AI on the Edge: Characterizing AI-based IoT Applications using Specialized Edge Architectures

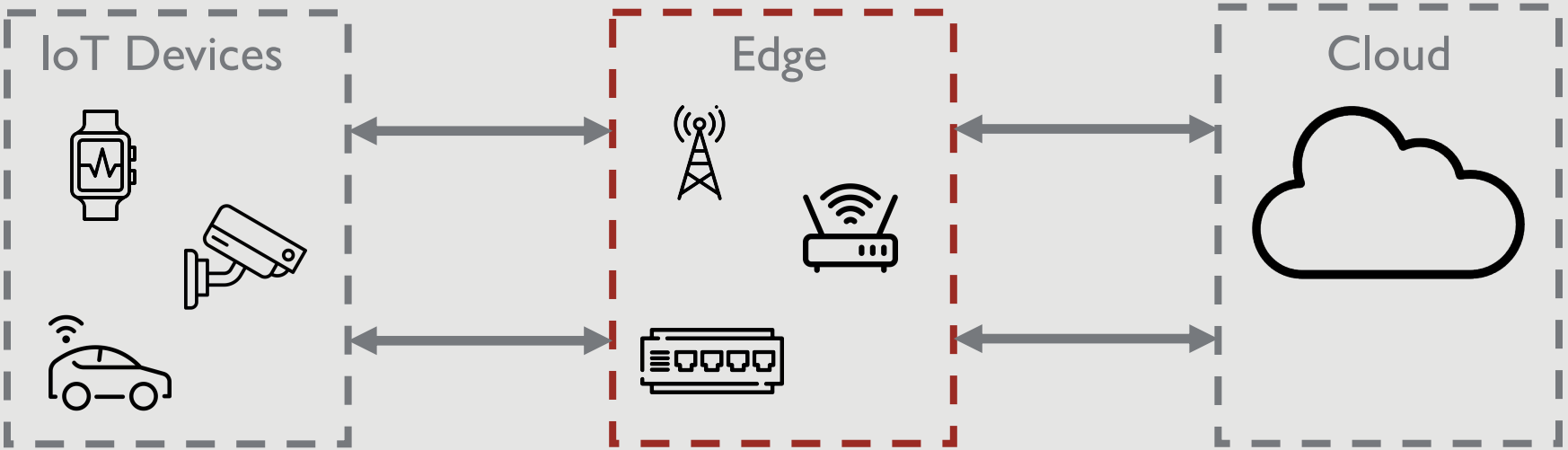
Qianlin Liang
qliang@cs.umass.edu

Prashant Shenoy
shenoy@cs.umass.edu

David Irwin
irwin@ecs.umass.edu

EDGE

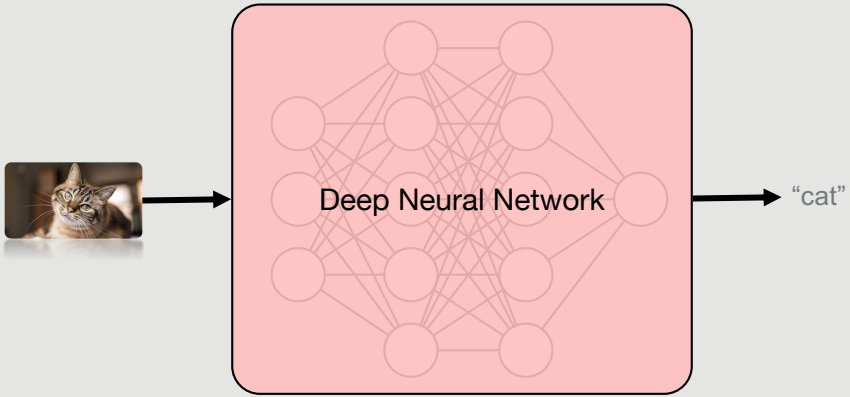
Computing infrastructure that is positioned between endpoint device and cloud.



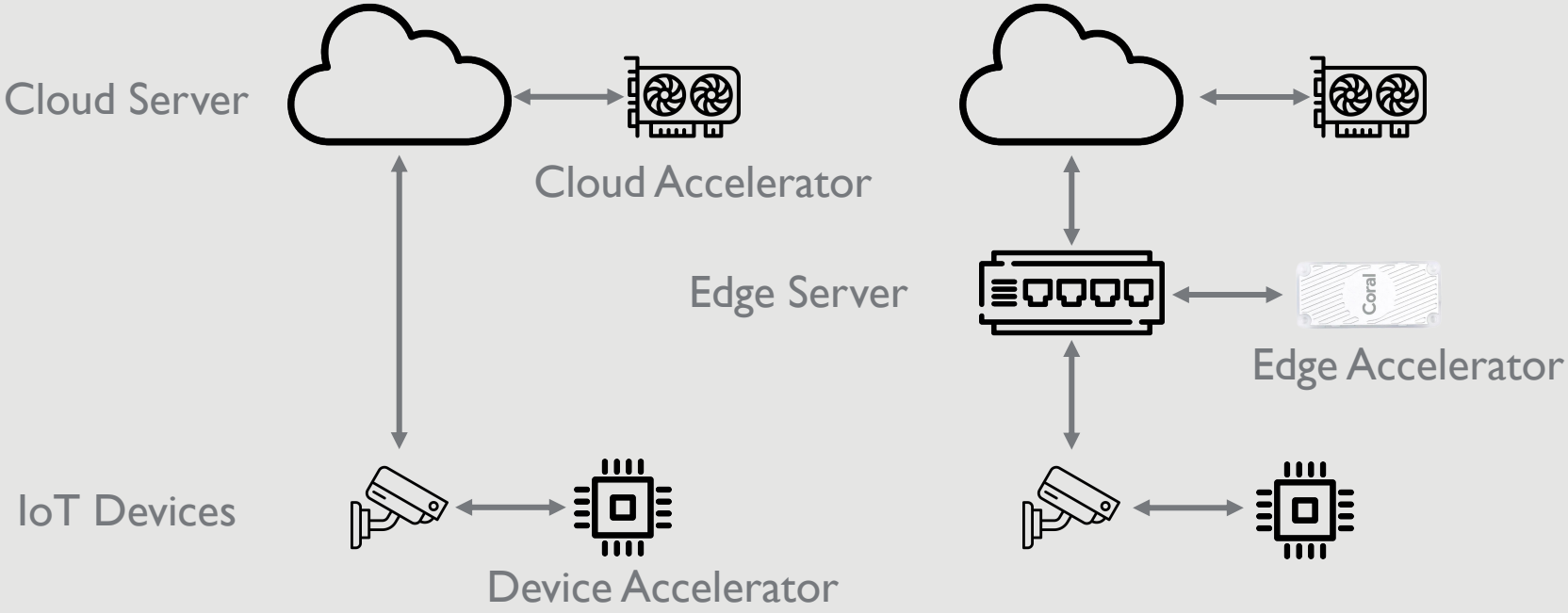
EDGE-BASED AI WORKLOADS

An emerging class of edge workloads:

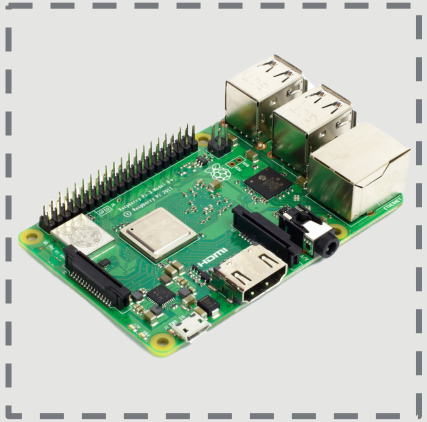
- Running deep learning inference on edge
- Computationally intensive



COMPUTING PARADIGMS FOR IOT APPLICATIONS



TRADITIONAL EDGE VS. ACCELERATORS



VS.

1) What are the price, performance, and energy benefits offered by edge hardware accelerators?
2) How should modern IoT applications exploit the distributed processing capabilities of specialized edge nodes and the cloud by using various types of split processing?
3) How suitable are edge accelerators for supporting concurrent edge applications from multiple tenants?

SPECIALIZED EDGE ACCELERARTORS

Intel NCS2

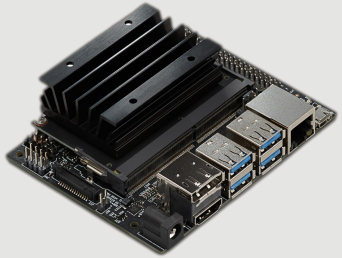
Power:	1-2 W
Memory:	512 MB
Price:	$99

Accelerate computer vision workloads

Google EdgeTPU

Power:	0.5-2 W
Memory:	8 MB
Price:	$75

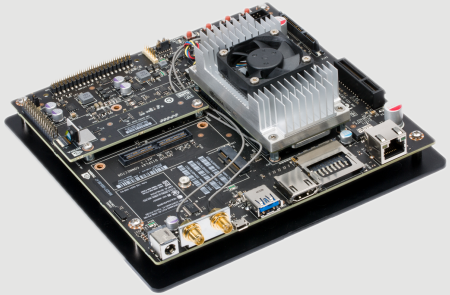
Accelerate 8-bit quantized models



Nvidia Jetson Nano

Power:	5-10 W
Memory:	4 GB
Price:	$99

Accelerate any GPU workloads



Nvidia Jetson TX2

Power:	7.5-15 W
Memory:	8 GB
Price:	$399

Accelerate any GPU workloads

METHODOLOGY

To ensure a fair comparison across hardware platforms, we run the same model on all platforms and subject it to the same inference workload.

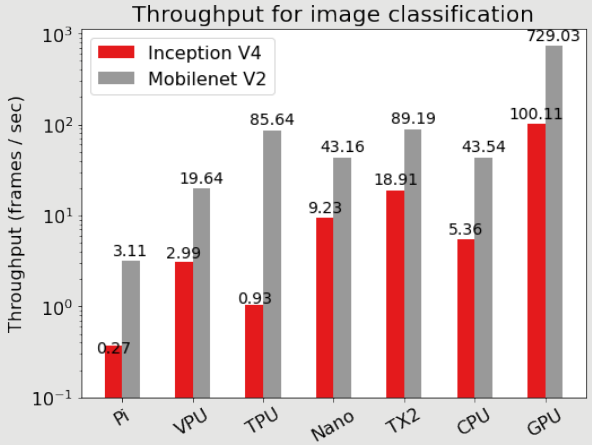
Workloads

- MobileNet V2 (Image classification)
- Inception V4 (Image classification)
- SSD MobileNet V1 (Object detection)
- SSD MobileNet V2 (Object detection)
- cnn-trad-fpool3 (Keyword spotting)

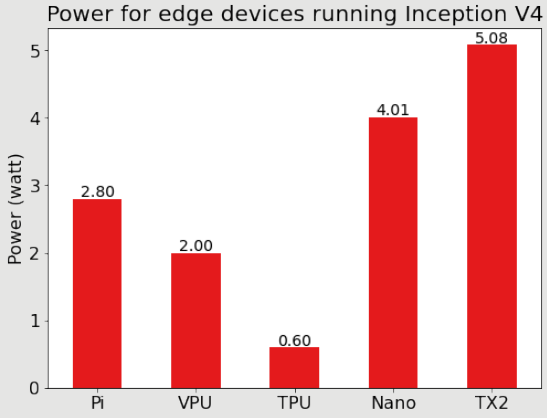
Platforms

- AWS p3.2xlarge (Server-class)
- Nvidia Tesla V100 GPU (Server-class)
- Raspberry Pi 3 B+ (Edge-class)
- Intel NCS 2 (Accelerator)
- Google EdgeTPU (Accelerator)
- Nvidia Jetson Nano/TX2 (Accelerator)

PERFORMANCE AND ENERGY MICROBENCHMARKS

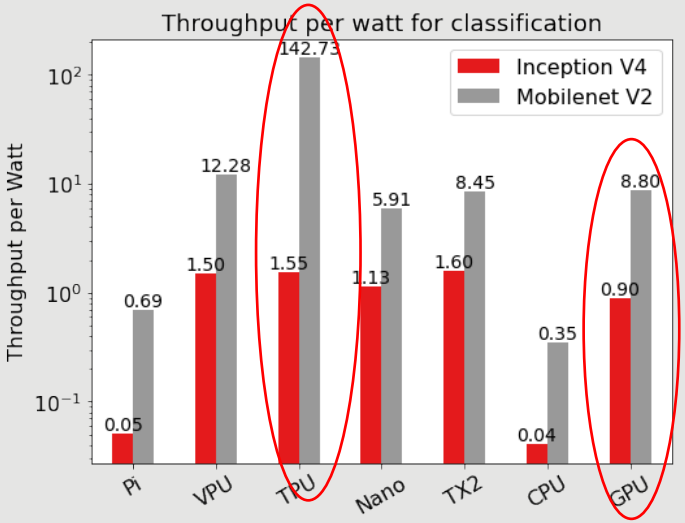


Edge Accelerators can achieve cloud CPU level throughput. Some of then can even outperform cloud CPU

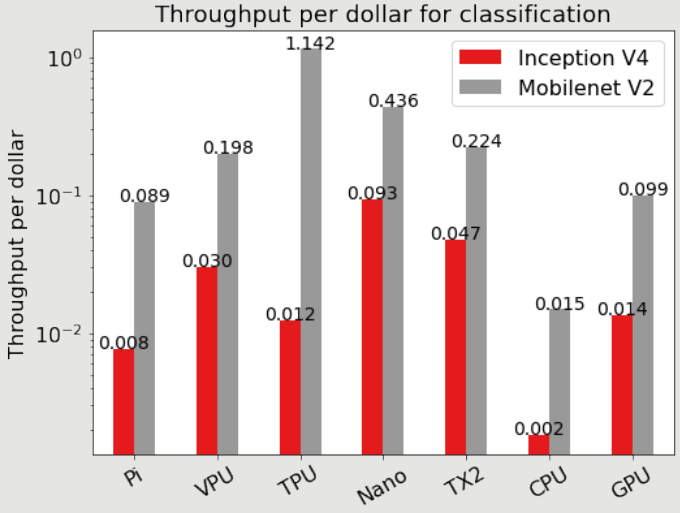


Edge accelerators exhibit very low power consumption compare to cloud CPU and cloud GPU, which consume 131.26W and 111.66W respectively.

PERFORMANCE AND ENERGY MICROBENCHMARKS



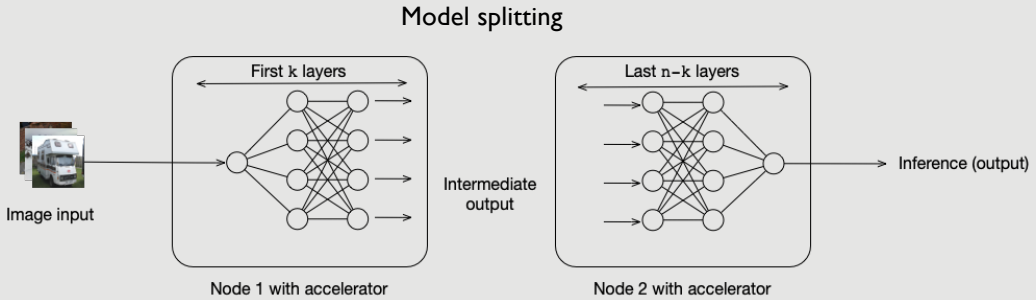
Edge accelerators have lower normalized power consumption than cloud CPU.

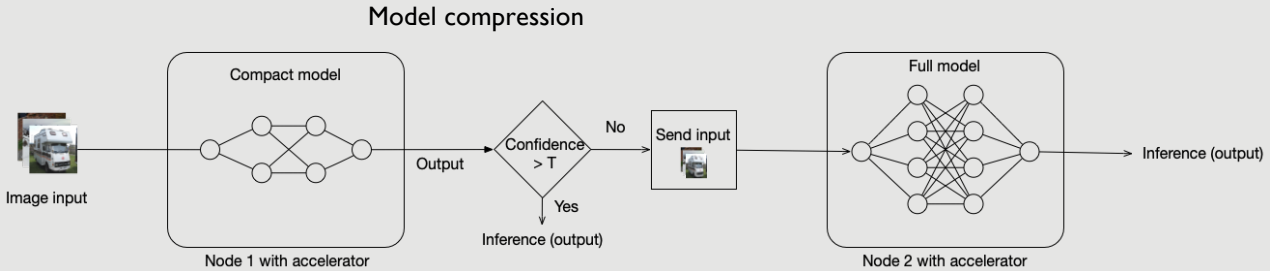


Edge accelerators have 10-100X higher throughput per dollar than cloud CPU and GPU.

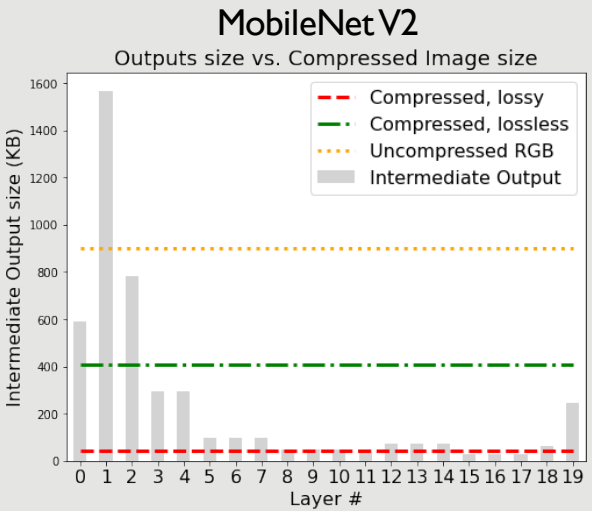
SPLIT PROCESSING ACROSS APPLICATION TIERS

How should IoT applications exploit distributed and split processing capabilities offered at various tiers?

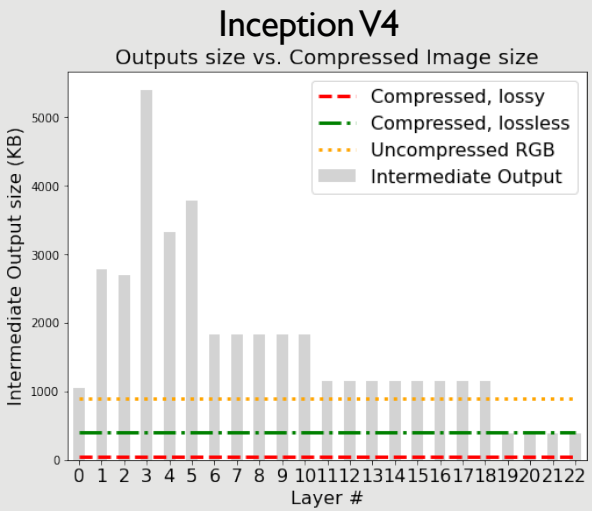




MODEL SPLITTING

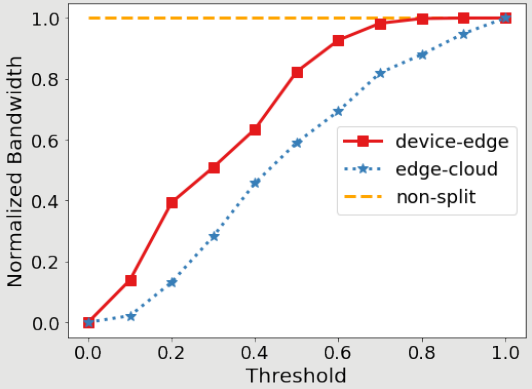


Splitting the model at layer 10 yields nearly 8x network saving over using lossless compression for a not-split model.

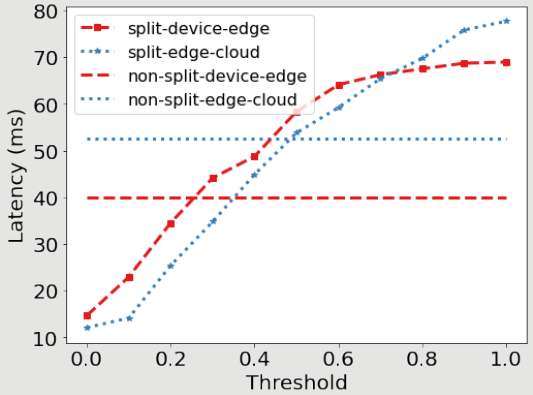


We cannot achieve any network bandwidth saving without splitting at the last 4 layers in this case.

MODEL COMPRESSION



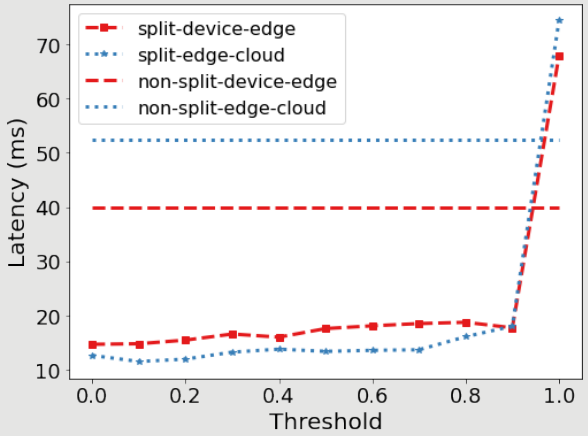
Model compression yield different level of network bandwidth saving depending on the threshold



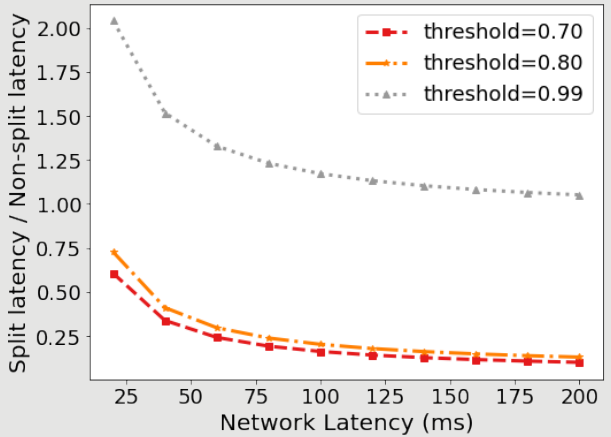
Model compression can also improve inference latency when the threshold is small

MODEL COMPRESSION – SKEWED WORKLOAD

Consider a scenario where the inputs are not random but skewed towards the common case (e.g. surveillance camera). The compressed model is well-trained for frequently occurring inputs.



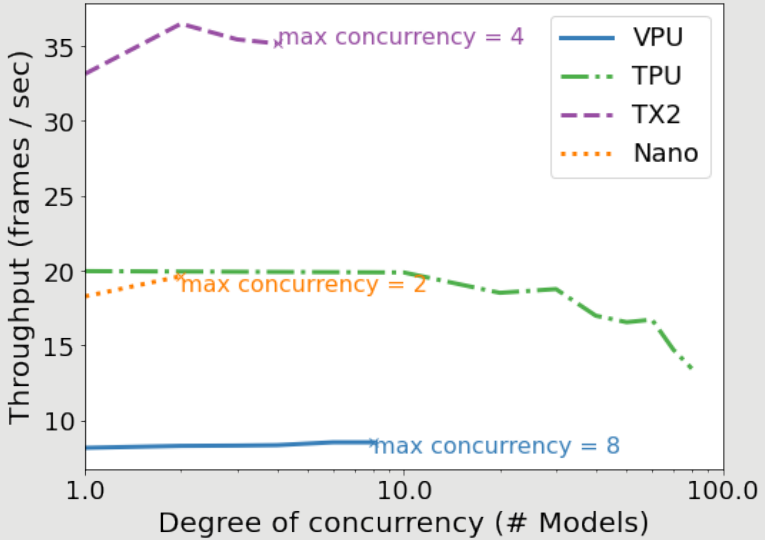
3x – 4x latency reduction



More latency reduction when network latency is high

CONCURRENCY AND MULTI-TENANCY

- For VPN, Nano and TX2, the degree of concurrency is bounded by device memory
- For Nano and TX2, memory are shared between host RAM and GPU. More RAM used by host process, less memory can be allocated by GPU
- For EdgeTPU, the degree of concurrency is unbounded as it automatically performs model swapping on-demand. However, this also result in switch overhead at run time if multiple models are loaded



CONCLUSIONS

1. Edge accelerators show promising performance
   - Higher throughput per watt
   - Higher throughput per dollar
2. Spiting processing paradigm with specialized edge accelerators can achieve considerable benefit
   - Model splitting for bandwidth saving and running large model
   - Model compression for both bandwidth saving and latency deduction
3. The degree of concurrency depends on the device memory, model size, framework software overheads, and system optimizations.

Thank you!!