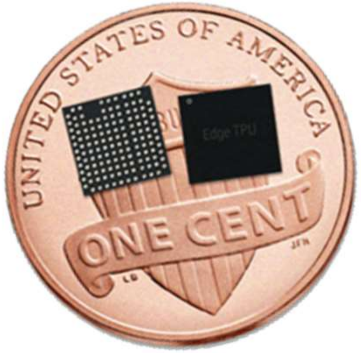
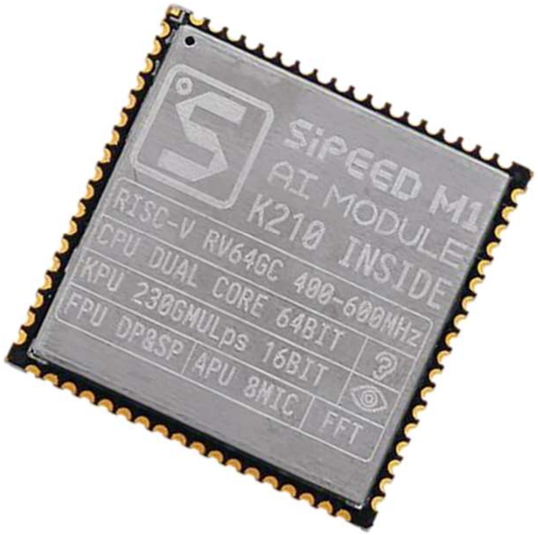
Artificial Neural Networks At the Edge

By Shaun Price

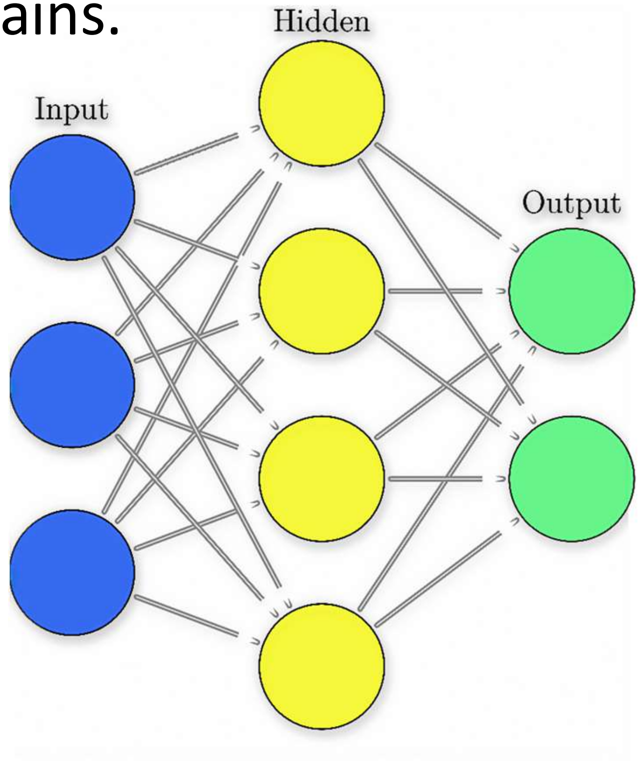
https://www.linkedin.com/in/shaunprice/





Artificial Neural Networks

- Inspired by biological neural networks in animal brains.
- Learn from example data to recognize patterns.
- They classify and cluster input data.
- They can classify by learning from labels people have given the training data – Supervised learning.
- They can cluster by detecting similarities between data - Unsupervised learning.



Why at the Edge Computing

- Bandwidth for communications is either limited or non-existent.
- Communications Latency is usually high.
- Privacy issues in sensitive environments (children, corporate, military, security devices).
- No centralised servers required.
- Devices can be independent and autonomous.

Why AI at the Edge?

- AI is learned and can infer complex decisions from prior training quickly and efficiently.
- AI can analyse images, sounds and other complex patterns not easily analyses using other computational technologies.
- Edge devices such as cameras, drones, robots and other edge sensors can use this inference to perform tasks such as detect behaviour, robot object avoidance, respond to speech.
- Specialised Inference at the edge devices allows for:
    - Quicker, low latency, decisions.
    - Low bandwidth or no-bandwidth inference.
    - Reduce the processing power requirements of smart edge devices.

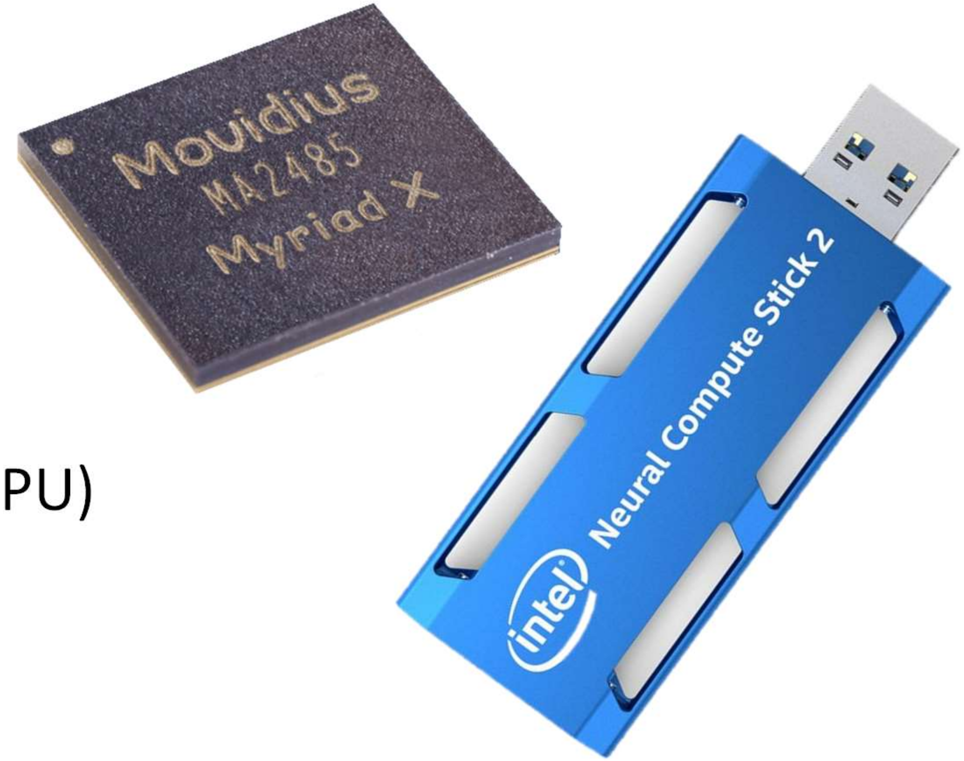
Issues with AI at the Edge?

- Power is normally limited
- Processing power is normally limited
- Size typically needs to be small
- Costs need to be low because there are typical many units

Currently Available AI at the Edge Devices

- The main AI at the Edge devices currently available include:
  - Intel Movidius Myriad Neural Compute
  - Google Edge TPU ML Accellerator Coprocessor
  - Kendryte K210 also used in the SiPEED M1 AI Module
  - Nvidia Jetson Nano
  - Lattice iCE40 UltraPLus FPGA
- Others are coming to market.

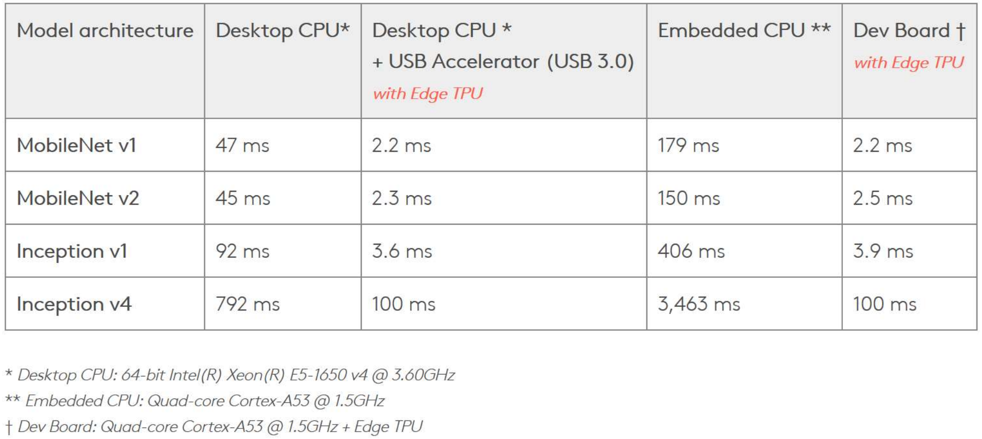
Intel Movidius Myriad



- ~AU$150 from Mouser Electronics
- Stated Performance: 4 TOPS
- A Neural Net Vision Processing Unit (VPU)
- Runs TensorFlow and Caffe models
- Development options include:
  - Intel NCS2 which uses the Myriad X VPU
  - Up-board AI Edge, Up Squared AI Vision X, Up Core Plus,
- Previous Generation was the Myriad 2 used in the Moviduis Neural Compute Stick

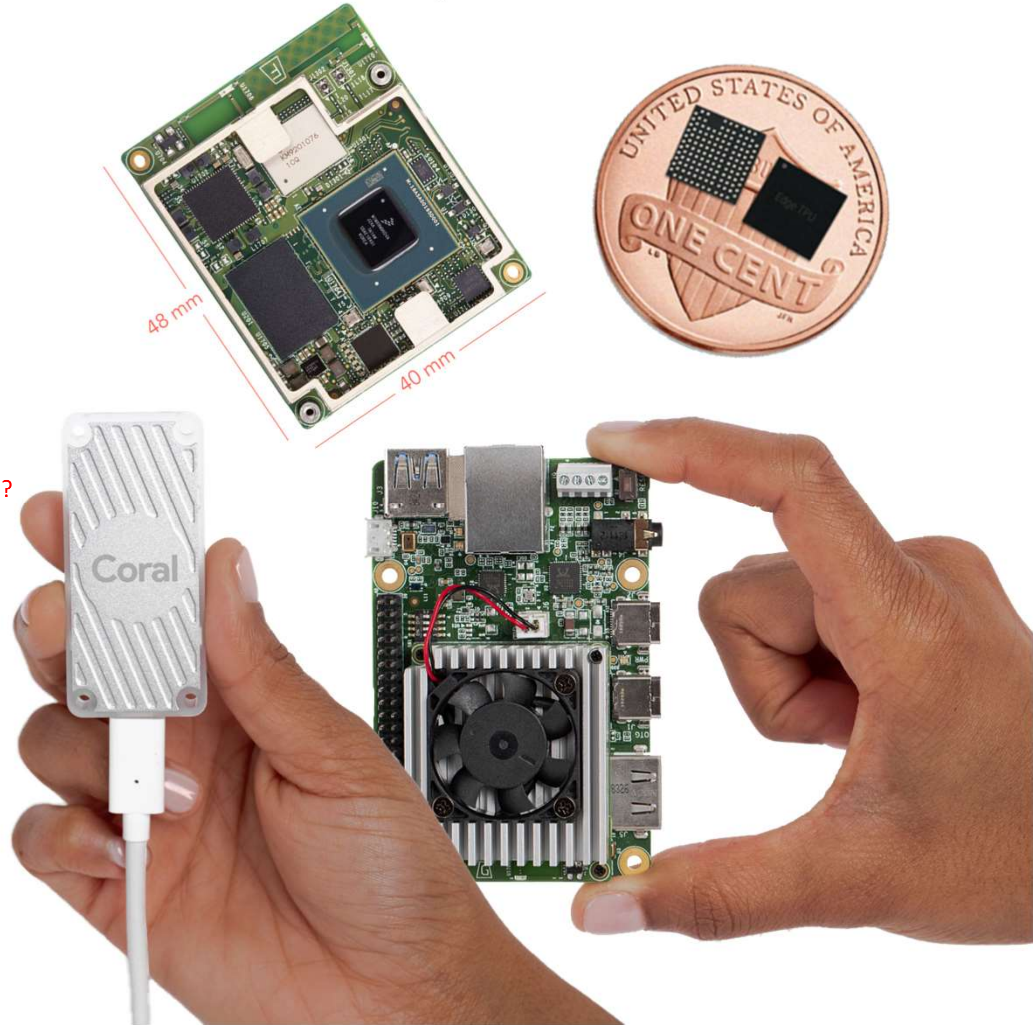
Google Edge TPU ML accelerator coprocessor

- ~US$75 for the Corel USB Stick
- ~US$150 for the Corel Dev Board
- SoM, Mini PCIe, and M.2 versions Coming Soon
- Supports TensorFlow lite only.
- Stated Performance: 4 TOPS at 2 TOPS/Watt
- There's shipping restrictions on these devices to Australia but you can buy them in Hong Kong???



Model architecture	Desktop CPU*	Desktop CPU * + USB Accelerator (USB 3.0) *with Edge TPU*	Embedded CPU **	Dev Board † *with Edge TPU*
MobileNet v1	47 ms	2.2 ms	179 ms	2.2 ms
MobileNet v2	45 ms	2.3 ms	150 ms	2.5 ms
Inception v1	92 ms	3.6 ms	406 ms	3.9 ms
Inception v4	792 ms	100 ms	3,463 ms	100 ms

** Desktop CPU: 64-bit Intel(R) Xeon(R) E5-1650 v4 @ 3.60GHz*
*** Embedded CPU: Quad-core Cortex-A53 @ 1.5GHz*
† Dev Board: Quad-core Cortex-A53 @ 1.5GHz + Edge TPU

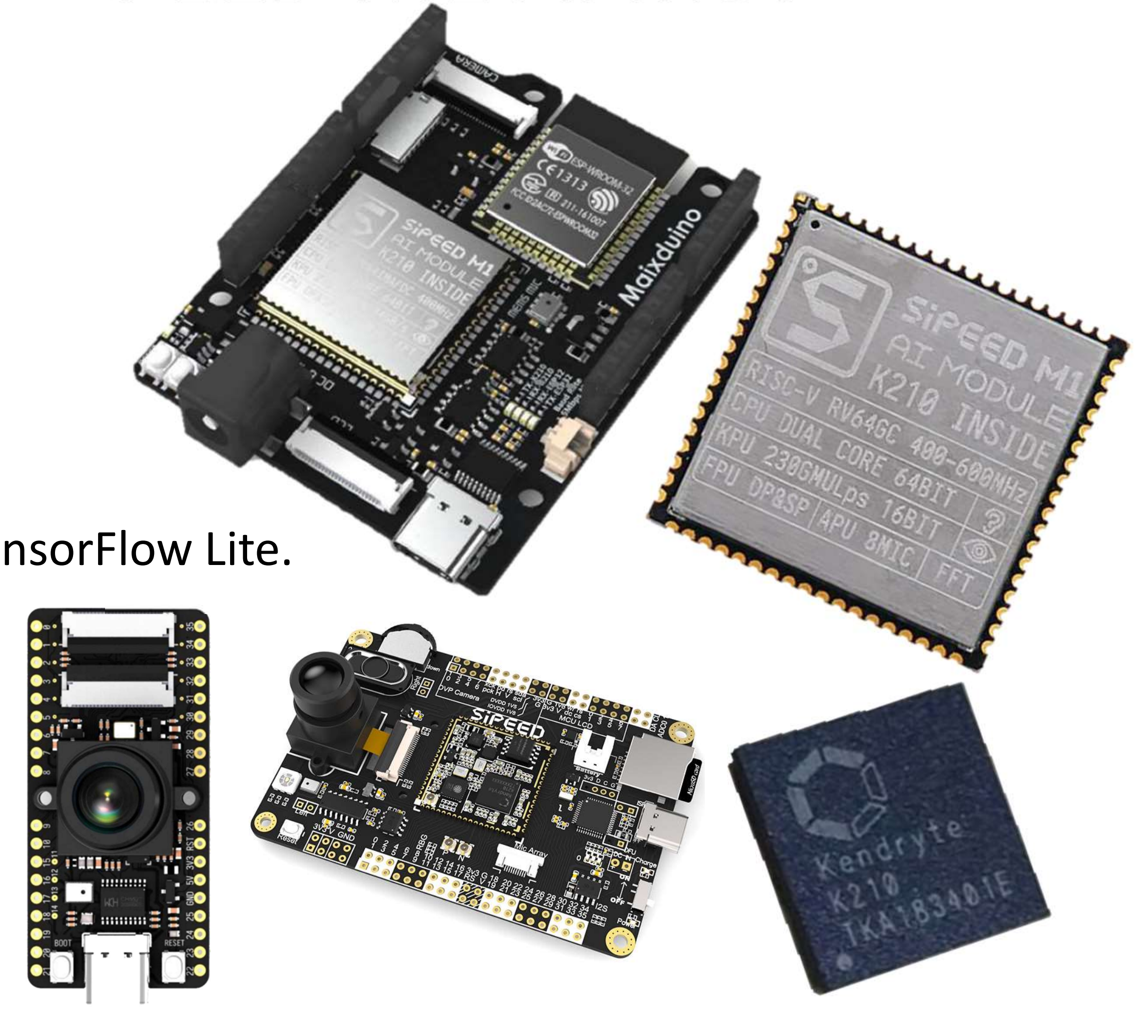


Nvidia Jetson Nano

- US$99 for the Development Kit.
- Uses 128-core NVIDIA Maxwell GPU for Inference.
- Stated Performance: 472 GOPS at 5-10 Watt
- Runs multiple deep learning models (It’s a GPU).
- Lots of examples including a 3D printed robot running ROS Melodic and using a Raspberry Pi Camera.
- Has a Raspberry Pi compatible connector.

Kendryte K210 and the SiPEED M1 AI Module

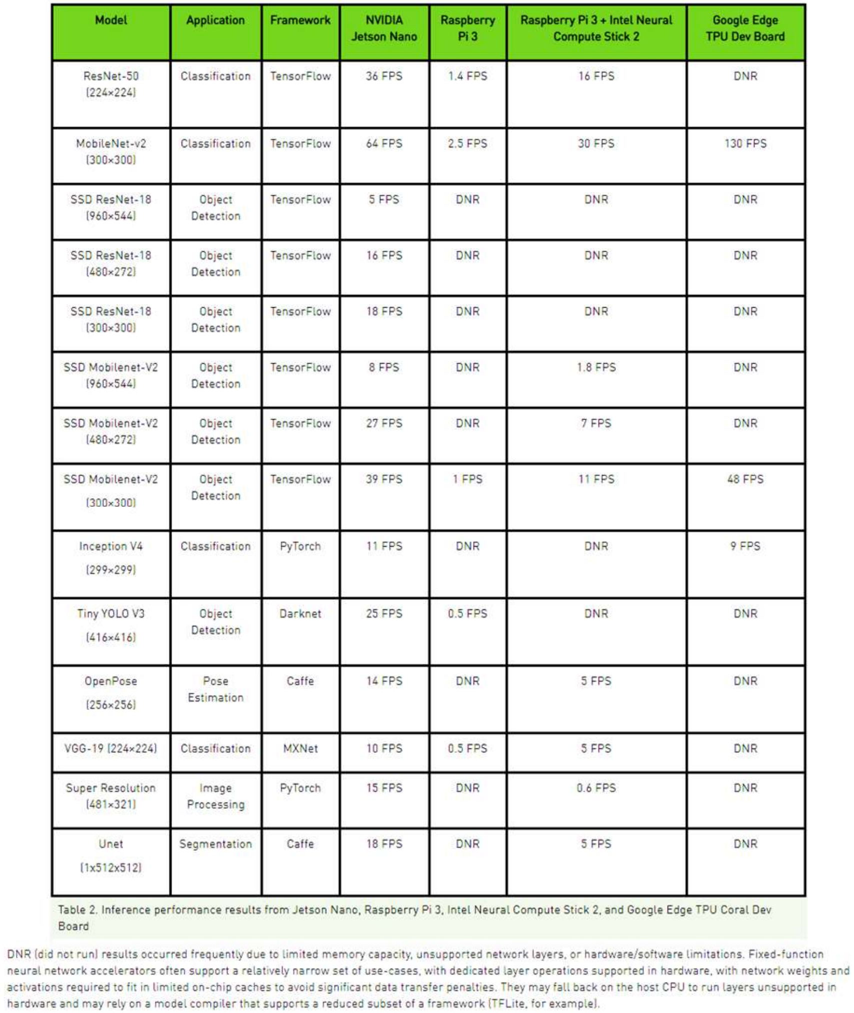
- ~US$9 for the SiPEED Maix-I Module
- ~US$24 for the MAixDuino which comes with an LCD and a camera module.
- ~US$13 for the MAIX Bit with a camera.
- Stated Performance:
  - 0.25 TOPS/0.3W @ 400MHz
  - 0.50 TOPS @ 800MHz
- Support for tiny-yolo, mobilenet-v1 and TensorFlow Lite.
- Camera and Sound Inputs with Audio Processor for 8 mics (192kHz) and an FFT Accelerator.
- ESP8285 Wi-Fi on the M1.
- The MAixDuino has an ESP32 on-board.



Lattice iCE40 UltraPlus FPGA

- AU$150 from Digi-Key or Mouser Electronics for the mobile dev board (see image).
- Integrated AI inference capability with sensAI and the Convolutional Neural Network (CNN) Compact Accelerator IP.
- FPGA's are software defined hardware.
- FPGA are low power (face detection ~800µW) and very high speed (face detection in 10mS or 100fps). approaching the speed of an ASIC.
- Currently supports TensorFlow, Caffe and Keras.
- Hardware can be put into production and updated models and firmware can be sent to upgrade functionality, patch bugs and fix security vulnerabilities in the field.
- Example Human Face Detection on Youtube: https://www.youtube.com/watch?v=YqtYiuPd6io

Comparison of Edge AI Devices by Nvidia



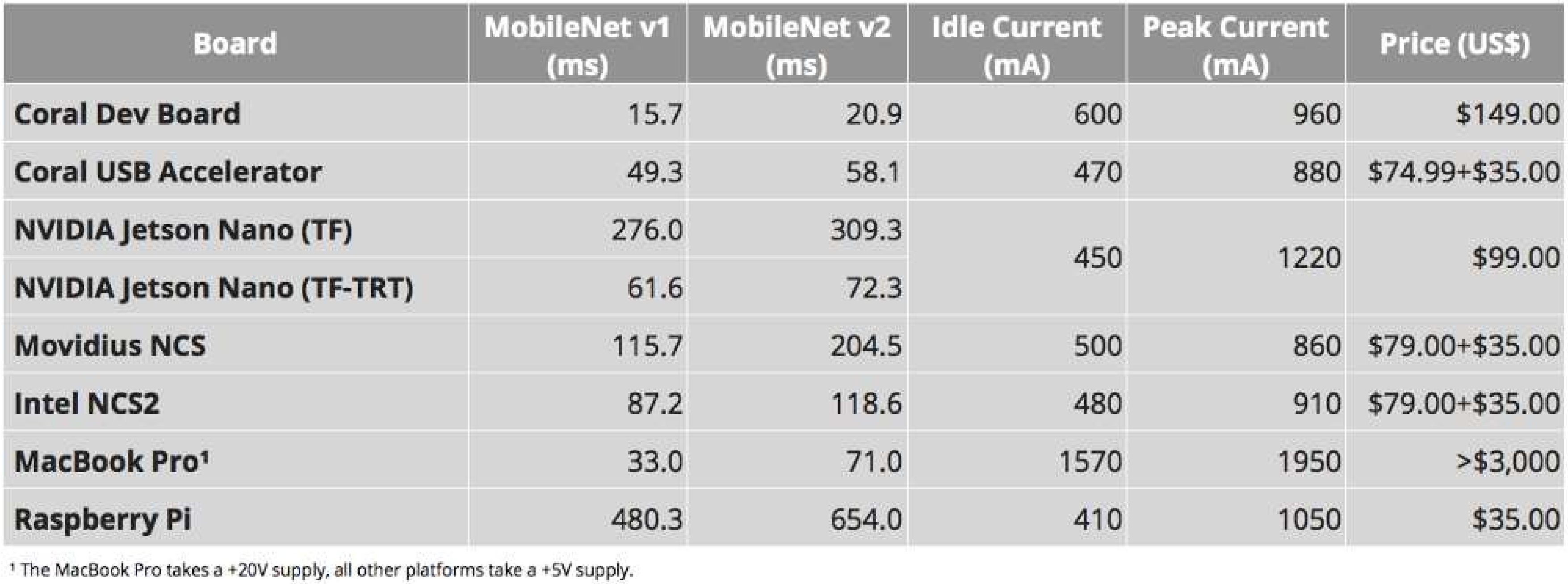
Model	Application	Framework	NVIDIA Jetson Nano	Raspberry Pi 3	Raspberry Pi 3 + Intel Neural Compute Stick 2	Google Edge TPU Dev Board
ResNet-50 (224×224)	Classification	TensorFlow	36 FPS	1.4 FPS	16 FPS	DNR
MobileNet-v2 (300×300)	Classification	TensorFlow	64 FPS	2.5 FPS	30 FPS	130 FPS
SSD ResNet-18 (960×544)	Object Detection	TensorFlow	5 FPS	DNR	DNR	DNR
SSD ResNet-18 (480×272)	Object Detection	TensorFlow	16 FPS	DNR	DNR	DNR
SSD ResNet-18 (300×300)	Object Detection	TensorFlow	18 FPS	DNR	DNR	DNR
SSD Mobilenet-V2 (960×544)	Object Detection	TensorFlow	8 FPS	DNR	1.8 FPS	DNR
SSD Mobilenet-V2 (480×272)	Object Detection	TensorFlow	27 FPS	DNR	7 FPS	DNR
SSD Mobilenet-V2 (300×300)	Object Detection	TensorFlow	39 FPS	1 FPS	11 FPS	48 FPS
Inception V4 (299×299)	Classification	PyTorch	11 FPS	DNR	DNR	9 FPS
Tiny YOLO V3 (416×416)	Object Detection	Darknet	25 FPS	0.5 FPS	DNR	DNR
OpenPose (256×256)	Pose Estimation	Caffe	14 FPS	DNR	5 FPS	DNR
VGG-19 (224×224)	Classification	MXNet	10 FPS	0.5 FPS	5 FPS	DNR
Super Resolution (481×321)	Image Processing	PyTorch	15 FPS	DNR	0.6 FPS	DNR
Unet (1x512x512)	Segmentation	Caffe	18 FPS	DNR	5 FPS	DNR

Table 2. Inference performance results from Jetson Nano, Raspberry Pi 3, Intel Neural Compute Stick 2, and Google Edge TPU Coral Dev Board

DNR (did not run) results occurred frequently due to limited memory capacity, unsupported network layers, or hardware/software limitations. Fixed-function neural network accelerators often support a relatively narrow set of use-cases, with dedicated layer operations supported in hardware, with network weights and activations required to fit in limited on-chip caches to avoid significant data transfer penalties. They may fall back on the host CPU to run layers unsupported in hardware and may rely on a model compiler that supports a reduced subset of a framework (TFLite, for example).

Independent Benchmark of Edge Computing

Testing by Alasdair Allan
Full article at: https://medium.com/@aallan/benchmarking-edge-computing-ce3f13942245



Board	MobileNet v1 (ms)	MobileNet v2 (ms)	Idle Current (mA)	Peak Current (mA)	Price (US$)
Coral Dev Board	15.7	20.9	600	960	$149.00
Coral USB Accelerator	49.3	58.1	470	880	$74.99+$35.00
NVIDIA Jetson Nano (TF)	276.0	309.3	450	1220	$99.00
NVIDIA Jetson Nano (TF-TRT)	61.6	72.3			
Movidius NCS	115.7	204.5	500	860	$79.00+$35.00
Intel NCS2	87.2	118.6	480	910	$79.00+$35.00
MacBook Pro[1]	33.0	71.0	1570	1950	>$3,000
Raspberry Pi	480.3	654.0	410	1050	$35.00

[1] The MacBook Pro takes a +20V supply, all other platforms take a +5V supply.

How Can I Use AI at the Edge

- Edge AI devices are used for:
  - Detecting objects (colors, signs, objects to pick up)
  - Avoiding objects (autonomous robots avoiding obstacles in their path)
  - Detecting people and identifying their mood
  - Staying within lanes or on a path
  - Monocular stereo vision for depth perception using a single camera
  - Voice recognition and command interpretation
  - Voice translation in real time anywhere
  - Smart Toys, Security Cameras, Drones, Robots, Medical devices, Manufacturing, Phones, Cameras..
- Examples where Edge AI hardware is used today:
  - DJI use Movidius VPU’s in their Phantom 4 and Spark for obstacle avoidance.
  - AWS DeepRacer has a Moviduis X VPU for vision processing to follow a track.
  - Samsung Galaxy S10 (with Exynos 9 Series 9820 processor)
  - Apple iPhone X A11 Bionic Chipset
  - Google Pixel 2 and newer

References

Where to buy from Australia

Google Corel:
Not available for export to Australia
https://aiyprojects.withgoogle.com/edge-tpu/

Intel Movidius:

au.mouser.com

SiPEED Maix:

www.seeedstudio.com

Lattice iCE40 UltraPlus:

au.mouser.com

www.digkey.com.au

Videos

Hardware Acceleration for AI at the Edge by Microsoft Developer
https://www.youtube.com/watch?v=OmOV_4MZ2aM

Hybrid Machine Learning: From the Cloud to the Edge (Cloud Next '18) by Google

https://www.youtube.com/watch?v=M-29naAVmI4

Articles

Comparison of Edge Compute Devices

https://medium.com/@aallan/benchmarking-edge-computing-ce3f13942245