# REFL: Resource-Efficient Federated Learning

**Ahmed M. Abdelmoniem**
**Queen Mary University of London, UK**

Recently distributed machine learning (ML) deployments have sought to push computation towards data sources in an effort to enhance privacy and security. Training models using this approach is known as Federated Learning (FL). FL presents a variety of challenges due to the high heterogeneity of participating devices, ranging from powerful edge clusters and smartphones to low-resource IoT devices (e.g., surveillance cameras, sensors, etc.). These devices produce and store the application data used to train a shared ML model. FL is deployed by large service providers such as Apple, Google, and Facebook to train computer vision (CV) and natural language processing (NLP) models in applications such as image classification, object detection, and recommendation systems. FL has also been deployed to train models on distributed medical imaging data, and smart camera images.

In Fl, system efficiency is primarily regulated by the time to complete a training round, which depends on which learners are selected and whether they become stragglers whose updates do not complete in time. It is common to configure a reporting deadline to cap the round duration, but if only an insufficient number of learners complete within this deadline, the entire round fails and is re-attempted from scratch. Since a tight deadline can yield more failed rounds, this can be mitigated by overcommitting the number of selected learners in each round to increase the likelihood that a sufficient number will finish by the deadline. Failed rounds and overcommitted participants lead to wasted computation, which has mostly been ignored in previous FL approaches.

All the above factors can lead to resource wastage---where learners perform training work that does not contribute to enhancing the model, whether due to updates that are ultimately discarded, or poor data distribution. We argue that this resource wastage deters users from participating in FL and makes the scaling of FL systems to larger deployments and more varied computational capabilities of learners problematic. We aim to optimize the design of FL systems for their resource-to-accuracy in a heterogeneous setting. This means the computational resources consumed to reach a target accuracy is reduced without a significant impact on time-to-accuracy. By considering heterogeneity at the heart of our design, we also intend to demonstrate improved robustness to realistic data distributions among learners.

In this presentation, we systematically address the question of resource efficiency in FL, showing the benefits of intelligent participant selection, and incorporation of updates from straggling participants. We demonstrate how these factors enable resource efficiency while also improving trained model quality [1].

[1] Ahmed M Abdelmoniem, Atal Narayan Sahu, Marco Canini, Suhaib A Fahmy, “REFL: Resource-Efficient Federated Learning”, ACM EuroSys, 2023