1. Dataset Creation
2. Training
3. Deploy/Infer

Notes based on
CS231n, Stanford University, and
EECS 498-007 / 598-005, University of Michigan

Deep Learning Hardware

Notes based on
CS231n, Stanford University, and
EECS 498-007 / 598-005, University of Michigan

Deep Learning Hardware

Hardware for training

- GPU
- TPU

Hardware for inferencing

- CPU
- Edge TPU
- VPU

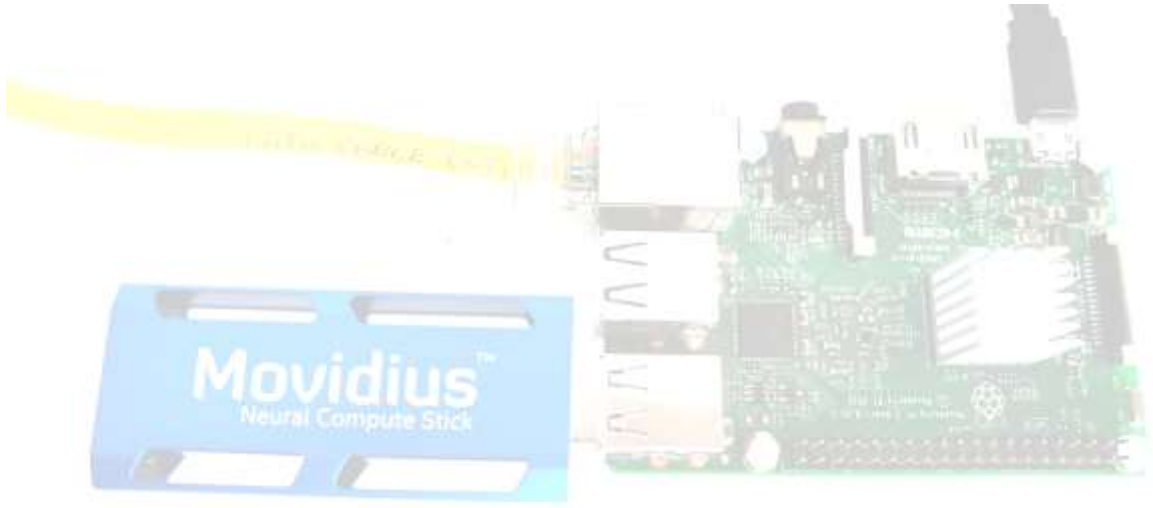
Deep Learning Hardware

Hardware for training

- GPU
- TPU

Hardware for inferencing

- CPU
- Edge TPU
- VPU



Inside a computer

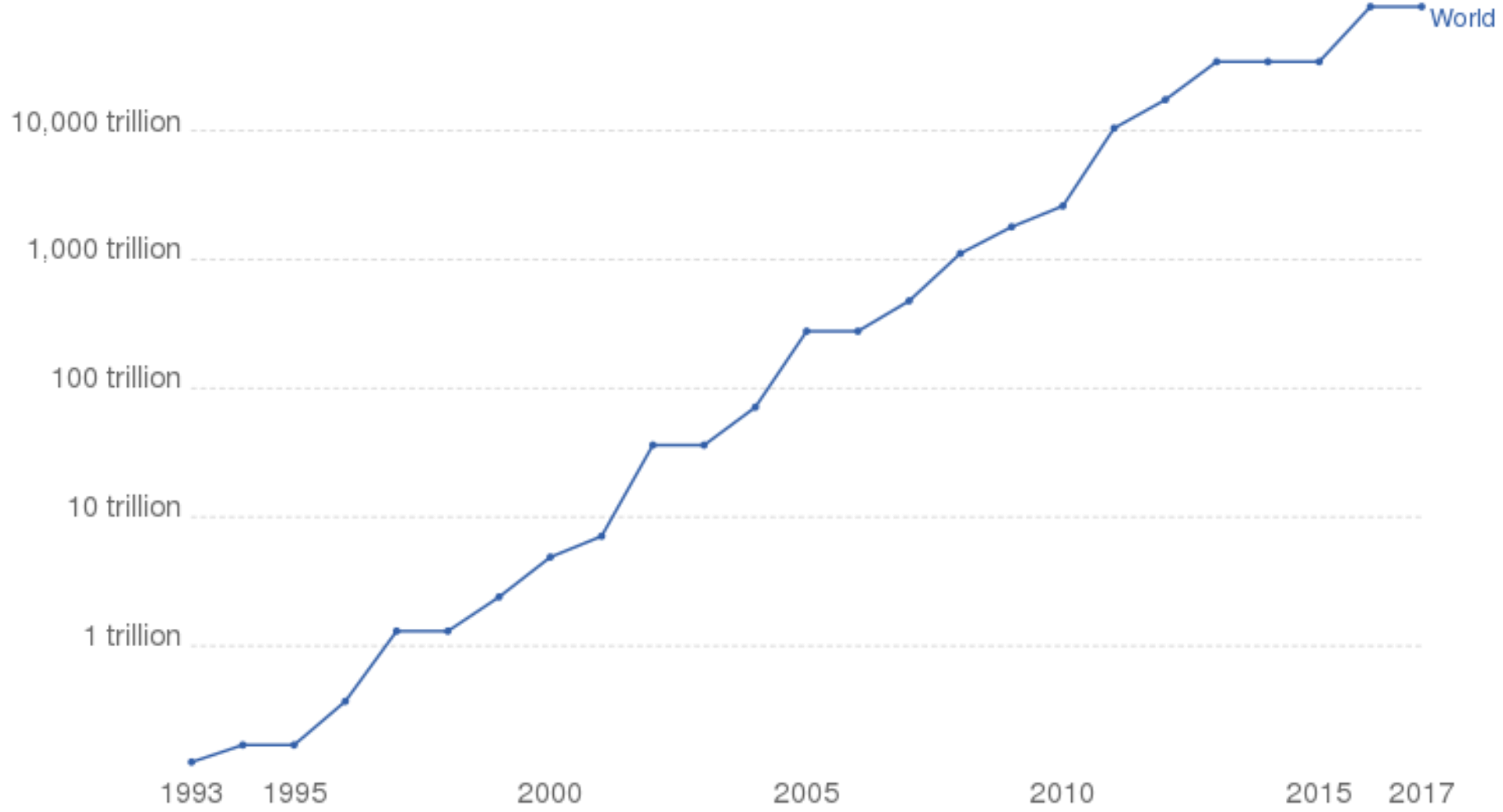
GPU: “Graphics Processing Unit”

CPU: “Central Processing Unit”

VS

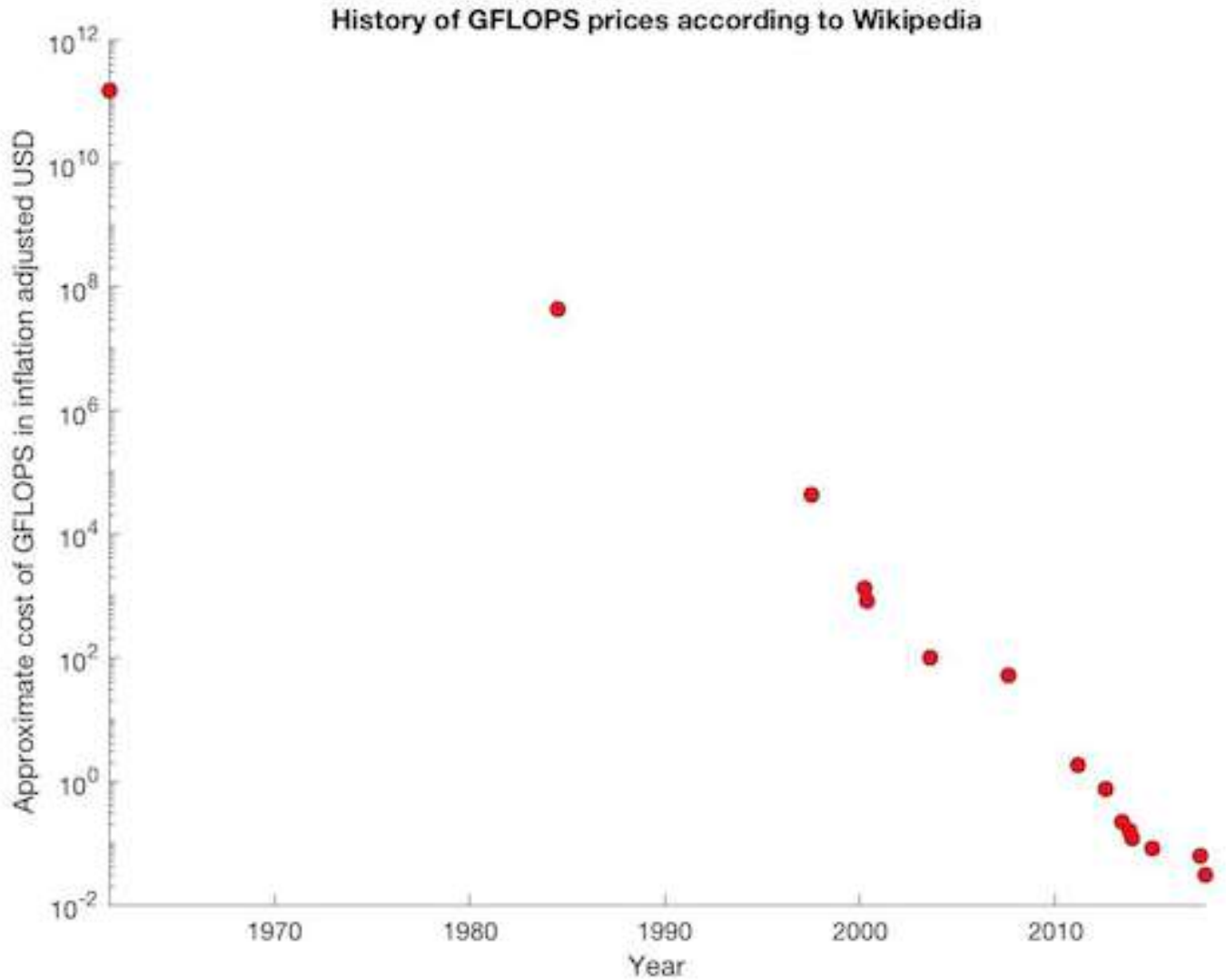
Supercomputer Power (FLOPS)

The growth of supercomputer power, measured as the number of floating-point operations carried out per second (FLOPS) by the largest supercomputer in any given year. (FLOPS) is a measure of calculations per second for floating-point operations. Floating-point operations are needed for very large or very small real numbers, or computations that require a large dynamic range. It is therefore a more accurate measured than simply instructions per second.



Source: TOP500 Supercomputer Database

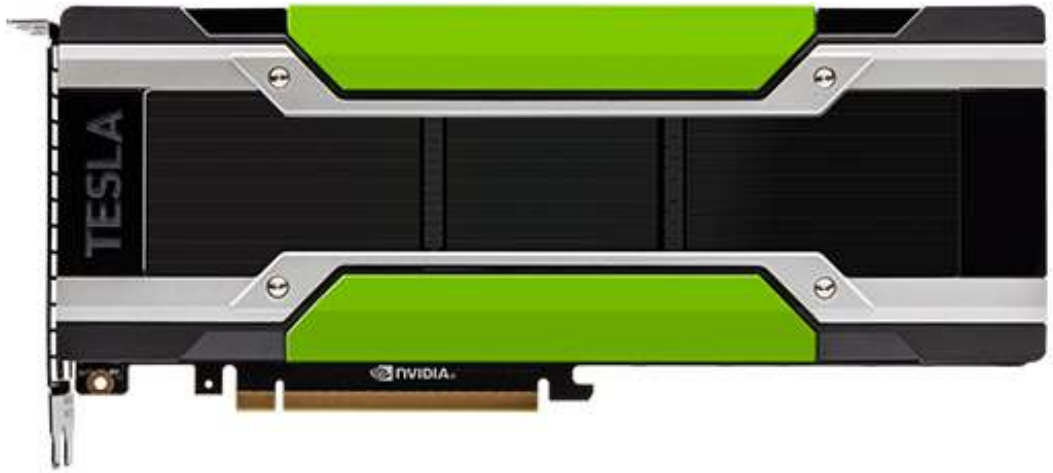
Higher GigaFLOPs at a lower cost

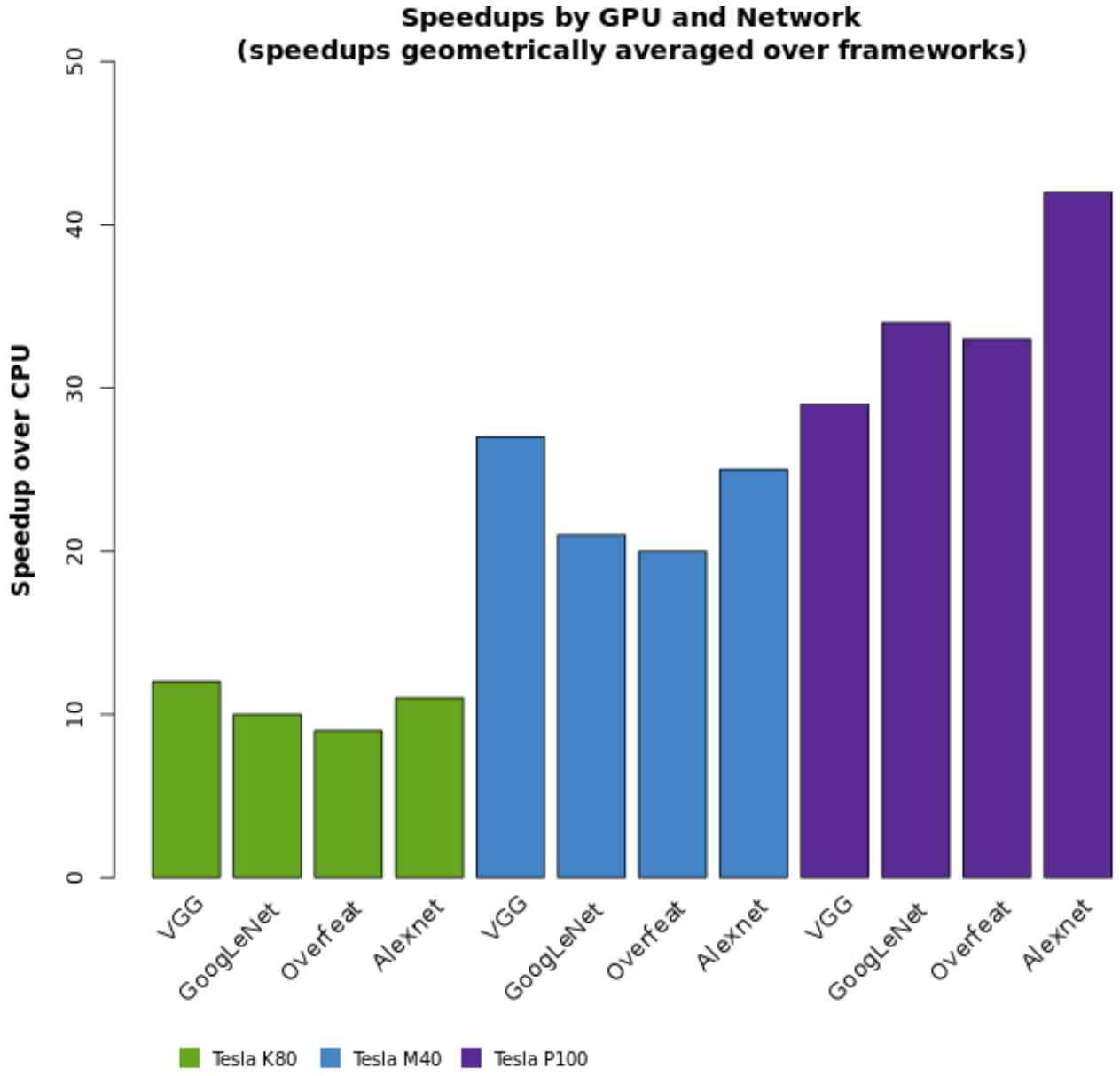


http://en.wikipedia.org/wiki/FLOPS#Hardware_costs

NVIDIA GPU Hardware

- Nvidia Tesla K80, T4, P4 and P100





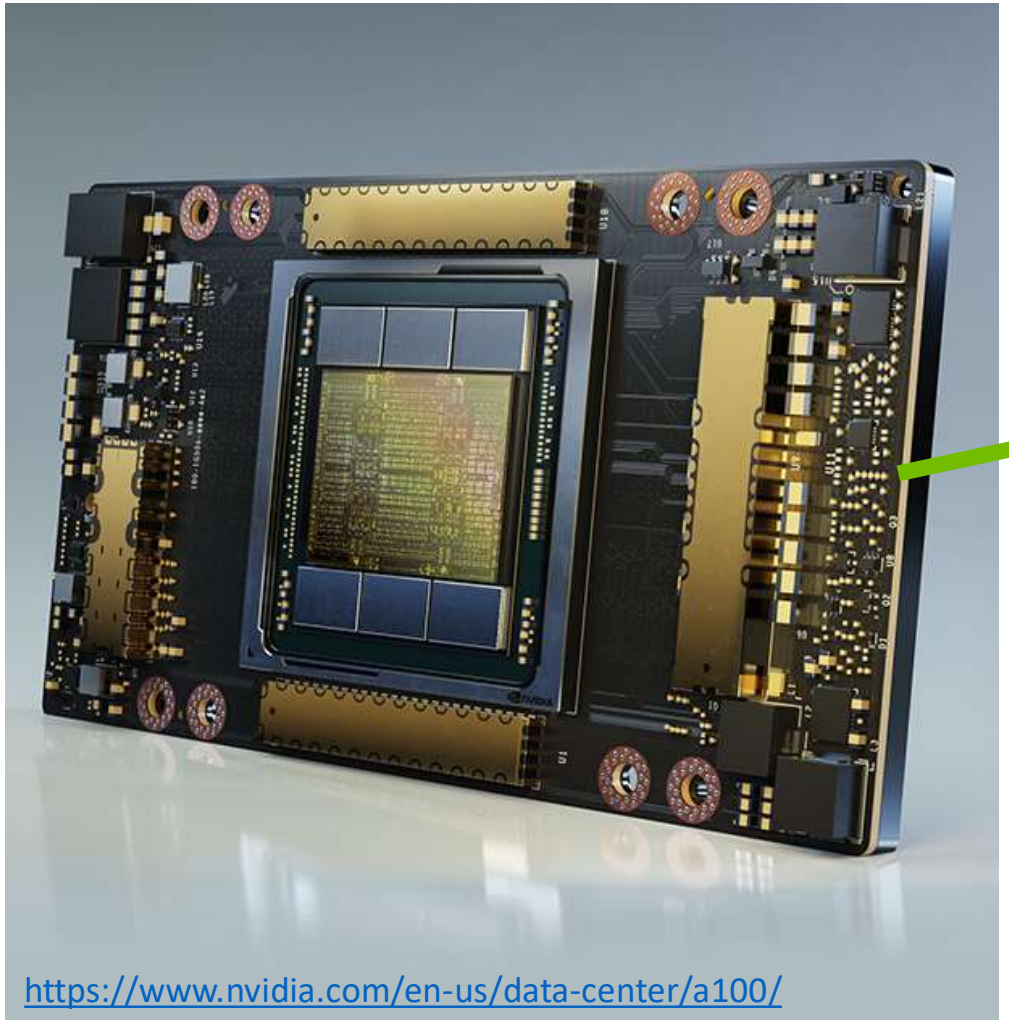
Programming GPUs

- CUDA (NVIDIA only)
  - Write C-like code that runs directly on the GPU
  - NVIDIA provides optimized APIs: cuBLAS, cuFFT, cuDNN, etc

- OpenCL
  - Similar to CUDA, but runs on anything
  - Usually slower on NVIDIA hardware

Scaling up: Typically 8 GPUs per server

NVIDIA A100



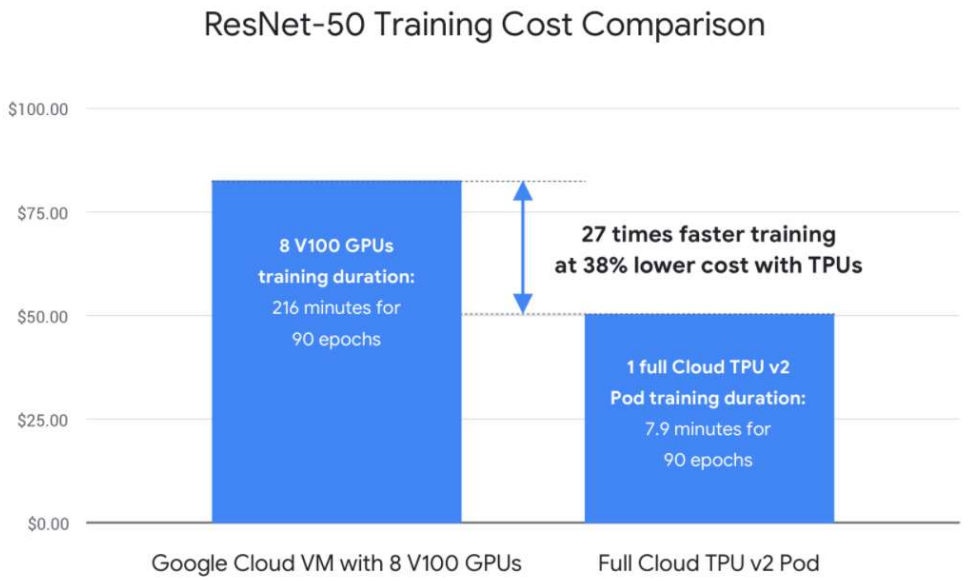
https://www.nvidia.com/en-us/data-center/a100/

NVIDIA DGX A100 - **8x A100 GPUs**

https://www.nvidia.com/en-us/data-center/dgx-a100/

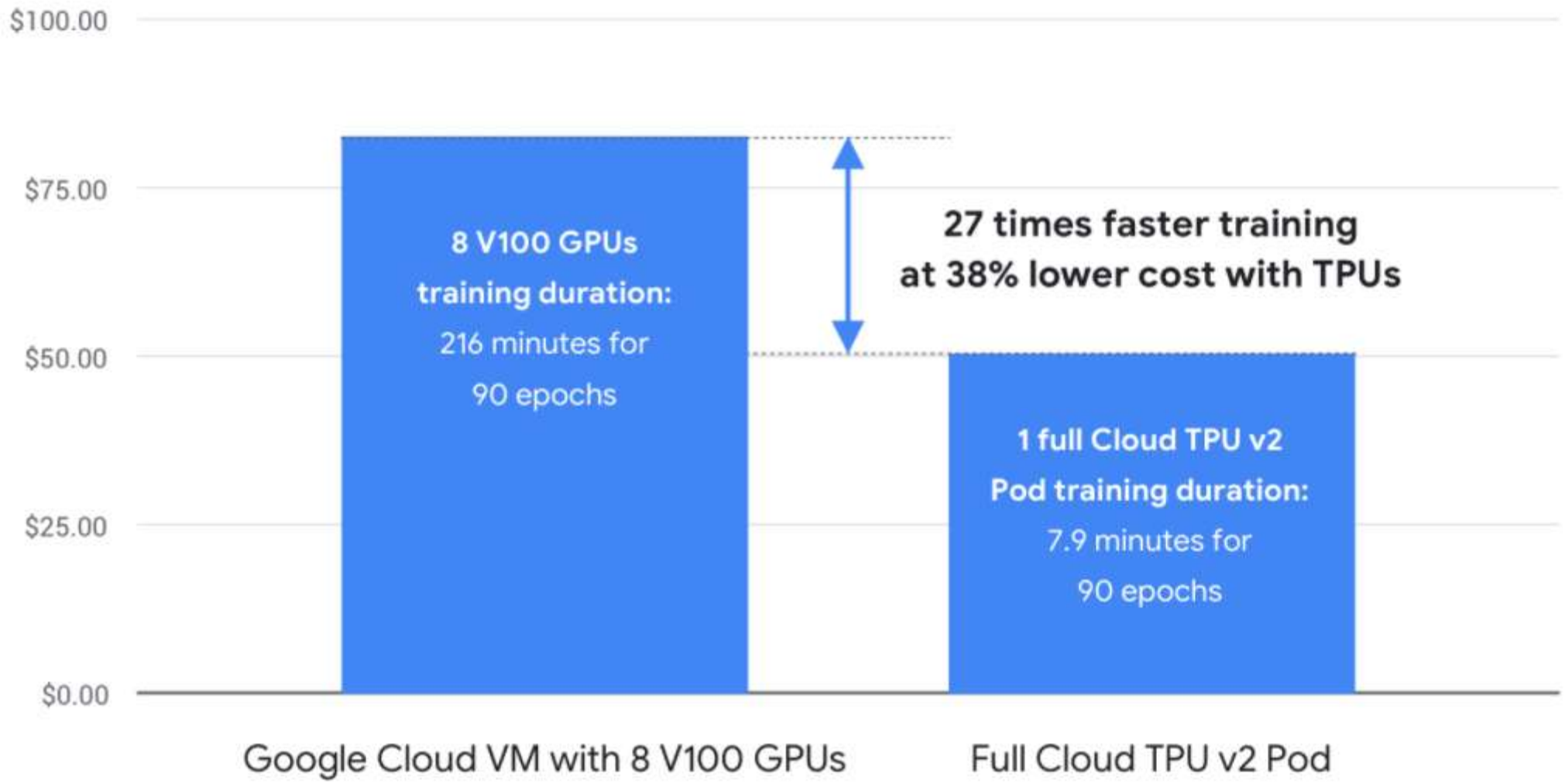
Google Tensor Processing Units (TPU) Servers

Tensor Processing Unit (TPU) is an AI accelerator application-specific integrated circuit (ASIC) developed by Google specifically for neural network machine learning



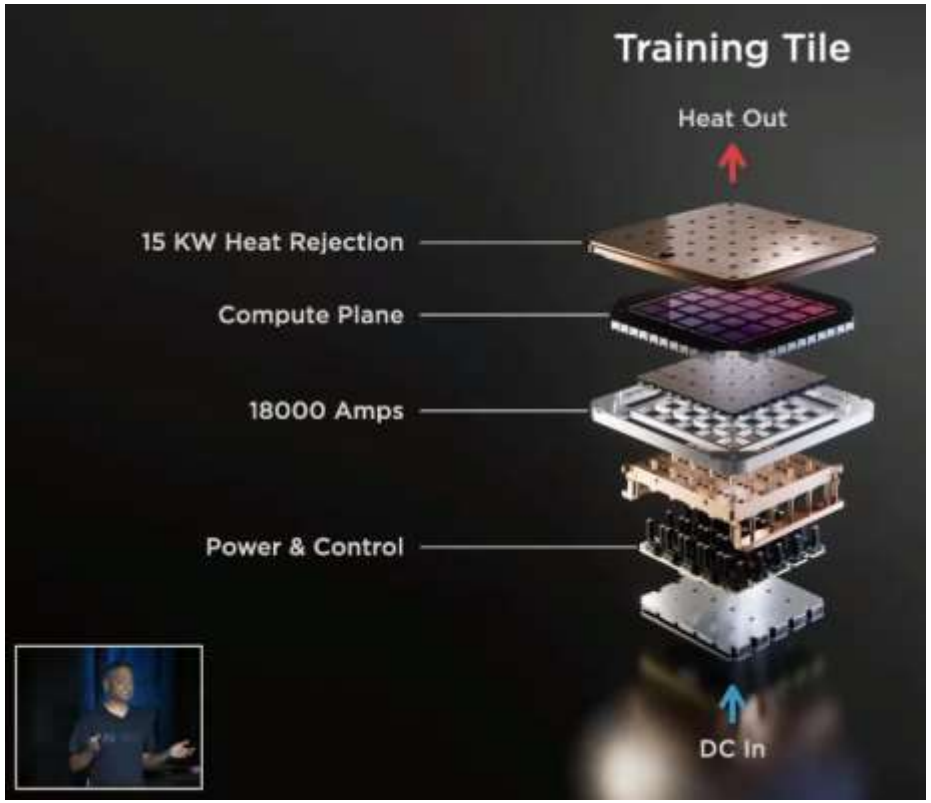
https://cloud.google.com/tpu

ResNet-50 Training Cost Comparison



Tesla Dojo

D1 chips as base
9 petaFLOPS
PFLOPS 10^{15}
9,000,000,000,000,000 FLOPS



Deep Learning Hardware

Hardware for training

- GPU
- TPU

Hardware for inferencing

- Edge TPU
- VPU
- CPU

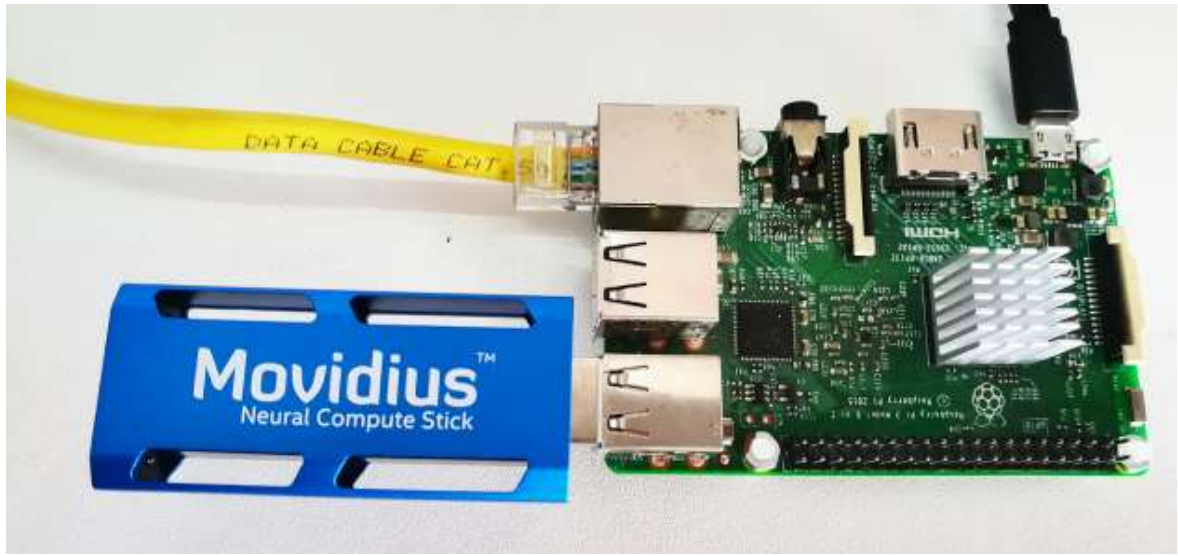
Deep Learning Hardware

Hardware for training

- GPU
- TPU

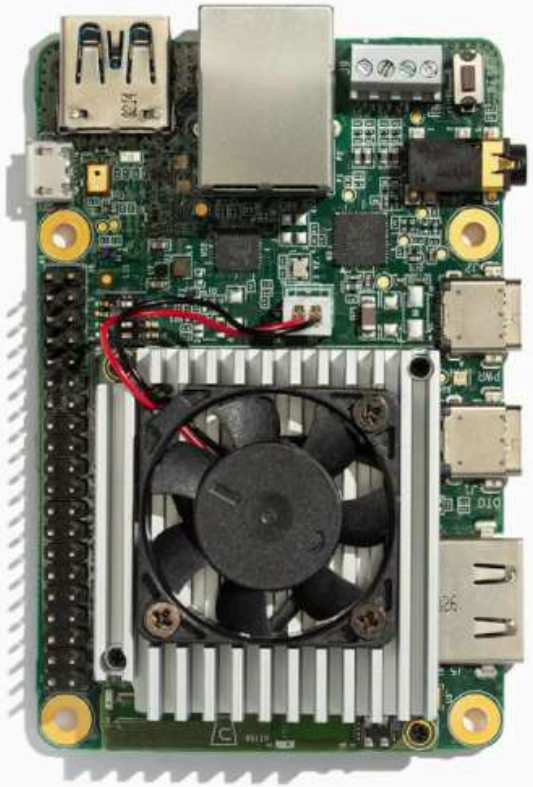
Hardware for inferencing

- Edge TPU
- VPU
- CPU



Coral Edge TPU

USB Accelerator



Dev Board

System-on-Module (SoM)

Mini PCIe Accelerator

M.2 Accelerator B+M key

M.2 Accelerator A+E key

Accelerator Module

M.2 Accelerator with Dual Edge TPU

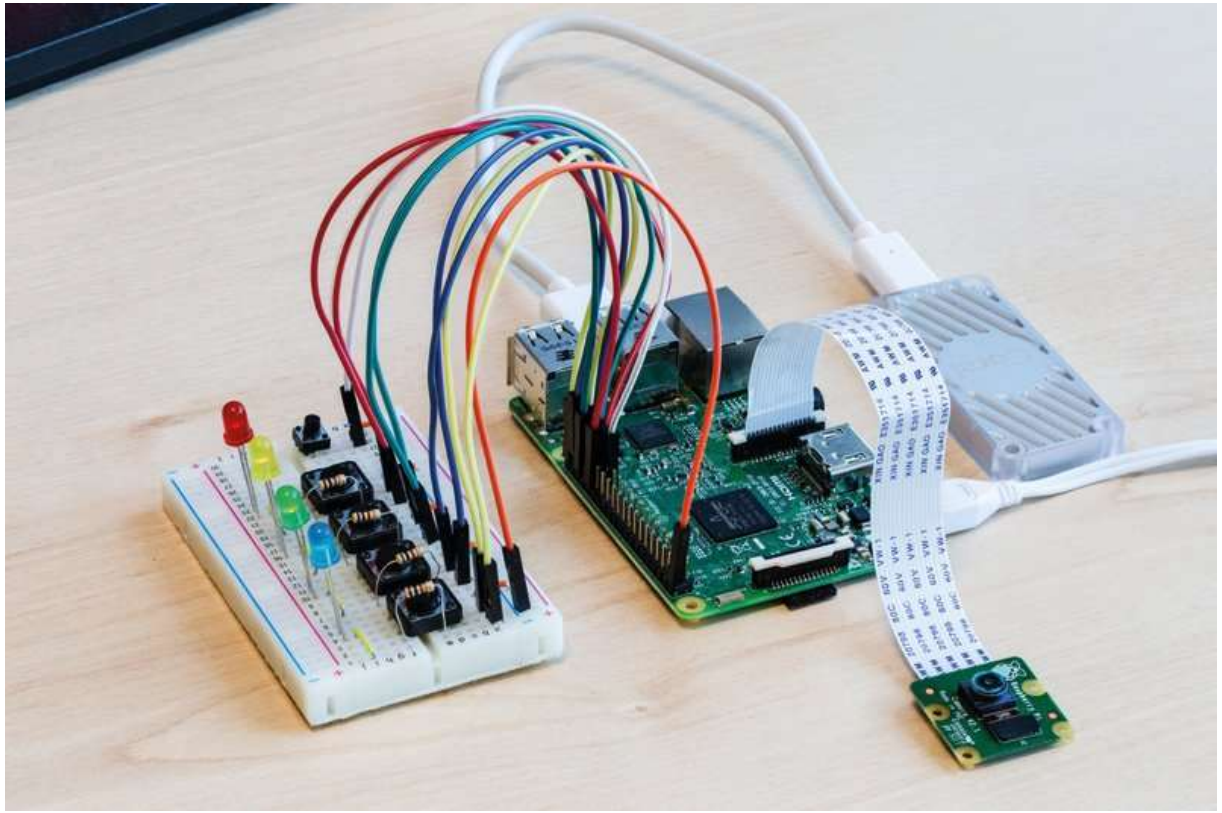
Dev Board Mini

https://coral.ai/products/

https://www.youtube.com/watch?v=bOYWx1jJCZo

CPU + Coral Edge TPU

Raspberry Pi with Coral Edge TPU



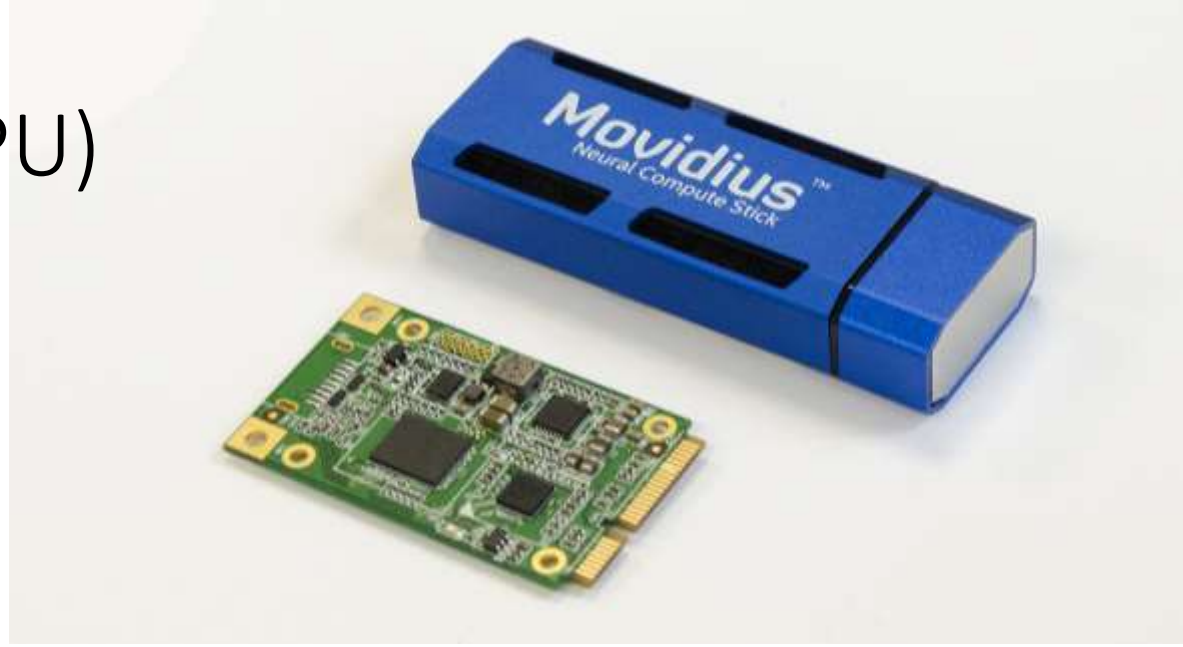
Only works with Tensorflow

https://magpi.raspberrypi.org/articles/teachable-machine-coral-usb-accelerator

https://www.youtube.com/watch?v=bOYWx1jJCZo

Intel Vision Processing Unit (VPU)

a.k.a. Myriad X, Movidius, Neural Compute Stick



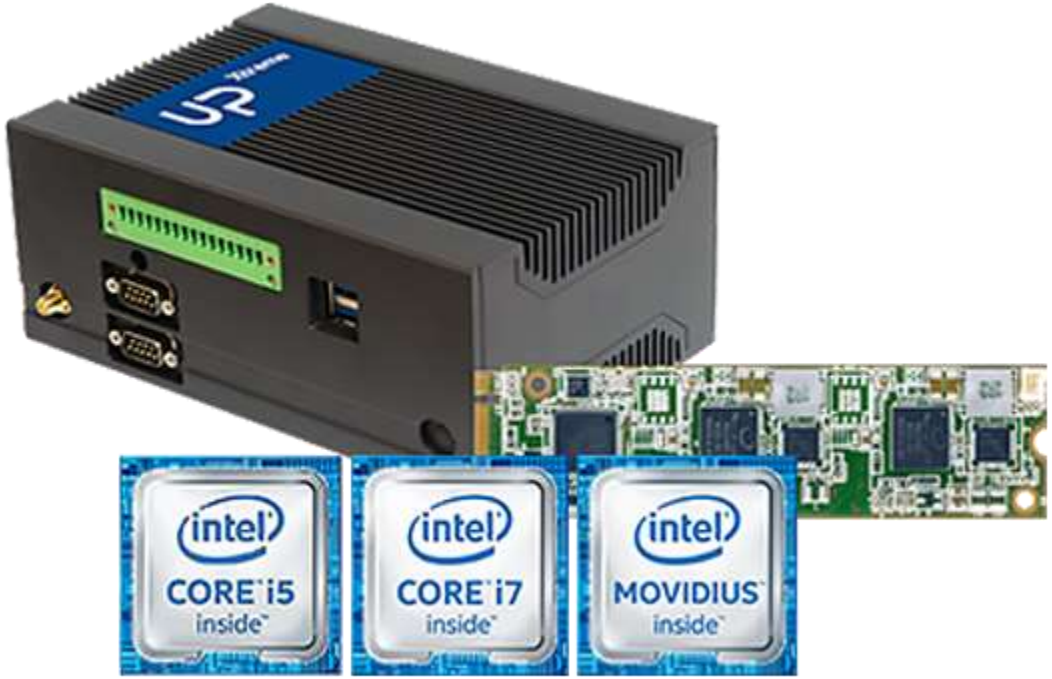
Specialized processors designed
to deliver high-performance
machine vision at ultra-low power.

- Supports up to 16 video streams per device
- Ideal for camera and network video recorder (NVR) use cases with power, size, and cost constraints
- Supports small memory footprint networks

- Only works with Intel OpenVINO

https://software.intel.com/content/www/us/en/develop/topics/iot/hardware/vision-accelerator-movidius-vpu.html

CPU + VPU



UpSquared with M.2 VPU

Raspberry Pi with Intel Movidius Stick

VPU on Edge Devices

[NEON-1000-MDX AI Smart Camera with Intel® Movidius™ Myriad™ X VPU]

[AI-Vue Series – WP1NNL0 2MP ANPR Bullet Camera Powered by Intel® Movidius™ Myriad™ X VPU]

OpenCV OAK-D

https://store.opencv.ai/products/oak-d

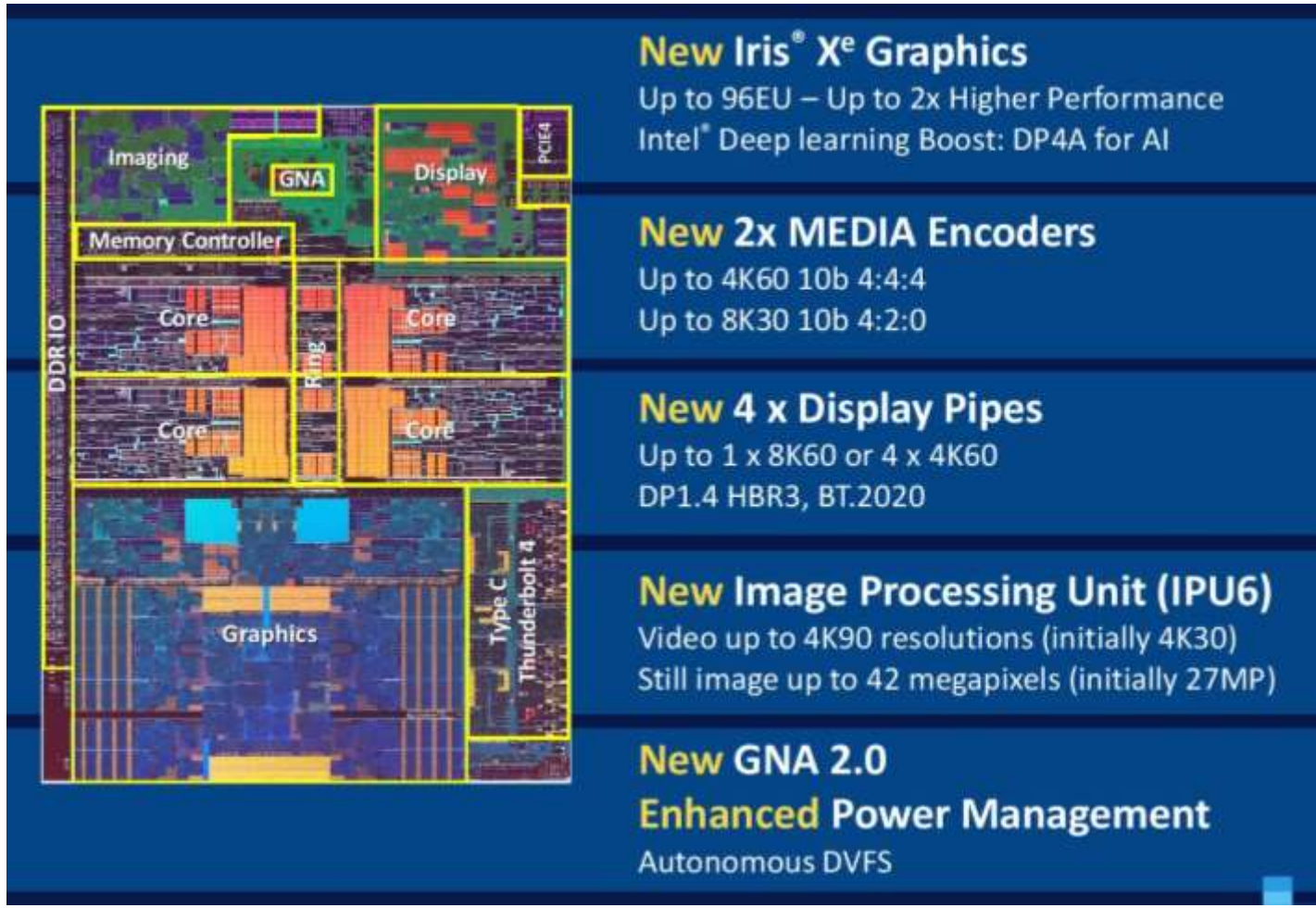
The OpenCV AI Kit (OAK) is a low-power hardware edge AI computing module based on Intel Movidius Myriad-X chip

CPU

Normal CPUs can still infer

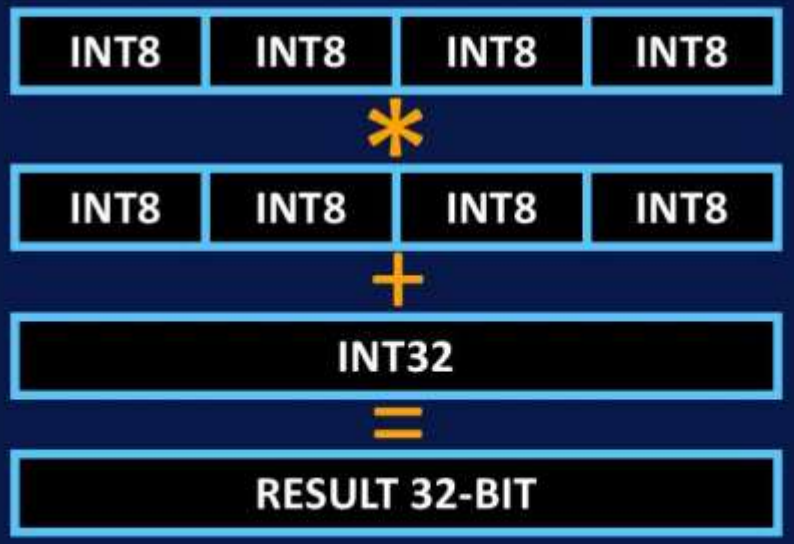
Intel® 11th Gen Core Processors

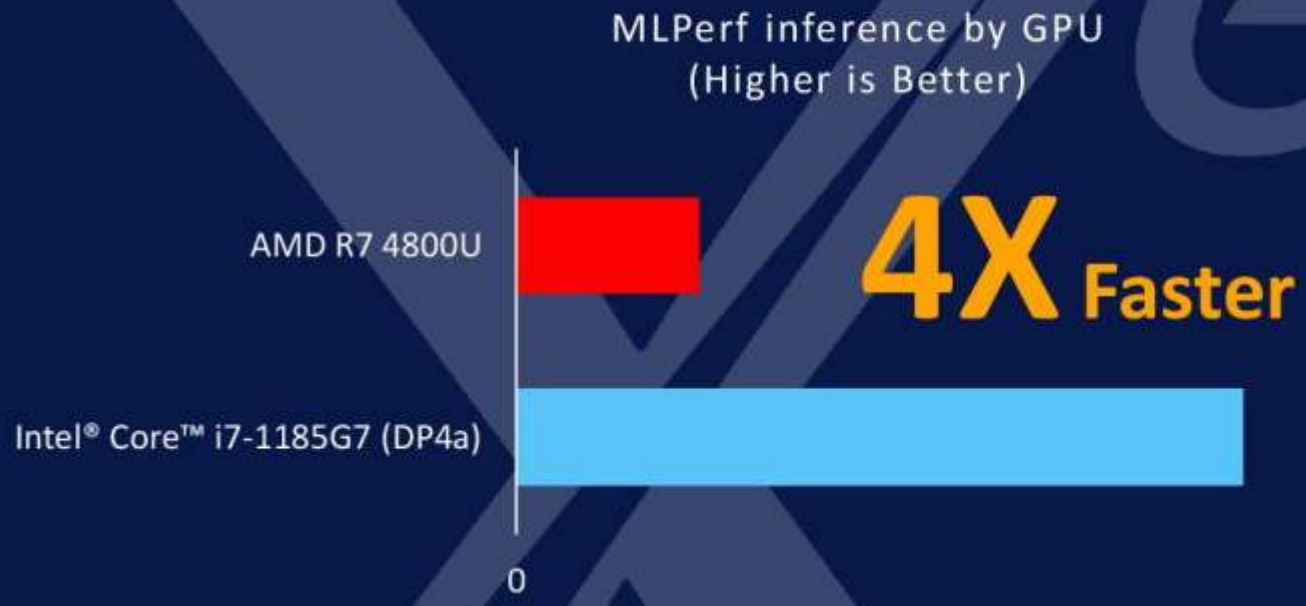
- Deep Learning Boost (Intel® DL Boost)



https://www.intel.com/content/www/us/en/products/docs/processors/core/intelligent-pc-overview.html

Intel® DL Boost: DP4a





- Introducing 8-bit INT acceleration on Intel® Iris® Xᵉ graphics
- Optimized with Intel® OpenVINO™ Toolkit & OneAPI
- 4X faster than competition (ML Perf)

For more complete information about performance and benchmark results, visit www.intel.com/11thgen (configuration details in section 2).

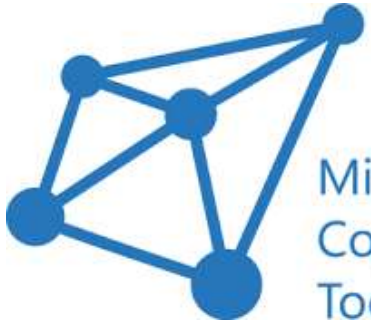
intel

Deep Learning Software

Notes based on
CS231n, Stanford University, and
EECS 498-007 / 598-005, University of Michigan

Linux and Python

A zoo of frameworks!



intel OpenVINO

TensorFlow

Amazon SageMaker

Keras

OpenCV

MATLAB

scikit learn

PyTorch

Caffe

theano

Tensorflow

TensorFlow is a free and open-source software library for machine learning.

You've been using it via 

Now TF 2.0

Can train on your own (Linux) computer or on Colab.

Can deploy anywhere.

https://www.tensorflow.org/learn

PyTorch

PyTorch is an open source machine learning library for Python, based on Torch. It is used for applications such as natural language processing and was developed by Facebook's AI research group.

Can train on your own (Linux) computer or on Colab.

Can deploy anywhere.

https://pytorch.org/

	PyTorch	TensorFlow	Keras
API Level	Low	High and Low	High
Architecture	Complex, less readable	Not easy to use	Simple, concise, readable
Datasets	Large datasets, high performance	Large datasets, high performance	Smaller datasets
Debugging	Good debugging capabilities	Difficult to conduct debugging	Simple network, so debugging is not often needed
Does It Have Trained Models?	Yes	Yes	Yes
Popularity	Third most popular	Second most popular	Most popular
Speed	Fast, high-performance	Fast, high-performance	Slow, low performance
Written In	Lua	C++, CUDA, Python	Python

https://www.simplilearn.com/keras-vs-tensorflow-vs-pytorch-article

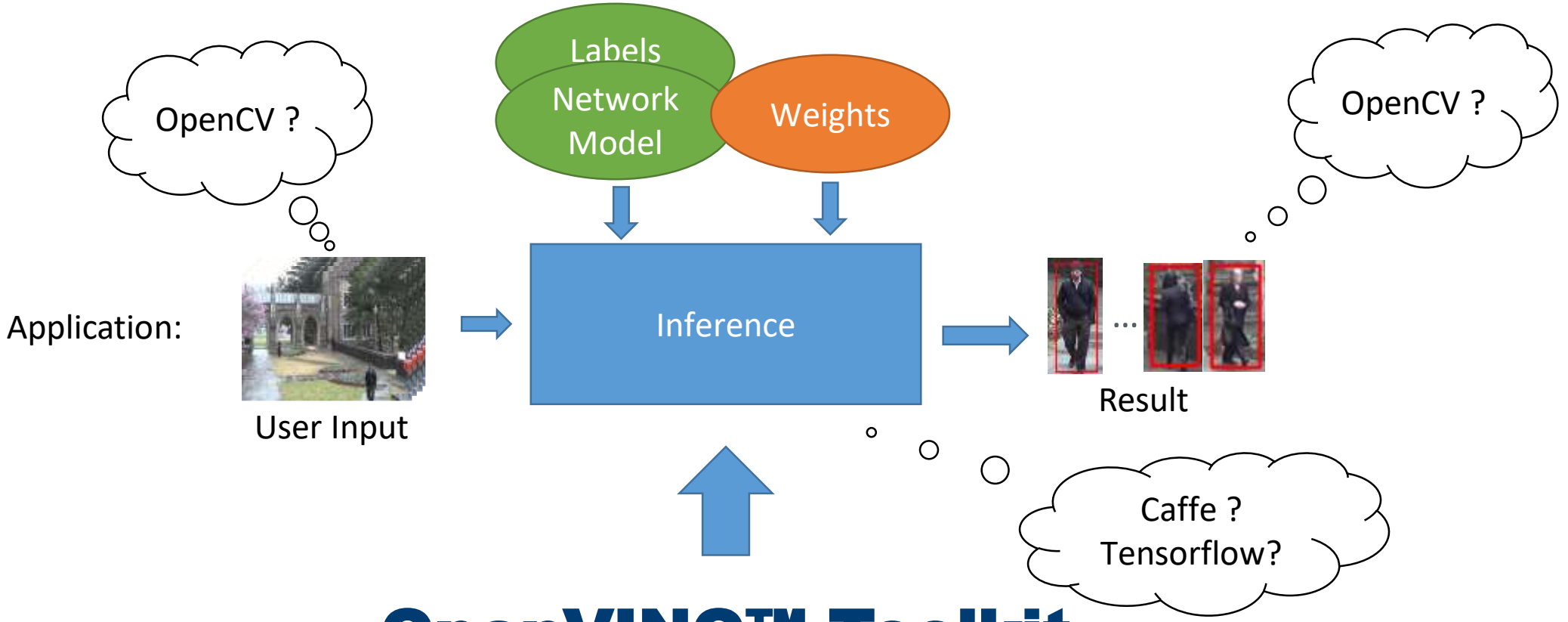
Colab

- Colab allows anybody to write and execute Python code through the browser.
- Get access to Nvidia Tesla K80, T4, P4 and P100 for free! Limited to 12GB memory.
- Connect for 12 hours a day.
- Did I mention it’s free?

- Want more power? There’s Colab Pro.

OpenVino

intel OpenVINO™



OpenVINO™ Toolkit

Open **V**isual **I**nference and **N**eural Network **O**ptimization

1. Dataset Creation
2. Training
3. Deploy/Infer

Notes based on
CS231n, Stanford University, and
EECS 498-007 / 598-005, University of Michigan

1. Dataset Creation

Existing datasets are best

- Kaggle
- A list on wikipedia

Otherwise:

1.1 Collect Images

- Take own photos
- Scrape the interwebs
- Synthetic/made up data

1.2 Annotate

- LabelImg
- CVAT
- Roboflow
- Perceptilabs

1.3 Data Augmentation

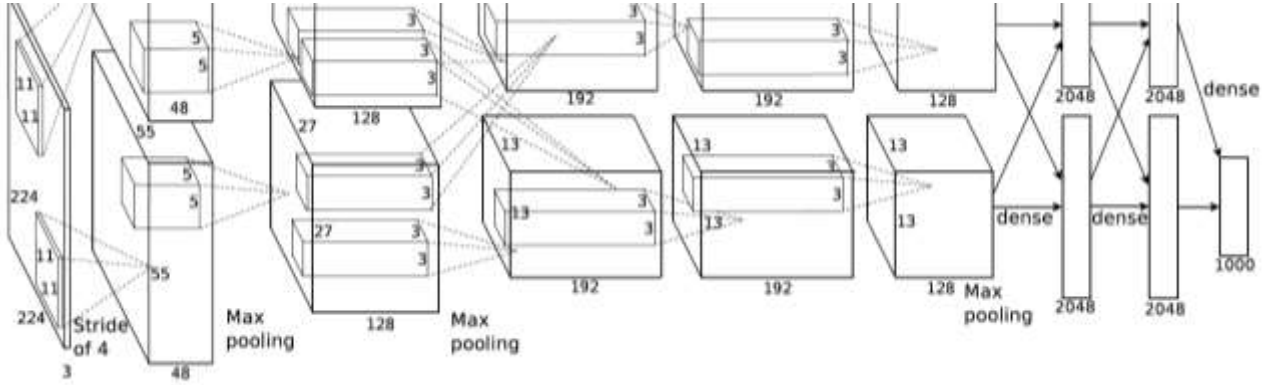
- Albumentations

1.4 Export to the correct format

2. Training

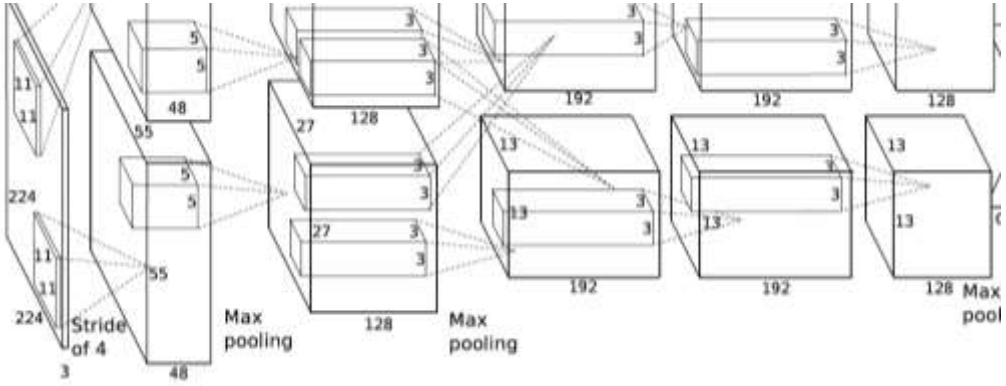
Train from scratch

- PyTorch, Tensorflow, Keras, etc.



Transfer learn – rewire the dense layer at the output

- PyTorch, Tensorflow, Keras
- Teachable machine
- Roboflow

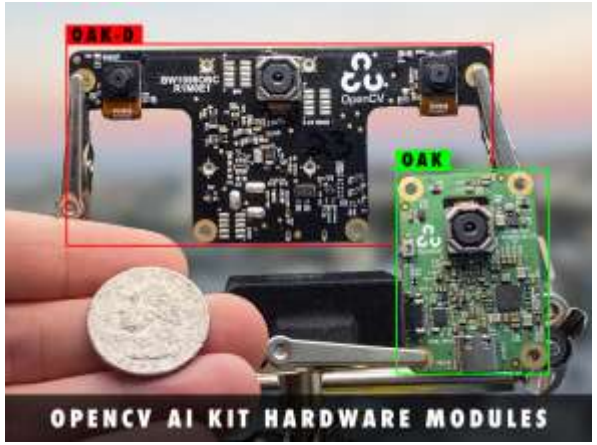


3. Deploy/Infer

- Edge devices

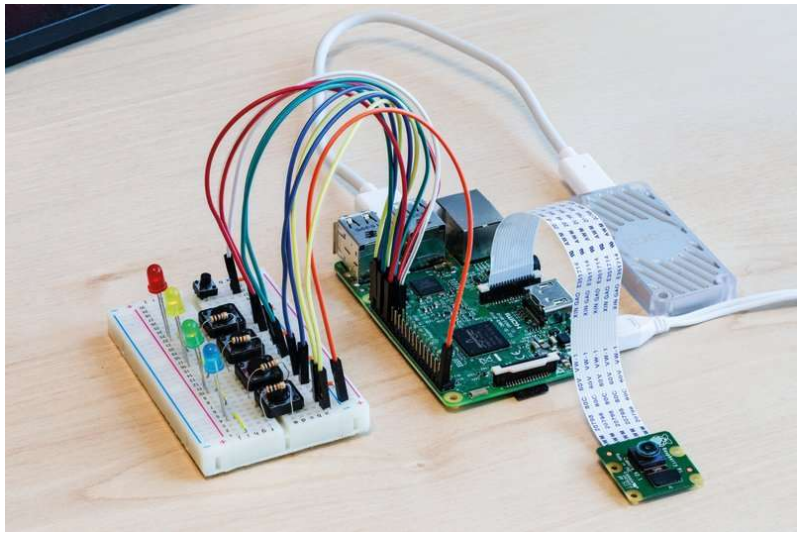
Ability AI Vue camera with VPU

Intel NUC with VPU

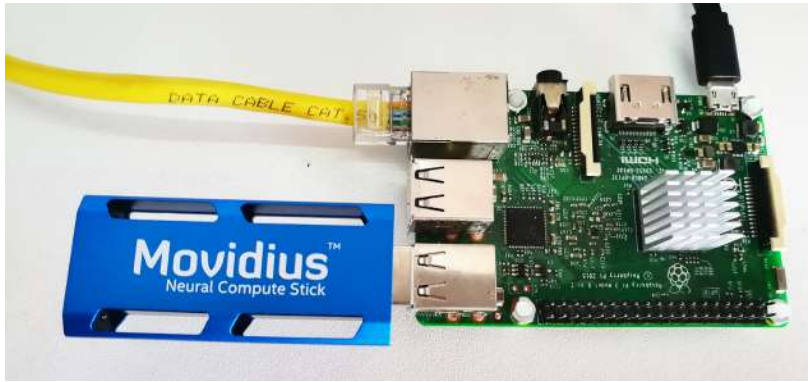


OpenCV AI Kit with Myriad X

Jetson Nano 2GB



Raspberry Pi with Coral Edge TPU



Raspberry Pi with NCS