KENNESAW STATE
UNIVERSITY

An Empirical Analysis and Resource Footprint Study of Deploying Large Language Models on Edge Devices

Nobel Dhar*, Bobin Deng* , Dan Lo*,
Xiaofeng Wu#, Liang Zhao*,
Kun Suo*
*: Kennesaw State University
#: City University of Macau

Outline

- Introduction
- Methodology
- Device Configuration
- Inference Demo
- Observations & Analysis
- Future Work

Introduction

Edge Computing

- Computation is performed at or near the data source
- It's faster, more efficient, and offers real-time analytics.
- It comes with limited computational resources (RAM, CPU Speed etc.) available

Large Language Models

- Large Language Models like GPT, LLaMA and BARD have become pivotal.
- Demands significant computational resources
- Often characterized by their substantial size

Why deploying LLMs on Edge?

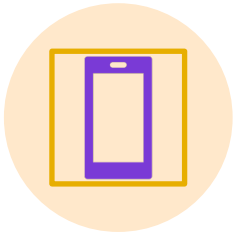
- Latency Reduction
- Enhanced Privacy and Security
- Offline Execution

Research Objectives

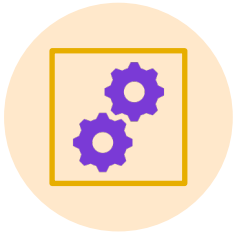
- Observe Performance
- Identify Bottlenecks
- Explain Bottlenecks

Methodology

Model Selection and Quantization



Edge Device Selection



Deployment and Text Generation

Data Collection

Table 1: Basic Specifications of Evaluated Edge Devices

Device Name	Memory	CPU Freq.	CPU #	Disk Size
Raspberry Pi 4B	1GB	1.8GHz	4	32GB
Raspberry Pi 4B	2GB	1.8GHz	4	32GB
Raspberry Pi 4B	8GB	1.8GHz	4	32GB
Jetson AGX Orin	32GB	2.2GHz	12	64GB

Table 2: Memory Bandwidth of Evaluated Edge Devices

Devices	Bandwidth (GB/s)
Raspberry Pi 4B 1GB	12.8
Raspberry Pi 4B 2GB	12.8
Raspberry Pi 4B 8GB	12.8
Jetson AGX Orin	204.8

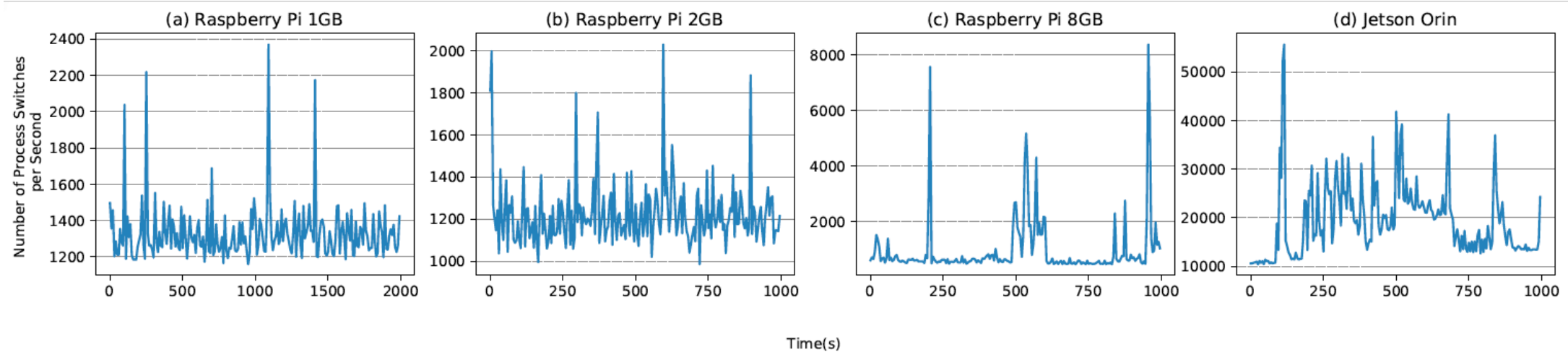
Edge Devices

- Nvidia Jetson AGX Orin
- Raspberry Pi 4B

Observations (Latency)

Devices	Performance (Tokens/Second)
Raspberry Pi 4B 1GB	0.01
Raspberry Pi 4B 2GB	0.01
Raspberry Pi 4B 8GB	0.11
Jetson AGX Orin	4.49

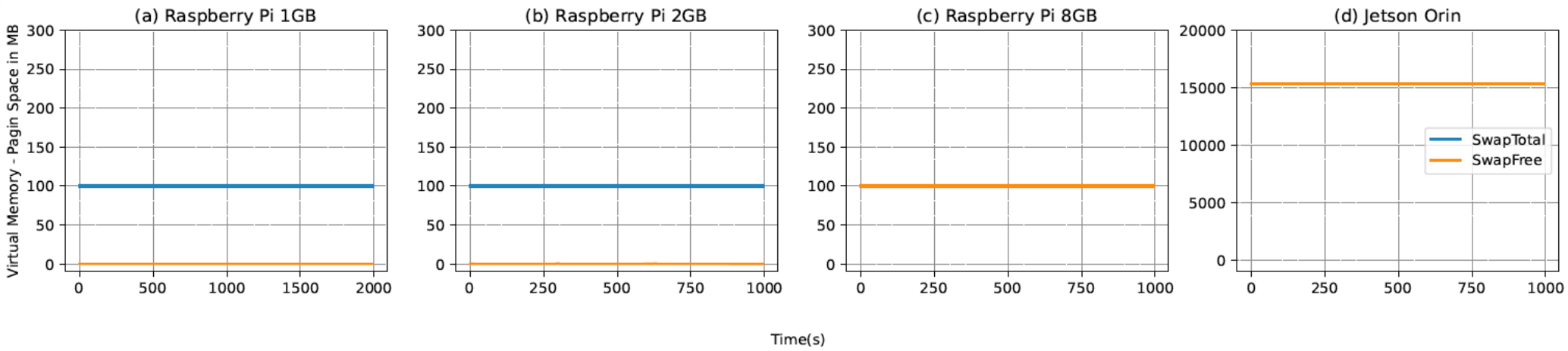
All the upcoming observations has contributed to this.



Observations (Process Switches)

- Tied to the concept of page faults
- CPU time is reassigned to processes with the necessary model portions residing in the RAM

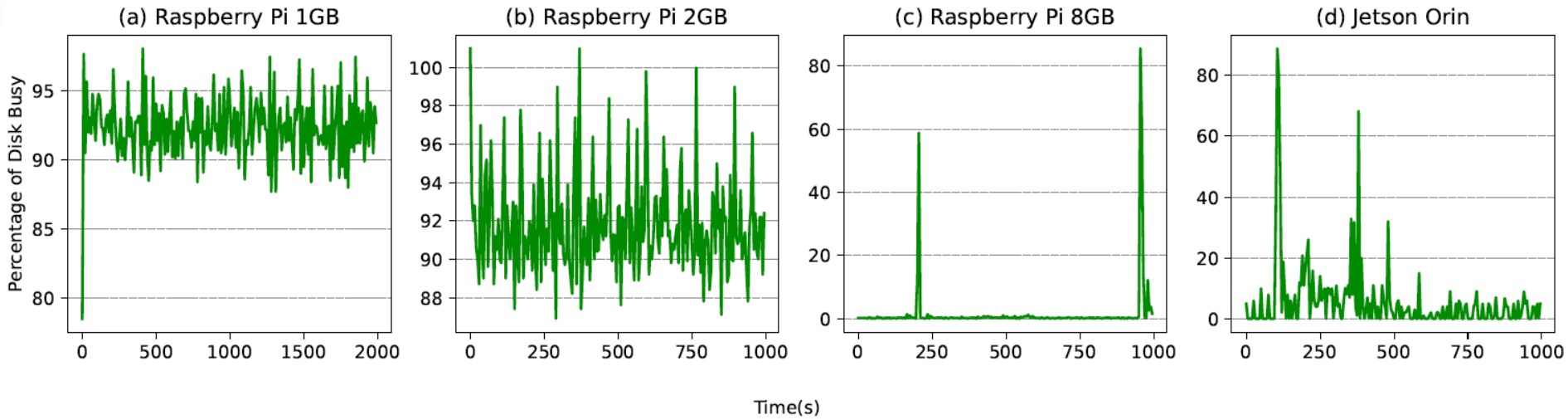
- **INSIGHT:** RAM size should be bigger than the model or model size should be reduced



Observations (Swaps)

- When the model cannot be accommodated within the device's RAM
- The System falls back on swapping different parts of the model
- Fits easily into the spacious 8GB and 32GB of RAM

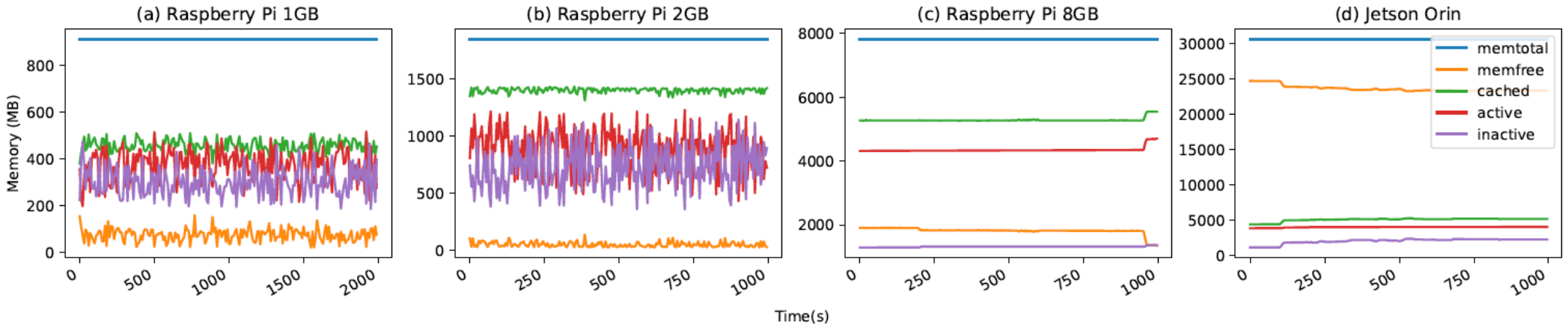
- **INSIGHT 1:** Enlarging the swap space may offer marginal improvements
- **INSIGHT 2:** It remains imperative to emphasize the need for augmenting the hardware memory size



Observations (Disk Activity)

- The system shuffles the parts of the model between RAM and disks when the RAM is limited.
- In the case of 1GB and 2GB configurations, the system does it more frequently

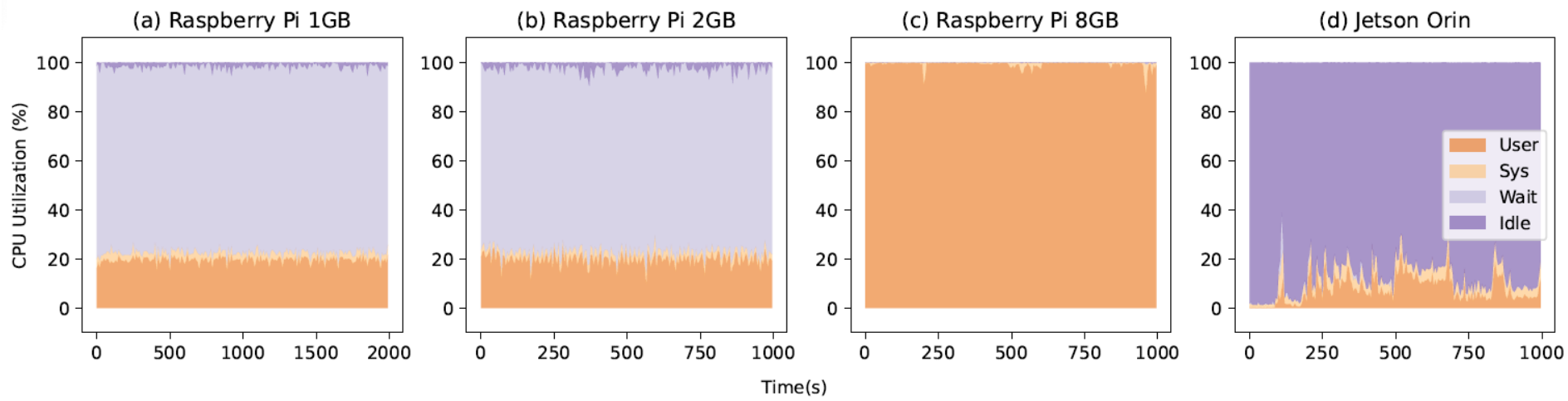
- **INSIGHT 1:** Faster Disks will improve the overall latency.
- **INSIGHT 2:** Increasing edge devices' memory size is the primary strategy to lower the disk busy rates



Observations (Memory Dynamics)

- Negligible Free Space in some devices
- In-and-out movement of model segments
- Frequent cache operations

- **INSIGHT 1:** Not only memory size but also bandwidth should be increased.
- **INSIGHT 2:** Memory efficiency of the model should be increase.



Observations (CPU Utilization)

- Lower overhead if smaller number of process switches and swaps
- Swapping with a slower disks
- CPU dedicates more time to managing these complexities, impacting individual process CPU times

- **INSIGHT 1:** If CPU time can be increased for Model by reducing waiting time, latency will be less.
- **INSIGHT 3:** CPU speed should be increased
- **INSIGHT 2:** Model parallelism of LLM to enhance the performance further

Future Work

- Activation Sparsity Utilization
- Better quantization
- DNN Pruning
- Model Parallelism
- Energy efficiency

Thank You!

Any Questions?