Real-World Federated Learning on Edge Devices with NVFLARE

George Igwegbe, Martins Awojide, Adetola Adetunji

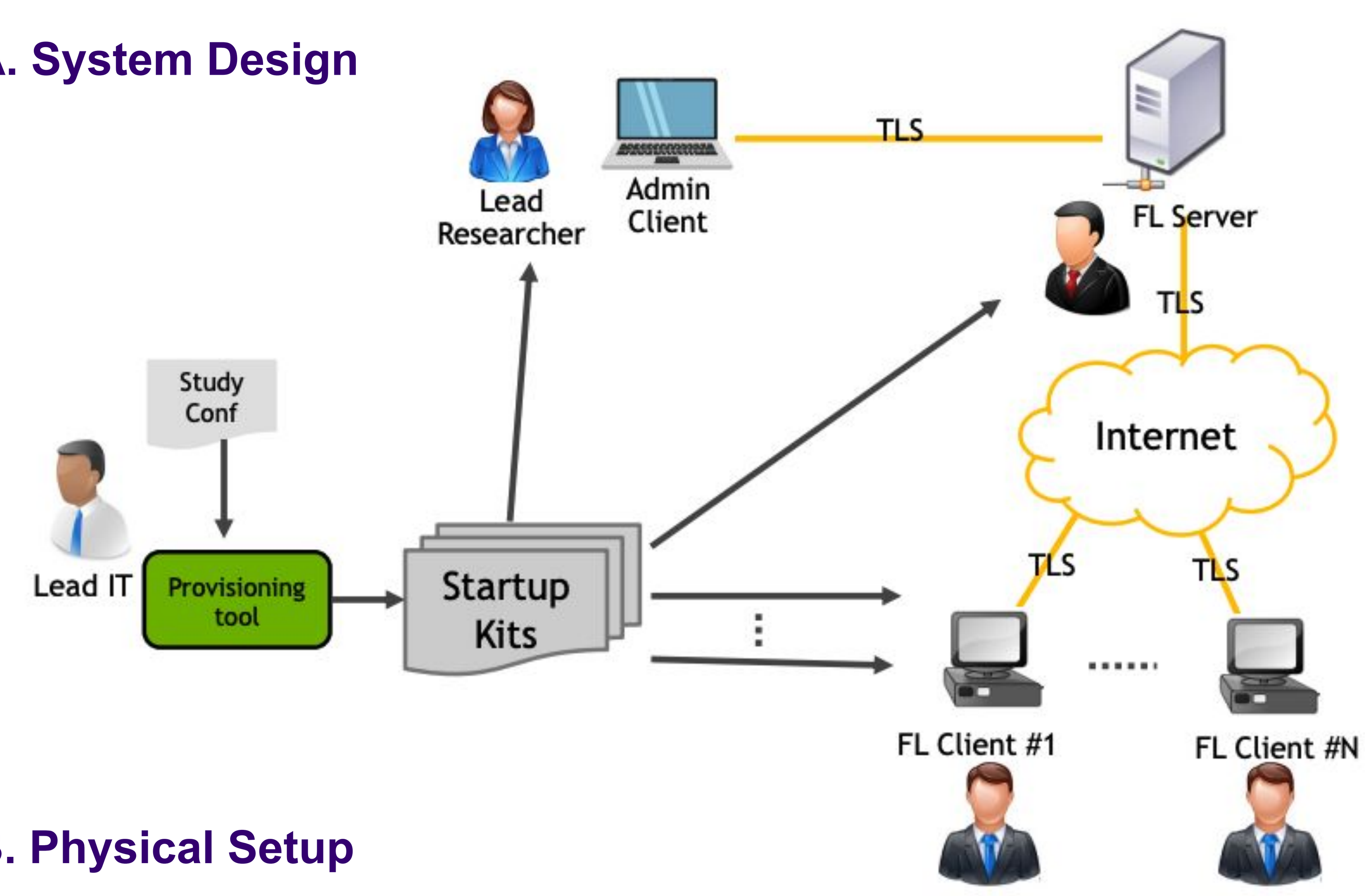
INTRODUCTION

Federated Learning (FL) has emerged as a paramount approach for advancing artificial intelligence beyond centralized models, particularly due to its ability to harness individuals' data while maintaining privacy. The largest hurdle toward evaluating FL algorithms is the difficulty of conducting real-world experiments. Most FL research results have been obtained through simulation, typically running on a single server, with few studies utilizing actual experimental testbeds. However, simulations often use simplified system models and cannot reliably assess certain performance aspects, such as individual device energy consumption or real-world behavior under various constraints like data, network, and device heterogeneity.

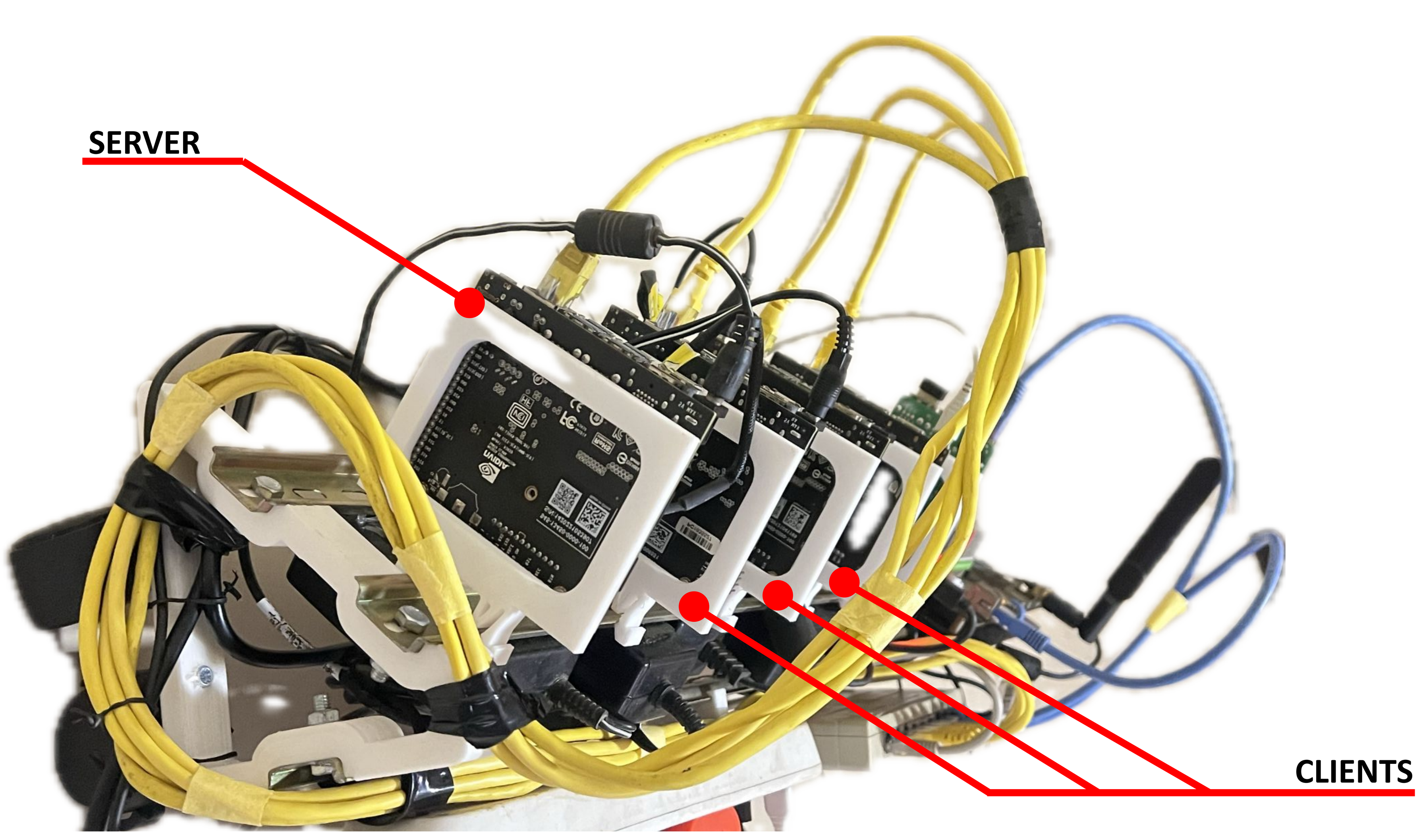
The limitations of simulation underscore the critical need for real-world testbeds that can account for the complexities of real-world FL environments. To address these challenges, our experiment will use NVFlare and our physical setup with edge devices to uncover previously unknown trade-offs and inefficiencies to provide a more holistic understanding of FL system performance.

APPROACH

A. System Design



B. Physical Setup

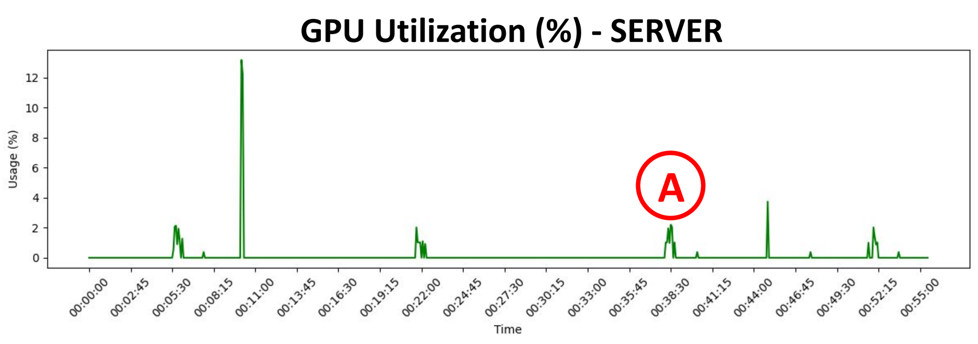


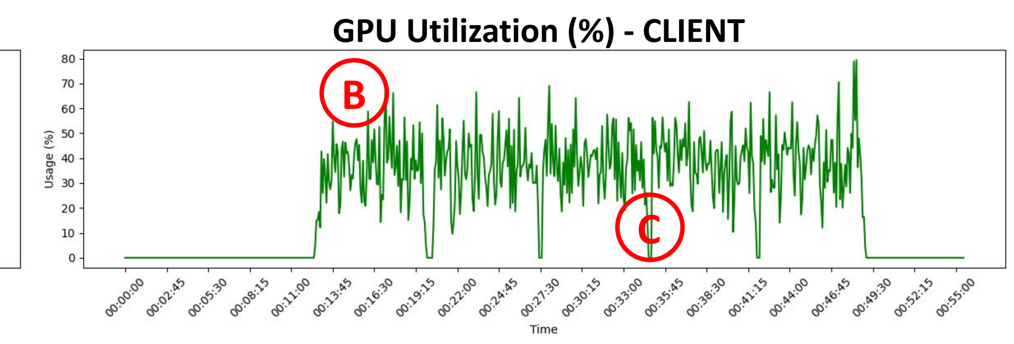
C. Specifications and Parameters

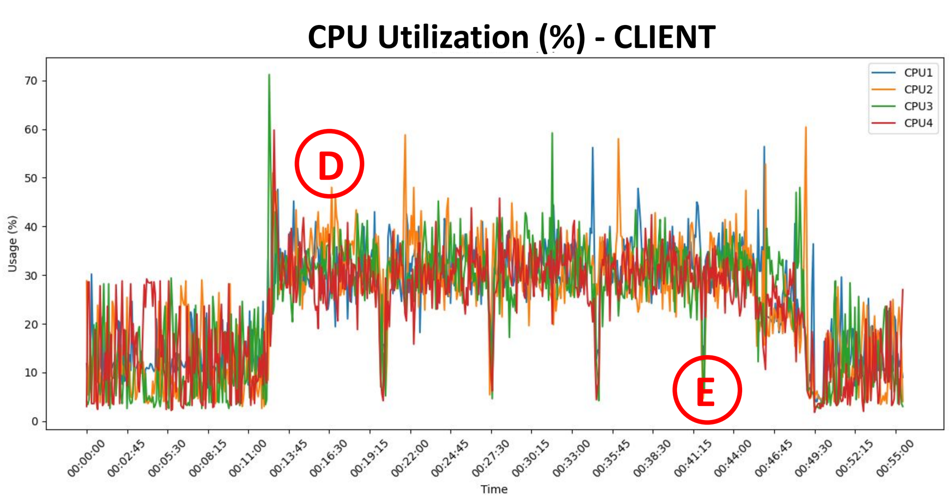
Device Specification	*Server*	*Client*
CPU	Quad-core ARM Cortex-A57	
GPU	128-core Maxwell GPU	
GPU Memory	2GB	4GB
Power Supply	5V⎓3A (Micro-USB)	5V⎓4A (Barrel Jack)
Software	PyTorch v1.10 and NVFlare 2.6.0	

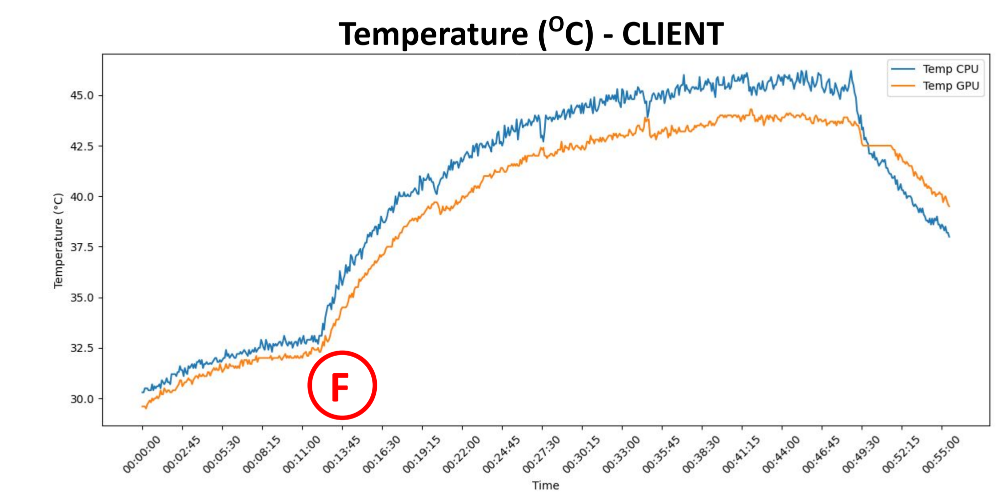
Number of Admin/Server	1/1
Number of Clients	3
Rounds	5
Epoch	5

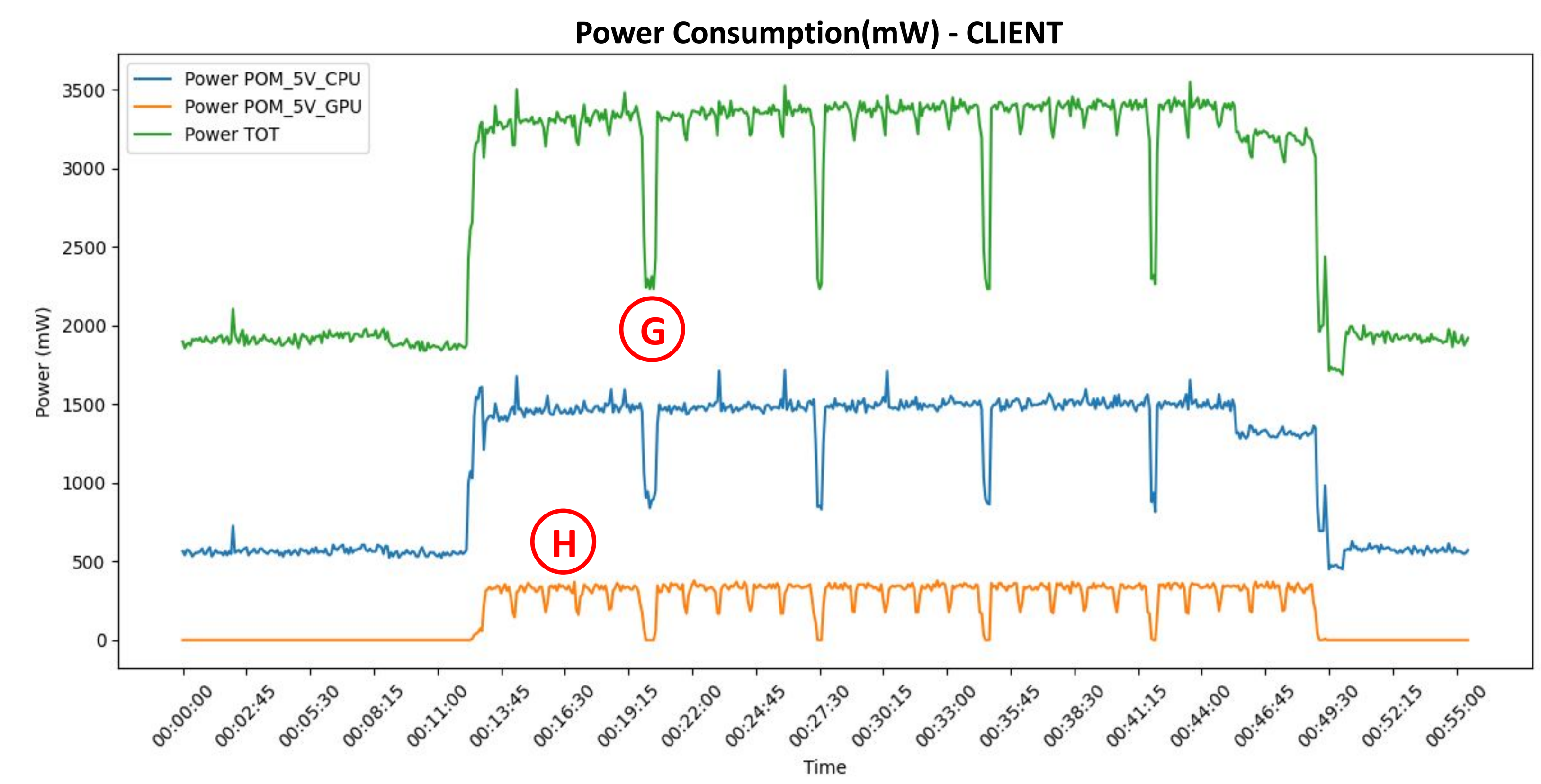
RESULTS

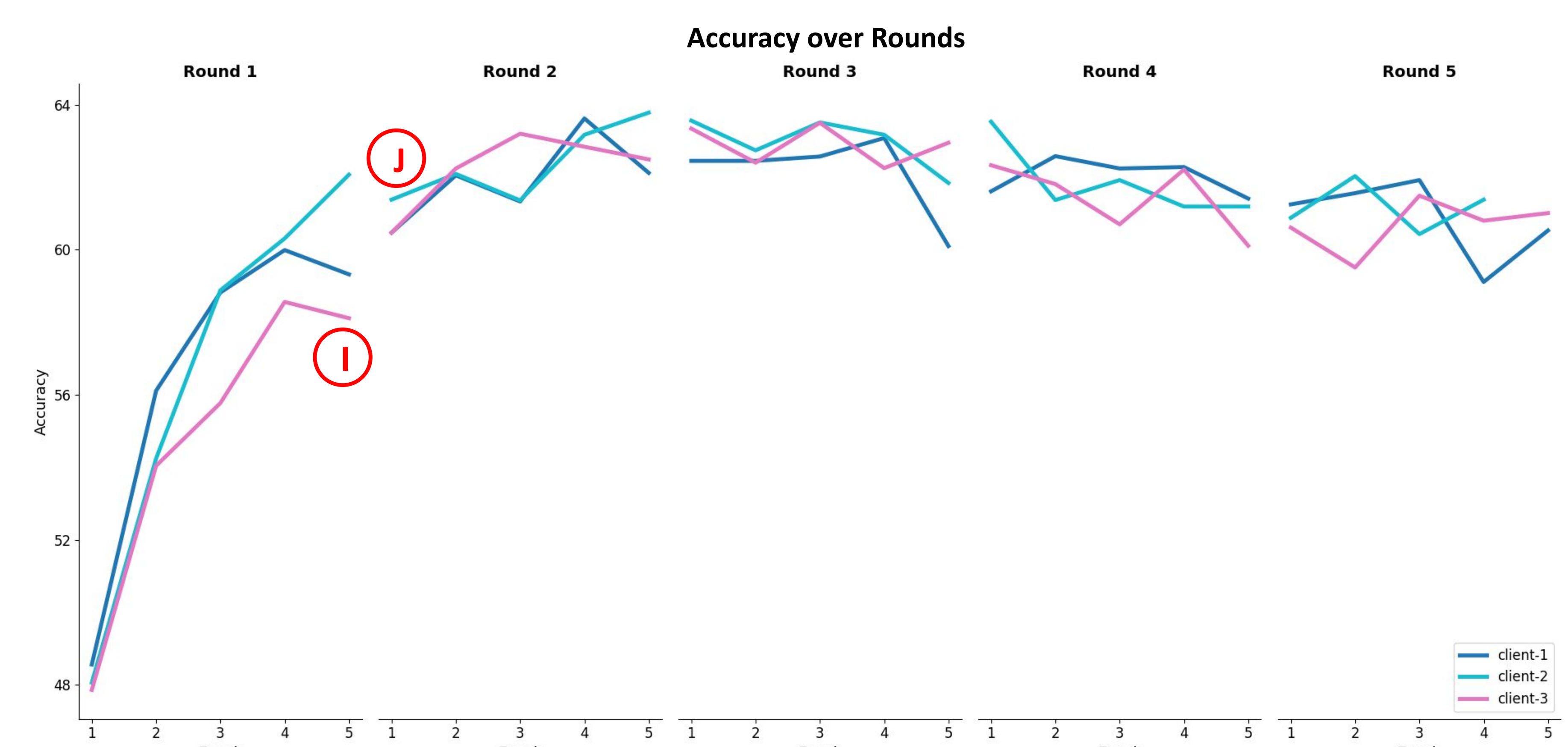












When training starts at **F**, there is an uptick in CPU (**D**) and GPU (**B**) utilization of the clients, during the training portion of the rounds, which drops after each epoch and drops even further after each round (**C**, **E**). This phenomenon is also evident in the temperature rise and power drawn by the CPU and GPU which clearly delineate the epochs (**H**) and rounds (**G**). The server's activity spike at **A** lines up with the end of each training round, when model averaging and transfer occurs.

The model accuracy across client is improved after each round as shown by the increase in the clients' average accuracy and drop in variance of the clients' accuracies between the last epoch of the prior round (**I**) and the first epoch of next round (**J**). The training time and power consumed for our experiment was **41 minutes** [for 5 rounds with 5 epochs each] and **0.012 kilowatt-hour** respectively.

FUTURE WORK

- Swapping out the Jetson Nano units with newer hardware such as Orin Nano and increasing the number of edge device units to enable experimentation with modern DL models and study system performance metrics in high resource utilization scenario
- Running FL experiments across heterogeneous devices with containerized software in place of native installations
- Studying the impact of advance privacy settings such as homomorphic encryption on system performance metrics

References:

- *H.R. Roth et al, "**NVIDIA FLARE: Federated Learning from Simulation to Real-World**" arXiv, Apr. 28, 2023. Accessed: July. 20, 2025. [Online]. Available: https://arxiv.org/abs/2210.13291*
- *Y. M. Boizic, A.R. Faustino, B. Radovic, M. Canini, and V. Pejovic, "**Where is the Testbed for my Federated Learning Research?**" arXiv, March 2, 2025. Accessed: July. 20, 2025. [Online]. Available: https://arxiv.org/abs/2407.14154*

Carnegie Mellon University AISaturdays Lagos