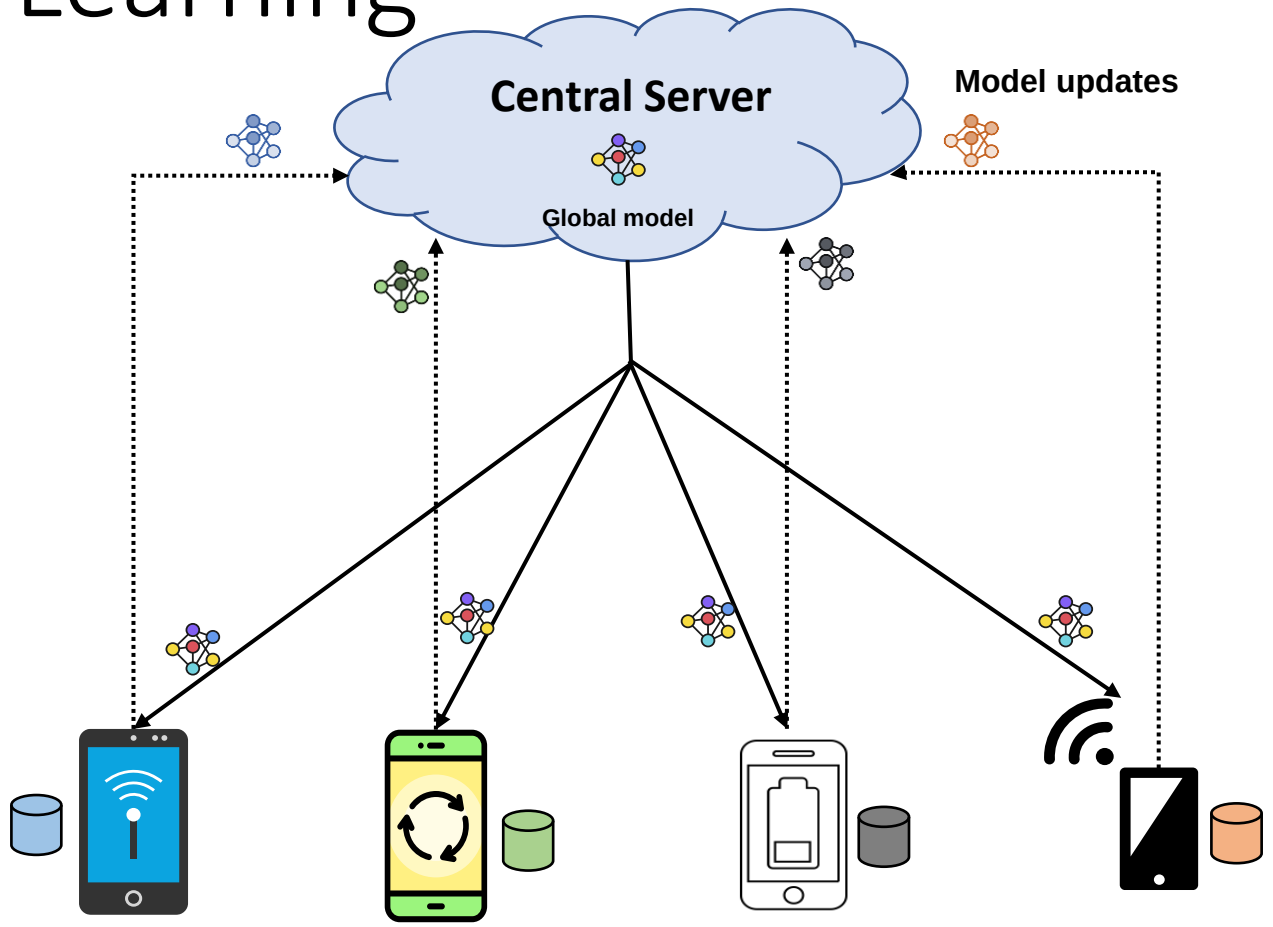
Towards Practical and Efficient Federated Learning

Ahmed M. Abdelmoniem

Lecturer/Assistant Professor @ School of EECS

Federated vs Centralized Learning

- Centralized Training:
  - Central (Data) server
  - Expensive data movement
  - Communication-intensive
  - **Privacy concerns**
- Federated Learning:
  - Central (Aggregation) server
  - Model exchange
  - Communication-efficient
  - Differential privacy + secure aggregation



Practical Use-cases of Federated Learning (FL)

What are good applications for FL?

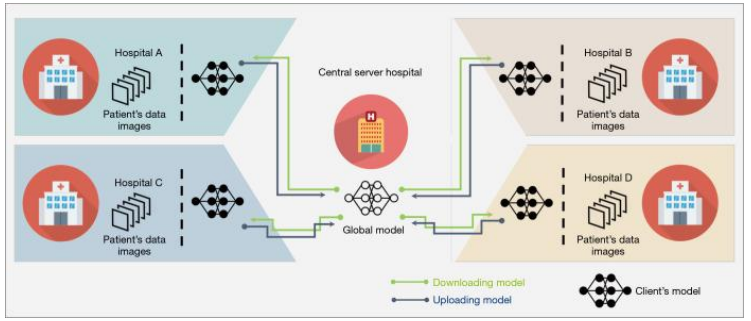
- Distributed on-device data is more relevant than server-side data (or lack of it)
- On-device data is privacy sensitive or large to communicate
- Labels can be inferred naturally from user interaction

Gboard: next-word prediction

Using FL, better next-word prediction accuracy: +24%

A. Hard, et al. Federated Learning for Mobile Keyboard Prediction. arXiv:1811.03604

Medical Imaging

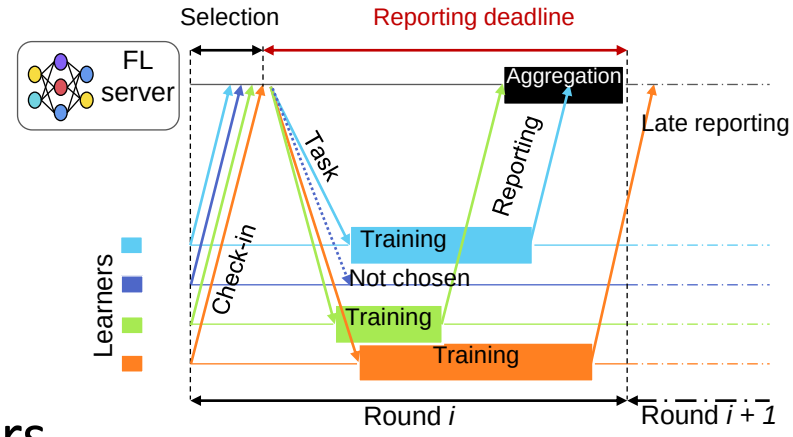


Ng D, Lan X, Yao MM, Chan WP, Feng M. Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. Quant Imaging Med Surg. 2021

Federated Learning Life-cycle

- FL server coordinates the learning stages

1. Online Learners check-in with the server
2. Participants Selection Stage:
   - The server selects a pre-set target number of learners
3. Updates Reporting Stage:
   - The selected participants perform the training task on local dataset
   - The server aggregates the updates to produce the global model



Heterogeneity is the main barrier to Practical and Efficient FL!

"Training in heterogeneous and potentially massive networks introduces novel challenges that require a fundamental departure from standard approaches for large-scale machine learning, distributed optimization, and privacy-preserving data analysis"
In Federated Learning: Challenges, Methods, and Future Directions, T. Li, A. K. Sahu, A. Talwalkar and V. Smith., IEEE Signal Processing Magazine, 2020

Impact of Heterogeneity on Federated Learning

ACM EuroMLsys 2022

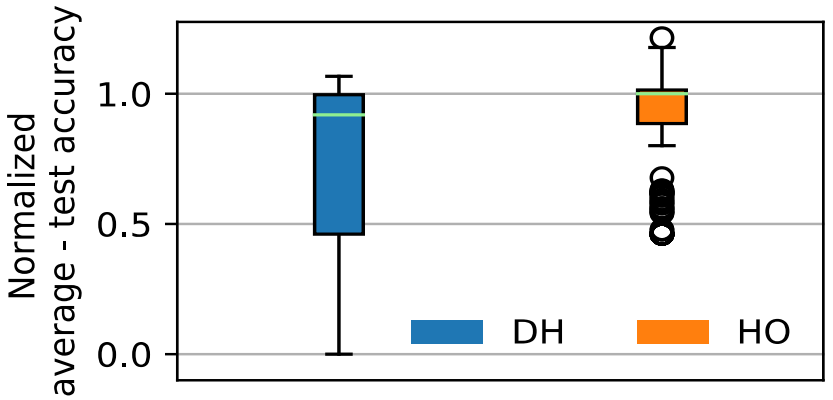
Work with Elton, Pantelis, Bilal, Marco @ KAUST

https://dl.acm.org/doi/10.1145/3517207.3526969
Full version: https://arxiv.org/abs/2102.07500

Impact of Heterogeneity

- Can heterogeneity impede the learning process?
  - Data, Device, Behavioral, Participation, Configuration, etc.
- Empirical study of the impact of heterogeneity on FL [1]
  - 1.5K experiments → varying models, datasets, FL & learning hyper-parameters, device proportions, aggregation algo, etc
  - We compare the following settings:
    - Homogenenous (HO): all devices are of the same configuration
    - Device Heterogenity (DH): Low-end, mid-range, high-end
- The impact is significant (**the model may not converge**)
  - Device heterogeneity degrades:

    Model quality by up to **4.9X**



More extensive analysis and detailed results in the paper

[1] Abdelmoniem et. al., Empirical analysis of federated learning in heterogeneous environments, ACM EuroMLSys'22

Mitigating the impact of Heterogeneity

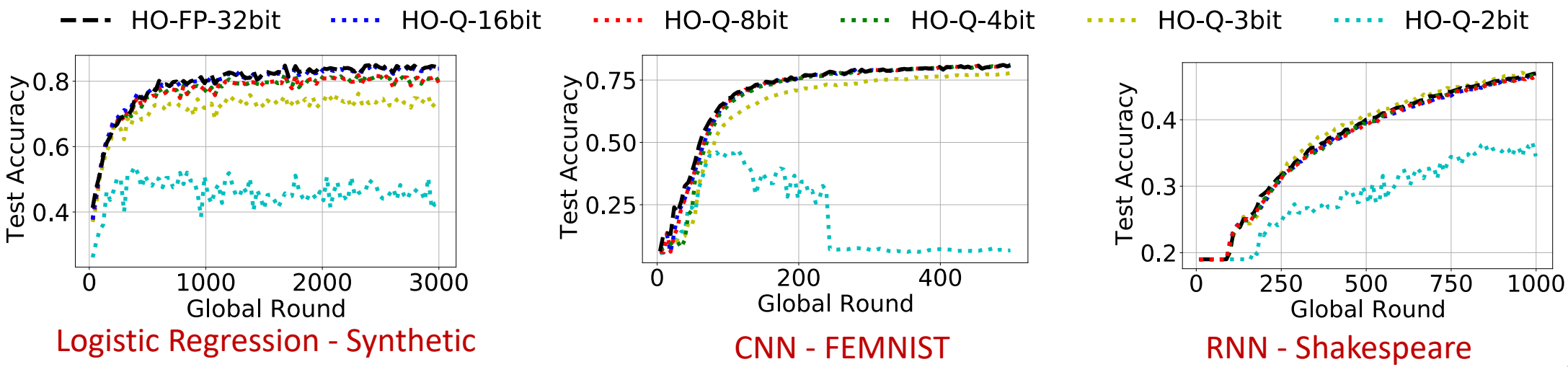
ACM EuroMLSys 2021

Work with Marco @ KAUST

https://dl.acm.org/doi/10.1145/3437984.3458839

Quantization-Aware Training

- Readily supported by wide-range of hardware and software
- Quantization vs Quality Tradeoff:
  - Evaluate the quality degradation in homogeneous FL setting
    - Model quantization using Tensorflow's quantize library:
      - Quantize function - Inserts quantization operations in the computational graph
  - Acceptable quality for bit-widths >= 4 bits



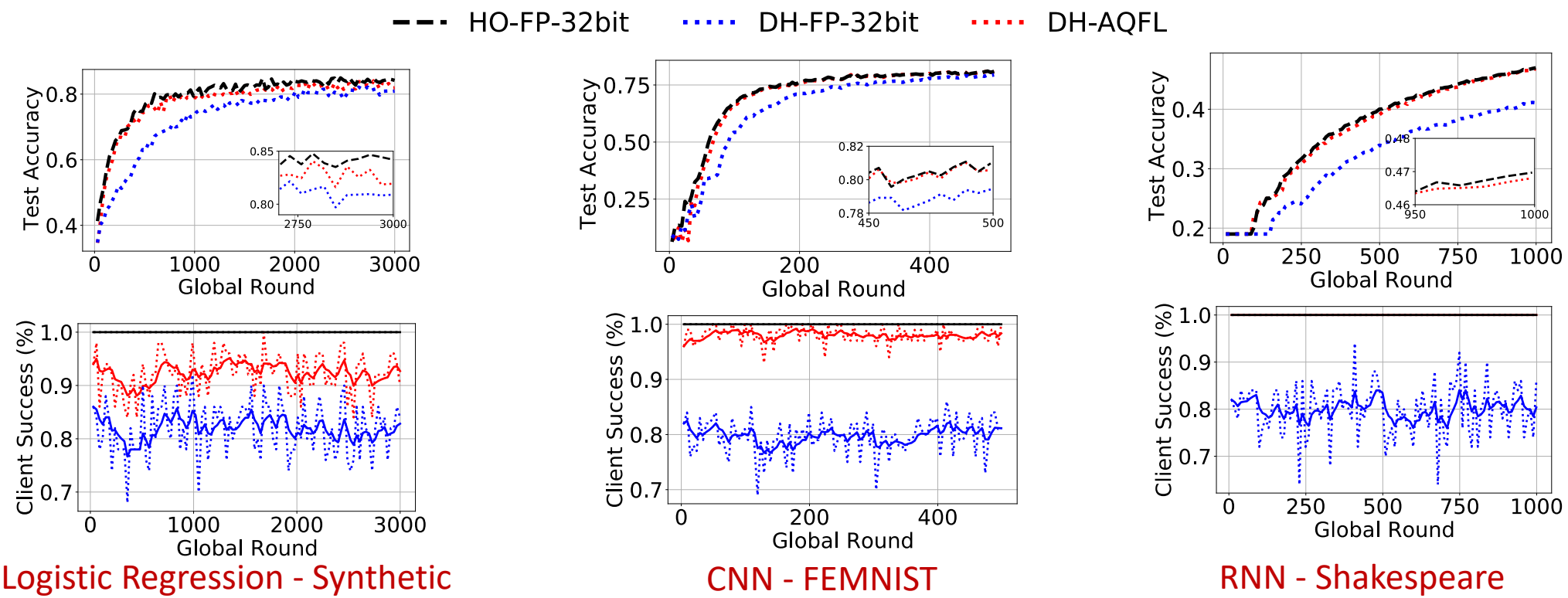
Logistic Regression - Synthetic

CNN - FEMNIST

RNN - Shakespeare

AQFL: Adaptive Model Quantization

- Server adaptively selects per-client quantization level to meet deadline
- Improves model quality and fairness:
  - High success rate of clients not missing the deadline



RELAY: Resource Efficient Federated Learning

ArXiv 2021, under review

Work with Atal, Marco and Suhaib @ KAUST

https://arxiv.org/abs/2111.01108

Status Quo

- Many approaches aim to improve the quality of FL trained models.
- Goal: reduce Time to Accuracy: Reduce Time ↓ + Improve Quality ↑

FedProx (MLSys 2020), Yogi (ICLR 2021)

- Improve the statistical efficiency of the learning process

Oort (USENIX OSDI 2021)

- Bias the client selection to Reduce the training time by exploiting the fast learners.

SAFA (IEEE ToC 2021)

- Invokes all learners for training and allows semi-synchronous model updates to boost the statistical efficiency

FLEET (ACM Middleware 2021)

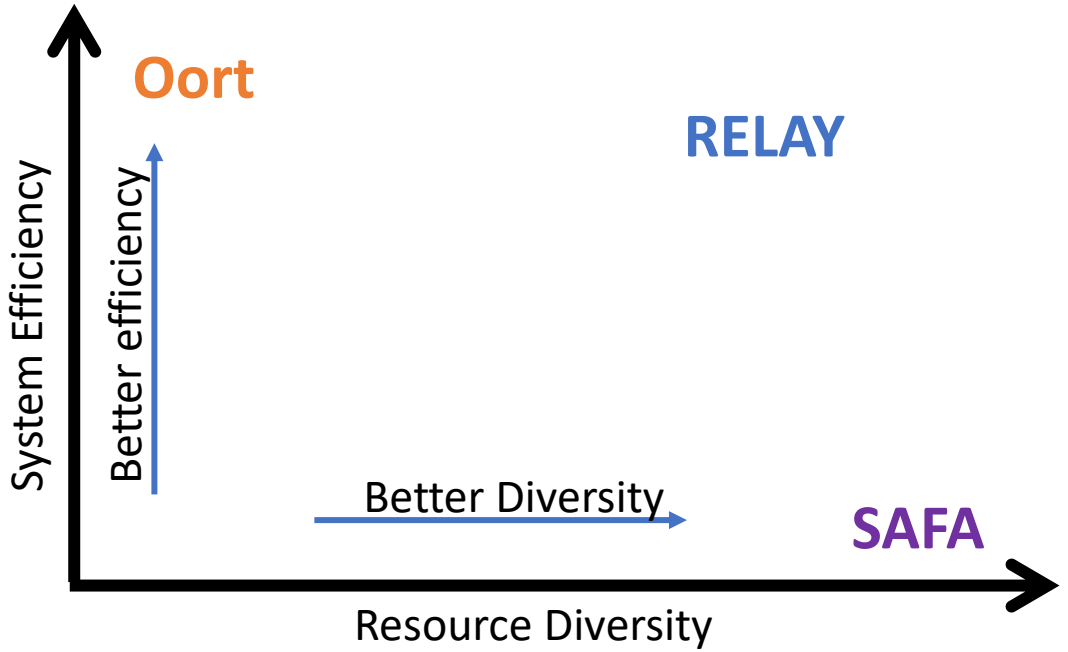
- Boost statistical efficiency via asynchronous updates with damping and boosting rules.

Motivation – **Resource Efficiency**

- We identify a Trade-off between System Efficiency vs Resource Diversity
  - Oort: favor fast learners over slow ones → high efficiency & low diversity
  - SAFA: select every learner → high diversity & low efficiency
  - Existing systems ignore resource consumption of the learners
- Our goal is to strike a balance between the two extremes

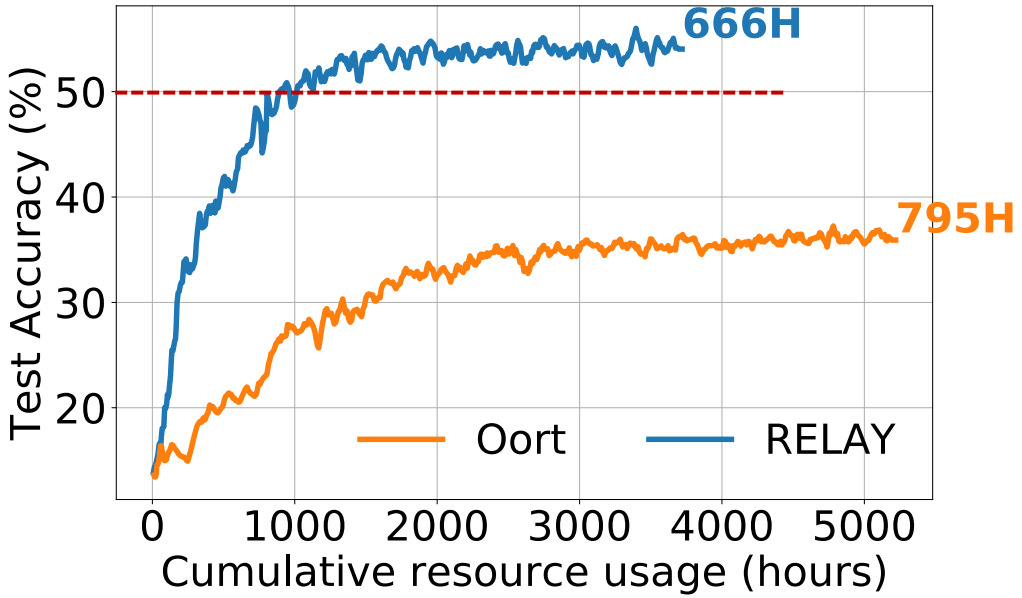
RELAY: Resource Efficient FL System

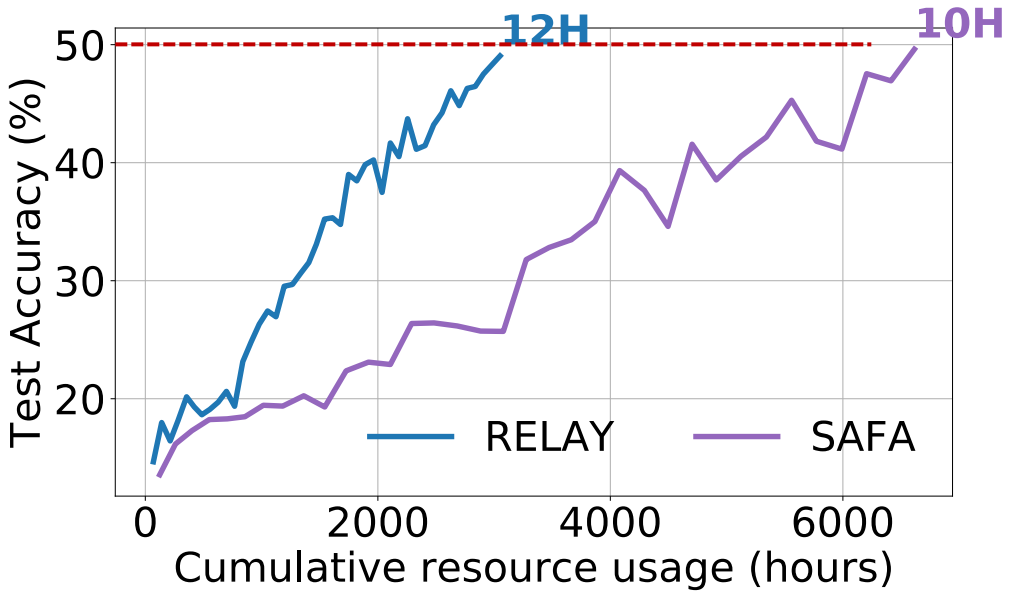
- ❑ Increases the **diversity** by prioritizing selection of least available learners
- ❑ Aggregates stale updates to boost the statistical **efficiency**
- ❑ Leads to reduced **Resource** consumption



Evaluation of RELAY

- RELAY → best model quality with least amount of resources and time
  - **Availability prioritization** leads to better diversity.
  - **Aggregation of stale updates** leads to resource savings.





Takeaways

- **Heterogeneity** is a major challenge for FL:
  - Model quality degradations are not acceptable, esp. with diverging cases
  - Heterogenous device settings impact the quality the most.

- To tackle heterogeneity → adapt to available system HW/configs/dynamics
  - **AQFL** leverages support of on-device quantized training and customizes the models to improve quality of the models.

- Need to strike balance between resource usage and diversity
  - **RELAY** innovates on client selection and stale updates aggregation to yield a resource efficient FL system.

Moving Forward

- More practical cases with **battery-powered devices** → Preliminary results under review
  - Optimize the selection strategy to maximize gains within power budgets.
- Leverage **recent ML methods** to improve FL performance → Work in Progress
  - Learning strategies (other than SGD) to boost the statistical gains of the training
- **Question: what is the right architectural design for decentralized learning?**
  - Novel learning ecosystem for decentralized learning on edge devices
  - Try to address key drawbacks in existing paradigms (Centralized learning, distributed learning, Federated Learning, split learning, gossip learning, etc).

Thanks

Q & A

To follow-up, please reach me at ahmed.sayed@qmul.ac.uk