Edge Computing

Lecture 07: Hardware and Accelerators

Recap

World of ML

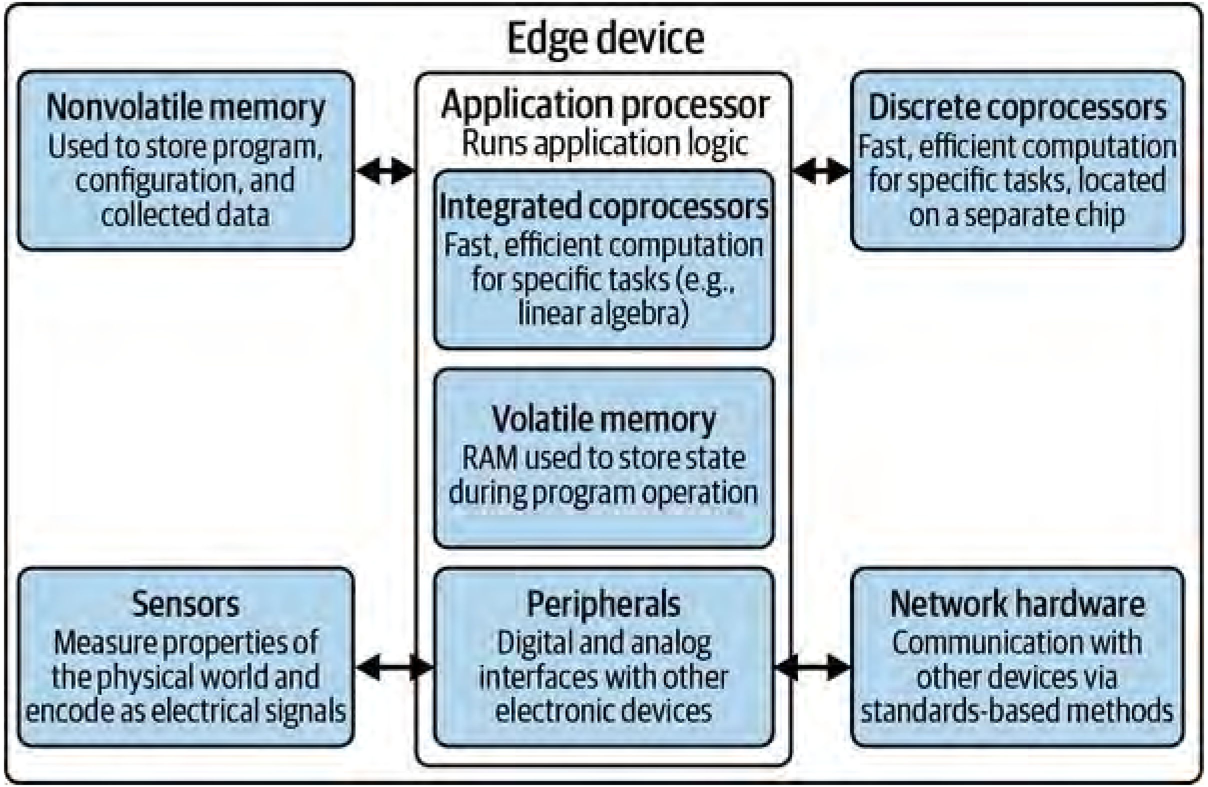
- Neural network: terminology
- Common building block: layer
- Convolution neural network
- Pruning
- Quantization

Agenda

ML on hardware

- CPU
- Memory
- Cache
- RISC vs CISC
- Special accelerators
- Sensors

Edge Device Hardware Architecture



Edge device

Nonvolatile memory
Used to store program, configuration, and collected data

Application processor
Runs application logic

Integrated coprocessors
Fast, efficient computation for specific tasks (e.g., linear algebra)

Volatile memory
RAM used to store state during program operation

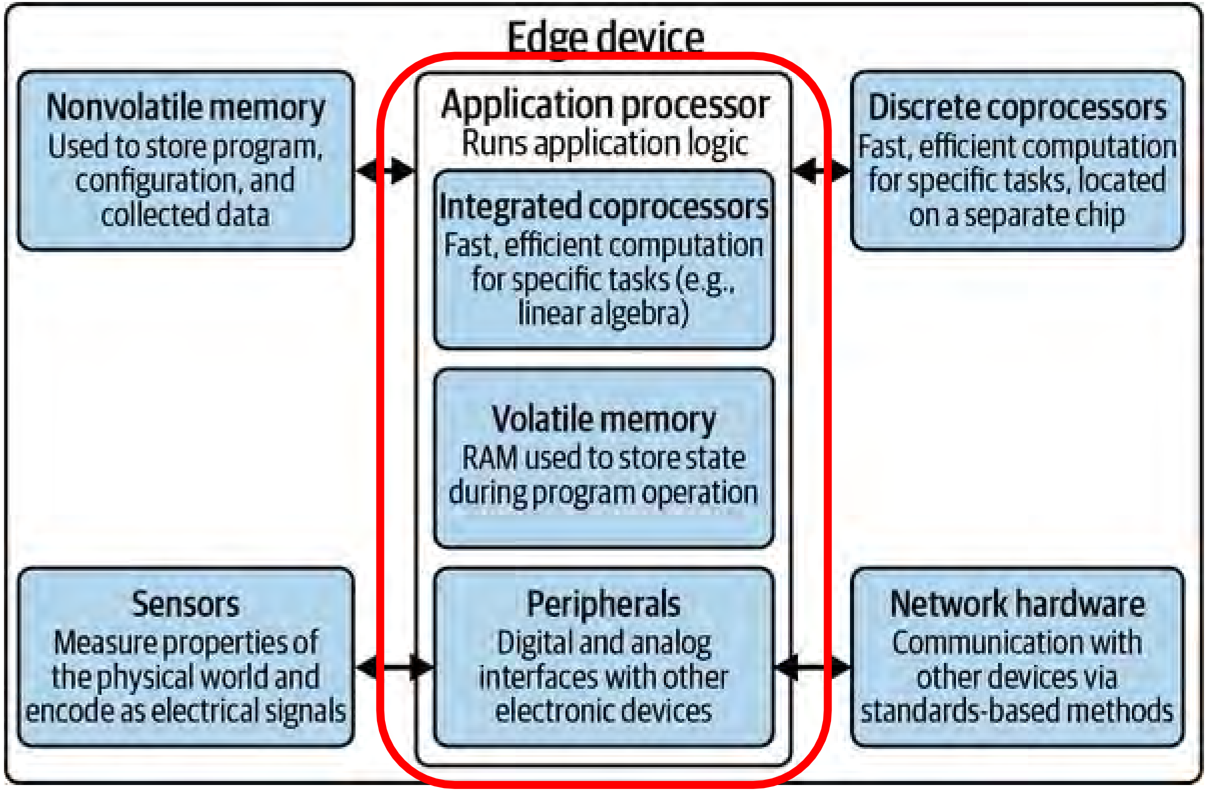
Peripherals
Digital and analog interfaces with other electronic devices

Discrete coprocessors
Fast, efficient computation for specific tasks, located on a separate chip

Sensors
Measure properties of the physical world and encode as electrical signals

Network hardware
Communication with other devices via standards-based methods

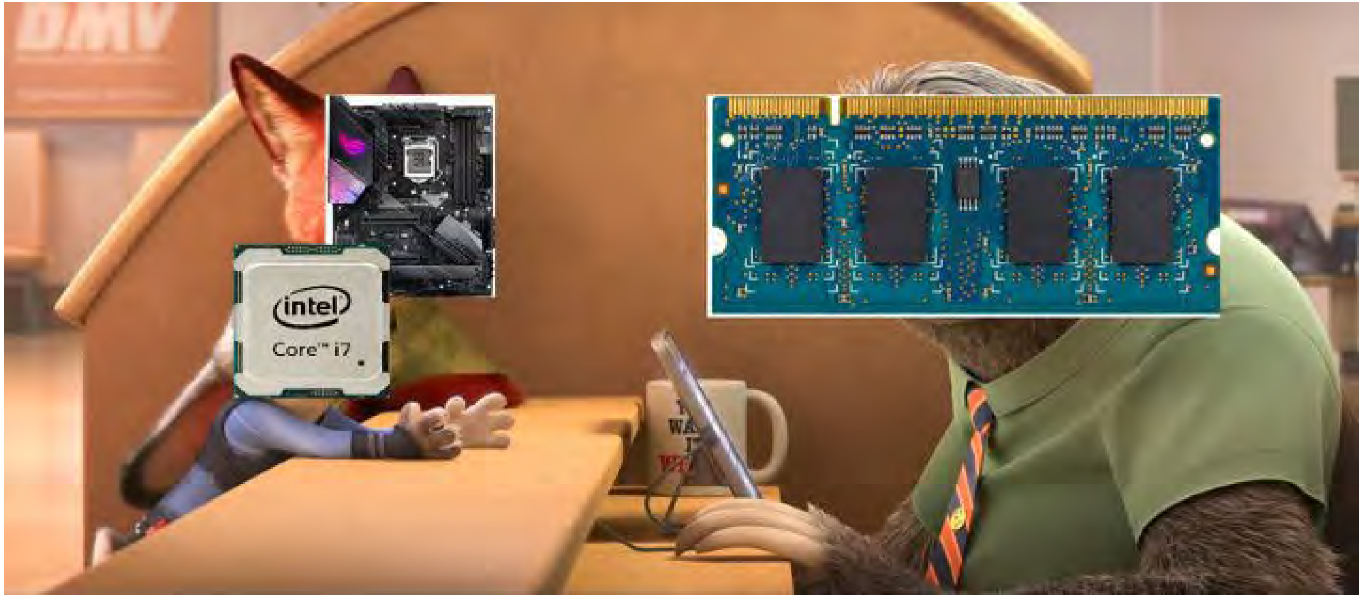
Edge Device Hardware Architecture



Edge device
Nonvolatile memory
Used to store program, configuration, and collected data
Application processor
Runs application logic
Integrated coprocessors
Fast, efficient computation for specific tasks (e.g., linear algebra)
Volatile memory
RAM used to store state during program operation
Peripherals
Digital and analog interfaces with other electronic devices
Discrete coprocessors
Fast, efficient computation for specific tasks, located on a separate chip
Sensors
Measure properties of the physical world and encode as electrical signals
Network hardware
Communication with other devices via standards-based methods

CPUs

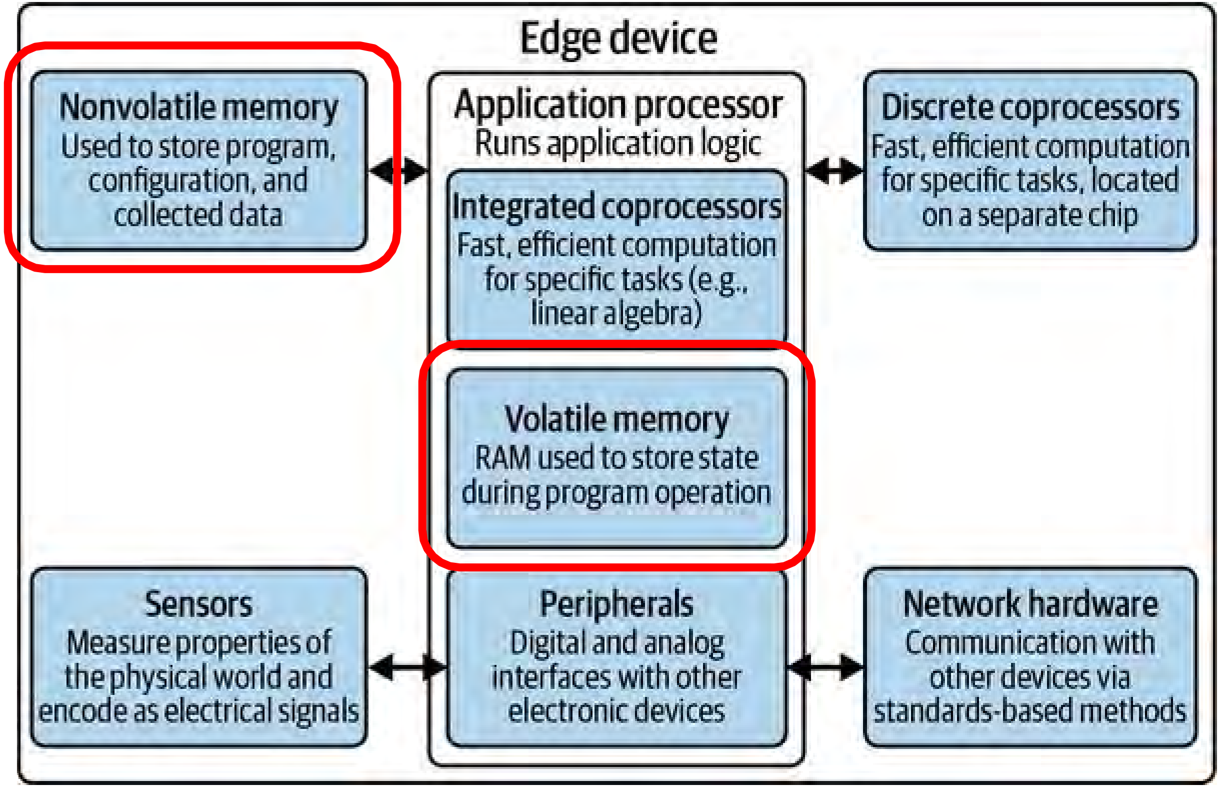
- Clockspeed
- Cores
- Threads
- Memory speed
- Cache size



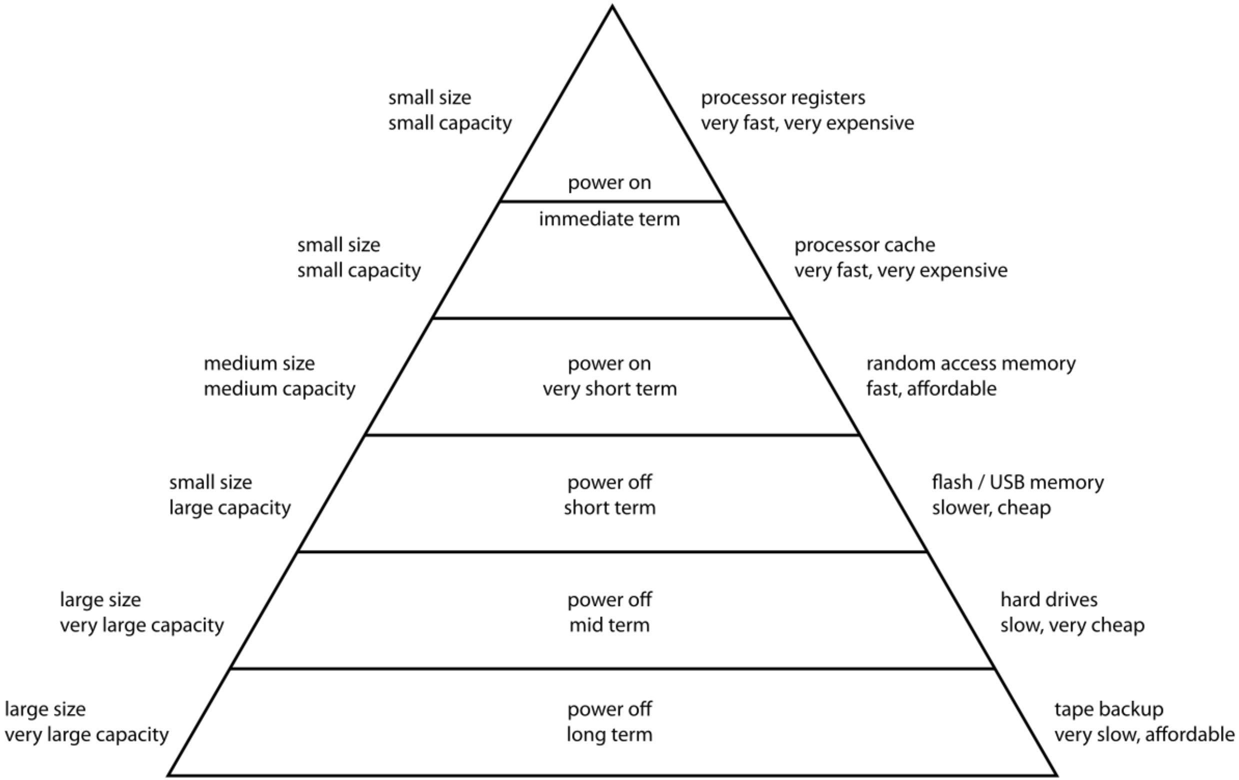
Example: Intel® Core™ i9 processor 14900K

Movie: Zootopia

Edge Device Hardware Architecture


Edge device
Nonvolatile memory
Used to store program, configuration, and collected data
Application processor
Runs application logic
Integrated coprocessors
Fast, efficient computation for specific tasks (e.g., linear algebra)
Volatile memory
RAM used to store state during program operation
Peripherals
Digital and analog interfaces with other electronic devices
Discrete coprocessors
Fast, efficient computation for specific tasks, located on a separate chip
Sensors
Measure properties of the physical world and encode as electrical signals
Network hardware
Communication with other devices via standards-based methods

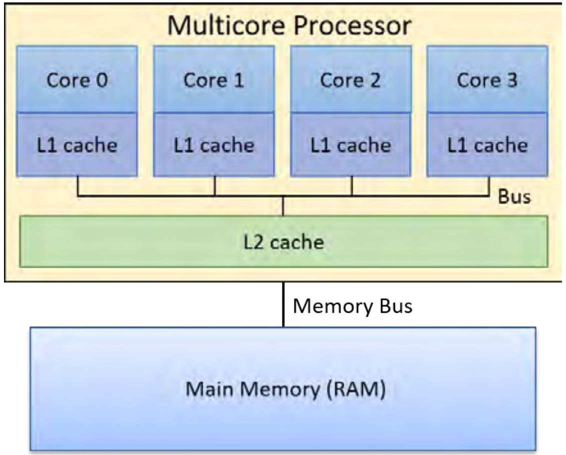
Computer Memory Hierarchy

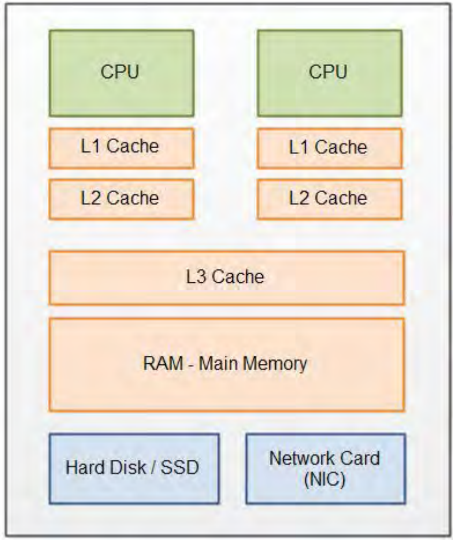


https://en.wikipedia.org/wiki/Memory_hierarchy

Cache

- L1 Cache is the fastest
- L1 & L2, bigger impact
- Cache level is limited
- Cache hit vs miss





https://diveintosystems.org/book/C11-MemHierarchy/coherency.html
https://jenkov.com/tutorials/java-performance/modern-hardware.html

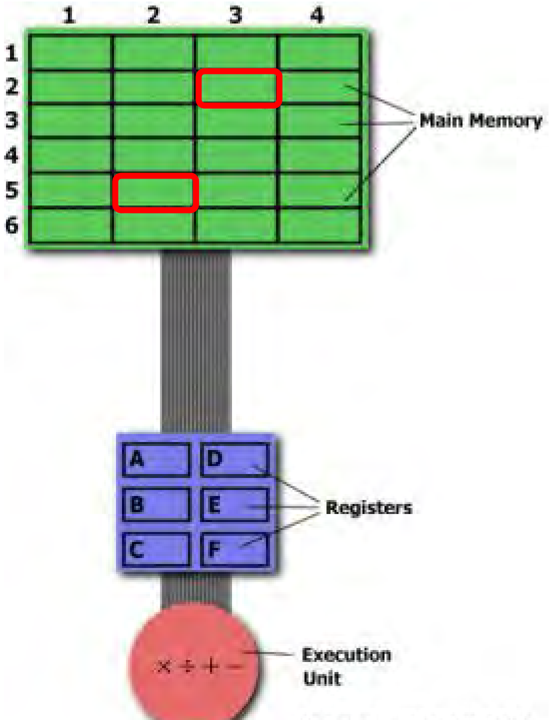
Heat

- Intel i7:
  - Operates at 165-195 W
  - Produces 45 W of heat
- ARM
  - Operates at 4-5 W
  - Produces 3W of heat

RISC vs CISC: Example

Reduced Instruction Set Computer (RISC)

```
LOAD A, 2:3
LOAD B, 5:2
PROD A, B
STORE 2:3, A
```



SOURCE:ARSTECHNICA

Complex Instruction Set Computer (CISC)

```
MULT 2:3, 5:2
```

https://cs.stanford.edu/people/eroberts/courses/soco/projects/risc/risccisc/

RISC vs CISC

Reduced Instruction Set Computer (RISC)

- Emphasis on software
- Single-clock, reduced instruction only
- Register to register:
  - "LOAD" and "STORE" are independent instructions
- Low cycles per second, large code sizes
- Spends more transistors on memory registers

Complex Instruction Set Computer (CISC)

- Emphasis on hardware
- Includes multi-clock complex instructions
- Memory-to-memory:
  - "LOAD" and "STORE" incorporated in instructions
- Small code sizes, high cycles per second
- Transistors used for storing complex instructions

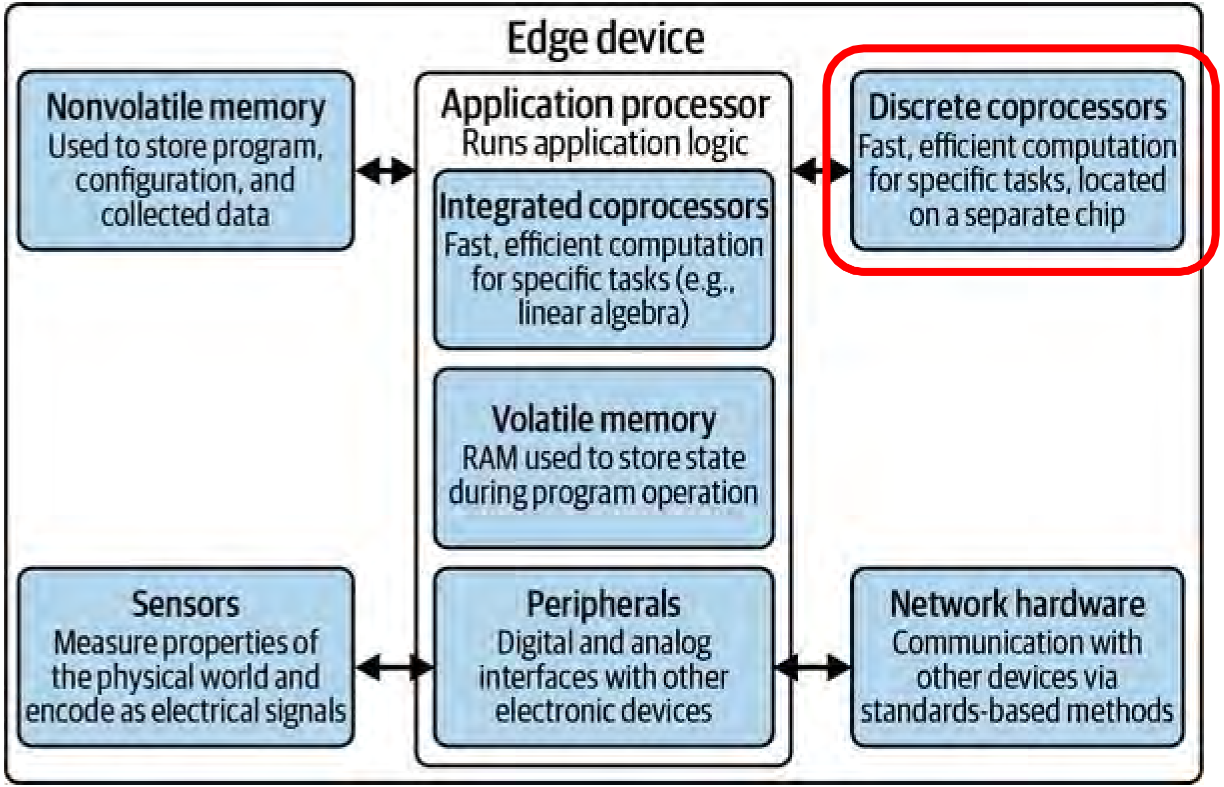
Why RISC?

- Fixed one-clock-cycle instruction -> structured for pipelining
- Less transistors storing only small set of simple instruction
- Lazy erasing of instruction from registers

ARM vs Intel

- Intel processors are performance oriented. They use dynamic size instructions and are about the state of the art for general purpose computing (along with AMD)
- ARM (Advanced RISC machine) are meant to be energy and space efficient.
- Intel is CISC (not really but it’s complicated) and ARM is RISC

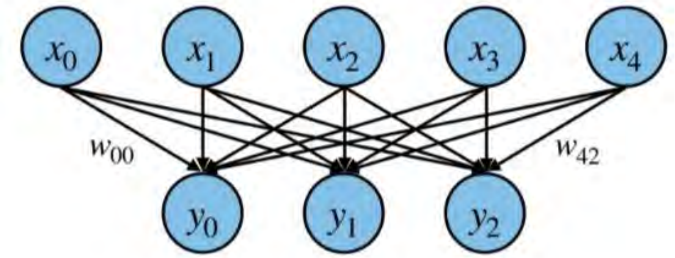
Edge Device Hardware Architecture



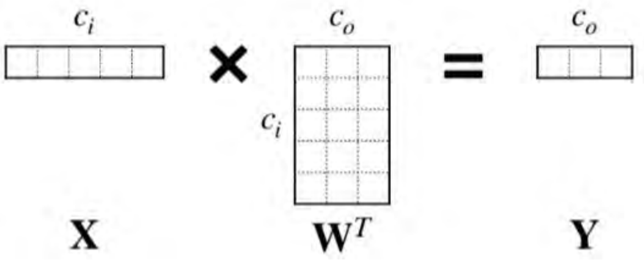
Edge device
Nonvolatile memory
Used to store program, configuration, and collected data
Application processor
Runs application logic
Integrated coprocessors
Fast, efficient computation for specific tasks (e.g., linear algebra)
Volatile memory
RAM used to store state during program operation
Peripherals
Digital and analog interfaces with other electronic devices
Discrete coprocessors
Fast, efficient computation for specific tasks, located on a separate chip
Sensors
Measure properties of the physical world and encode as electrical signals
Network hardware
Communication with other devices via standards-based methods

GPU

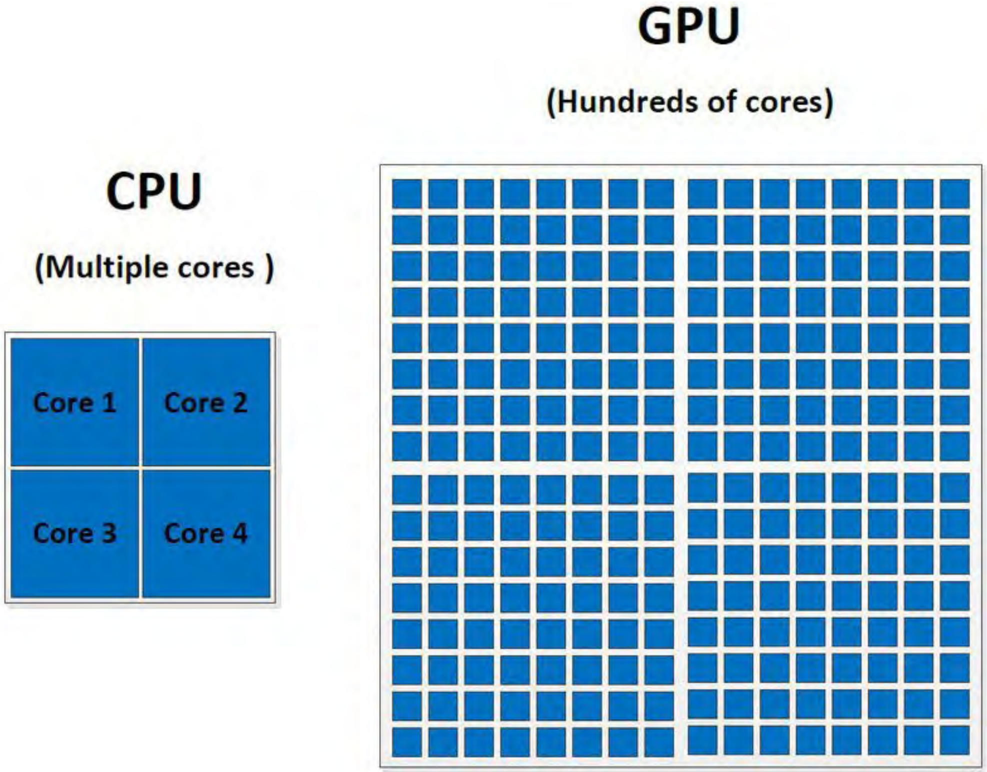
- Graphical processing unit
- Designed for display and rendering
- Highly parallel computation
  - suitable for *general purpose* computation that is parallelizable (GP-GPU)
  - E.g. matrix multiplication



$$y_i = \sum_j w_{ij} x_j + b_i$$

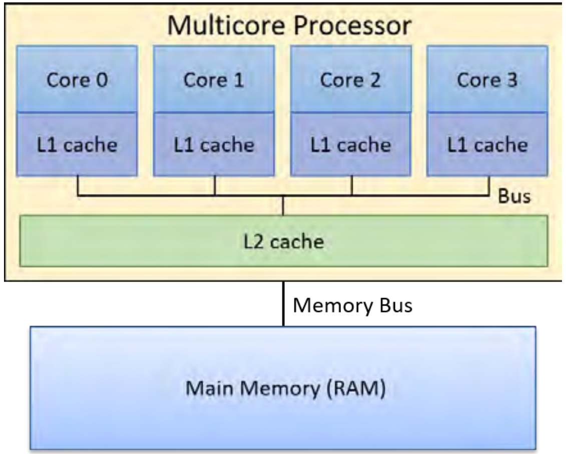


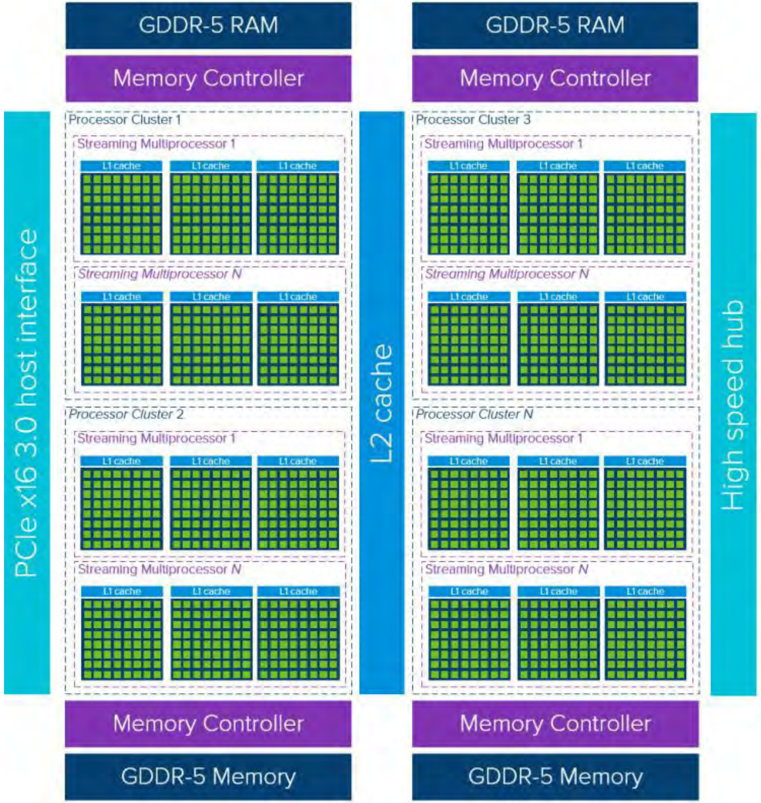
GPU vs CPU



https://www.researchgate.net/figure/CPU-vs-GPU-architecture-each-blue-square-represents-one-core_fig1_323281068

GPU vs CPU: Memory architecture

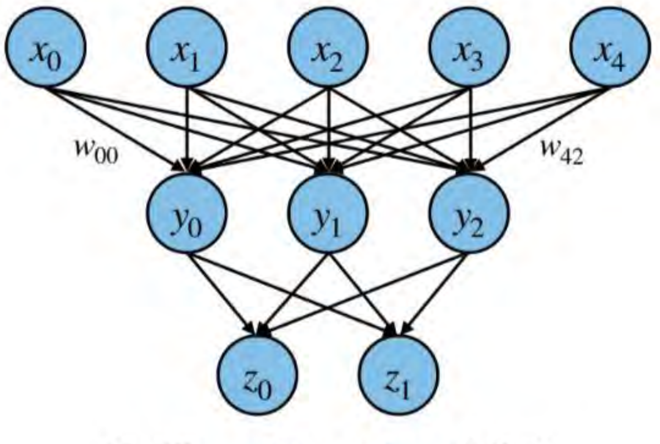




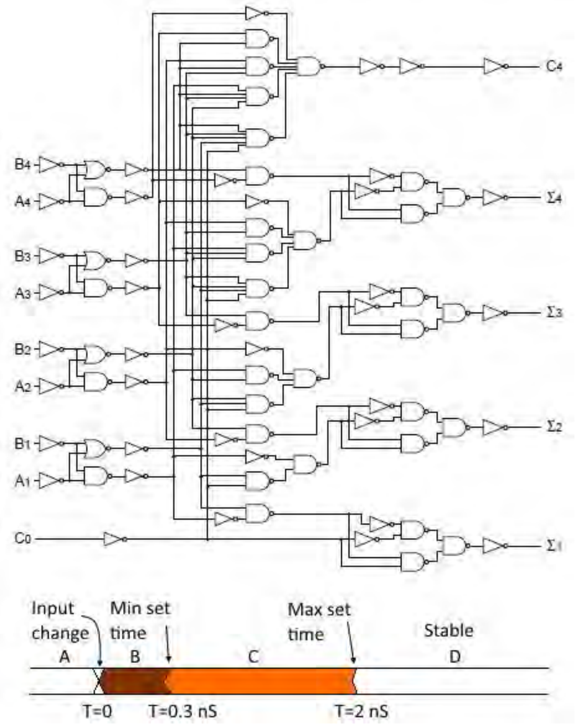
https://nielshagoort.com/2019/03/12/exploring-the-gpu-architecture/

Systolic Array

- Static circuit
- Bake a DNN into silicon



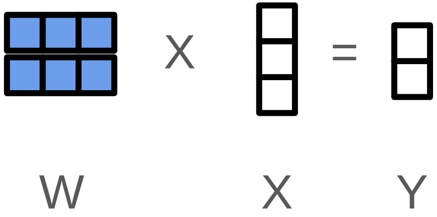
Multilayer Perceptron (MLP)

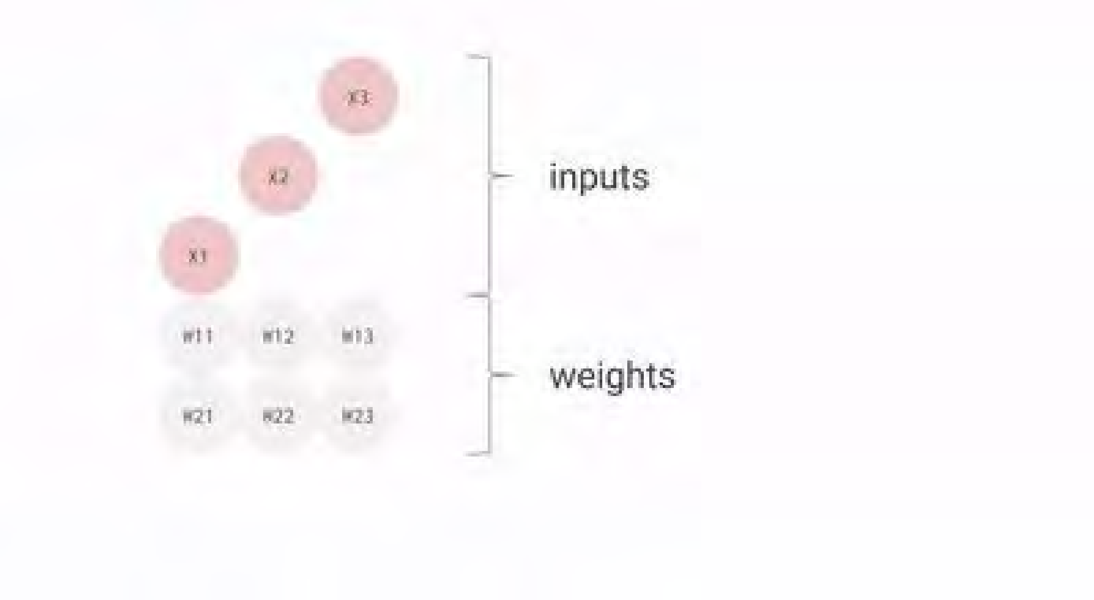


An in-depth look at Google's first Tensor Processing Unit (TPU)

Systolic Array

- Data flow in waves (heart pumps blood)

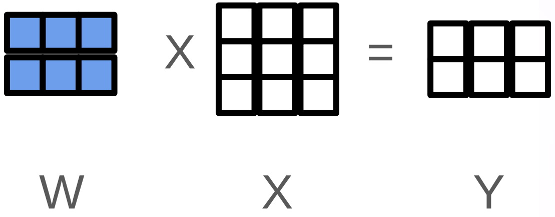


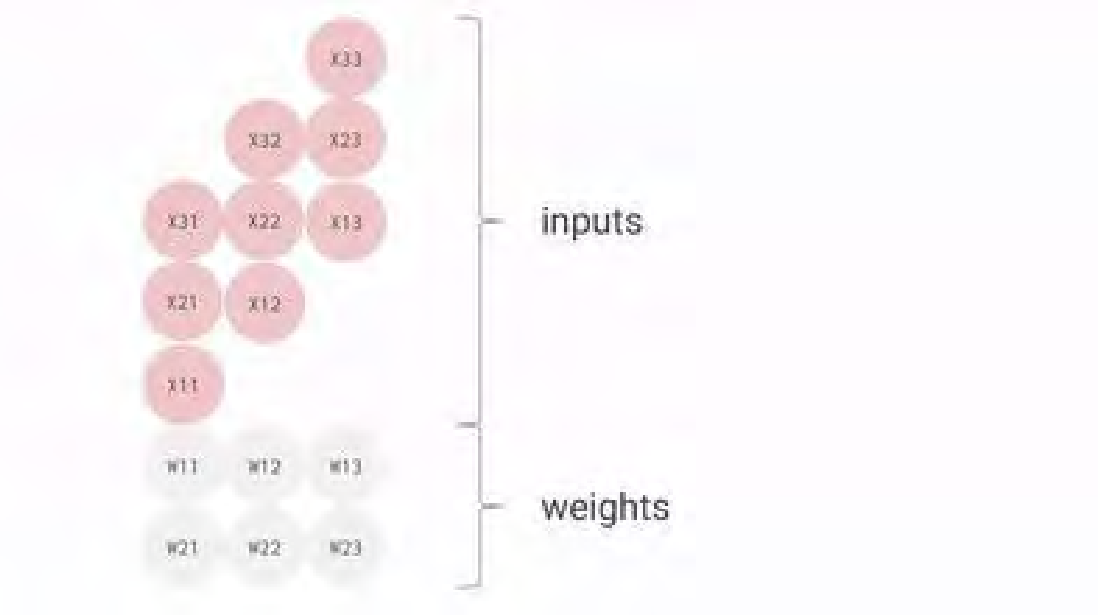


An in-depth look at Google's first Tensor Processing Unit (TPU)

Systolic Array

- Data flow in waves (heart pumps blood)





An in-depth look at Google’s first Tensor Processing Unit (TPU)

Why Systolic Array?

Why Systolic Array?

- Fixed add/mul cells
- High throughput
- Power efficient

FPGA vs ASIC

Field Programmable Gate Arrays (FPGAs)

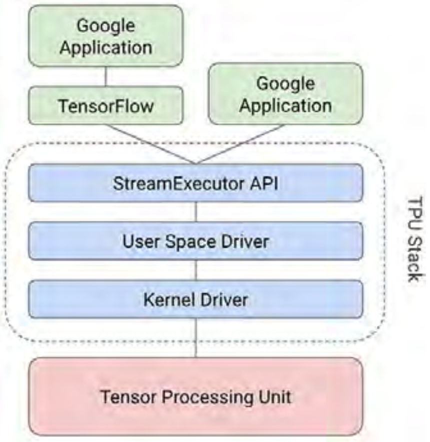
- Programmable hardware fabric
- Flexible for different functions

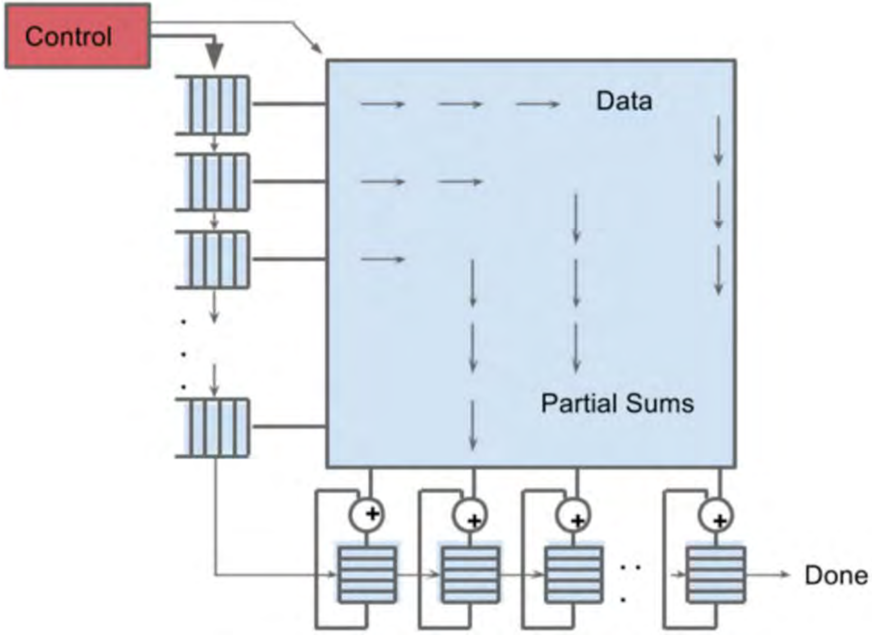
Application-specific integrated circuits (ASICs)

- Static IC for specific applications
- Power efficient, mass produced

TPU

- Tensor processing unit
  - An in-depth look at Google's first Tensor Processing Unit (TPU)
- Specifically designed
  - Structured ASIC
  - Systolic arrays for DNN

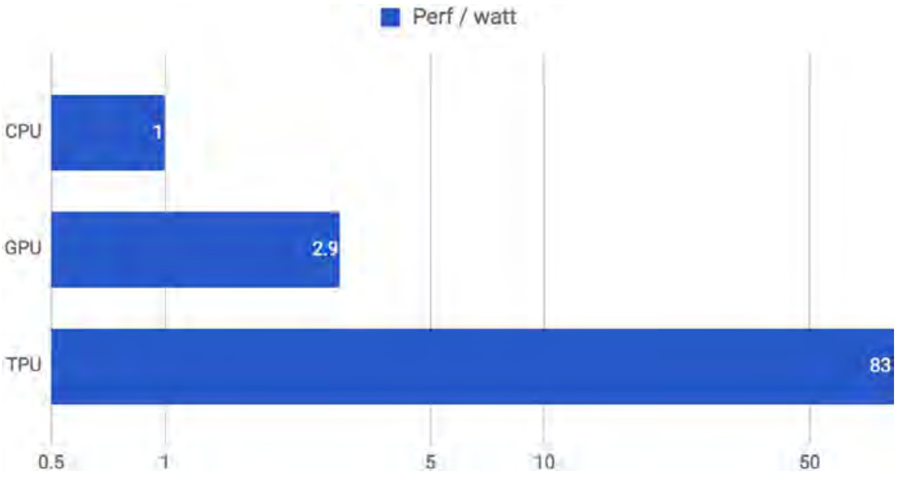


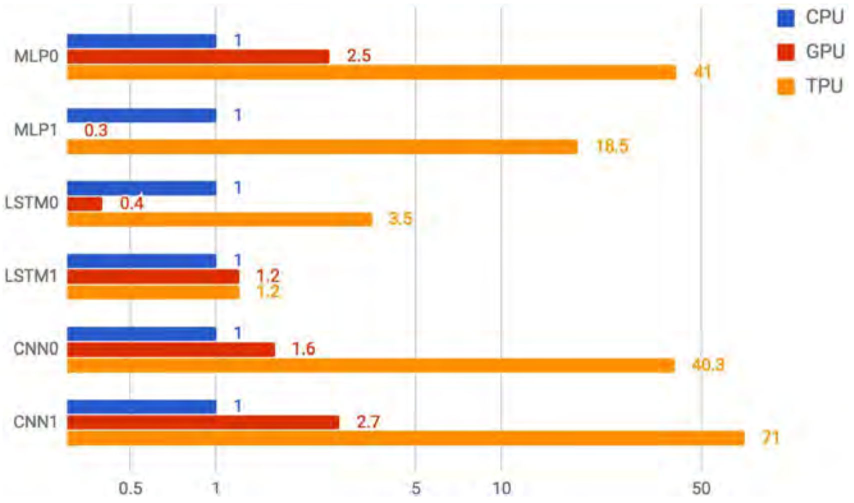


Matrix Multiplier Unit (MXU) of TPU

An in-depth look at Google's first Tensor Processing Unit (TPU)

TPU Performance

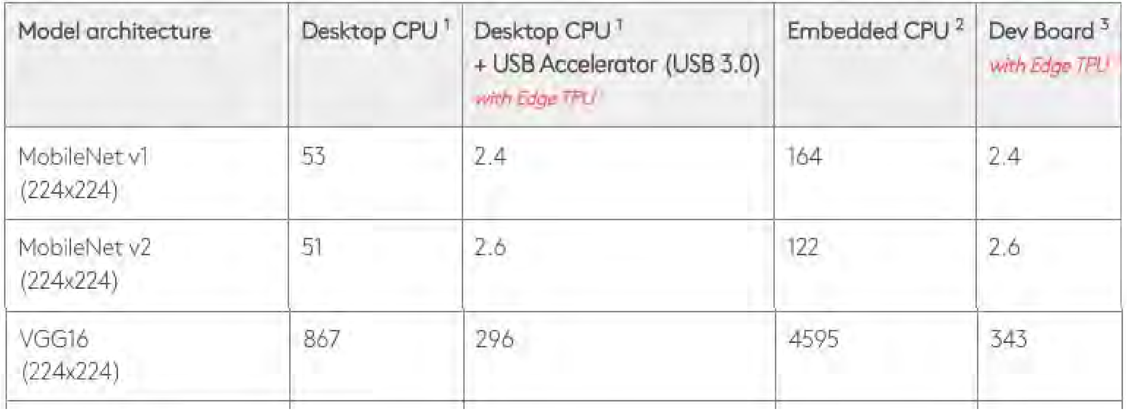




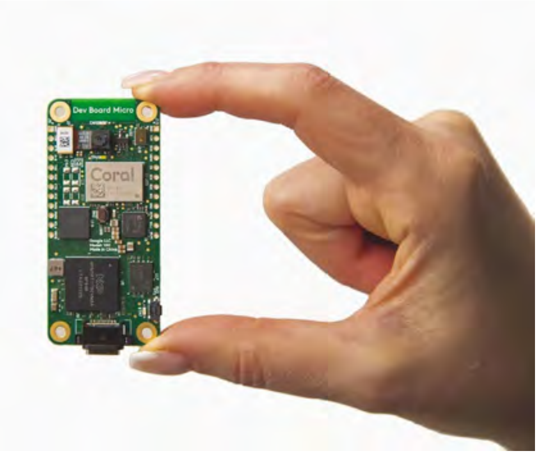
An in-depth look at Google’s first Tensor Processing Unit (TPU)

Coral Edge TPU

- Products | Coral
- Smaller Matrix Multiplier Unit (MXUs)

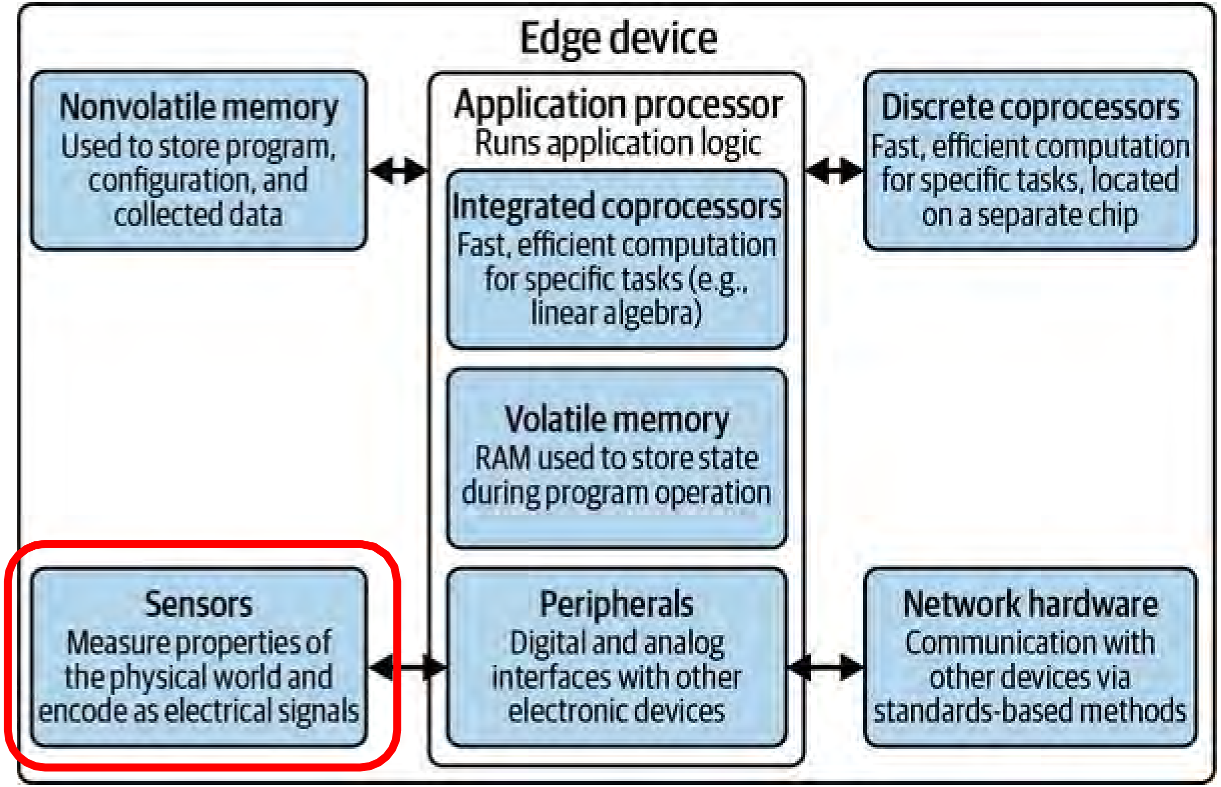


Model architecture	Desktop CPU [1]	Desktop CPU [1] + USB Accelerator (USB 3.0) *with Edge TPU*	Embedded CPU [2]	Dev Board [3] *with Edge TPU*
MobileNet v1 (224x224)	53	2.4	164	2.4
MobileNet v2 (224x224)	51	2.6	122	2.6
VGG16 (224x224)	867	296	4595	343



https://coral.ai/docs/edgetpu/benchmarks/

Edge Device Hardware Architecture



Edge device

Nonvolatile memory
Used to store program, configuration, and collected data

Application processor
Runs application logic

Integrated coprocessors
Fast, efficient computation for specific tasks (e.g., linear algebra)

Volatile memory
RAM used to store state during program operation

Peripherals
Digital and analog interfaces with other electronic devices

Discrete coprocessors
Fast, efficient computation for specific tasks, located on a separate chip

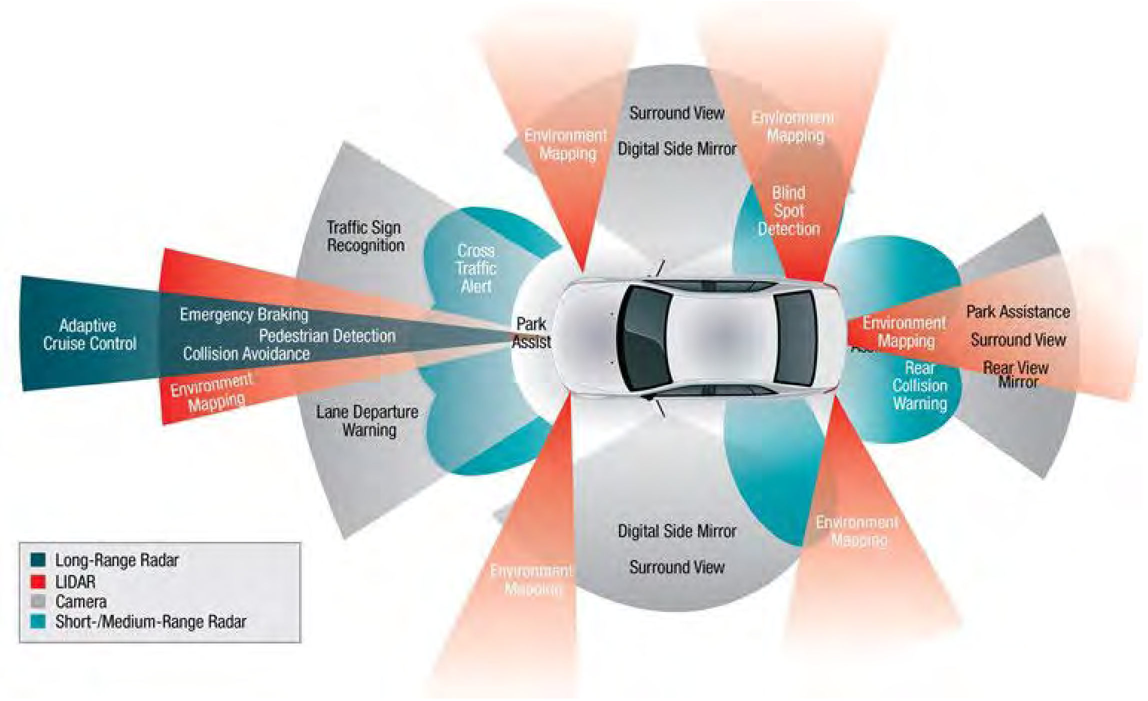
Sensors
Measure properties of the physical world and encode as electrical signals

Network hardware
Communication with other devices via standards-based methods

Sensors

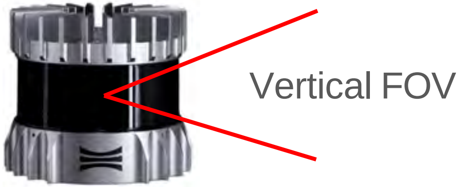
- Acoustic and vibration
- Visual and scene
- Motion and position
- Force and tactile
- Optical, electromagnetic, and radiation
- Environmental, biological, and chemical

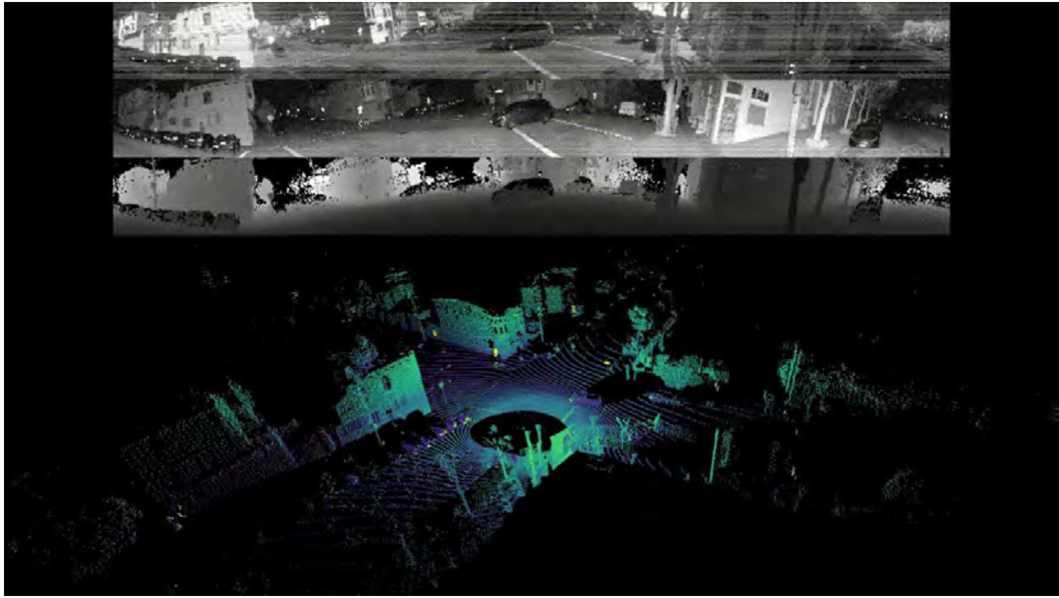
Sensors on Self-driving Cars



LiDAR

- Light Detection and Ranging
  - Range
  - Vertical resolution (beams)
  - Horizontal resolution
  - Range resolution
  - Angular accuracy
  - FPS
  - Field of View (FOV)
    - Vertical
    - Horizontal
- Example
  - [Ouster OS1 Mid-Range Lidar]()
  - [Ouster OS2 Long-Range Lidar]()





https://ouster.com/insights/blog/the-camera-is-in-the-lidar

Stereo Camera

- Intrinsics
  - Aperture
  - Focal length
  - FOV
- Sensor size
- Shutter (rolling, global)
- FPS
- Resolution
- Baseline (Stereo)

Example

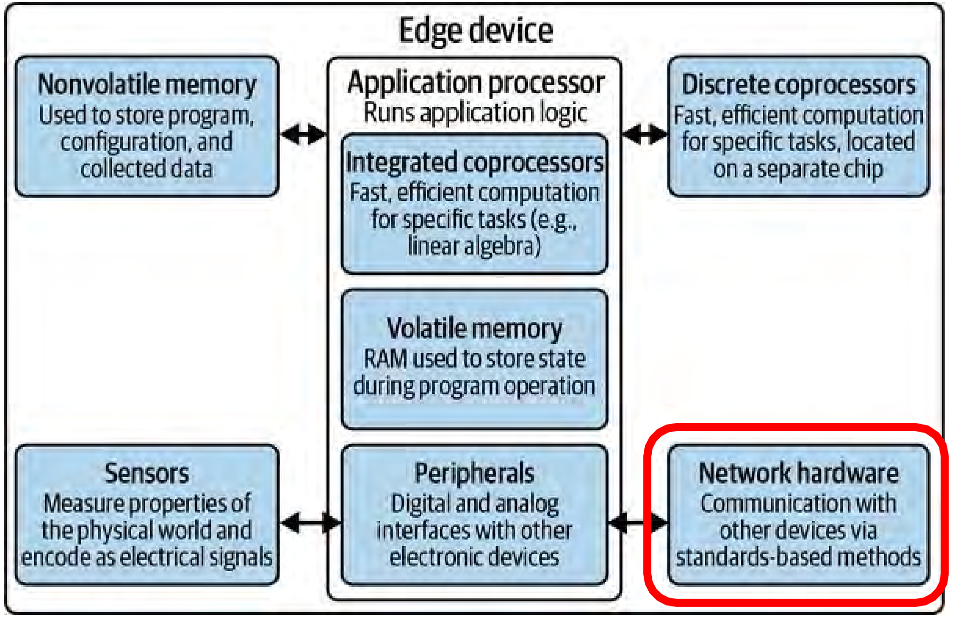
- ZED 2 - AI Stereo Camera
- Intel® RealSense™ Depth and Tracking Cameras

Summary

- CPU
  - Clock speed, memory speed, cores, threads, cache size
- Memory
- Cache
  - L1, L2, L3, cache hit & miss
- RISC vs CISC
- Special accelerators
  - GPU, TPU, systolic array, FPGA, ASIC
- Sensors
  - LiDAR, Stereo Camera

Next Lecture

- Middleware
- Lab 4: Networking with Cloud



Edge device

Nonvolatile memory
Used to store program, configuration, and collected data

Application processor
Runs application logic

Discrete coprocessors
Fast, efficient computation for specific tasks, located on a separate chip

Integrated coprocessors
Fast, efficient computation for specific tasks (e.g., linear algebra)

Volatile memory
RAM used to store state during program operation

Sensors
Measure properties of the physical world and encode as electrical signals

Peripherals
Digital and analog interfaces with other electronic devices

Network hardware
Communication with other devices via standards-based methods