Keep Clear of the Edges

An Empirical Study of Artificial Intelligence Workload Performance and Resource Footprint on Edge Devices

Kun Suo, Tu N. Nguyen, Yong Shi,
Jing (Selena) He, and Chih-Cheng Hung

Kennesaw State University

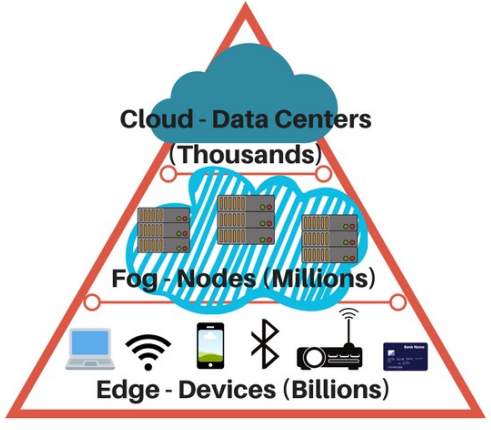
IPCCC 2022

From Cloud to Edge

- **In recent years, computing moved from cloud to the edges**

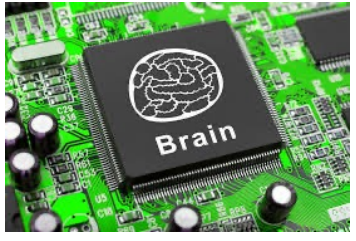
- **Cloud computing**
  - All resources at one place
  - Data computing, transmission, storage at cloud

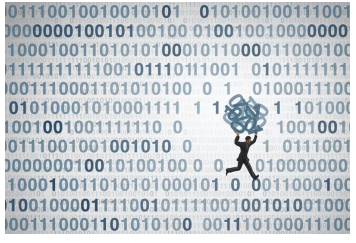
- **Edge computing**
  - Local data is processed nearby
  - Save network bandwidth, reduce the transmission delay, and improve quality-of-service (QoS)



Artificial Intelligence at the Edge

- **AI is being increasingly deployed on edge computing**
  - The rapid innovation on hardware such as various AI dedicated chips enables to deploy artificial intelligence models on modern edges
  - Various new devices are constantly emerging and generating massive data which traditional cloud cannot handle
  - Privacy is another concern for edge computing as many data (i.e., pictures, audios, videos, etc.) contain a lot of personal information





AI Workload v.s. Edge

- AI Model #1: Linear Regression
- AI Model #2: Deep Neural Networks
- AI Model #3: Logistic Regression
- AI Model #4: Decision Trees
- AI Model #5: Linear Discriminant Analysis
- AI Model #6: Naive Bayes
- AI Model #7: Support Vector Machines
- AI Model #8: Learning Vector Quantization
- AI Model #9: K-nearest Neighbors
- AI Model #10: Random Forest

... ...

Raspberry Pi 3B+

BeagleBone Black

NVIDIA Jetson Nano

LattePanda Alpha

Khadas Edge-V

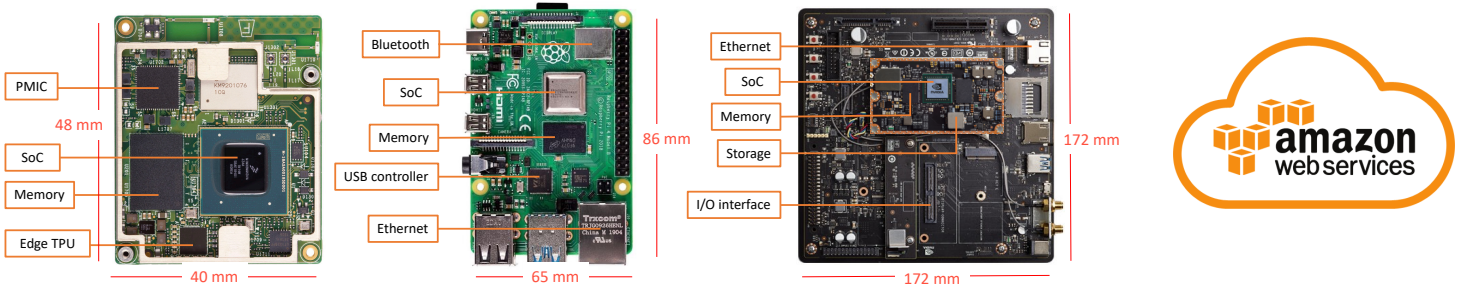
The Purpose of Our Research

- **Why this research?**
  - Lacks research on the resource consumption and performance analysis of different AI models or apps on various edge computing platforms
  - Important for the improvement of edge platforms and the development & deployment of AI workloads and services

- **What we did in this paper**
  - A detailed analysis of representative AI workloads on edges
  - An empirical study of their performance and resource footprints
  - The first to explore these aspects of AI workloads on edges

Our Experiment

- **Two edge devices and one cloud instance were used in our research**
  - Raspberry Pi 4B with quad-core CPU and 4GB of RAM
  - Nvidia Jetson TX2, a quad-core A57 processor, a dual-core Denver processor, a 56-core GPU, and 8GB of RAM
  - VM (quad-core vCPU and 16GB of memory) and PM (24-core Intel Xeon E5-2670 v3 CPU and 504 GB memory)



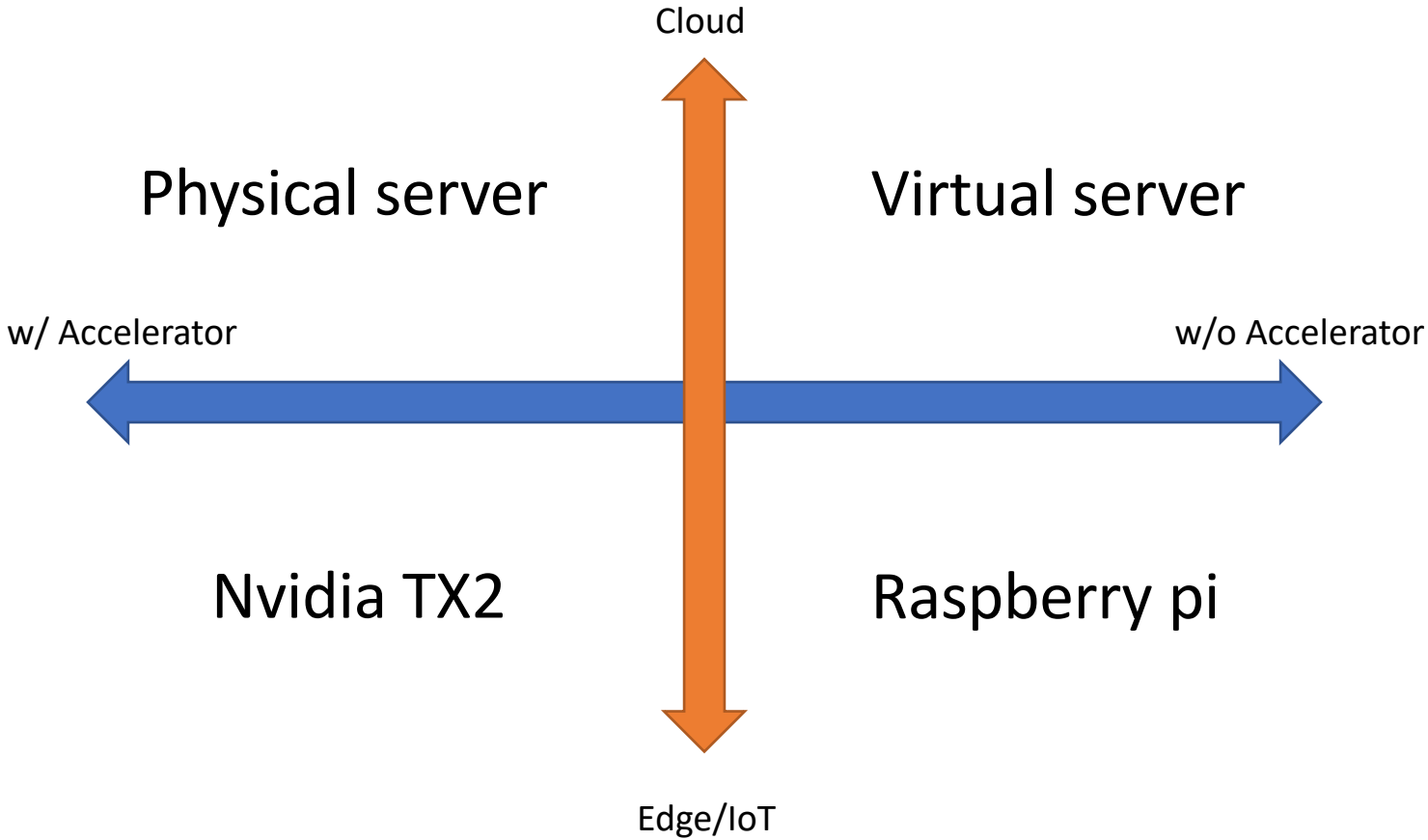
Processor with Accelerator

Specifications of AI workloads

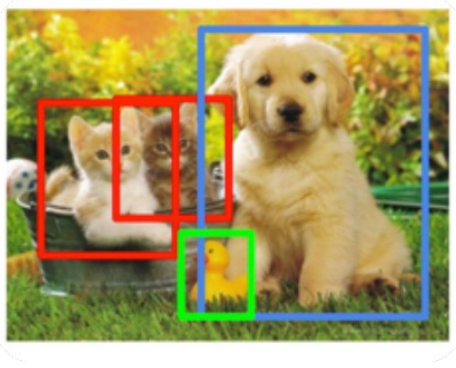
- **Software**
  - TensorFlow 1.14
  - AI Benchmark: 42 tests and 16 sections

Task	Neural Network	Description	Resolution(px)
Object Recognition	MobileNet v2 [31]	A depthwise separable convolution network with linear bottlenecks between the layers and linear bottlenecks between the layers	224x224
Classification	Inception v3 [39]	A widely-used image recognition model that has been shown to attain greater than 78.1% accuracy on the ImageNet dataset	346x346
Classification	Inception v4 [37]	Built on Inception v3 but with specialized grid reduction	346x346
Facial Recognition	Inception-ResNet v2 [14]	A variation of Inception V3 with residual networks	346x346
Classification	ResNet-50 v2 [14]	A 50 layers deep convolutional neural network, consists of 5 stages each with a convolution and Identity block. Each convolution block has 3 convolution layers and each identity block also has 3 convolution layers	346x346
Classification	ResNet-152 v2 [14]	Similar as ResNet-50 v2 but contains 152 layers	256x256
Classification	VGG-16 [34]	A CNN with 16 layers including 13 convolutional layers and 3 dense layers	224x224
Super-Resolution	VGG-19 [21]	Similar as VGG-16 but with 19 layers including 16 convolutional layers and 3 dense layers	512x512
Super-Resolution	ResNet-SRGAN [22]	Super-resolution Using a Generative Adversarial Network (GAN). Residual block designed using ResNet	512x512
Image Deblurring	SRCNN 9-5-5 [11]	A deep learning method for single image super-resolution	512x512
Image Enhancement	ResNet-DPED [17]	A residual CNN improving color resolution and image sharpness	256x256
Bokeh Simulation	U-Net [30]	A convolutional network for fast and precise segmentation of biomedical images	512x512
Semantic Image Synthesis	Nvidia-SPADE [28]	A semantic image synthesis system with spatially-adaptive normalization	128x128
Image Segmentation	ICNet [43]	A network for real-time semantic segmentation on high-resolution images	1024x1024
Image Segmentation	PSPNet [44]	Pyramid scene parsing network with pixel level prediction	720x720
Image Segmentation	DeepLab v1 [9]	A deep convolutional network for semantic image segmentation	512x512

Results of AI workloads comparison Evaluation



Results of AI workloads comparison Evaluation

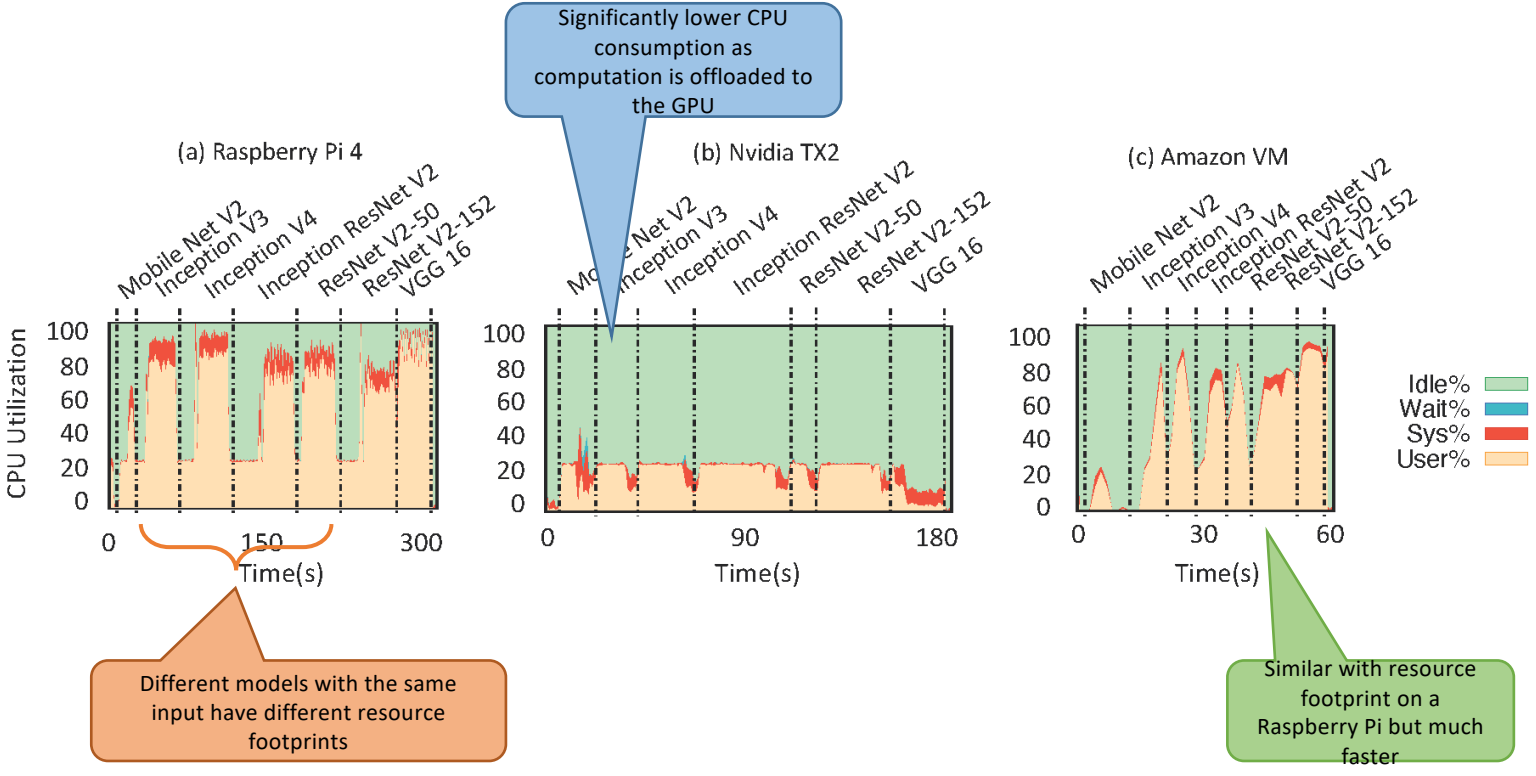


object recognition
and classification

image-to-image

segmentation

CPU Utilization of Object Recognition Applications

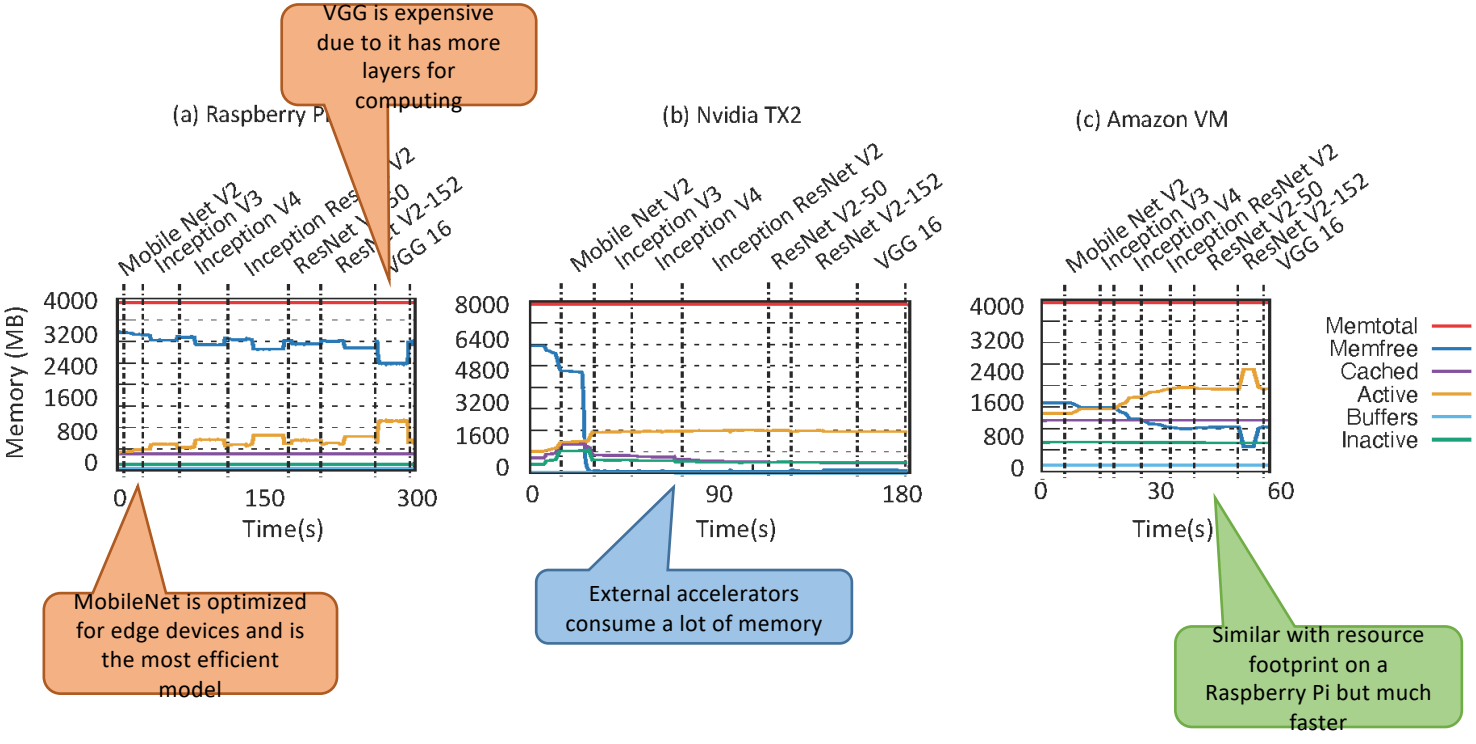


Edge
w/o Accelerator

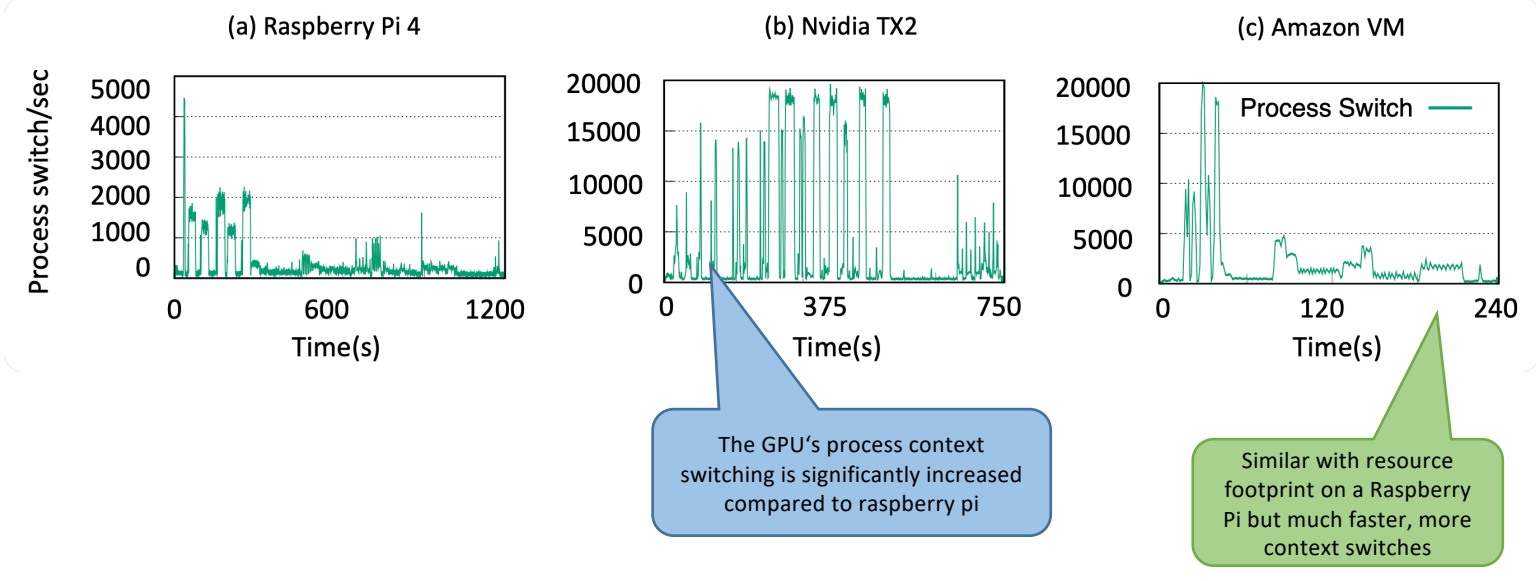
Edge
w/ Accelerator

Cloud
server

Memory Utilization of Object Recognition Applications



Process Switching Results

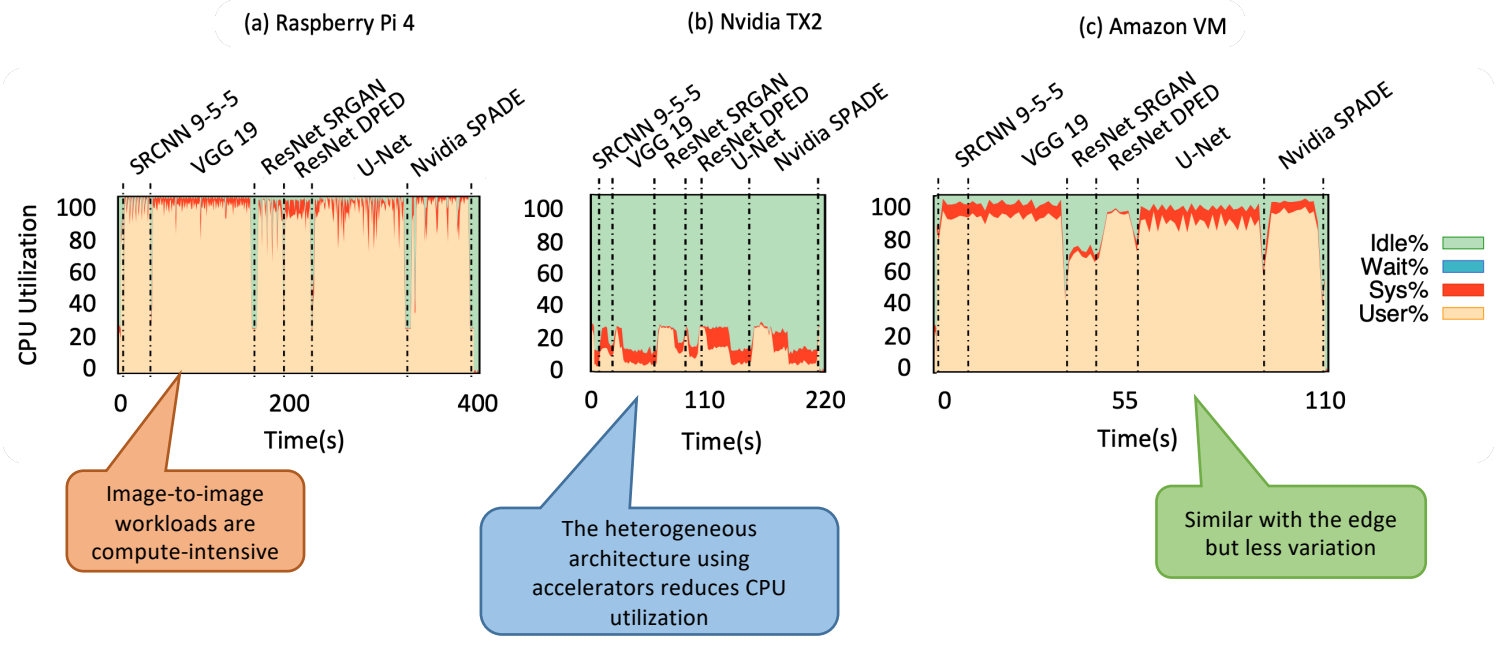


Edge
w/o Accelerator

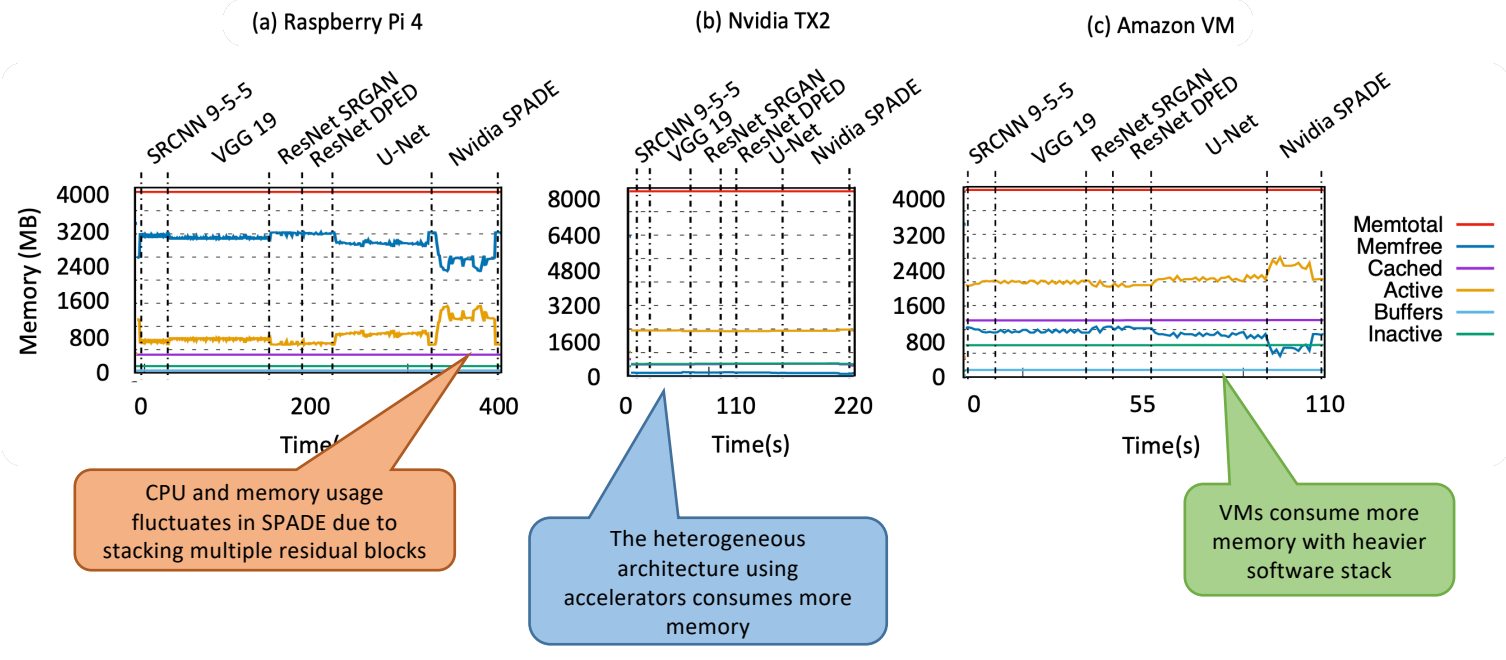
Edge
w/ Accelerator

Cloud
server

CPU Utilization of Image-to-image Applications



Memory Utilization of Image-to-image Applications

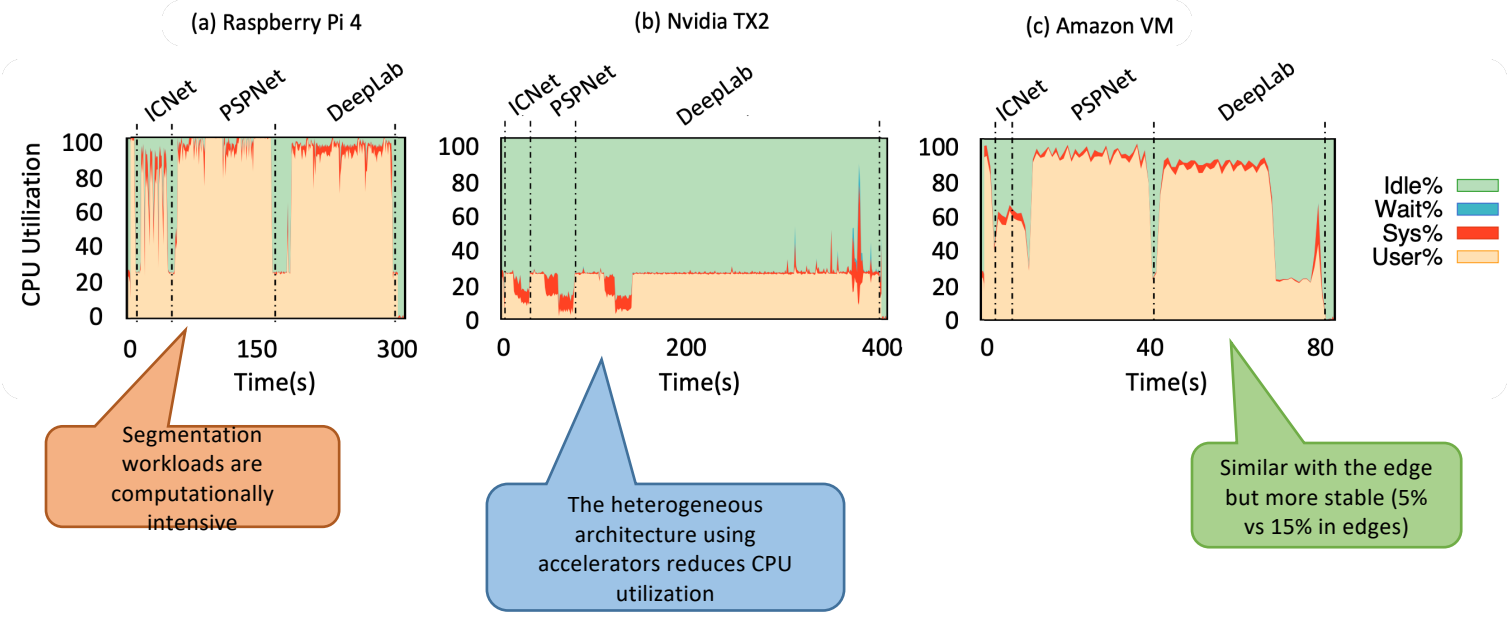


Edge
w/o Accelerator

Edge
w/ Accelerator

Cloud
server

CPU Utilization of Segmentation Applications

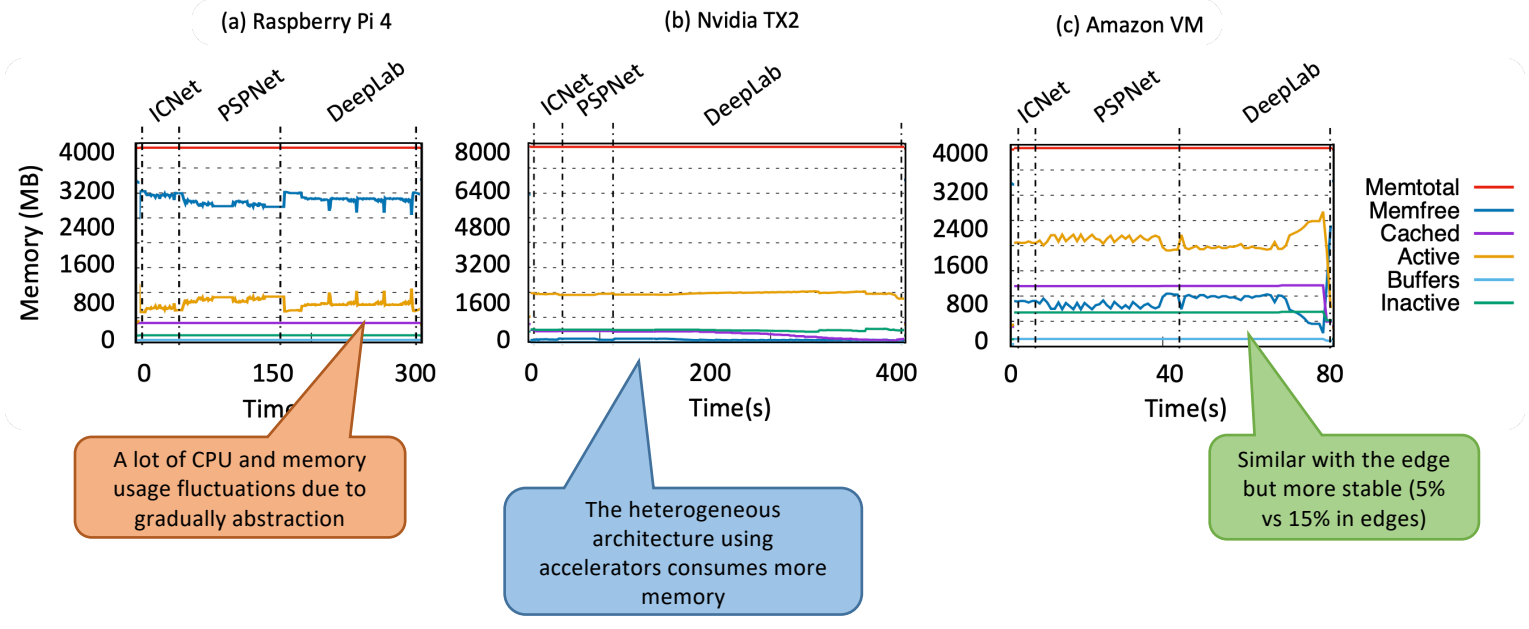


Edge
w/o Accelerator

Edge
w/ Accelerator

Cloud
server

Memory Utilization of Segmentation Applications



Edge
w/o Accelerator

Edge
w/ Accelerator

Cloud
server

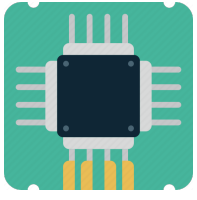
Conclusion

- We perform a comprehensive empirical study and quantitatively evaluate the representative AI applications with different neural network models
- We compare the performance discrepancy on different types of edge hardware as well as cloud platforms
- Our analysis and findings could help the users deploy the appropriate models and workloads on their scenarios, and guide the optimization of AI workloads on modern edge environments

Our Findings & Insights

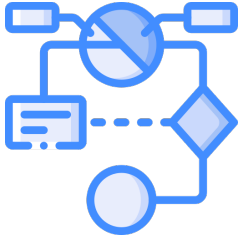
- General and heterogeneous edge architecture with accelerators present a significant performance difference of AI workloads. Good performance can be attained by offloading computation to accelerators while it might sacrifice with massive memory consumption.

- For the same type of workload, cloud and general edge perform similarly and have analogous resource footprints. For AI applications on edges, local hardware such as SoC and memory determine its performance while the cloud is faster and more stable.



Our Findings & Insights

- Same types of workloads perform comparably but algorithm models such as neural networks could impose certain overhead on resources. This is due to complex interactions between the AI workloads and the software stacks in edge devices



- AI processing in edge devices is also concerned with other factors such as response time, model volume and cost. Overall, only a few AI workloads and related models are suitable to execute or deploy on edges

Thank you !

Questions?