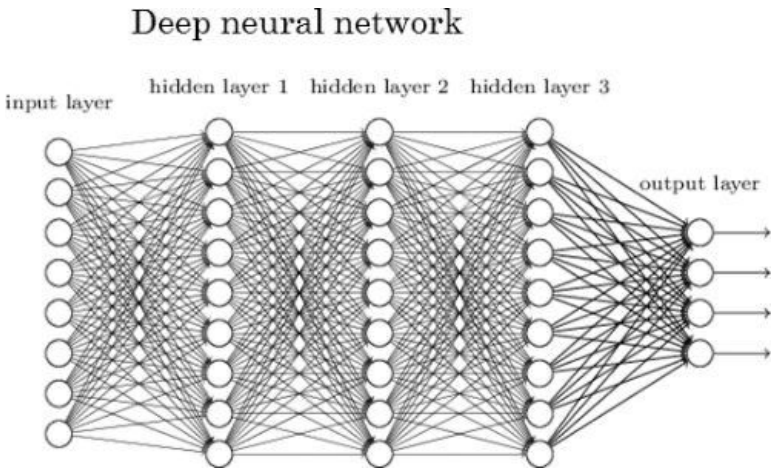
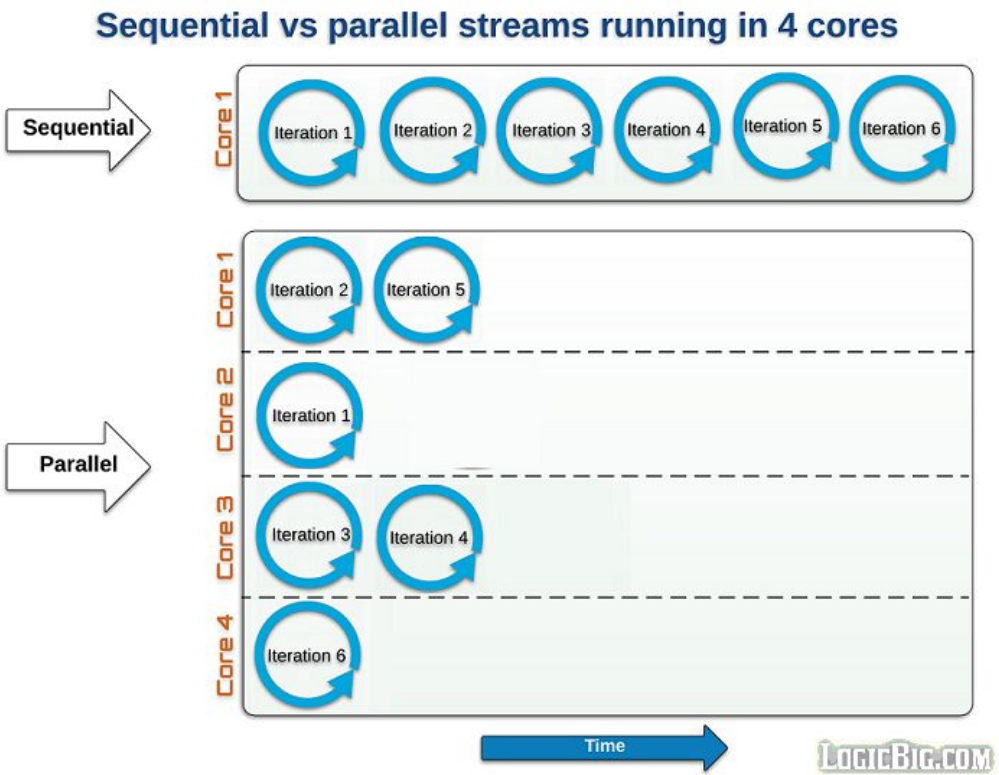
Machine Learning Hardware

Areeb Gani, Michael Ilie, Vijay Shanmugam

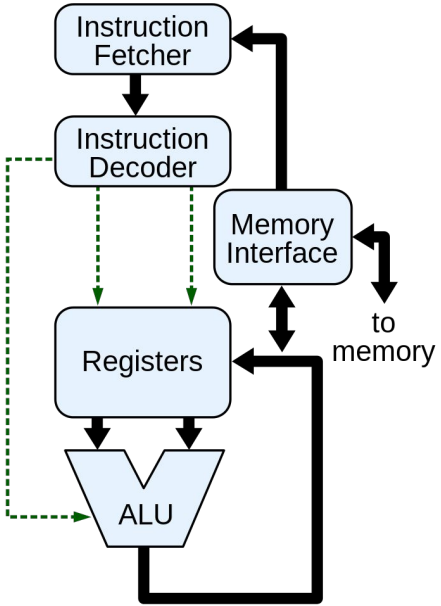
Quick Refresh

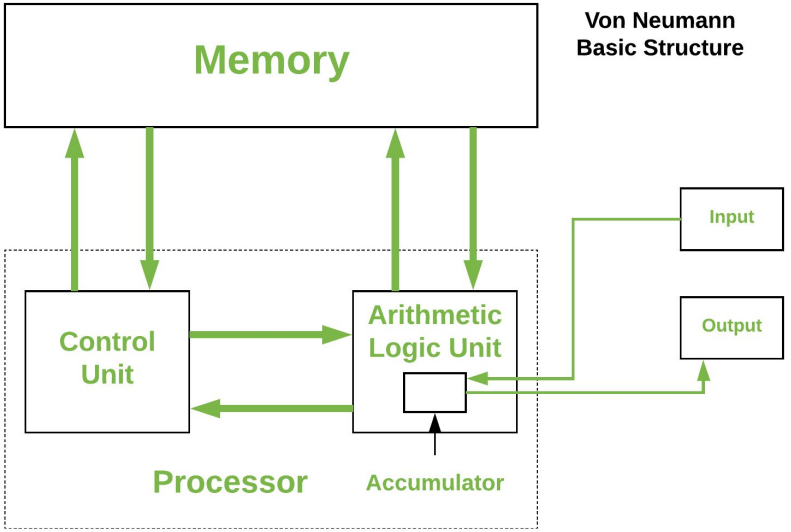
- Neural networks look like this:
  - 

Sequential vs Parallel



What Is a Processor

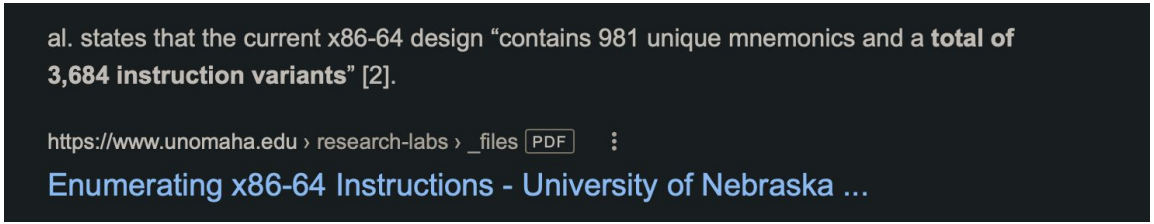




What is a CPU

- General is comprised up of cores
- Central processing unit aka its very general
- Different kinds of functions and use cases, meant to be general
- Different architectures and ISAs (arm64, x86, amd_64, riscv-rv32i, etc)
- Each different architecture has different *architecture*
  - Risc vs cisc
  - Risc = reduced instructions
- Caches (levels), memory, Decode width, ipc, ghz, vliw, simd, mimd, etc

What is x86/amd64?

- Dominant architecture on pc desktops and (for now) servers
- Strong single core performance
- very very very general
- Several extensions
- BLOAT
  - 

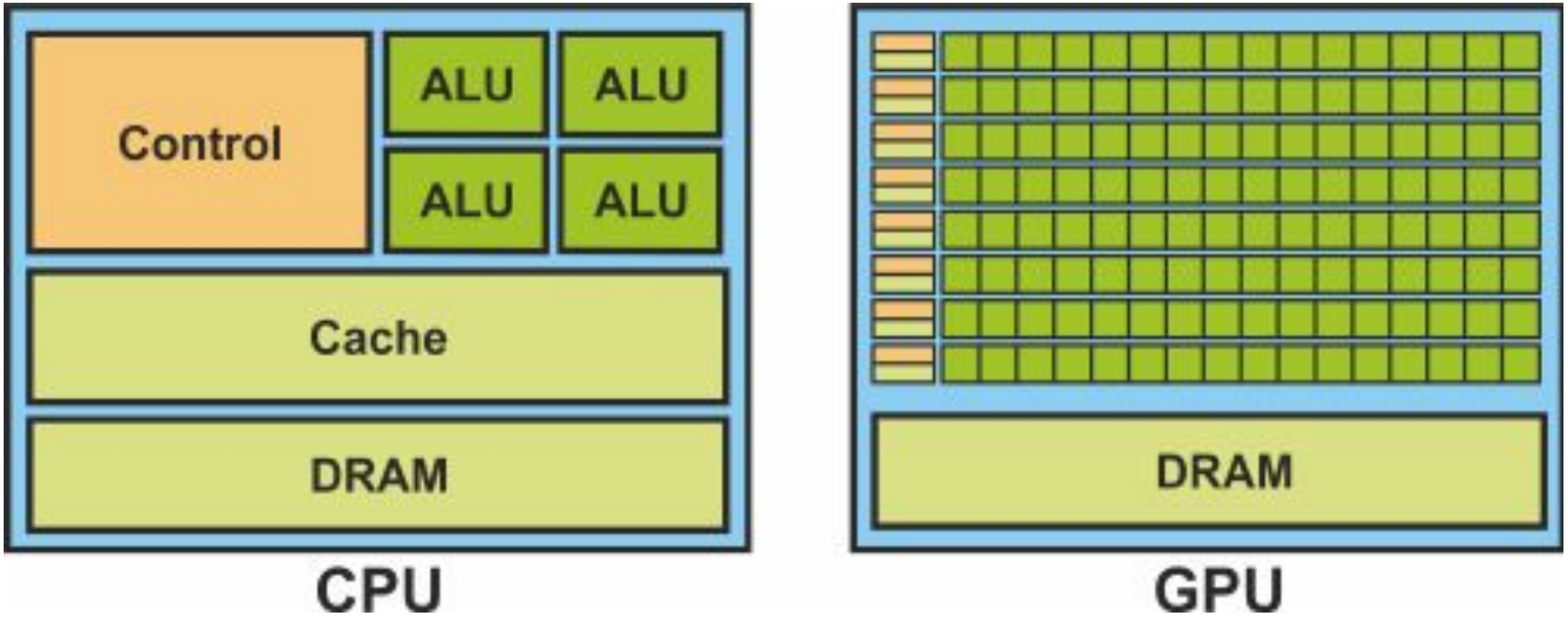
Custom Silicon/Hardware

- Can use an open ISA or some alternative to x86
- Can just be even a purely custom ISA and IC
- Does not have to adhere to x86/amd64 requirements
  - Does not have to have all the same instructions
  - Does not have to be so general
  - Can be more specialized into one task
-

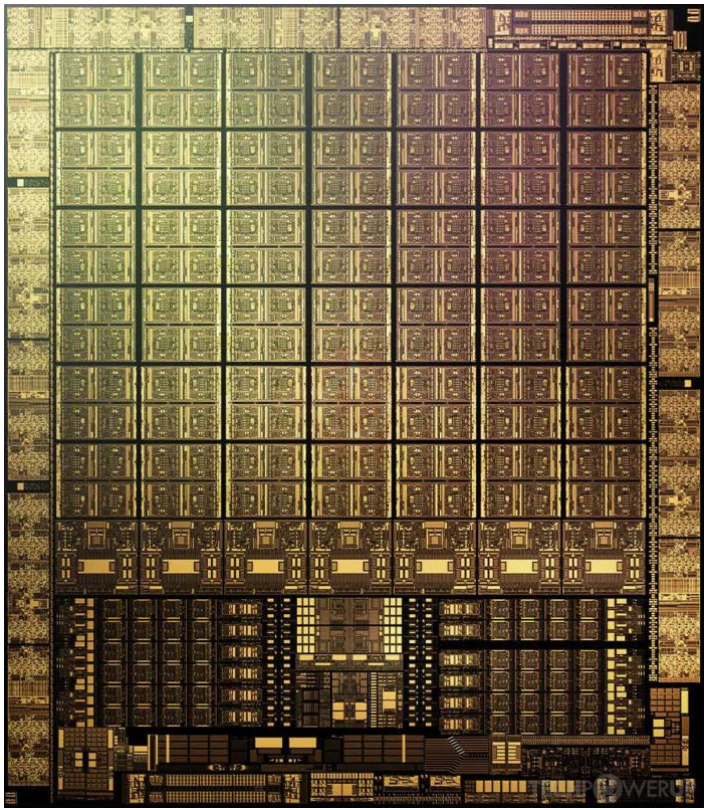
What is a GPU

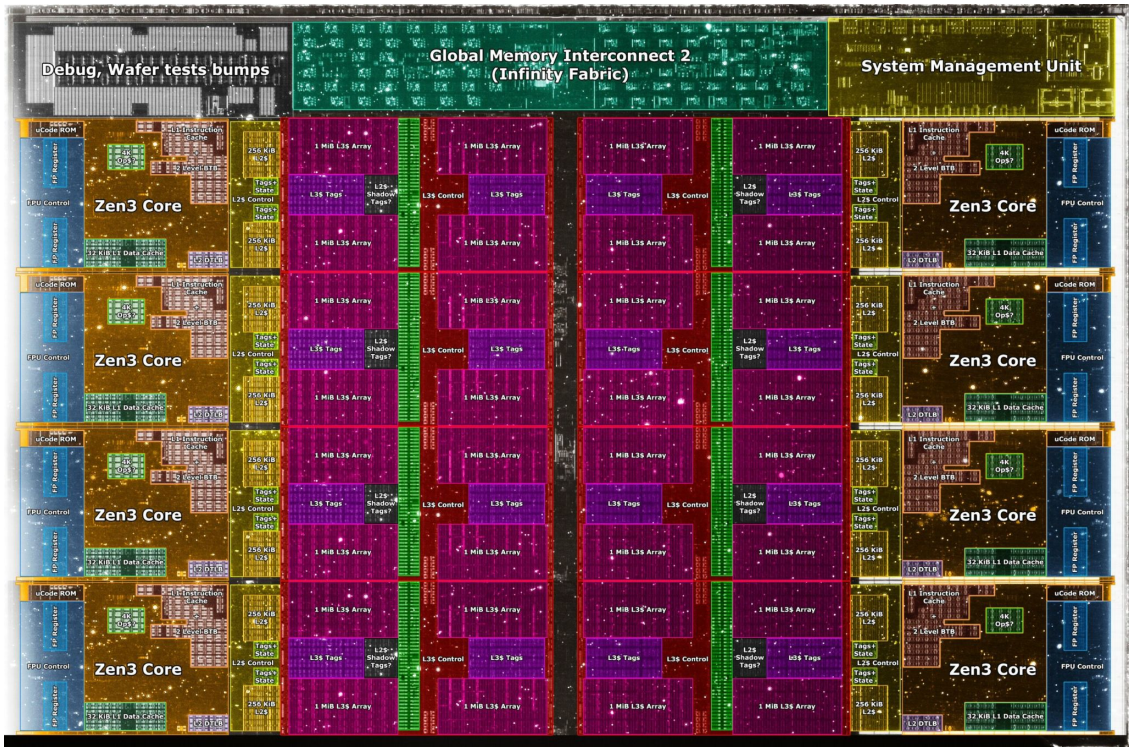
- Graphics processing unit
- Used for games and graphical processing (ie rendering)
- People realised that cpus don't really work so fast with images
- Images are giant matrixes of pixels
- Thus were created custom ICs and silicon that were really good at handling large matrices
- (spoiler) this turns out to be useful in machine learning
- But why are gpus better than cpus?

GPU vs CPU



GPU vs CPU Dieshot



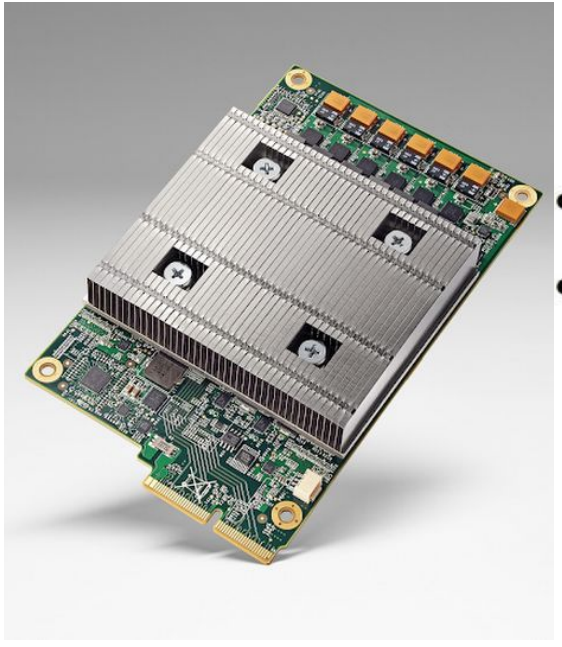


Specific GPU Example

(look at int32/fp32)

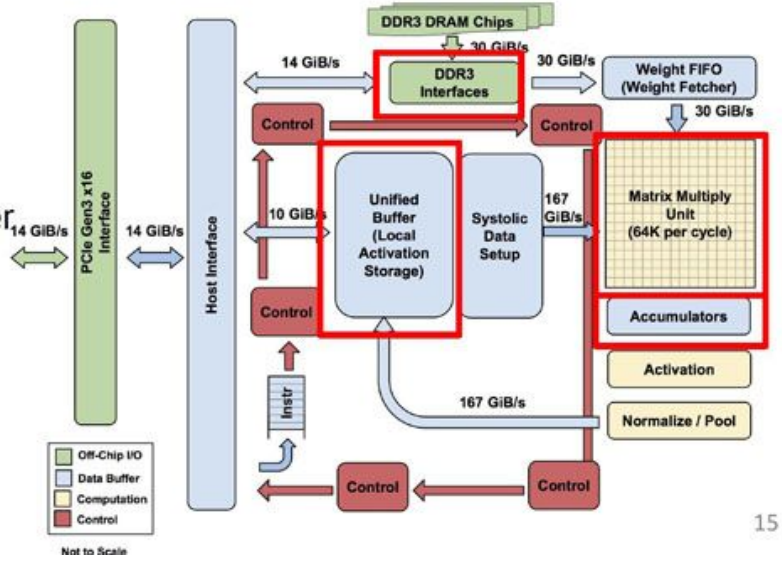
Sometimes gpus are too general even

Custom Silicon



- The Matrix Unit: 65,536 (256x256) 8-bit multiply-accumulate units
- 700 MHz clock rate
- Peak: 92T operations/second
  - 65,536 * 2 * 700M
- >25X as many MACs vs GPU
- >100X as many MACs vs CPU
- 4 MiB of on-chip Accumulator memory
- 24 MiB of on-chip Unified Buffer (activation memory)
- 3.5X as much on-chip memory vs GPU
- Two 2133MHz DDR3 DRAM channels
- 8 GiB of off-chip weight DRAM memory

TPU: High-level Chip Architecture



Fpgas