EVD 3

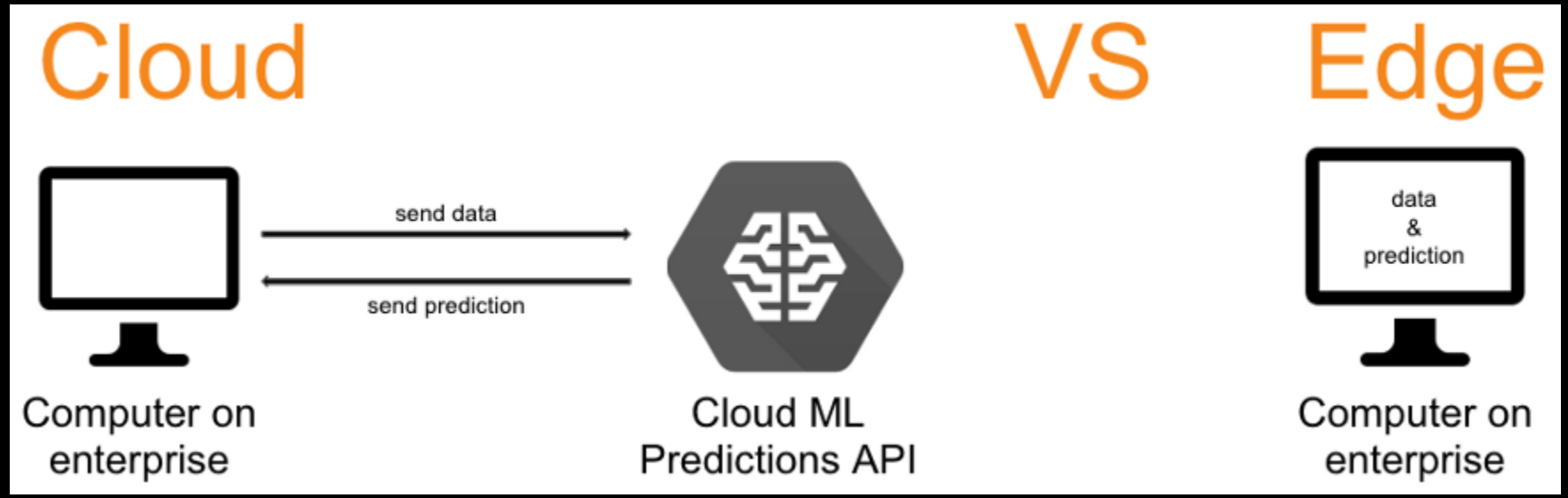
EDGE COMPUTING

JEROEN VEEN

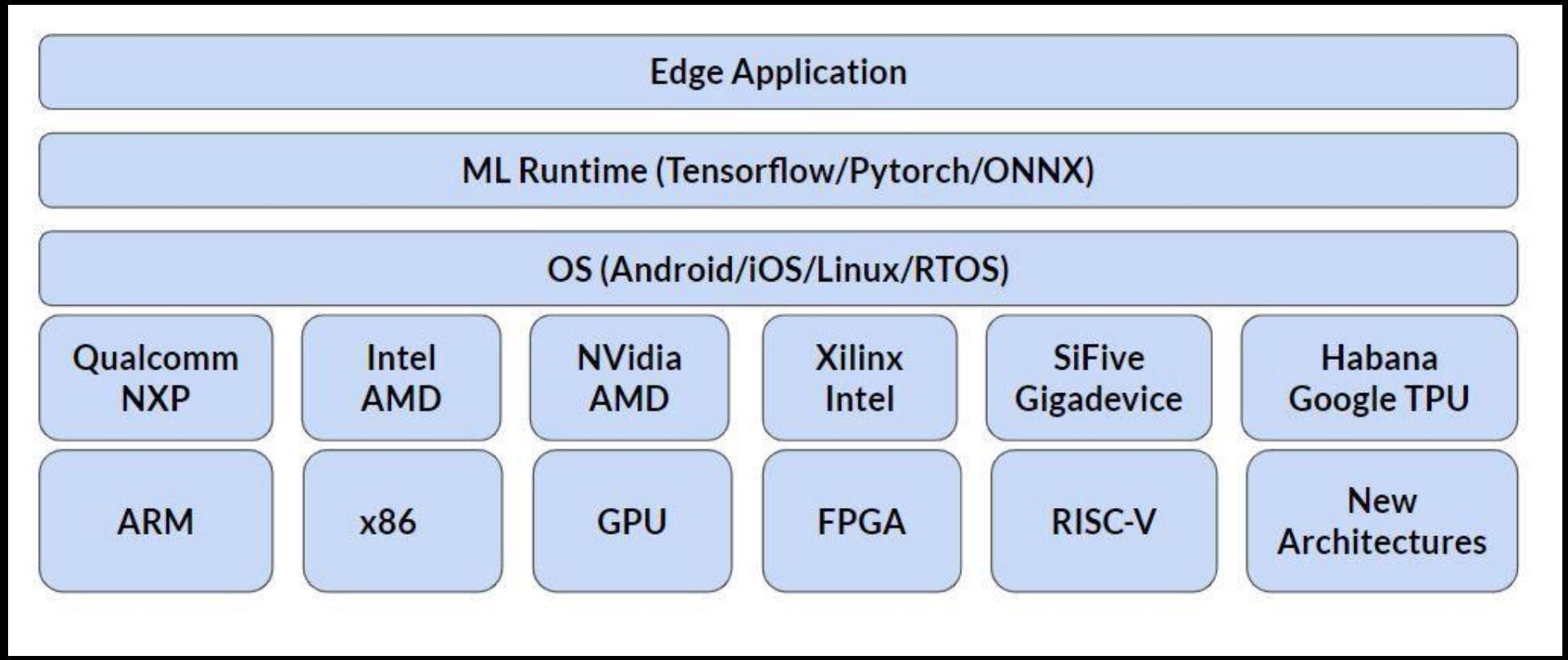
HAN_UNIVERSITY
OF APPLIED SCIENCES

WHY?

- Network latency and bandwidth
- Security and decentralization
- Distributed swarms and redundancy



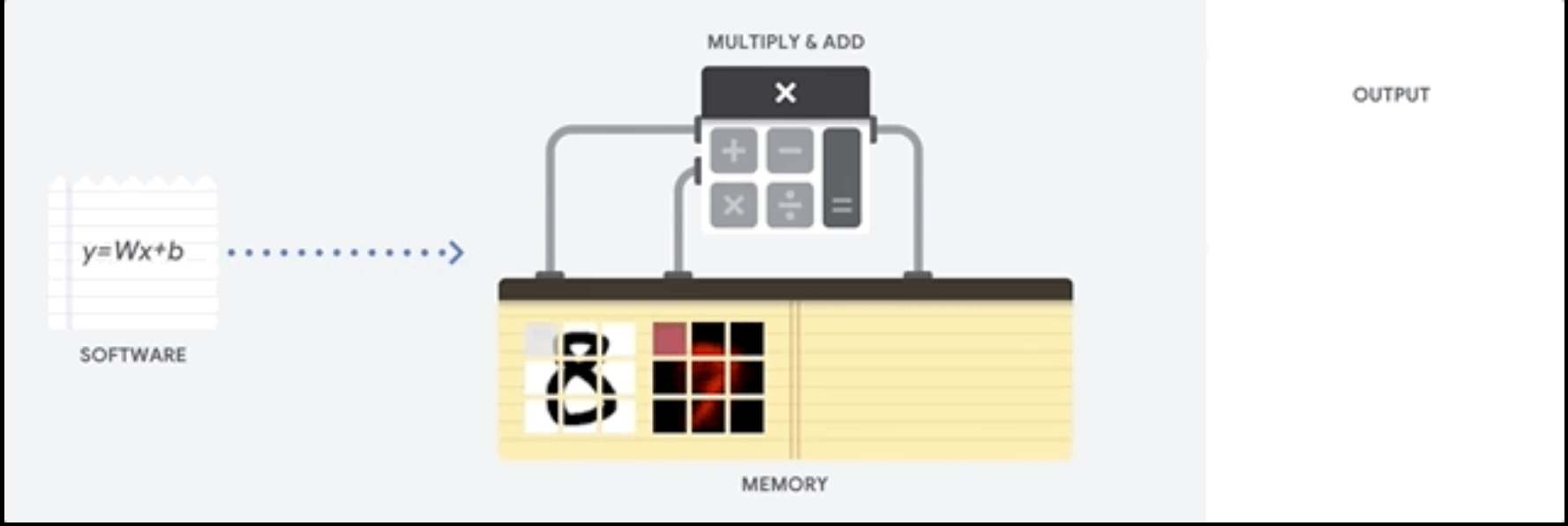
EDGE ARCHITECTURES



NEURAL PROCESSING UNITS (NPU)

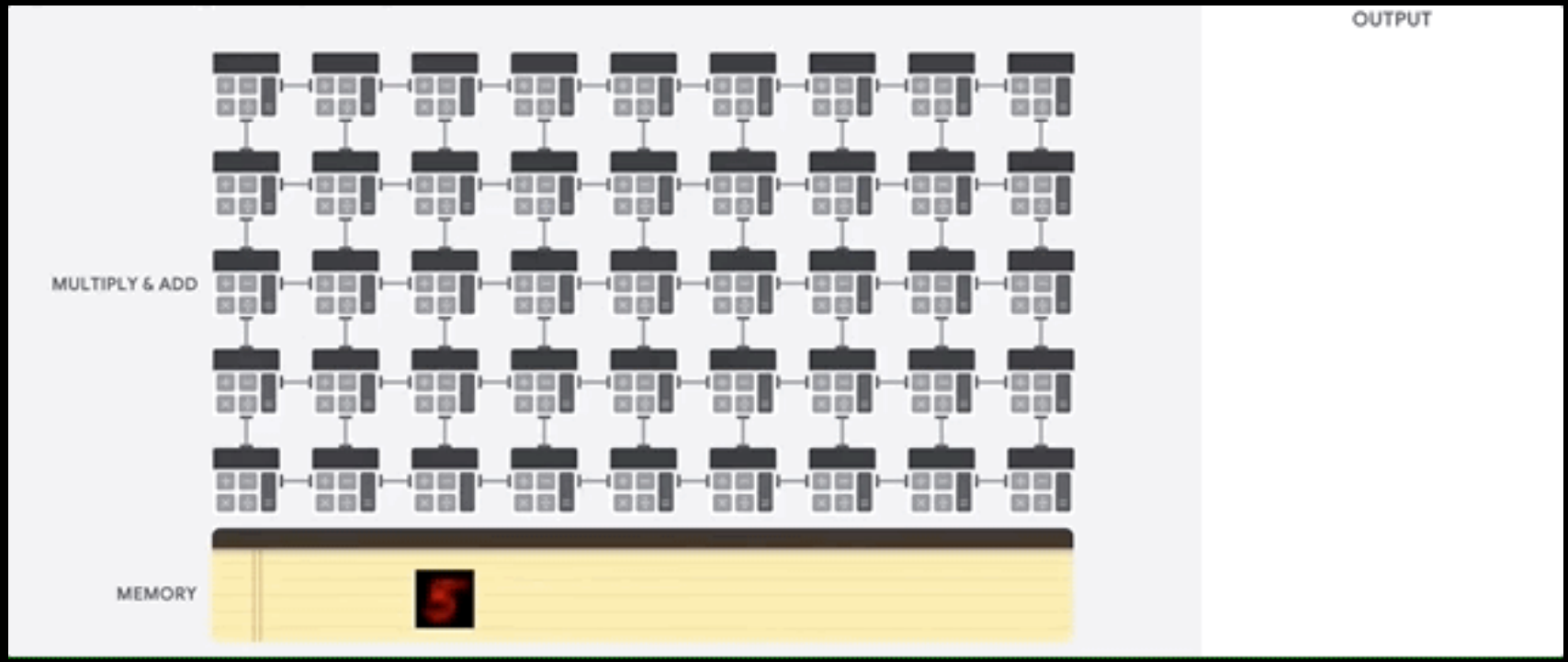
- Aka tensor processing unit (TPU), vision processing unit (VPU)
- Typically optimized for ANNs, CNNs or random forests (RFs)
- Lots of parallel multiply-accumulate operations

MULTIPLY-ADD OPERATION ON CPU

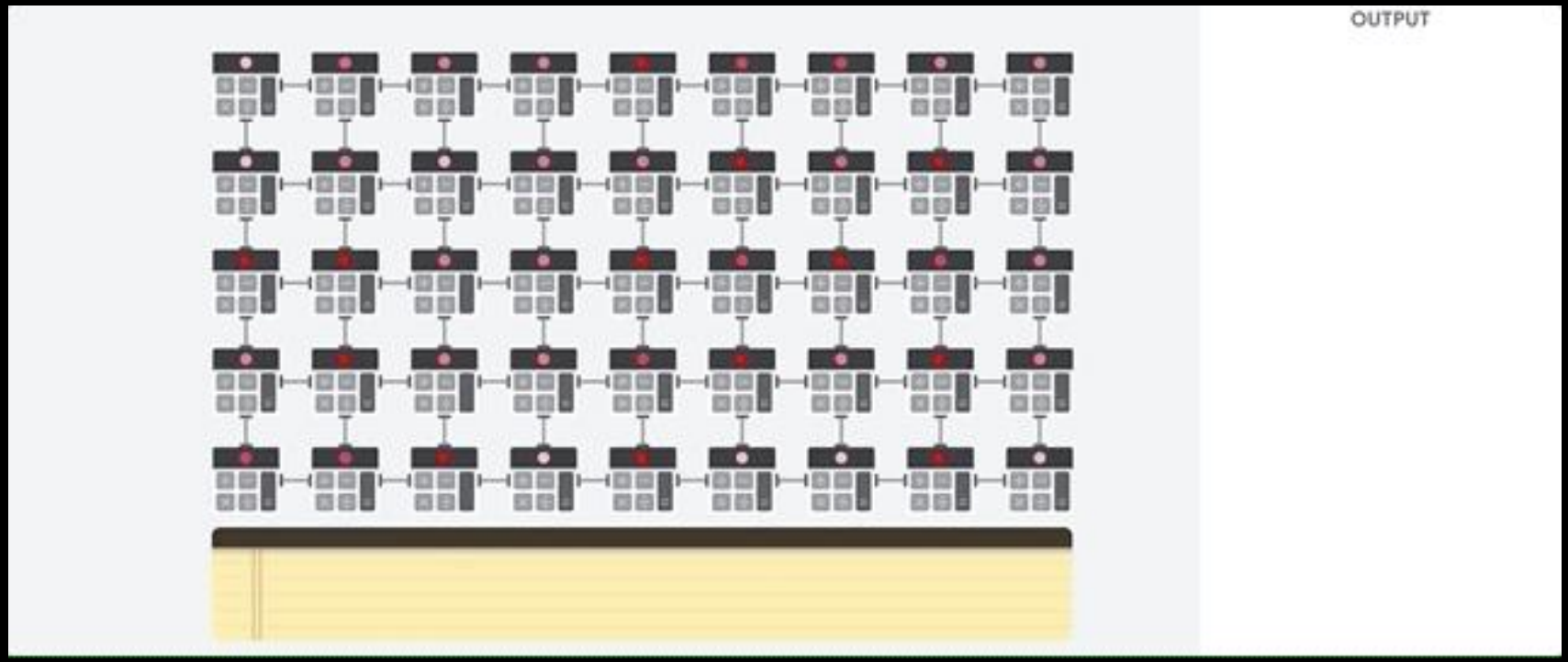


MULTIPLY-ADD OPERATION ON GPU

READING WEIGHTS FOR THE MULTIPLY-ADD OPERATION ON TPU

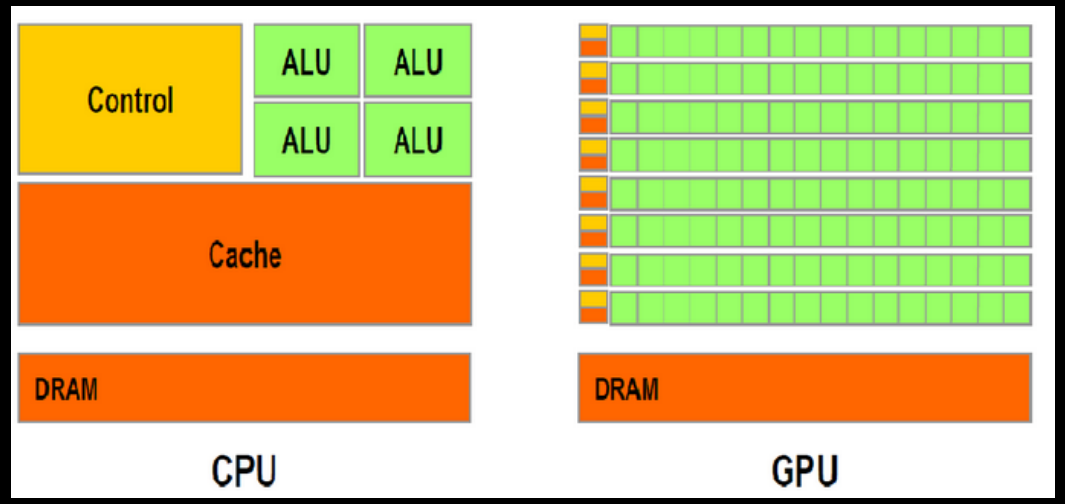


MULTIPLY-ADD OPERATION ON TPU



CPU VS GPU

- https://youtu.be/-P28LKWTzrI

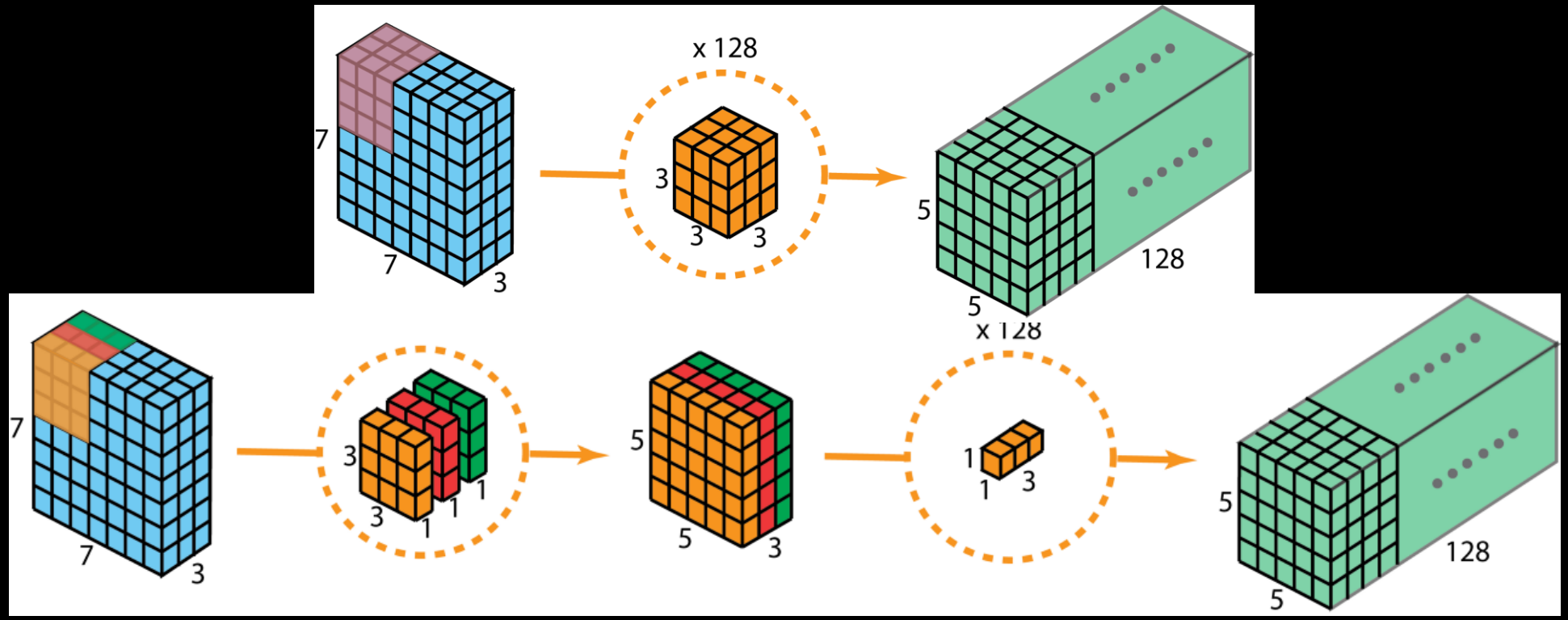


REDUCING COMPUTATIONAL COMPLEXITY

- Quantization
- Pruning of trees
- Separable convolutions

MOBILENET FOR EDGE-COMPUTING

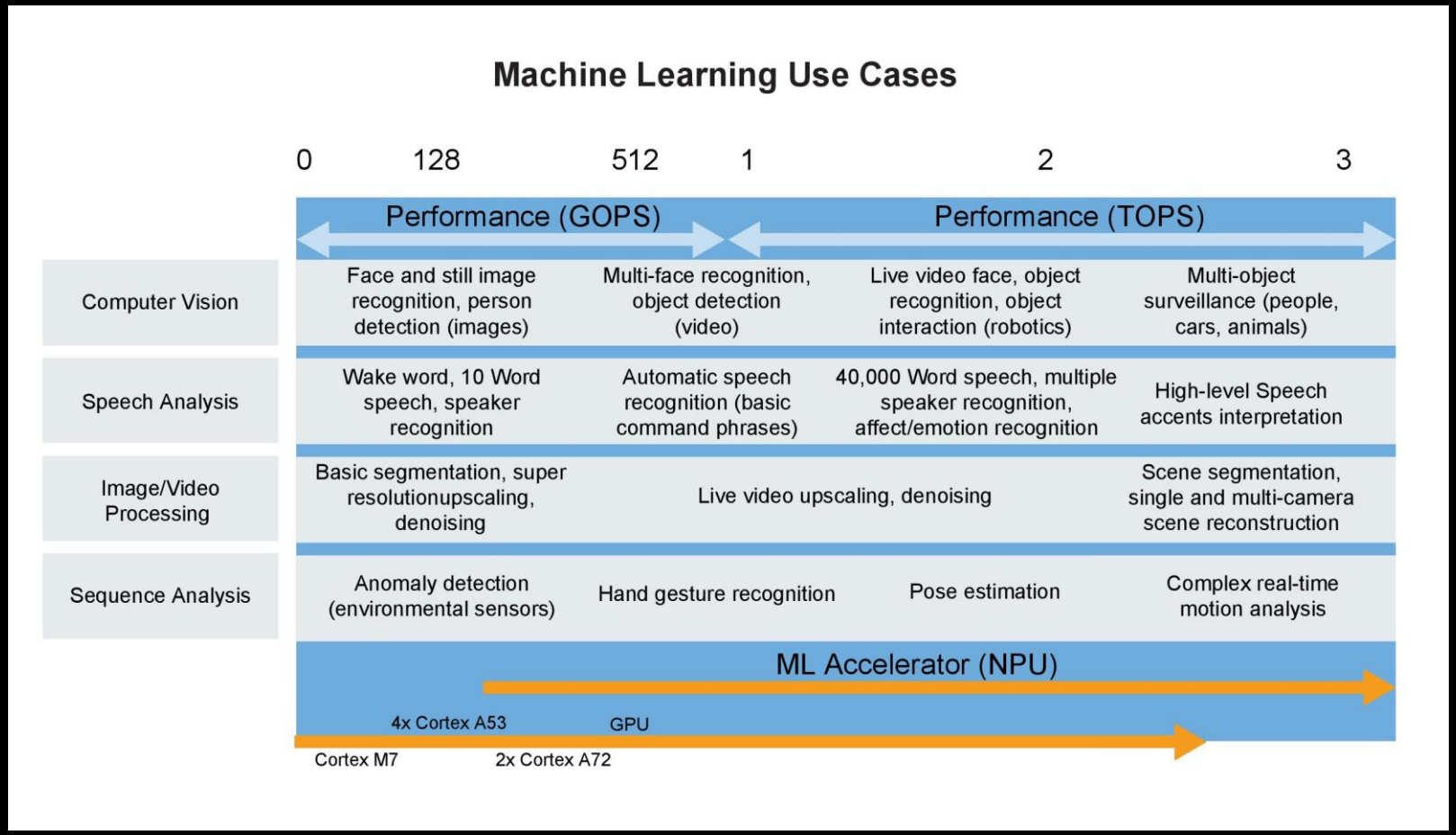
- https://iq.opengenus.org/separable-convolution/



UTENSOR, TENSORFLOW LITE

- Open-source frameworks to bring machine learning onto microcontrollers by a micro-inference engine
- Support for Arm Cortex-M, Arduino nano 33 BLE, ESP32, etc.
- Optimized pre-trained models for common mobile and edge use cases e.g., using mobilenet
- TensorFlow Lite converter: API that converts trained TensorFlow models into the TensorFlow Lite format
-

NEXT: TINYML


Machine Learning Use Cases
0
128
512
1
2
3
Performance (GOPS)
Performance (TOPS)
Computer Vision
Face and still image recognition, person detection (images)
Multi-face recognition, object detection (video)
Live video face, object recognition, object interaction (robotics)
Multi-object surveillance (people, cars, animals)
Speech Analysis
Wake word, 10 Word speech, speaker recognition
Automatic speech recognition (basic command phrases)
40,000 Word speech, multiple speaker recognition, affect/emotion recognition
High-level Speech accents interpretation
Image/Video Processing
Basic segmentation, super resolutionupscaling, denoising
Live video upscaling, denoising
Scene segmentation, single and multi-camera scene reconstruction
Sequence Analysis
Anomaly detection (environmental sensors)
Hand gesture recognition
Pose estimation
Complex real-time motion analysis
ML Accelerator (NPU)
4x Cortex A53
GPU
Cortex M7
2x Cortex A72

POPULAR EDGE DEVELOPMENT PLATFORMS

- Jetson family of NVIDIA
- Google Coral edge TPU or stick
- Intel Movidius
  https://www.intel.com/content/www/us/en/products/processors/movidius-vpu.html
- Intel FPGAs for AI
- Kendryte K210

DEMOS

- https://youtu.be/kZrjmy_UeQw
- https://www.youtube.com/watch?v=Ou-gulnNkaE&feature=youtu.be&t=117