM big data accelerator

- UPMEM DIMMs
  - **Replacing standard DIMMs**
  - **DDR4** R-DIMM modules
    - 8GB+128DPUs (16 PIM chips)
- UPMEM PIM-DRAM chips
  - **4Gb DDR4 2400 DRAM** + **8 DPUs** @500MHz
  - **Single die**, standard 2x nm DRAM process
- Massive additional compute & bandwidth
  - **2TB/s** DRAM-DPU BW for **128GB+2048 DPUs config**
- **Easily programmable SDK:** C-programmable



*PIM server: Typically with 128GB DRAM/2048 DPUs*

**Take away**

Scalable as compatible with
- Current servers
- Unmodified DRAM process
- Programmers ;)

**Samples & apps available**



HOT CHIPS 31

up mem

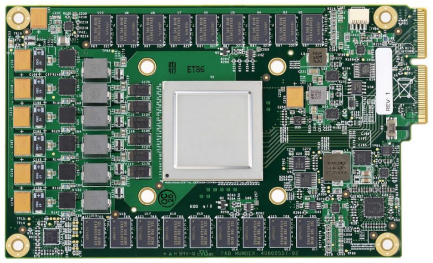



**Figure 3.** TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

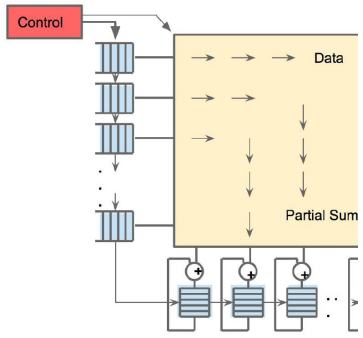
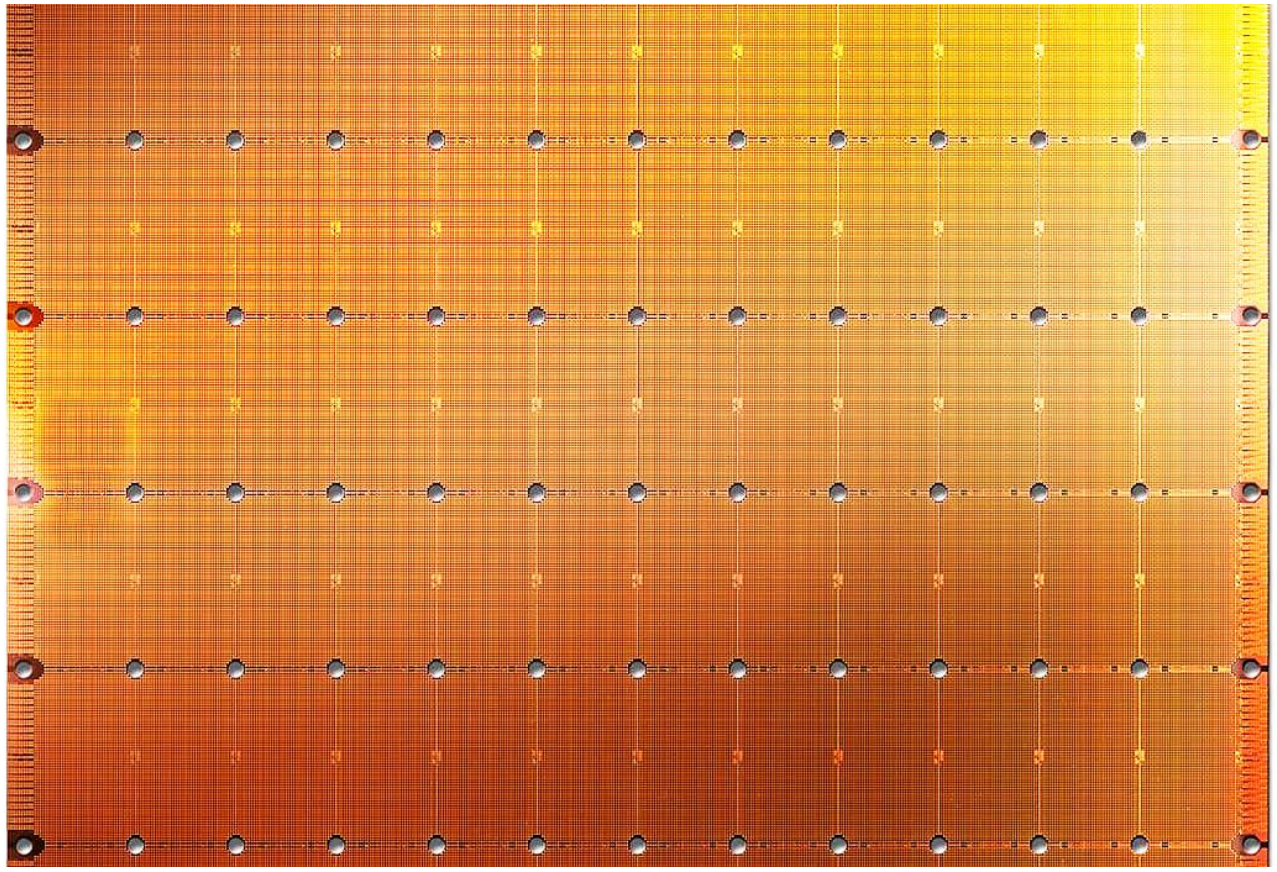



**Figure 4.** Systolic data flow of the Matrix Multiply U has the illusion that each 256B input is read at once, a update one location of each of 256 accumulator RAM

# The News

# Intel Optane

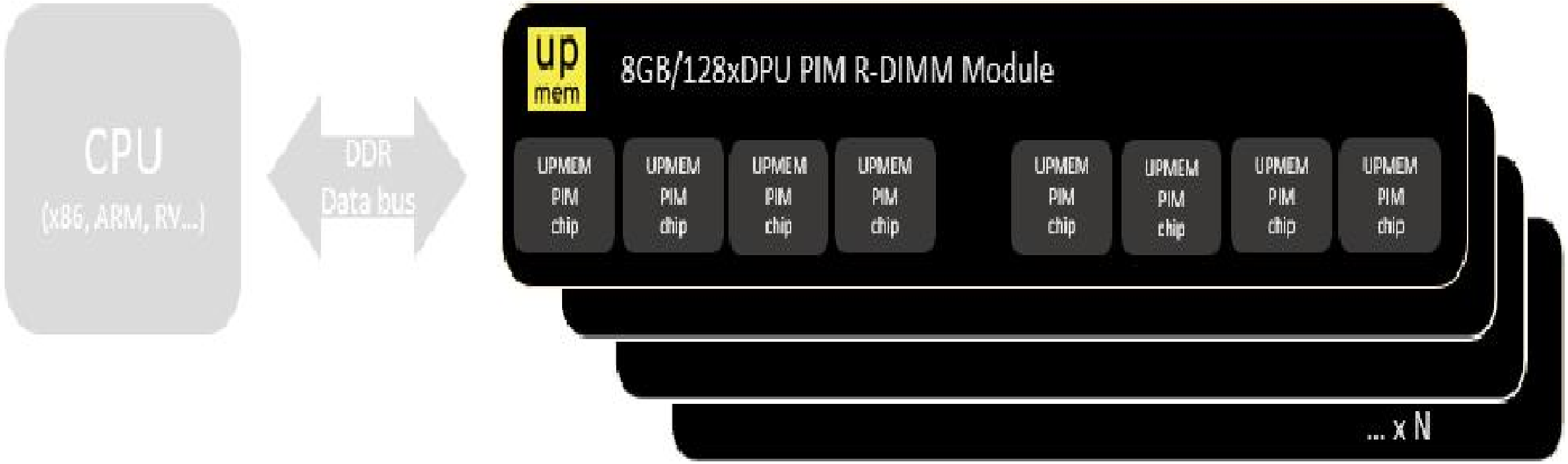


# Cerebras's Wafer Scale Engine

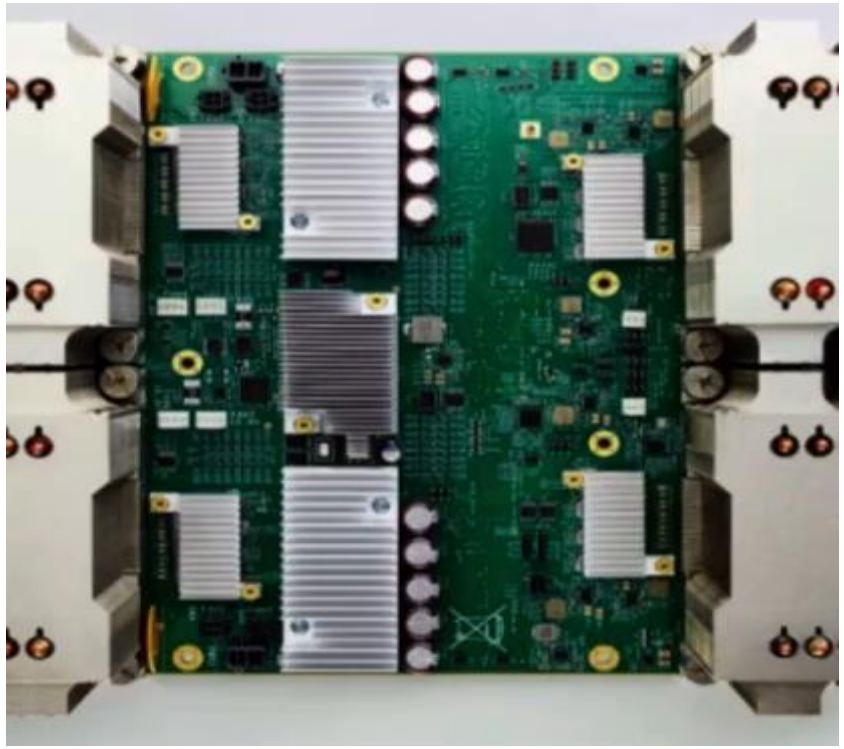
- *The largest ML accelerator chip*
- *400,000 cores*
-

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

# Processing in Memory



# Google TPU



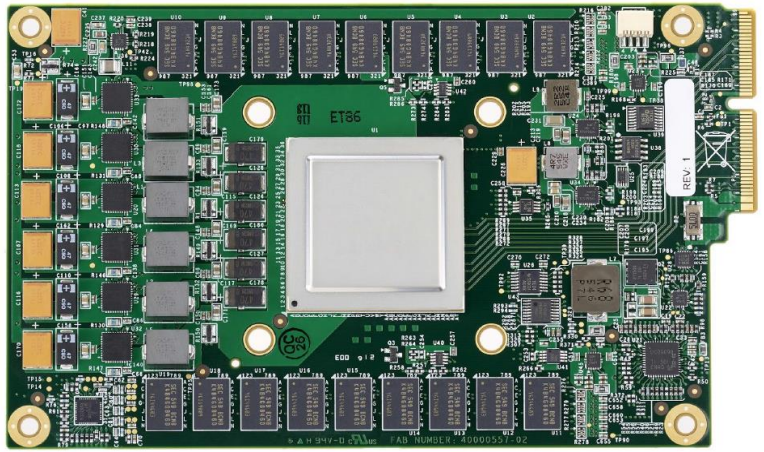


**Figure 3.** TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

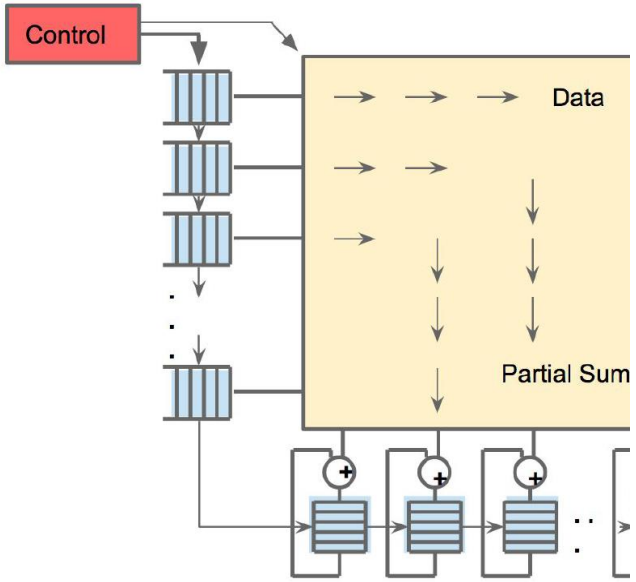
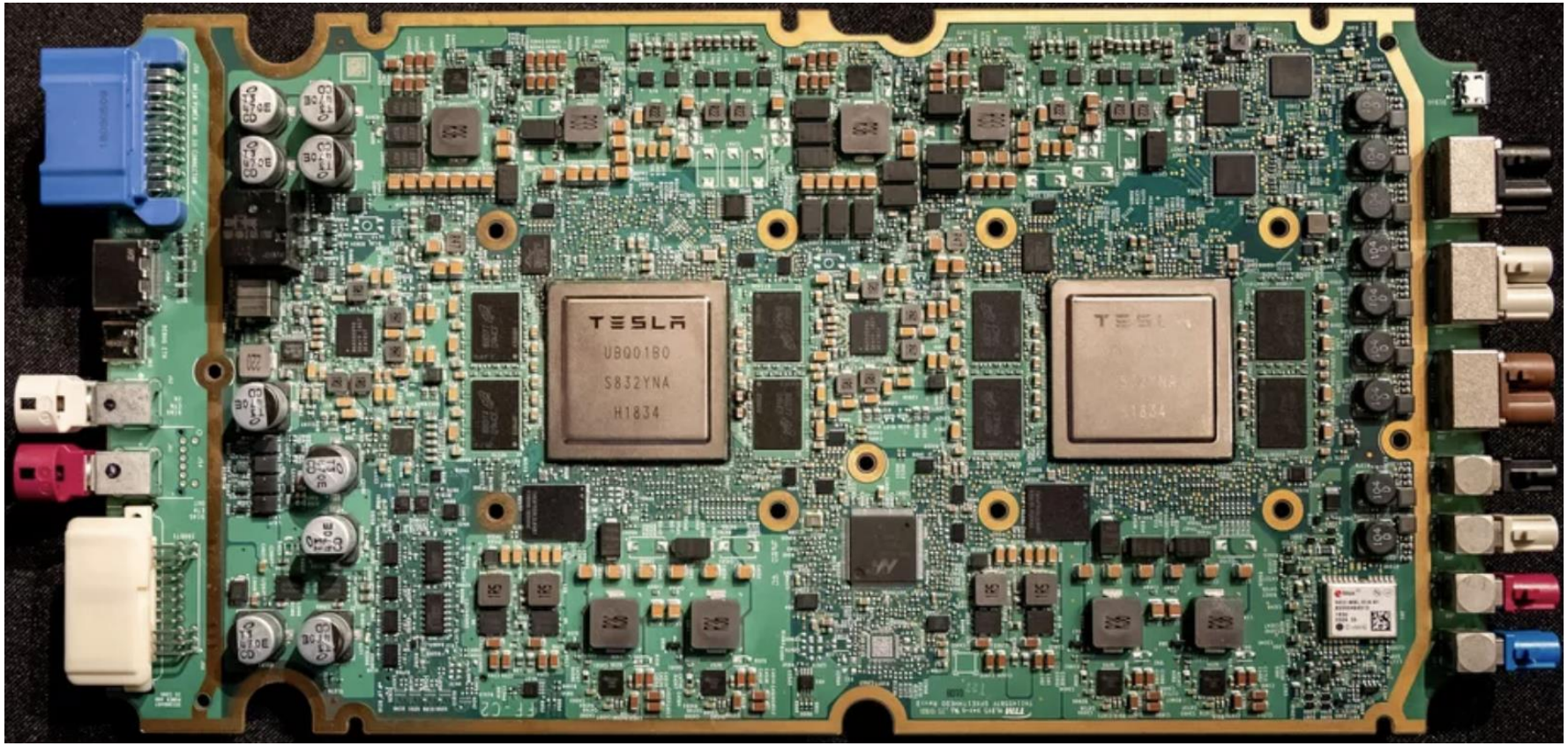


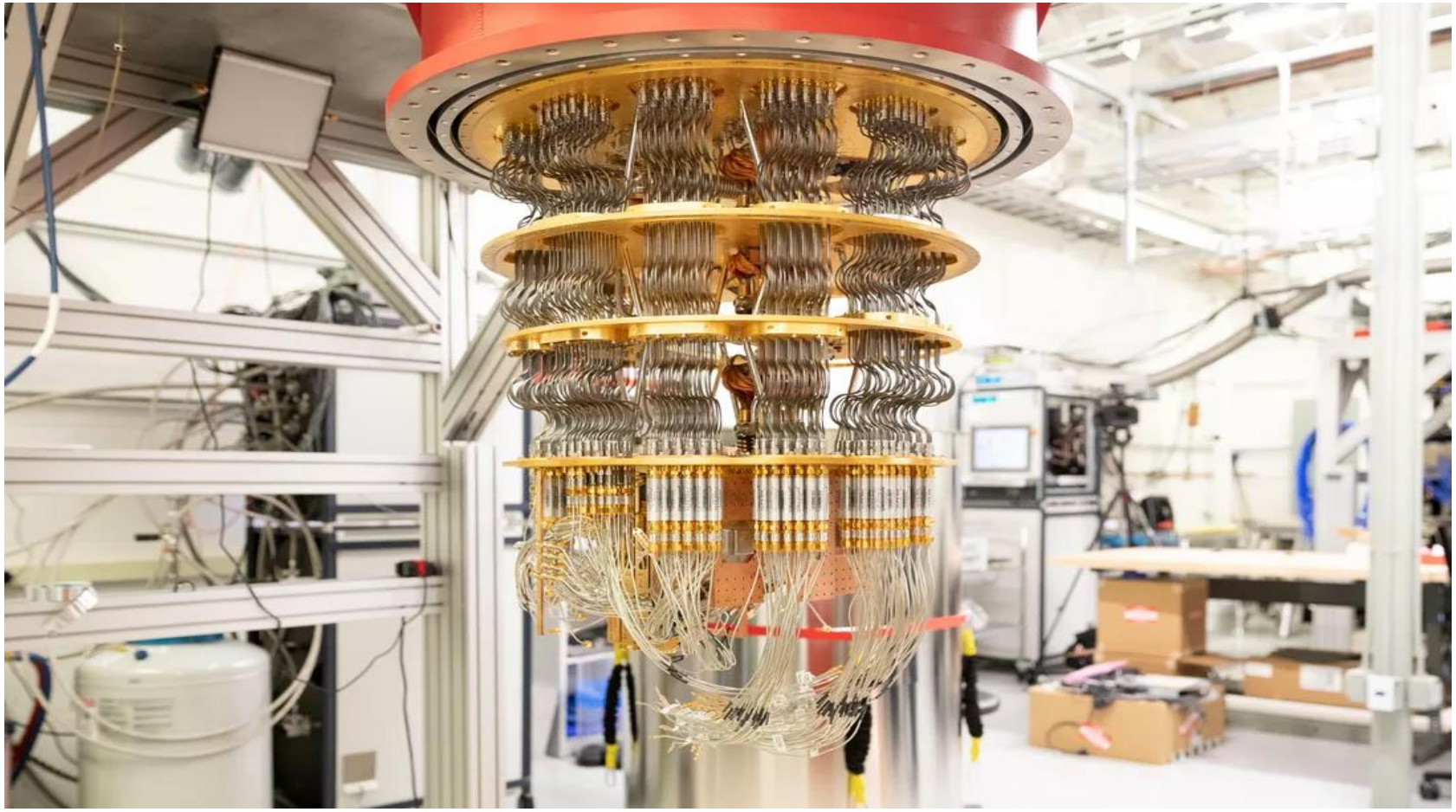

**Figure 4.** Systolic data flow of the Matrix Multiply U has the illusion that each 256B input is read at once, a update one location of each of 256 accumulator RAM

# Tesla Self-driving Car

https://youtu.be/Ucp0TTmvqOE?t=4236



# Google’s Quantum Computer



# Micro-architecture Attacks

# Abstraction? Good/bad?

It is good if you don't care about the performance of underlying entities.

**What?**

**ABSTRACTION BARRIER**

**How? Why?**

How many of you can drive a bike?

How many of you know how a bike works?

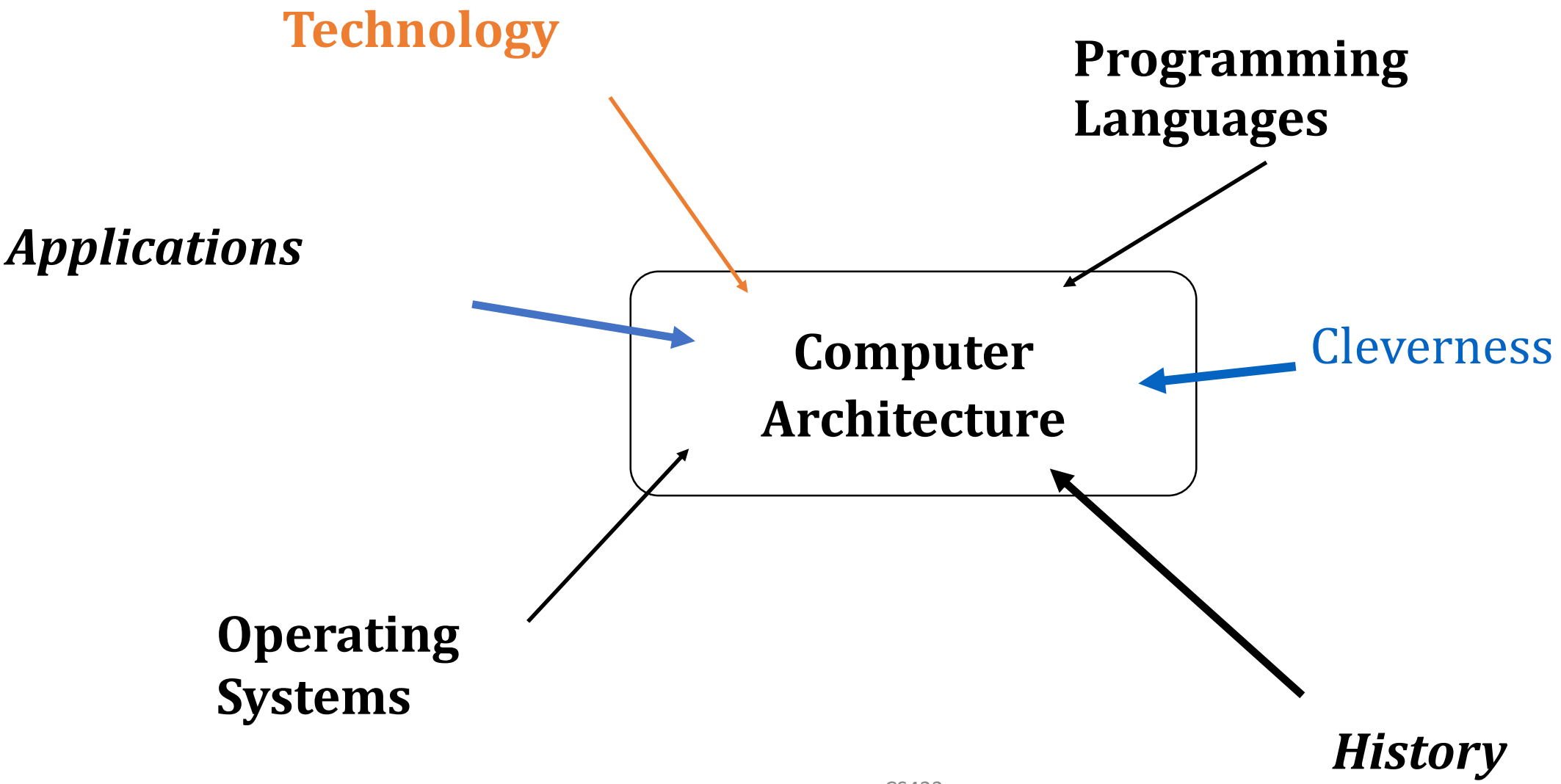
# Good until it gets bad !

What if? What if? And what if?

# CS422

Let's break our abstraction barriers

# What is Computer Architecture?



# Computer Architecture?

*VLSI++ or ++VLSI*
*Writing Verilog/VHDL code for designing a processor*
*Understanding how transistors work: Nah ☹*

*Computer theorists propose algorithms that solve important problems and analyze their asymptotic behavior (e.g., O(NlogN), O(N)). Computer architects (applicable to computer systems) set the constant factors of these algorithms* – Christos Kozyrakis, Stanford

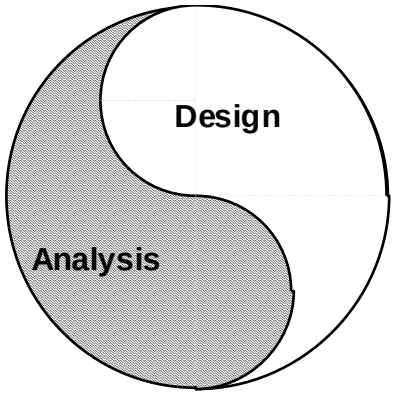
# My View

For non-CS/EE minds: Abstraction layer that enables computation in (running a C program and getting an output) hardware. The layer decides how/when/why of the enabler.

For CS/EE minds: Study of design trade-offs of different components (five) that are part of the abstraction layer. Trade-offs can be in terms of performance, power, energy, area, security, ……..

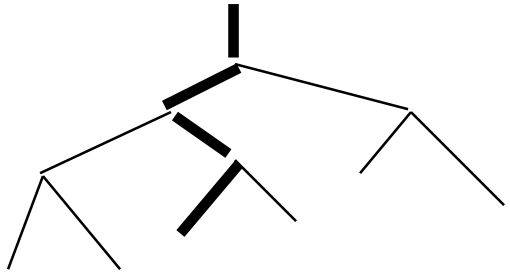
*A blog if that helps: https://medium.com/@ biswa/two-cents-on-computer-architecture-research-101-4f00957c312a*

# Design Process



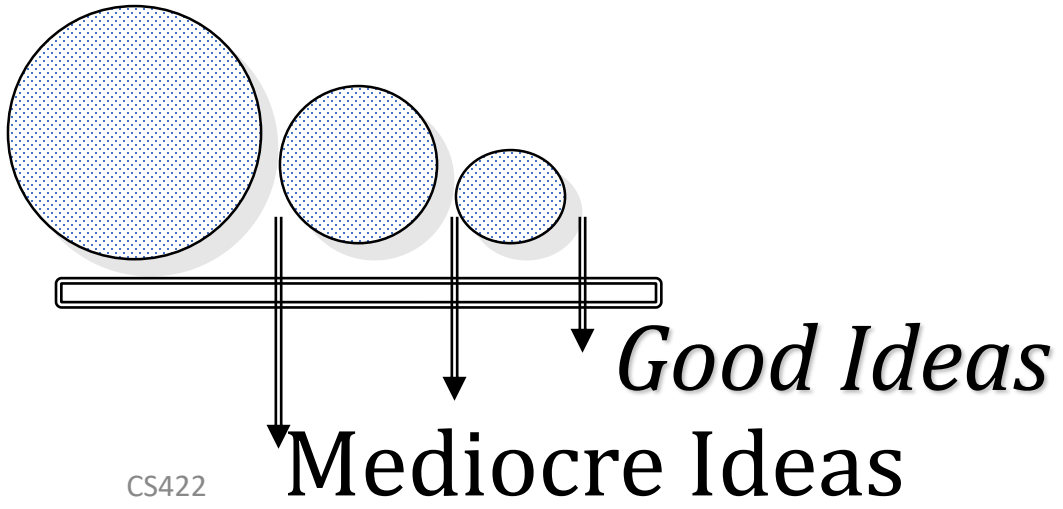

Architecture is an iterative process:

- Searching the space of possible designs
- At all levels of computer systems

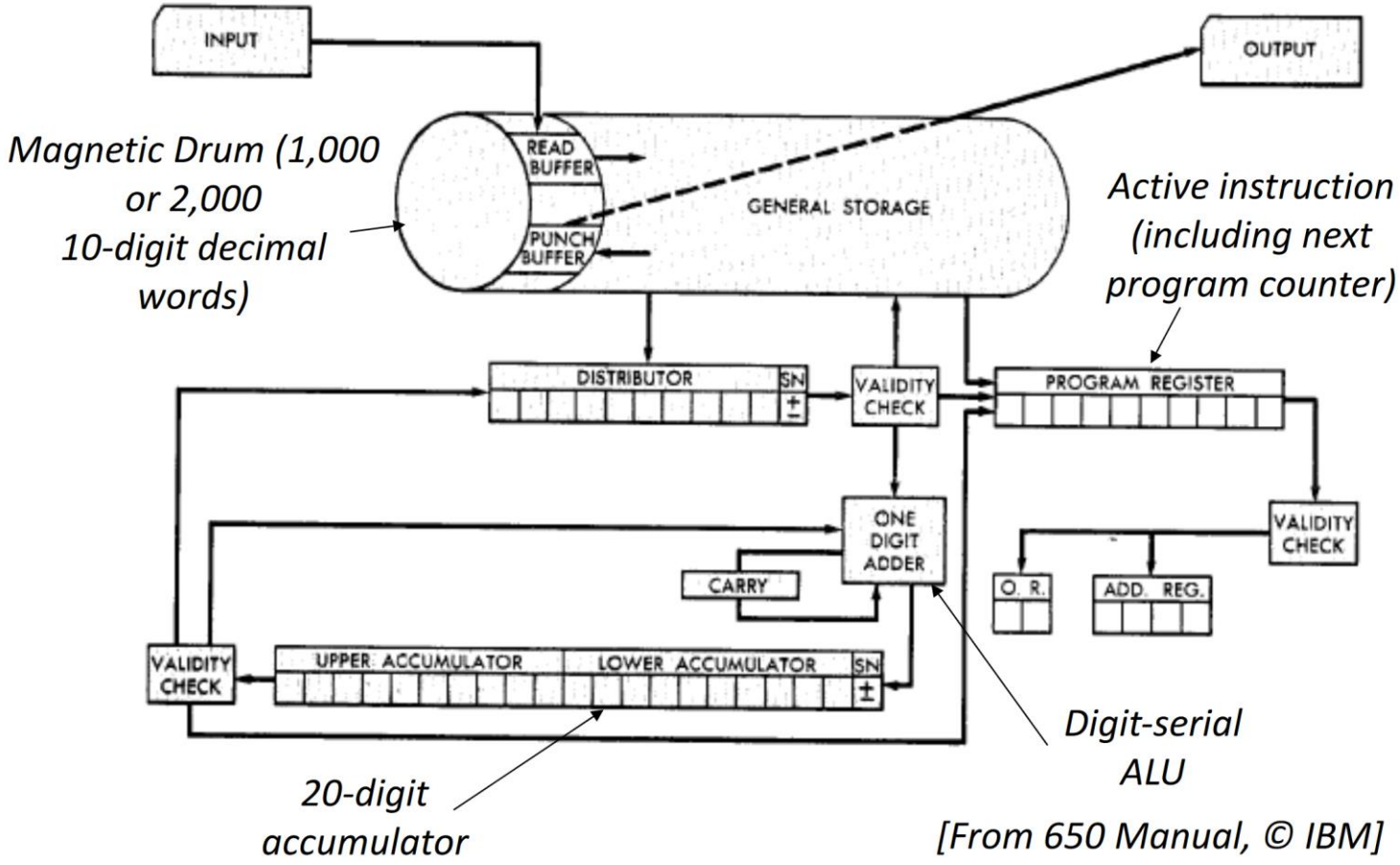


Creativity

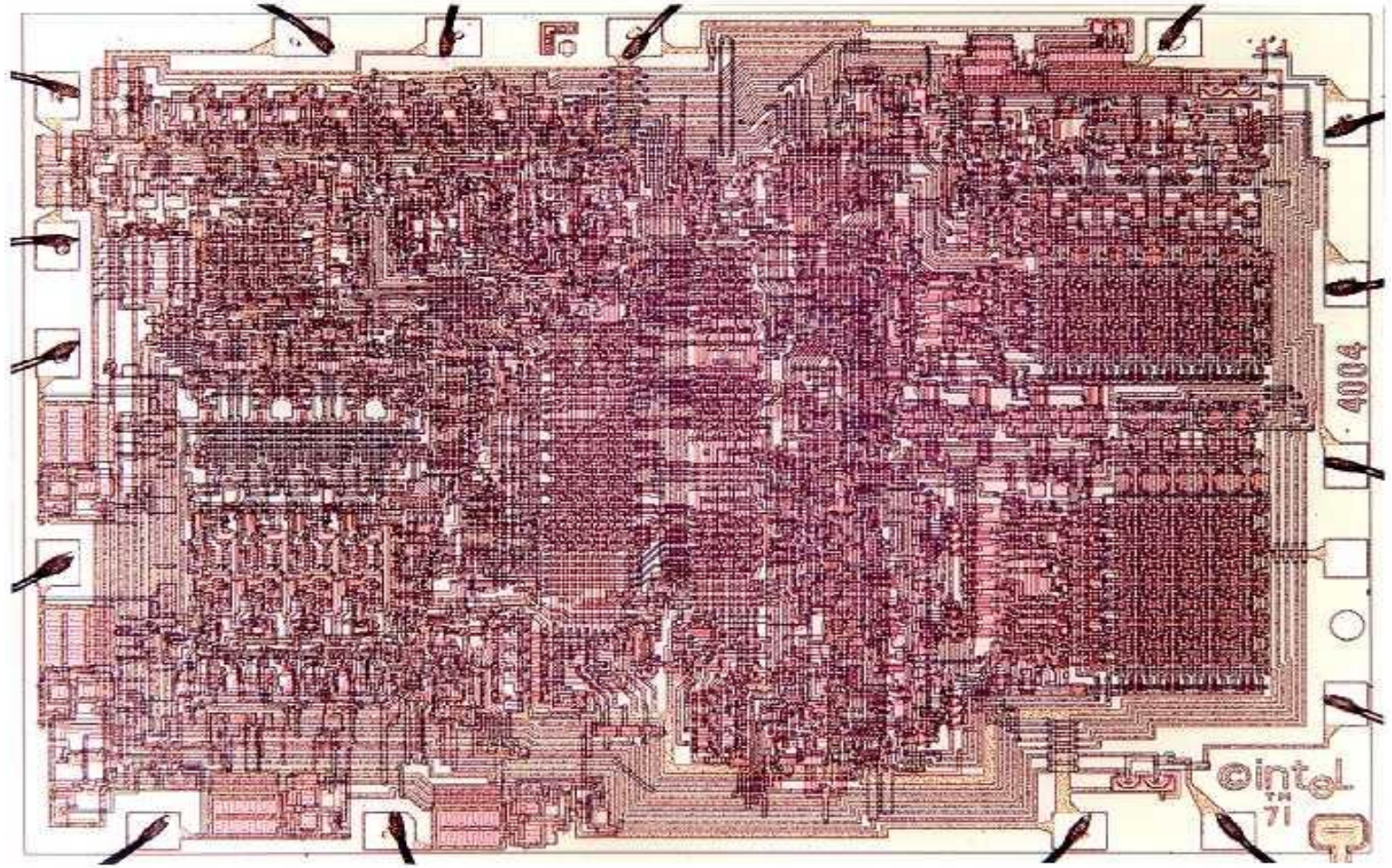
Cost /Performance Analysis



# IBM 650 (~1950s)



# 1st Microprocessor (1971, http://www.intel4004.com/)




- 4-bit accumulator architecture
- 8µm pMOS
- 2,300 transistors
- 3 x 4 mm2
- 750kHz clock
- 8-16 cycles/inst.

# x86


1978: Around 50 instructions

2020: Around 700 instructions

Time for RISC-V

# Design? Good/Bad?

# Design? Good/Bad?

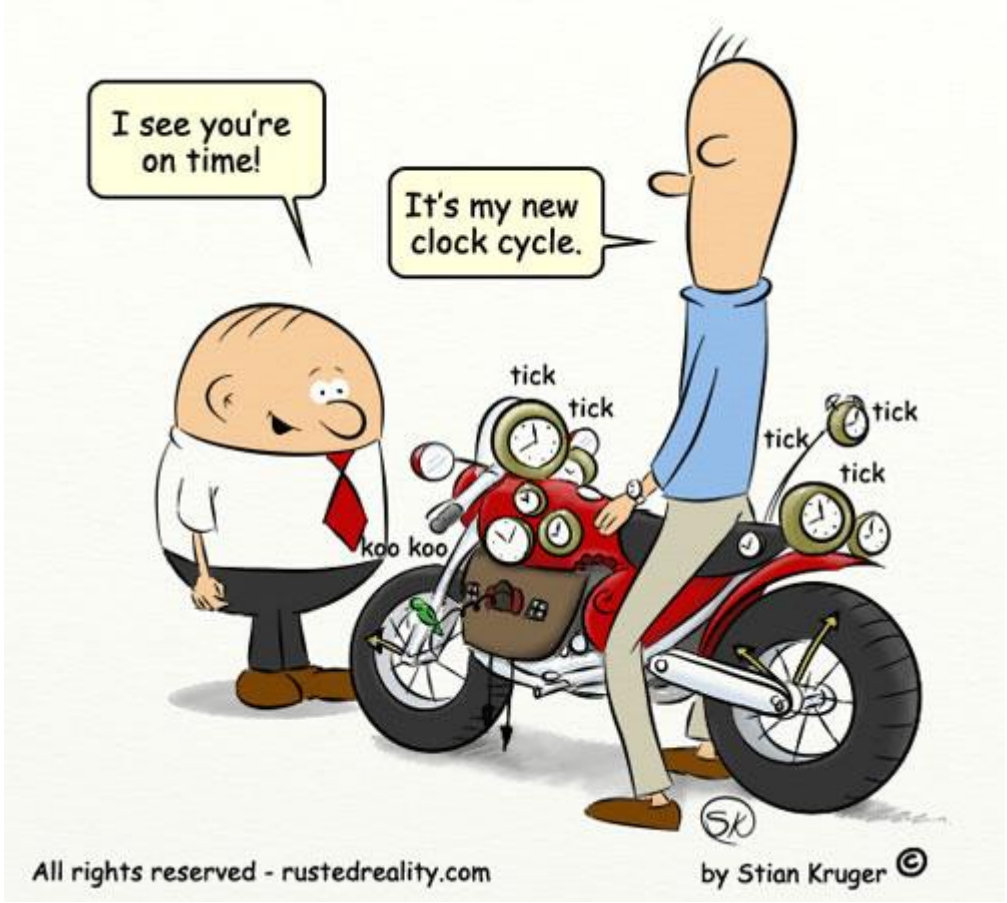
# Assignment-1 (one point): Deadline Jan. 12th

*Compare and contrast KD vs RM building*

*[Put your perspective on Piazza]*

*Assignment 0:*
*https://docs.google.com/forms/d/e/1FAIpQLSeofKOH0a9oktzGdf7zW-9GpMmcVfnxqsbgLuspFGeJ-MfRlg/viewform*

# Let’s start with performance




- Single core
- Multi core
- Evaluation
- Benchmarks
- Metrics
- Simulators
- Latency and bandwidth

# Iron Law

# Todos

*Assignment 0:* *https://docs.google.com/forms/d/e/1FAIpQLSeofKOH0a9oktzGdf7zW-9GpMmcVfnxqsbgLuspFGeJ-MfRlg/viewform*

*Assignment 1: RM vs KD building Design*

*Reading Assignment 1:* *https://medium.com/@__biswa/two-cents-on-computer-architecture-research-101-4f00957c312a*