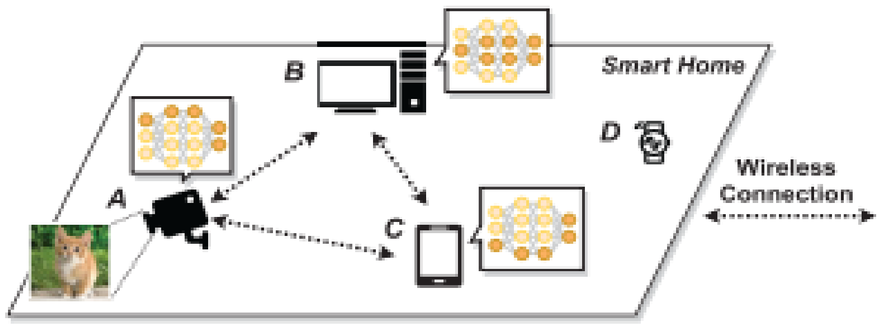
Optimizing Distributed DNN Inference for Low Latency and Energy Efficiency on Edge Devices

Adiba Masud, Maryam Tabar
The University of Texas at San Antonio

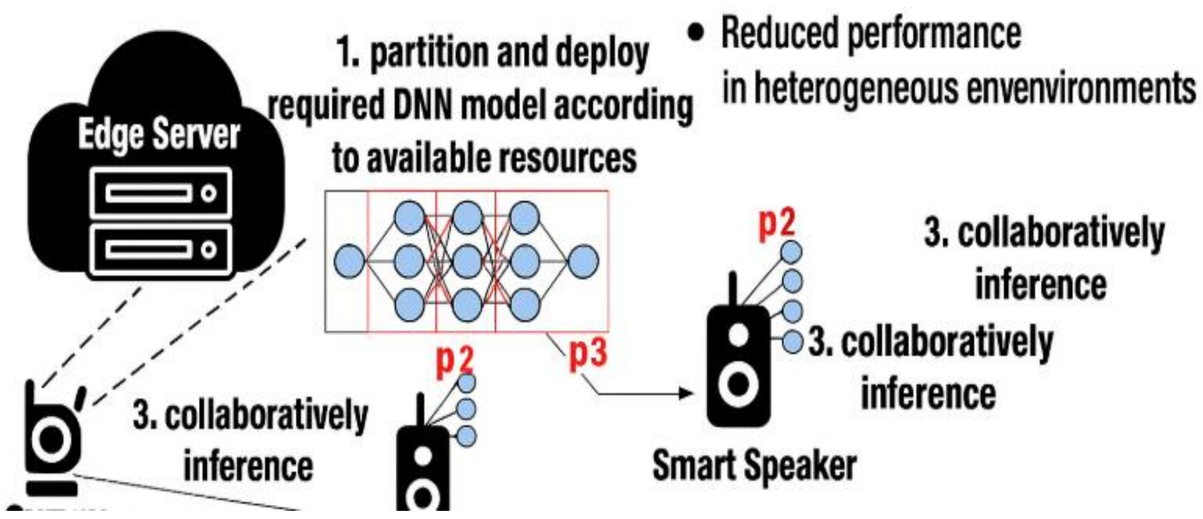
Introduction

- The rise of **deep learning (DL)** and **IoT technologies** has led to a rapid increase in **AI-driven edge applications** (e.g., smart surveillance, autonomous vehicles, healthcare wearables).
- These applications depend on **deep neural networks (DNNs)** to provide **real-time, intelligent decision-making** at the edge.
- Performing DNN inference on **resource-constrained devices** like **Raspberry Pi** and **NVIDIA Jetson Nano** is challenging due to:
  - Limited computing power
  - Fluctuating network conditions
  - Strict latency and energy constraints
- Traditional **static partitioning** divides DNNs across devices without adapting to runtime changes, causing:
  - Poor adaptability to dynamic workloads
  - Latency increases
- These issues **motivate the need for adaptive systems** that:
  - Dynamically reallocate DNN layers
  - Improve **latency**, **energy efficiency**, and **scalability**



Research Goals:

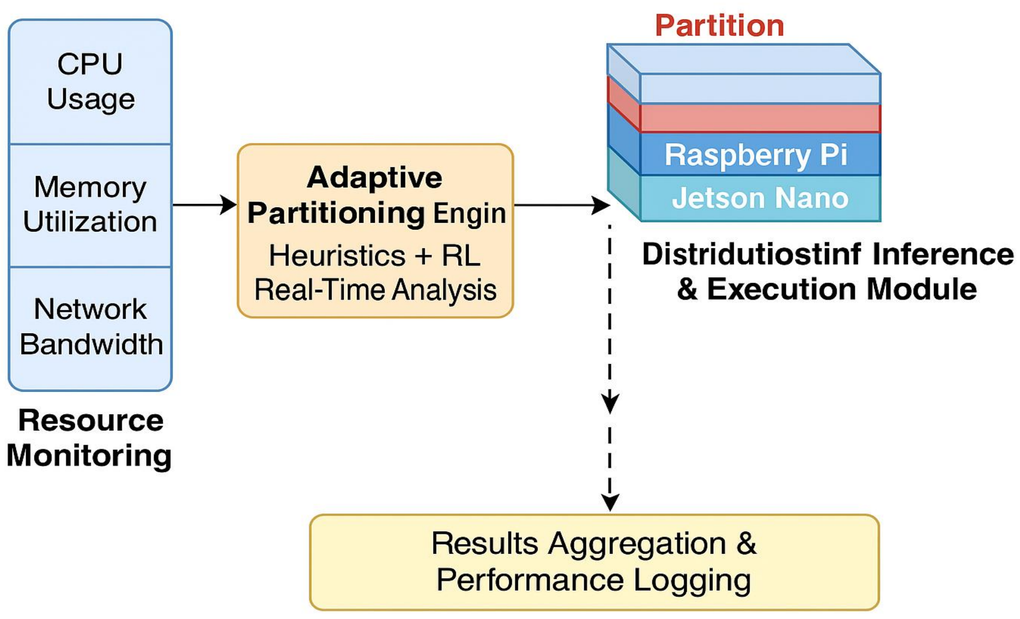
- Develop a **real-time adaptive DNN partitioning framework** for edge computing environments.
- Enable **dynamic redistribution** of DNN layers across heterogeneous edge devices based on live resource metrics.
- Improve **inference latency**, **energy efficiency**, and **resource utilization** in dynamic and constrained environments.
- Integrate **fine-grained monitoring tools** to track CPU, memory, and network usage on edge devices.
- Leverage **heuristic and reinforcement learning-based algorithms** for intelligent task scheduling.
- Validate the framework using **heterogeneous testbeds** (e.g., Raspberry Pi and Jetson Nano).
- Facilitate **scalable, real-time AI applications** in edge environments such as healthcare, autonomous systems, and smart cities.



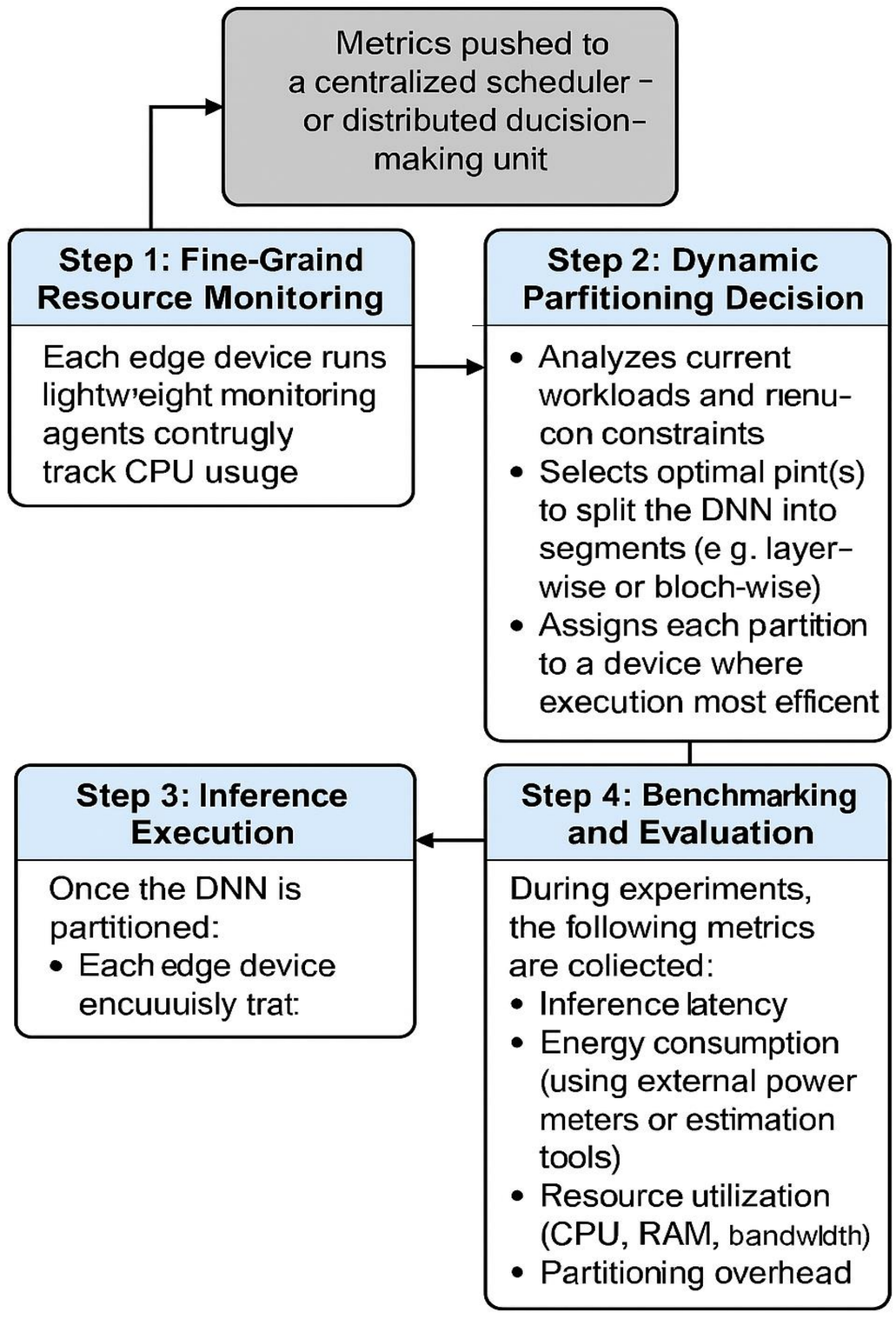
Proposed Methodology

The proposed methodology introduces an adaptive framework for distributed DNN inference on heterogeneous edge devices. The system aims to dynamically allocate computational tasks (DNN layers or blocks) across devices like Raspberry Pi and NVIDIA Jetson Nano based on real-time resource availability, such as CPU utilization, memory usage, and network bandwidth.

System Architecture



- **Workflow**



Expected Contribution

•Develop a **real-time adaptive DNN partitioning framework** for edge computing environments.
•Enable **dynamic redistribution** of DNN layers across heterogeneous edge devices based on live resource metrics.
•Improve **inference latency**, **energy efficiency**, and **resource utilization** in dynamic and constrained environments.
•Integrate **fine-grained monitoring tools** to track CPU, memory, and network usage on edge devices.
•Leverage **heuristic and reinforcement learning-based algorithms** for intelligent task scheduling.
•Facilitate **scalable, real-time AI applications** in edge environments such as healthcare, autonomous systems, and smart cities
•Contributes to Distributed Deep Learning Research.

Conclusions

- Proposed a **real-time adaptive DNN partitioning framework** for efficient edge inference.
- Addresses limitations of **static partitioning** in dynamic edge environments.
- Utilizes **resource monitoring** and **adaptive layer redistribution** based on system conditions.
- Validated on **Raspberry Pi and Jetson Nano**, showing improved performance and adaptability.
- Enhances **latency, energy efficiency**, and **system responsiveness**.

Broader Impact

- Energy-Efficient Edge AI
- Enhanced Real-Time Responsiveness
- Democratization of AI Technology
- Educational and Research Advancement
- Alignment with NSF Broader Impacts

Ongoing and Future Work

- **Support for Complex Models:** Extend the framework to handle deeper architectures like Transformers and EfficientNet.
- **Scalability:** Test the system on larger, more diverse edge networks.
- **Edge–Cloud Integration:** Enable dynamic task offloading between edge and cloud.
- **Energy Optimization:** Integrate precise energy profiling tools for fine-tuned efficiency.

Acknowledgement: This material is based upon work supported by the U.S Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR), Reaching a New Energy Sciences Workforce (RENEW) under Award Number DE-SC0024322

UTSA