University of St Andrews

Rethinking Federated Learning Systems

Di Wu and Blesson Varghese

dw217@st-andrews.ac.uk

Seventh Annual UK System Research Challenges Workshop

Federated Learning (FL)

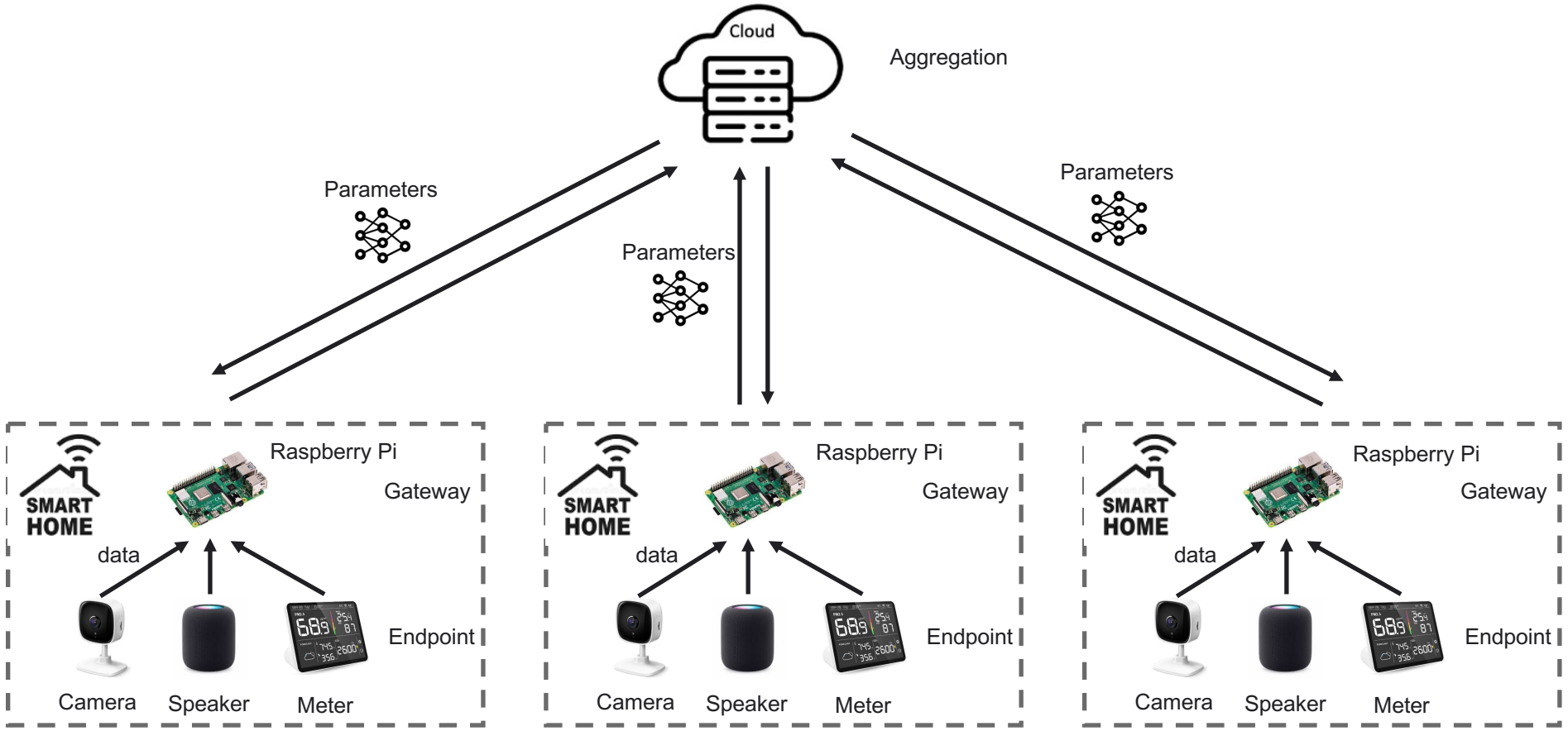
Cross-device FL

- Data owned by end users
- Smart phones and IoT devices
- Thousands to millions

Cross-Silo FL

- Data own by individual organizations.
- Banks and hospitals
- Tens to hundreds

A Typical FL system in Smart Home



Practical Challenges

- Computation
- MobileNet – 8 hours training of a Raspberry Pi per round on CIFAR-10 (10 K samples)

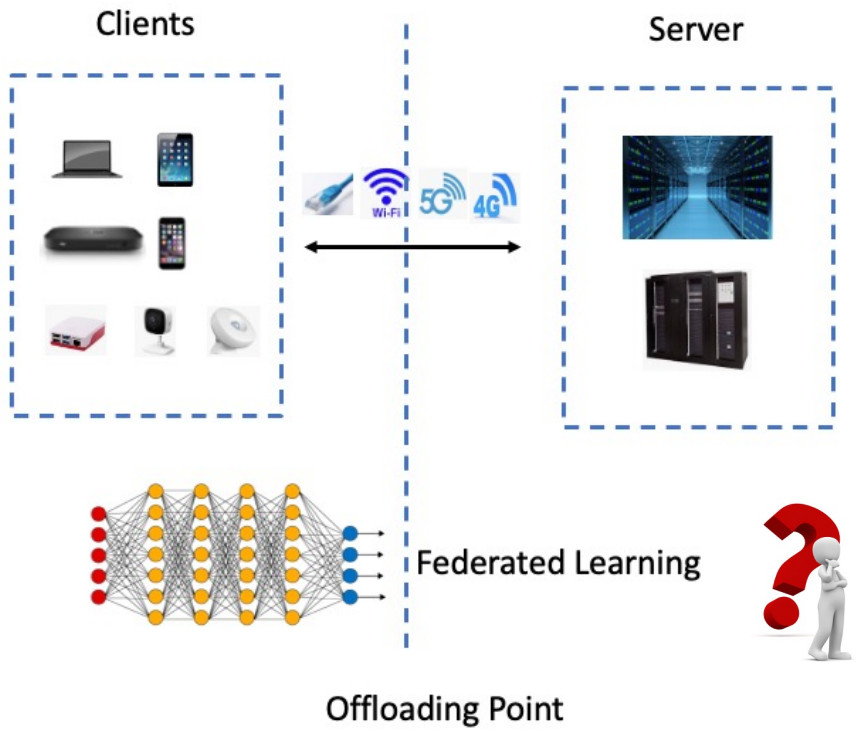
- Communication
- VGG11 – 25 GB data communication of 100 rounds per device on CIFAR-10 (10 K samples)

FL **can not** be directly applied on IoT devices.

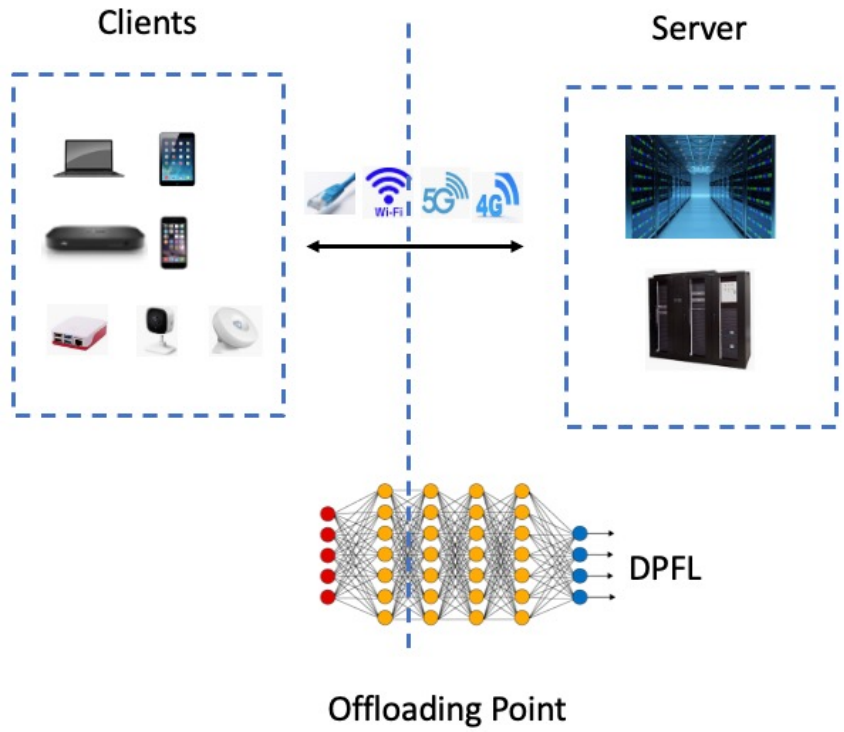
Gao, Y., Kim, M., Abuadbba, S., Kim, Y., Thapa, C., Kim, K., Camtepe, S.A., Kim, H. and Nepal, S., 2020. End-to-end evaluation of federated learning and split learning for internet of things. *arXiv preprint arXiv:2003.13376*.

DNN Partitioning-based FL (DPFL)

Device-native Training



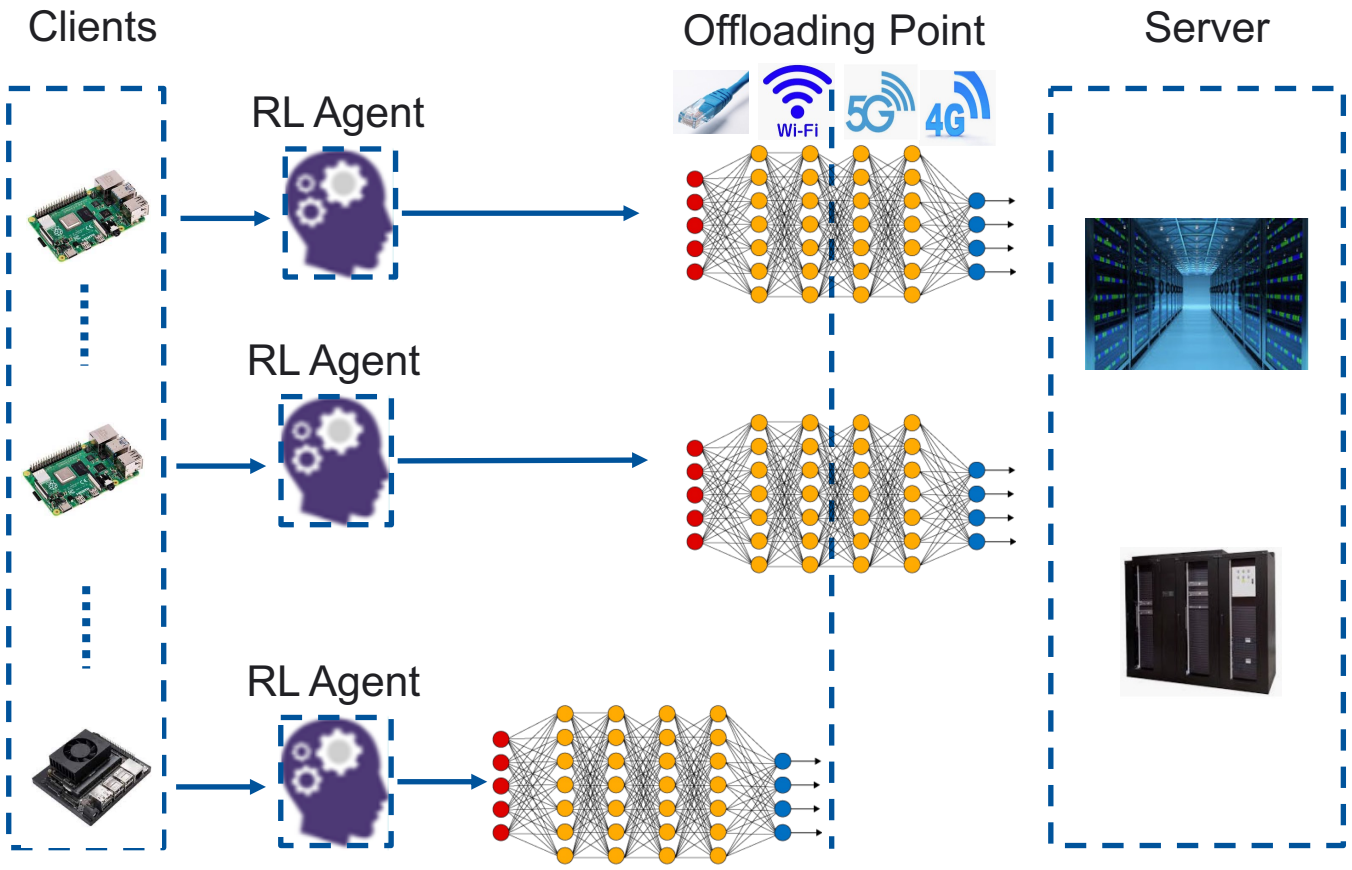
DPFL Training



FedAdapt: Adaptive Offloading for IoT Devices in FL

Q1: To what extent, can DPFL accelerate training?

Q2: How to decide the offloading point for various devices?

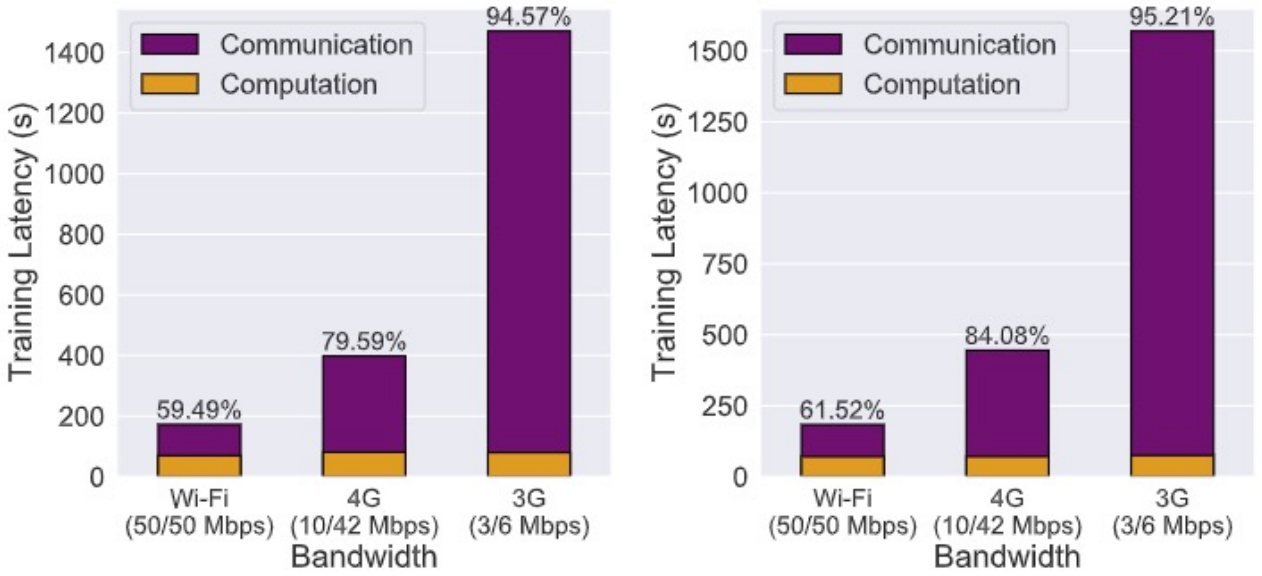


Wu, D., Ullah, R., Harvey, P., Kilpatrick, P., Spence, I. and Varghese, B., 2022. Fedadapt: Adaptive offloading for iot devices in federated learning. *IEEE Internet of Things Journal*, *9*(21), pp.20889-20901.

Communication Bottleneck in DPFL

- Communication latency is a new bottleneck in DPFL as the intermediate results need to be transferred for each batch of training samples.

- The communication requires up to 60% of the overall training time under Wi-Fi conditions and around 95% for 3G bandwidth.

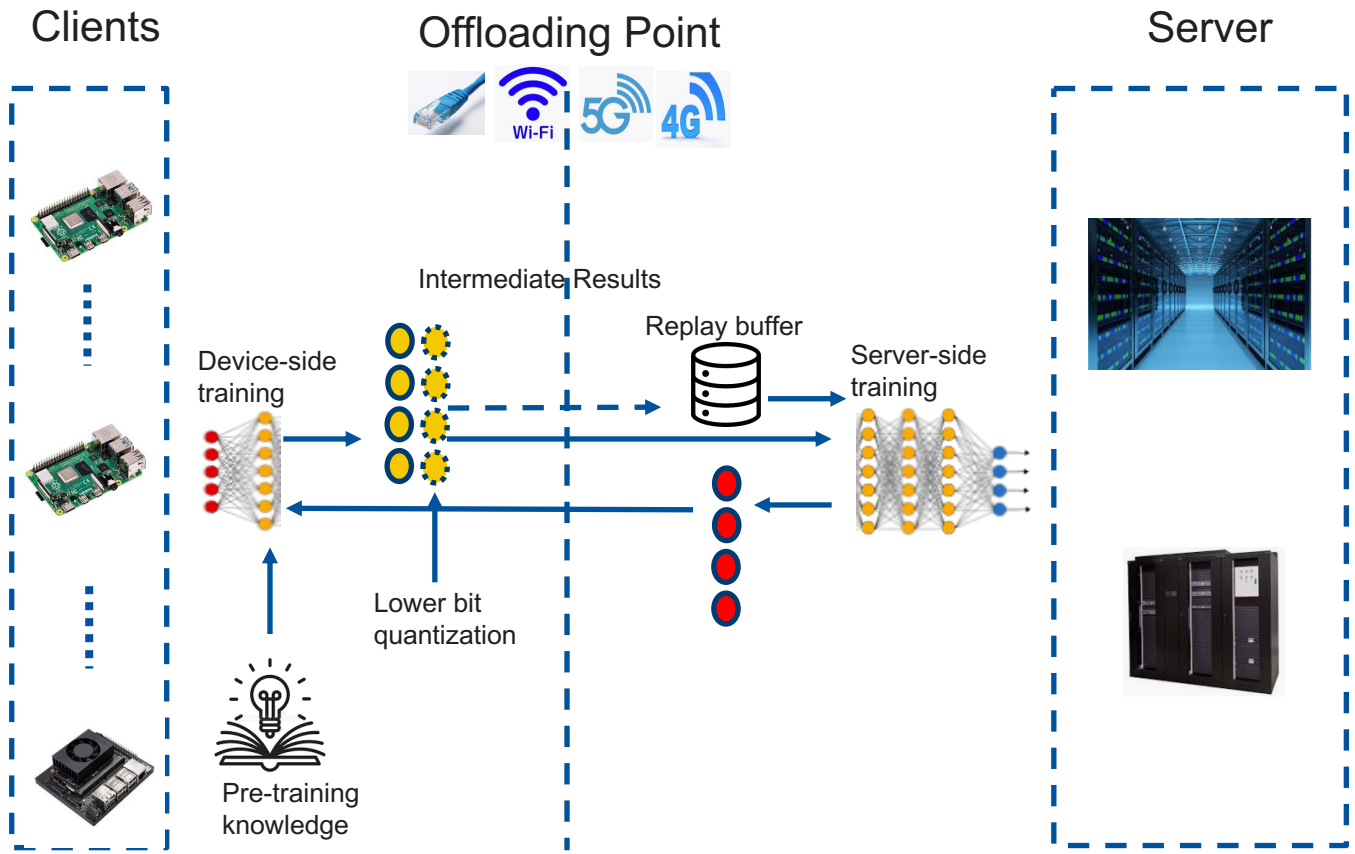


(a) VGG11 (b) ResNet9

Fig. 1: Computation and communication latency in DPFL training under different network bandwidths. Numerical value above the bars is the percentage of communication latency.

ActionFed: A Communication Efficient DNN DPFL Approach

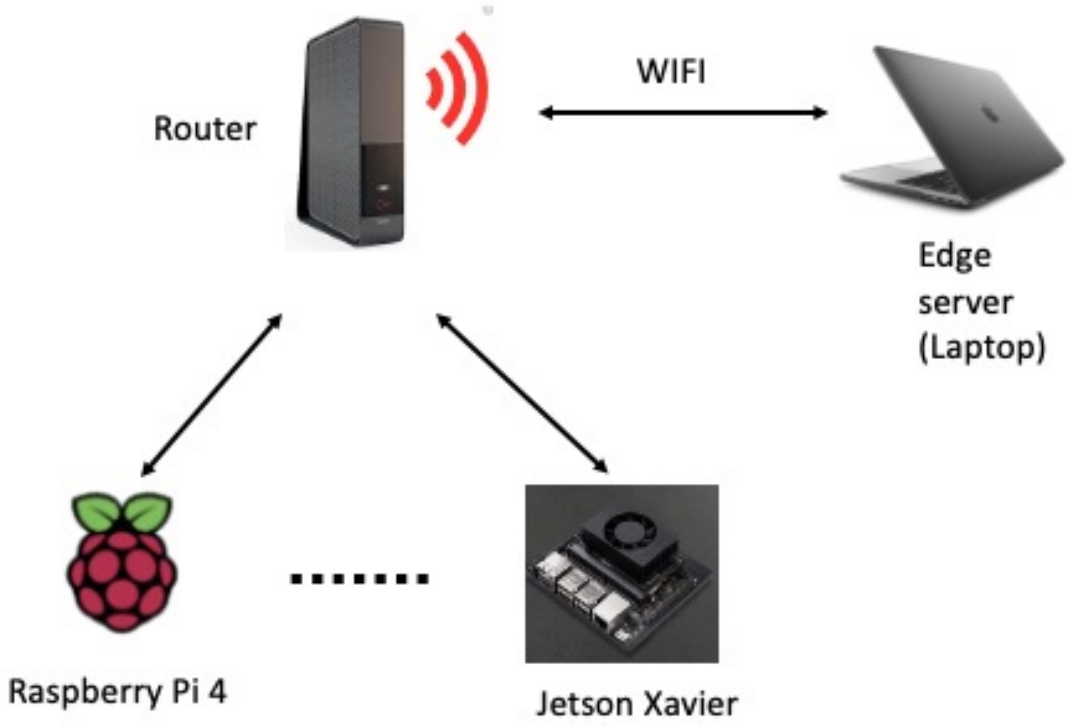
- Pre-trained knowledge (eliminating the need for gradient)
- Replay Buffer (reducing the communication frequency of activation)
- Quantization (Intermediate data compression)



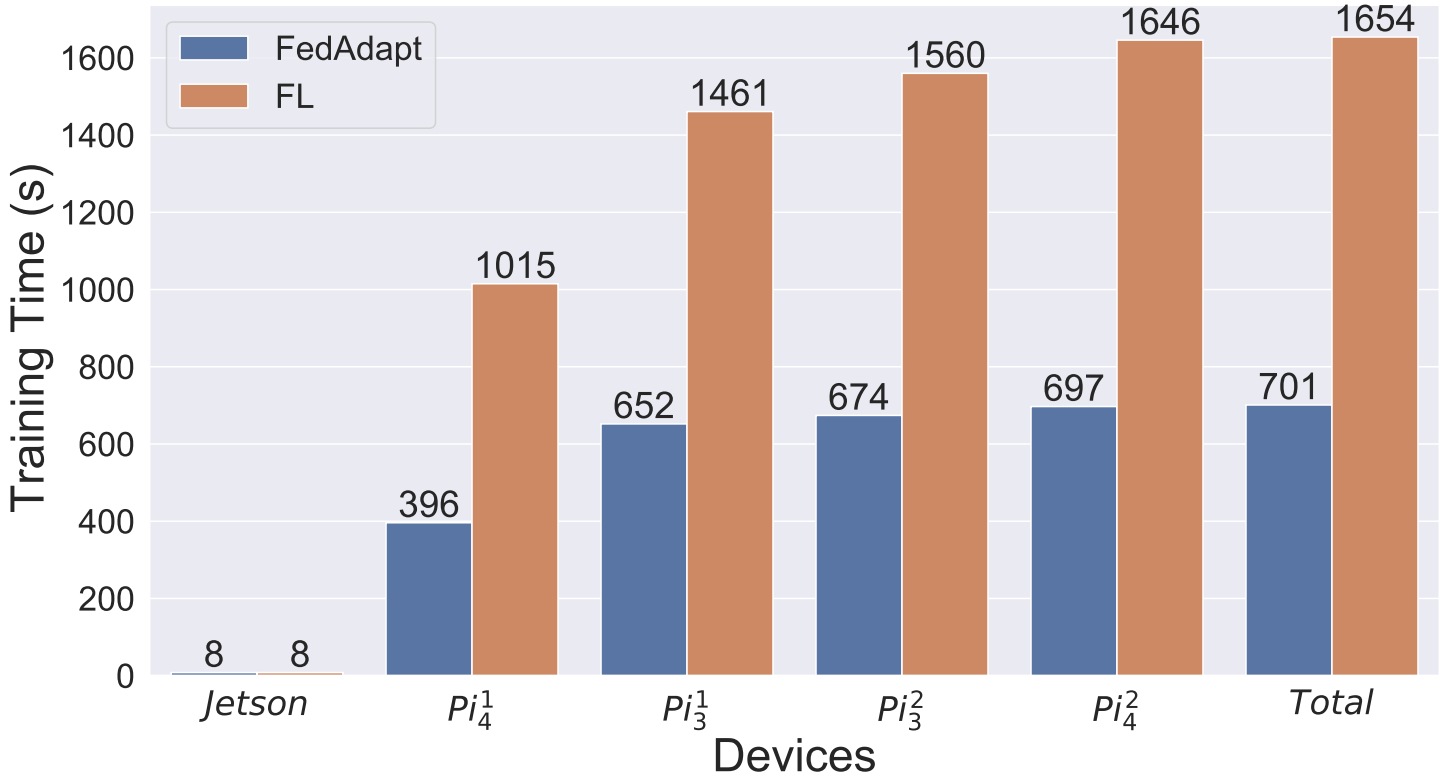
Wu, D., Ullah, R., Rodgers, P., Kilpatrick, P., Spence, I. and Varghese, B., 2023. Communication Efficient DNN Partitioning-based Federated Learning. *arXiv preprint arXiv:2304.05495*.

Testbed

- One edge server (Laptop)
- Five IoT devices: 4 Raspberry Pis and 1 Jetson Xavier
- We manually lower down the maximum CPU frequency to 1.2GHz and 0.7Ghz.
- TC commands are used to adjust the network bandwidths.



Experiment Results

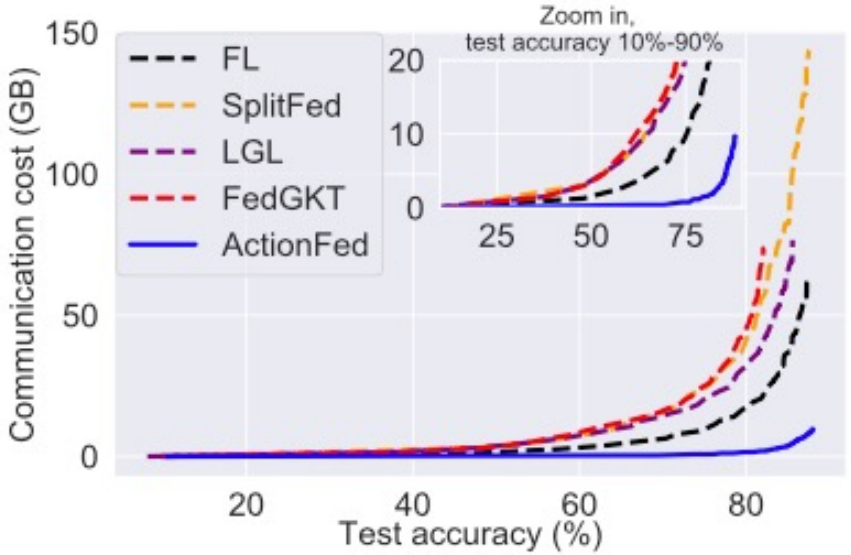


A total speed up of 2.35x for one round of FL training considering the heterogeneity of devices.

Experiment Results

TABLE VI: Communication cost for one training round.

Methods	Communication cost	
	VGG11	ResNet9
FL	1.28 GB	0.35 GB
SplitFed	3.05 GB	3.05 GB
LGL	1.52 GB	1.52 GB
FedGKT	1.53 GB	1.53 GB
ActionFed **w/o buffer**	**0.39 GB**	**0.39 GB**
ActionFed **w buffer**	**0 GB**	**0 GB**



(b) VGG11 on CIFAR-10

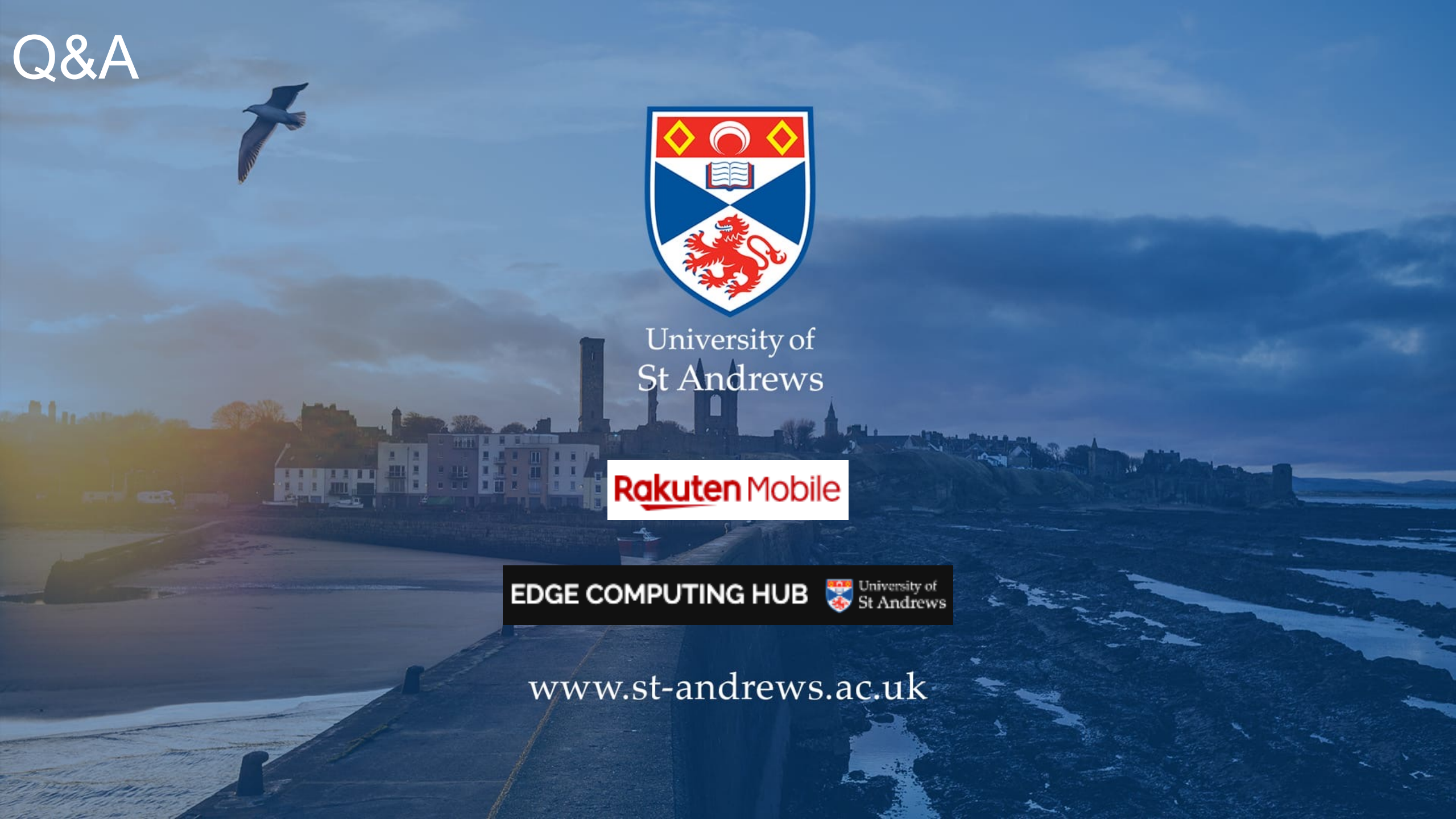
ActionFed can reduce the communication cost by up to **15.77x** compared to classical DPFL and can achieve the same learning performance with much less communication cost.

Next steps

- To further accelerate the device-side training, we will investigate the advantages of applying on-device ML methods like network slimming.

- A comprehensive framework that facilitates federated learning at the edge will be developed by integrating the techniques.

Q&A


University of
St Andrews
Rakuten Mobile
EDGE COMPUTING HUB
University of
St Andrews
www.st-andrews.ac.uk