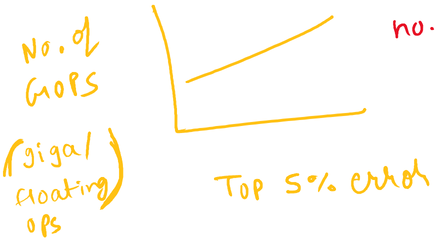
Ai accelerators 0

29 July 2025 00:18

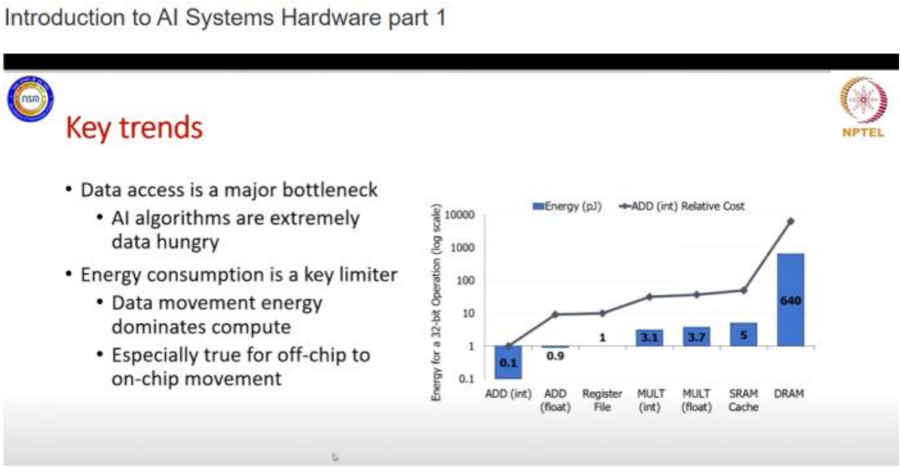
Top 5% error: depending on diff. params learned by DL algo. they produce diff. probabilities like regression/classification => whether they represent true class/not is measure of %.

Top 5% => how much % of time they can guess correct

with ↑ in accuracy no. of ops ↑ exponentially.



more energy spend to access data => more heat generate => more cooling sys. deployed => ↑ cost


Introduction to AI Systems Hardware part 1

Key trends

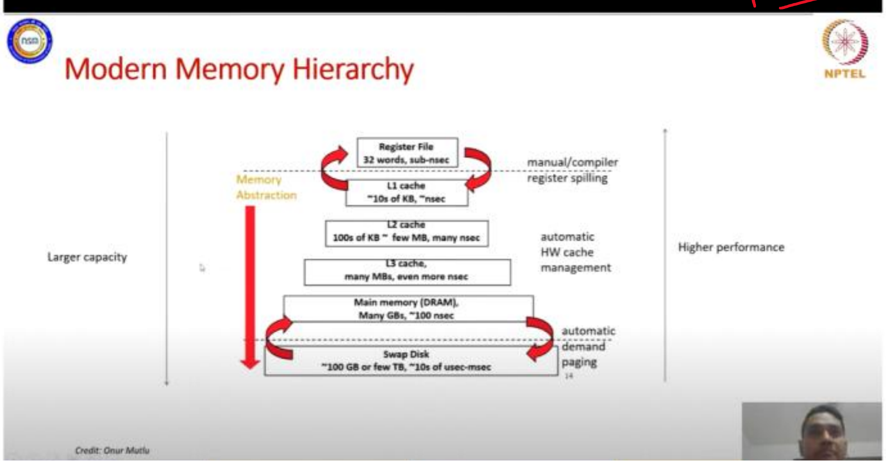
- Data access is a major bottleneck
  - AI algorithms are extremely data hungry
- Energy consumption is a key limiter
  - Data movement energy dominates compute
  - Especially true for off-chip to on-chip movement

Traditional Computation system

memory → Communication unit ↔ chip

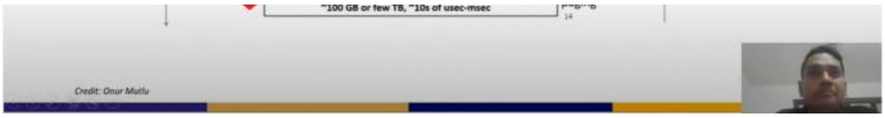
↳ on-chip memory

↳ off-chip memory, need offchip interconnect to access memory system.


Introduction to AI Systems Hardware part 1

Modern Memory Hierarchy

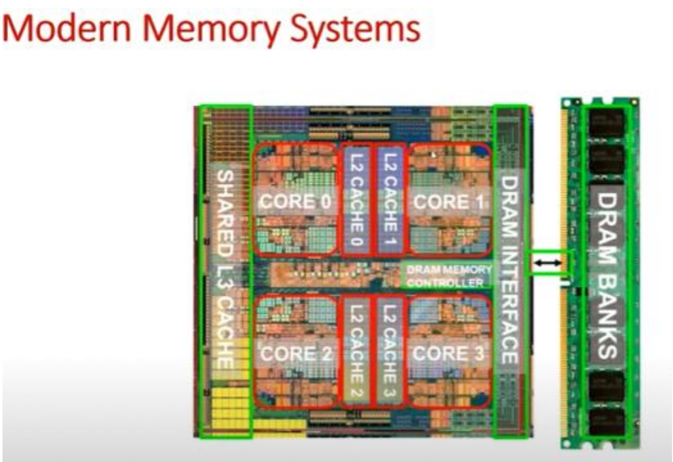
Credit: Onur Mutlu



~100 GB or few TB, ~10s of usec-msec

Credit: Onur Mutlu

modern computation systems



Modern Memory Systems

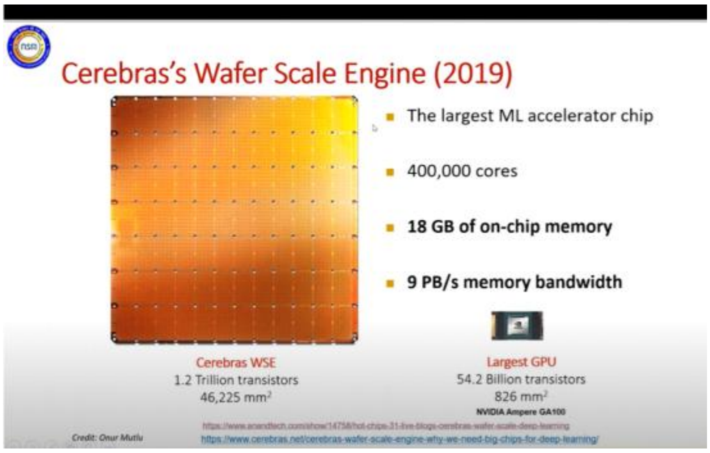
↳ swap-disk extendable to remote storage through diff- gate-ways

why we need !

↳ @ core level access time is very slow

↳ to ↑ performance &

↳ faster data access. } Major bottleneck



Cerebras's Wafer Scale Engine (2019)

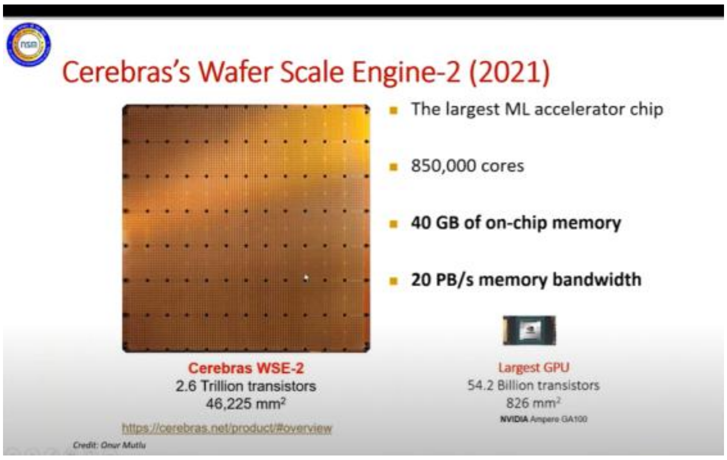
- The largest ML accelerator chip
- 400,000 cores
- **18 GB of on-chip memory**
- **9 PB/s memory bandwidth**

Cerebras WSE
1.2 Trillion transistors
46,225 mm²

Largest GPU
54.2 Billion transistors
826 mm²
NVIDIA Ampere GA100

https://www.anandtech.com/show/14758/hot-chips-31-live-blogs-cerebras-wafer-scale-deep-learning
https://www.cerebras.net/cerebras-wafer-scale-engine-why-we-need-big-chips-for-deep-learning/

Credit: Onur Mutlu



Cerebras's Wafer Scale Engine-2 (2021)

- The largest ML accelerator chip
- 850,000 cores
- **40 GB of on-chip memory**
- **20 PB/s memory bandwidth**

Cerebras WSE-2
2.6 Trillion transistors
46,225 mm²

Largest GPU
54.2 Billion transistors
826 mm²
NVIDIA Ampere GA100

https://cerebras.net/product/#overview

Credit: Onur Mutlu

↳ Processors are basic computation engines.

↳ one cannot ↑ clk speed anymore due to

↳ Processors [illegible] basic [illegible]

↳ one cannot ↑ clk speed anymore due to denard scalling. & power-wall heating.

↳ that`s why one cannot get more than a particular fixed clk freq^n i.e. available