An Empirical Study of Artificial Intelligence Performance on Edge Devices

Justin Duchatellier, Haley Granger, Yong Shi, Kun Suo

Kennesaw State University

CCERP 2020 – Kennesaw State University

The Ubiquity of Artificial Intelligence

- Artificial intelligence (AI) is the capability of a machine to make its own decisions without explicit commands
- Applications of AI
  - Autonomous vehicles – drones
  - Search engines – Google Search, Microsoft Bing
  - Intelligent virtual assistants – Siri, Cortana, Alexa
- Benefits of AI
  - Facial recognition
  - Smart manufacturing

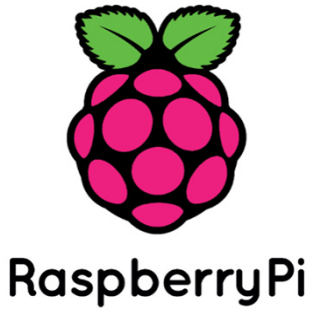
KENNESAW STATE UNIVERSITY

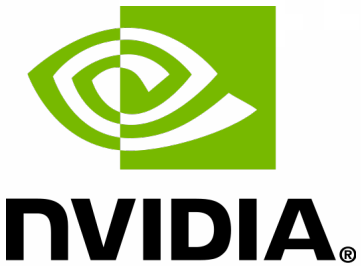
AI in Computing

	Cloud Computing Devices (e.g. AWS p3dn.24xlarge)	Terminal Devices (e.g. Raspberry Pi 4)
Network performance	100 Gbps	1 Gbps
CPU	Up to 96 vCPUs @ 3.1 GHz (turbo)	Quad core Cortex-A72 @ 1.5GHz
GPU	Up to 8 NVIDIA Tesla V100 GPUs (8x 5,120 CUDA Cores and 8x 640 Tensor Cores, double precision performance rated at 7 TFLOPS)	Broadcom VideoCore IV @ 250 MHz
Memory	768 GB, 256 GB (GPU)	8GB LPDDR4-2400 SDRAM
Cost	High $26.928/hour	Low $55 one-time purchase
Real-time processing	Lower (~10s latency AWS Aurora)	Higher (~3600us latency)
Main uses	Training	Inference

KENNESAW STATE UNIVERSITY

AI Applications and Hardware Manufacturers





The Purpose of Our Research

- We aimed to analyze the performance and resource utilization of AI workloads on edge devices
- The global market for edge computing is expected to reach over $16.5 billion by 2025 [1]
- Top three segments of edge computing expected by 2025 [1]
  - Connected cars
  - Smart grids
  - Security and surveillance

[1] https://www.alliedmarketresearch.com/edge-computing-market

KENNESAW STATE UNIVERSITY

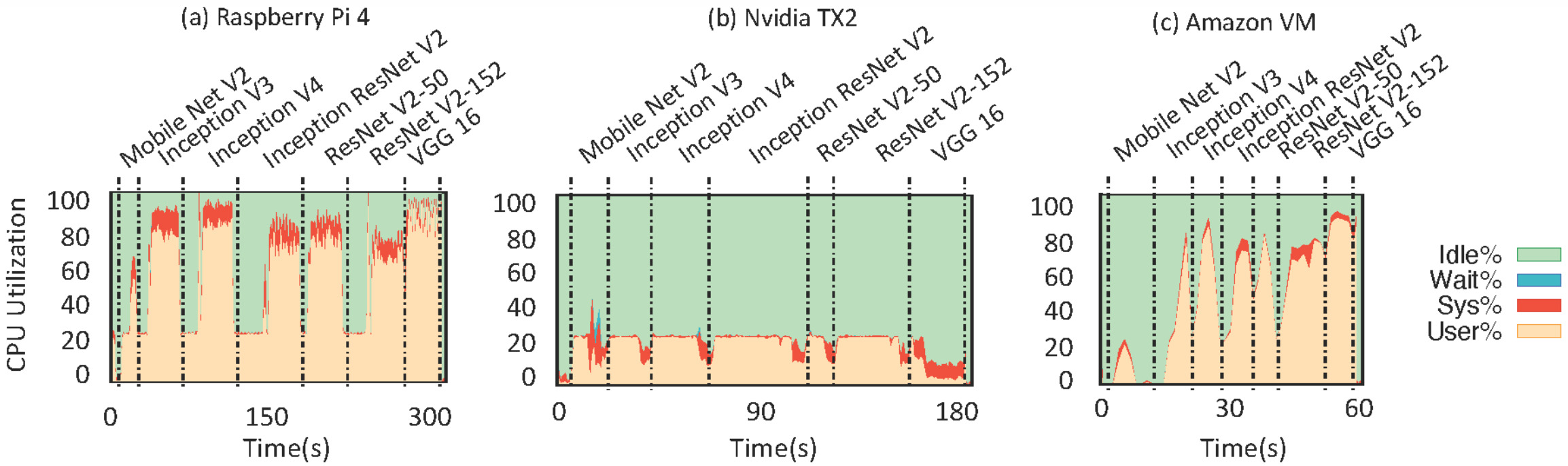
Our Experiment

- Two edge devices and one cloud instance were used in our research
  - Raspberry Pi 4
  - NVIDIA Jetson TX2
  - Amazon EC2 t2.xlarge
- AI Benchmark and MLMark were the benchmarks used to assess the performance of the devices

- AI Workloads that were evaluated
  - MobileNet v2, MobileNet v1
  - SSD-MobileNet v1
  - Inception v3
  - Inception v4
  - Inception-ResNet v2
  - ResNet-50 v2, ResNet-50 v1
  - ResNet-152 v2
  - VGG-16

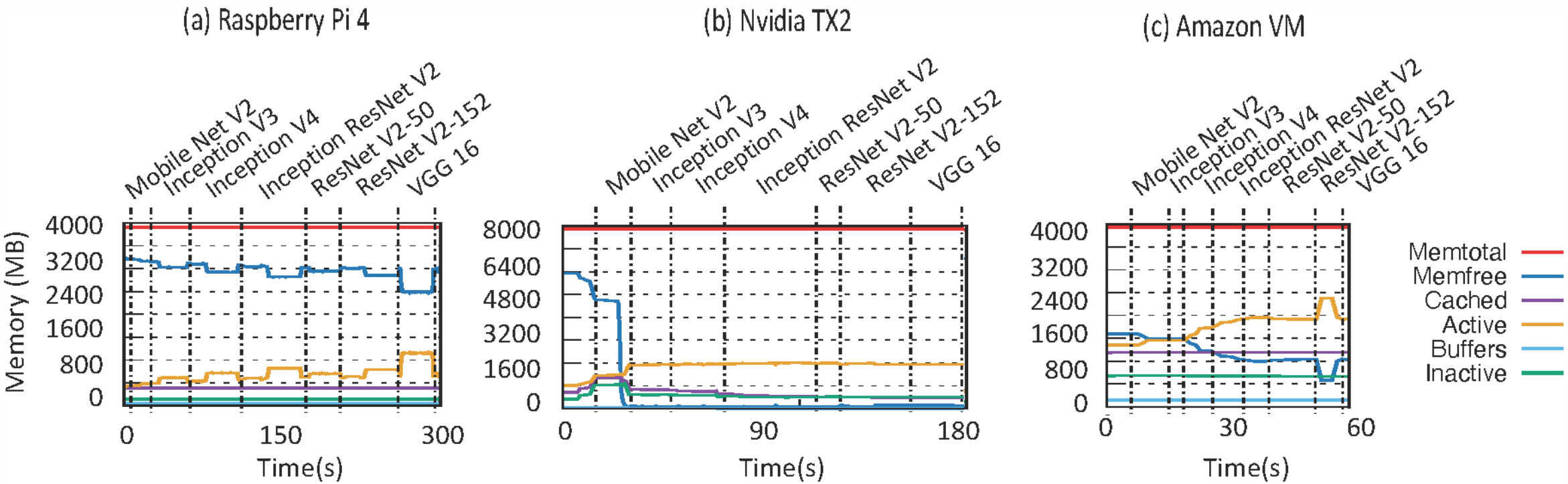
KENNESAW STATE UNIVERSITY

Our Results for CPU Utilization of Object Recognition Applications with AI Benchmark



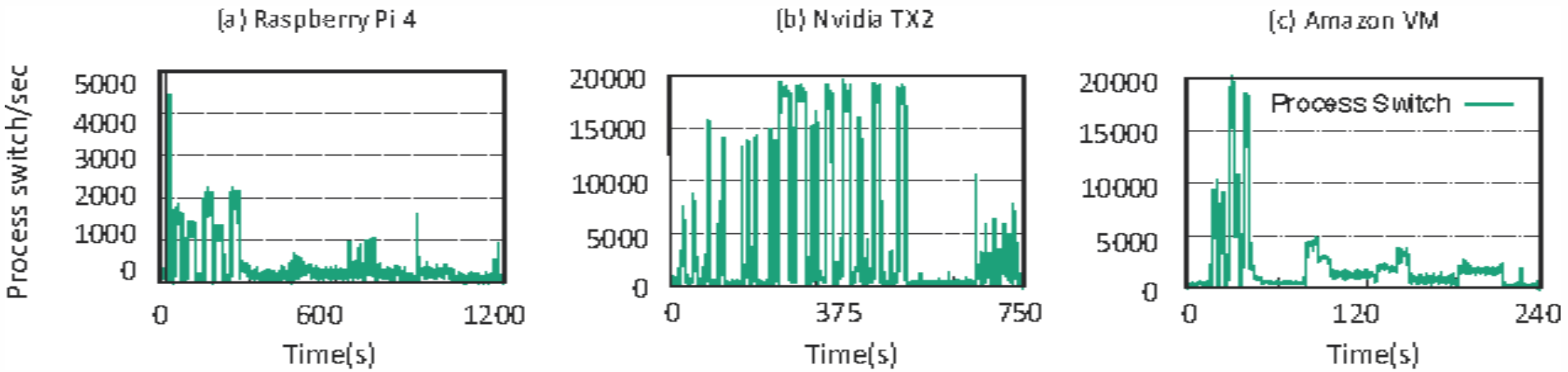
KENNESAW STATE UNIVERSITY

Our Results for Memory Utilization of Object Recognition Applications with AI Benchmark



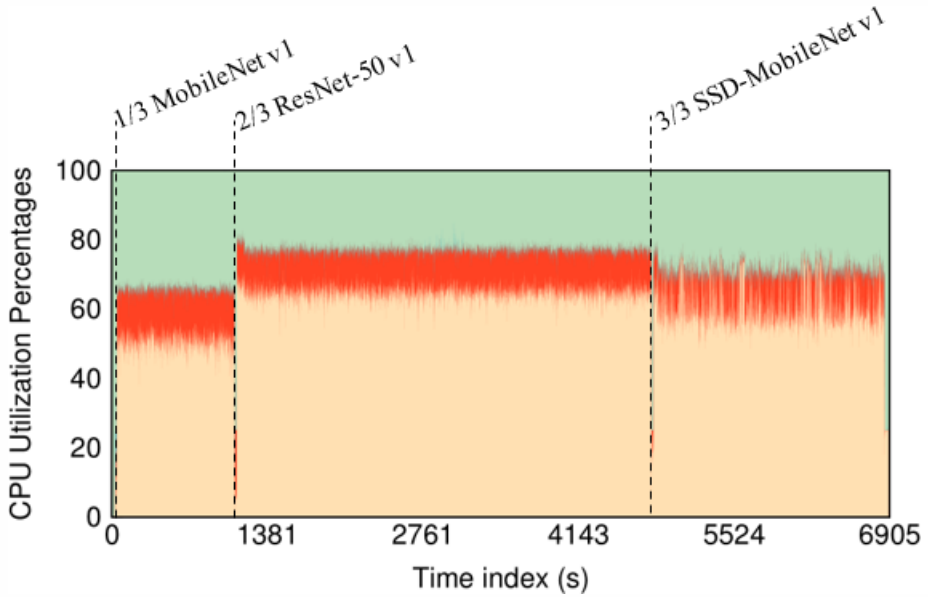
KENNESAW STATE UNIVERSITY

Our Results for Process Switching of Devices with AI Benchmark

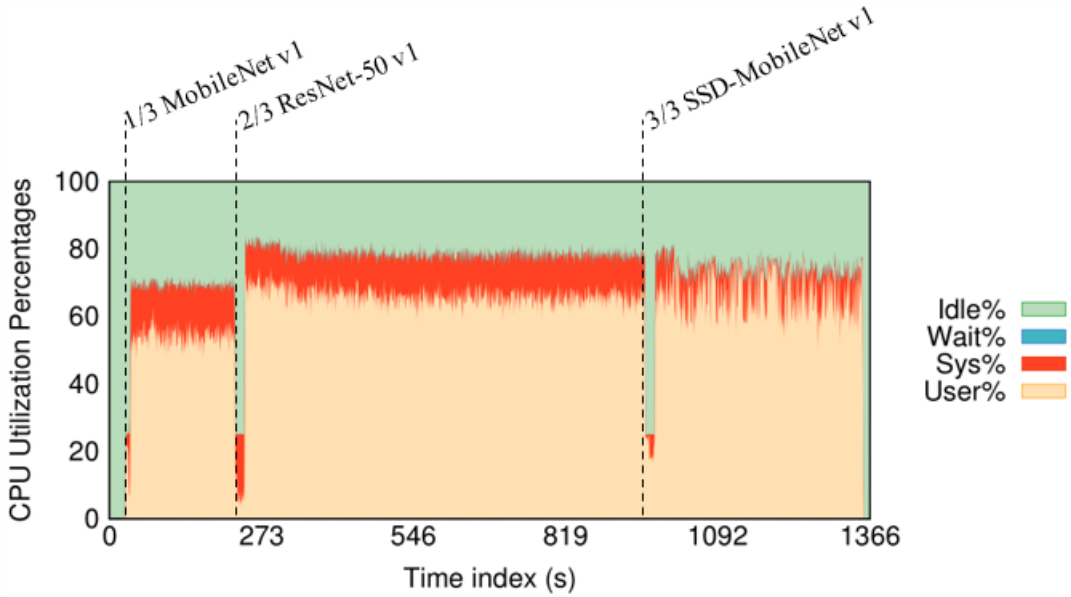


KENNESAW STATE UNIVERSITY

Our Results for CPU Utilization with MLMark on the Raspberry Pi 4



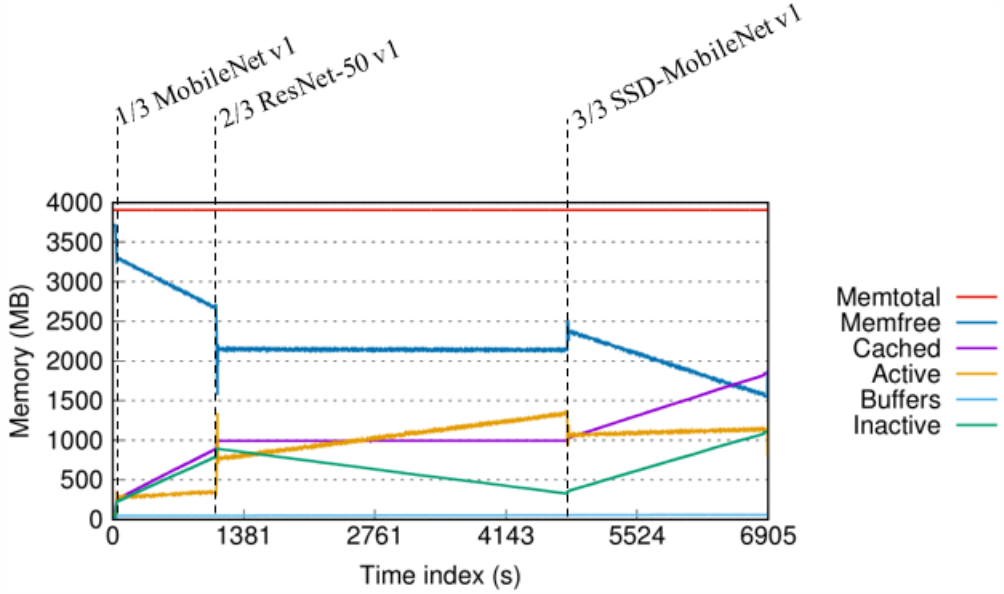
(a) CPU accuracy option



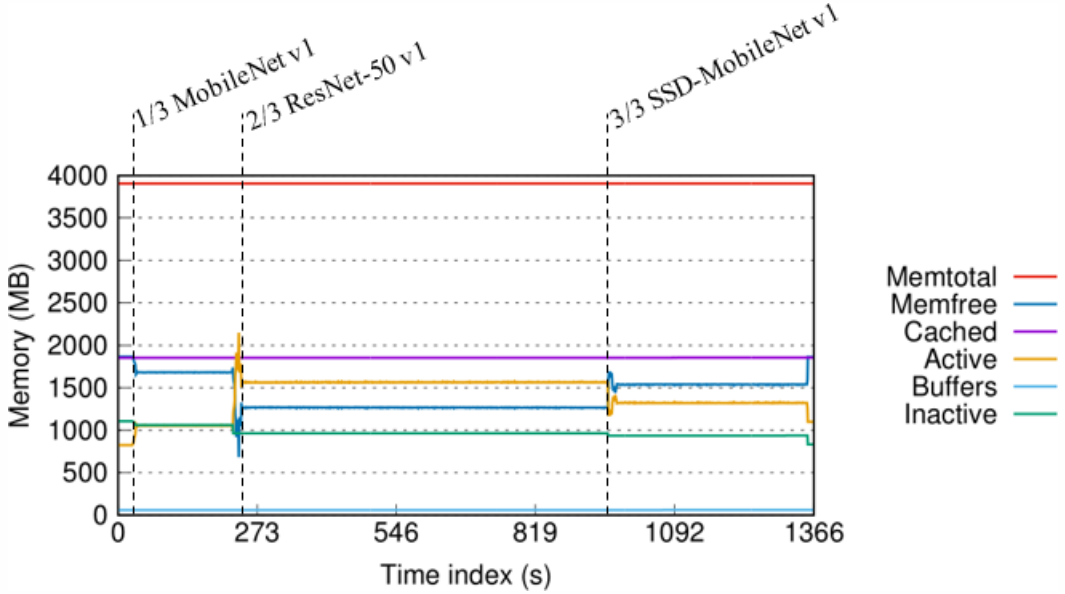
(b) CPU throughput option

KENNESAW STATE UNIVERSITY

Our Results for Memory Utilization with MLMark on the Raspberry Pi 4



(a) Memory accuracy test



(b) Memory throughput test

KENNESAW STATE UNIVERSITY

Our Analysis

- Depending on the system's needs, older AI models may perform better than their predecessors
  - MLMark (MobileNet v1) has a lower CPU utilization than AI Benchmark (MobileNet v2)
  - MobileNet v2 is not supported for some of its layers with certain GPUs
  - Newer models consume more memory
- VGG is not recommended to be deployed on edge devices because of its memory requirements
- Edge devices with more powerful GPUs utilize a smaller CPU percentage than cloud instances (NVIDIA Jetson TX2 vs Amazon EC2 tx2.large)

KENNESAW STATE UNIVERSITY

Security Concerns with Edge Devices

- Default configurations have least restrictive options and minimal security features
  - WEPS/WPA enabled
  - FTP/HTTP enabled
  - Default credentials that are publicly published
- Outdated firmware
- Challenges in managing large-scale systems
- Confusing web user interfaces on networked devices
- Installed backdoors for legal reasons

KENNESAW STATE UNIVERSITY

Thank You!

I would like to answer your questions

KENNESAW STATE UNIVERSITY