GlueFL: Reconciling Client Sampling and Model Masking for Bandwidth Efficient Federated Learning

Shiqi He, Qifan Yan, Feijie Wu, Lanjun Wang, Mathias Lécuyer, Ivan Beschastnikh

To appear at MLSys 2023!

Cross-device federated learning (FL) is a distributed machine learning setting where **many edge clients** communicate with a **central server** to collaboratively train a global model while keeping their training data **local**

Challenges

- Enormous number of clients (up to $N=10^{10}$)
  - Necessitates the use of **client sampling**
- Highly heterogeneous client network speed
  - Susceptible to the **straggler effect** and need to decrease communication volume
- Local client data is non-I.I.D.
  - **Unbiasedness** of updates is critical

Existing work

- Sparsification – STC [1] and parameter freezing – APF [2] employ **masking**
- Masked updates are the **largest *q*% of value changes** where *q* is the compression ratio

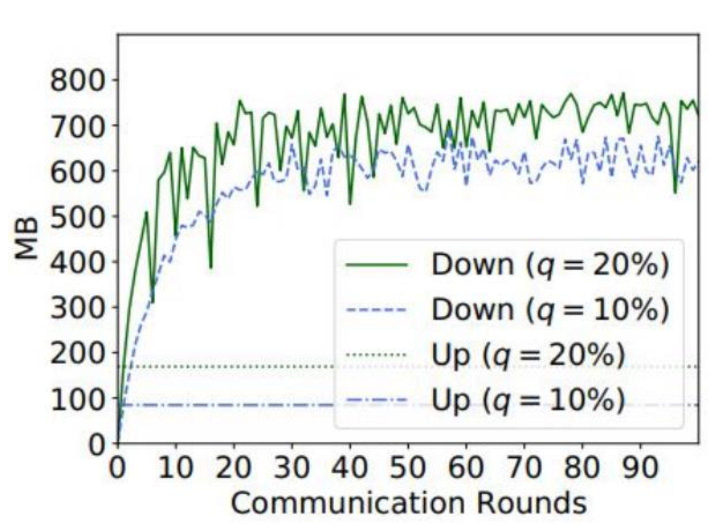


Fig 3. Downstream and upstream bandwidth usage of all clients per round

- Masking + sampling **saves upstream bandwidth** **but not downstream bandwidth**

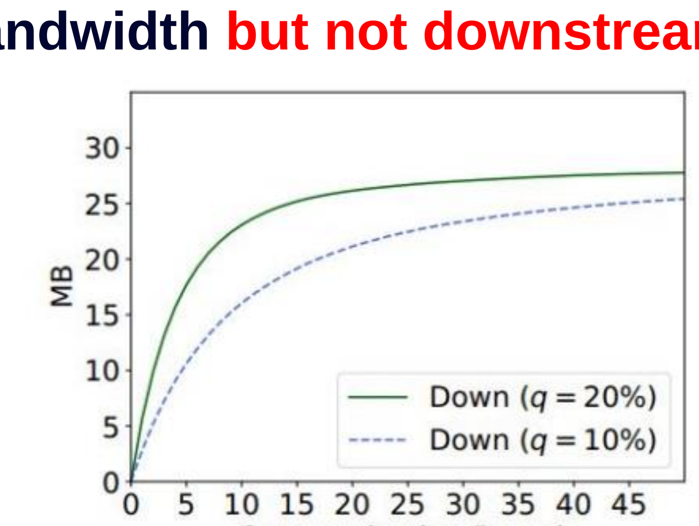


Fig 4. Model size a client must download when being re-sampled after a certain number of rounds

Reason: client local states become stale

- A client may skip many rounds by not being sampled in cross-device FL
- Server updates across two successive rounds have low overlap

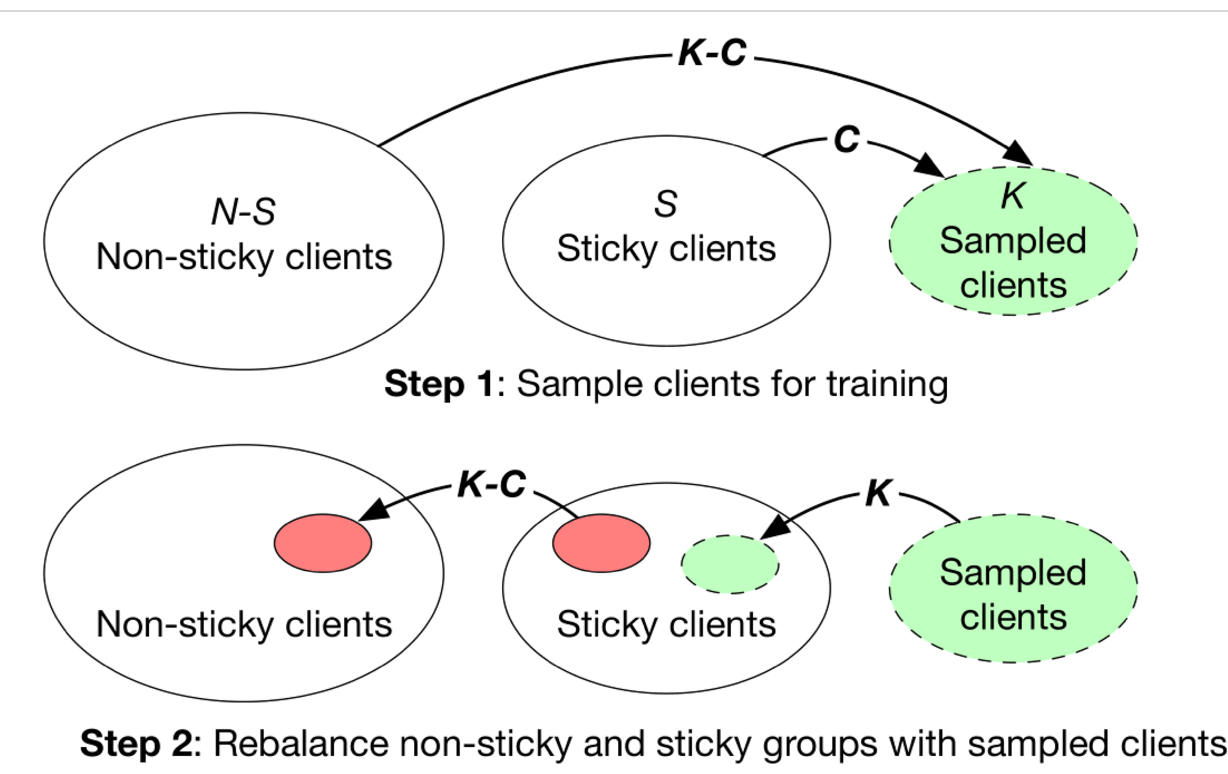
Our contributions

1. First work to combine client sampling (**sticky sampling**) with masking (**mask shifting**) in FL
2. Theoretical guarantees for preserving **unbiasedness** of updates and **convergence**
3. Empirically evaluated in realistic environments across three non-I.I.D. datasets to show a **29%** and **27%** reduction in training time and downstream bandwidth compared to state-of-the-art

GlueFL design

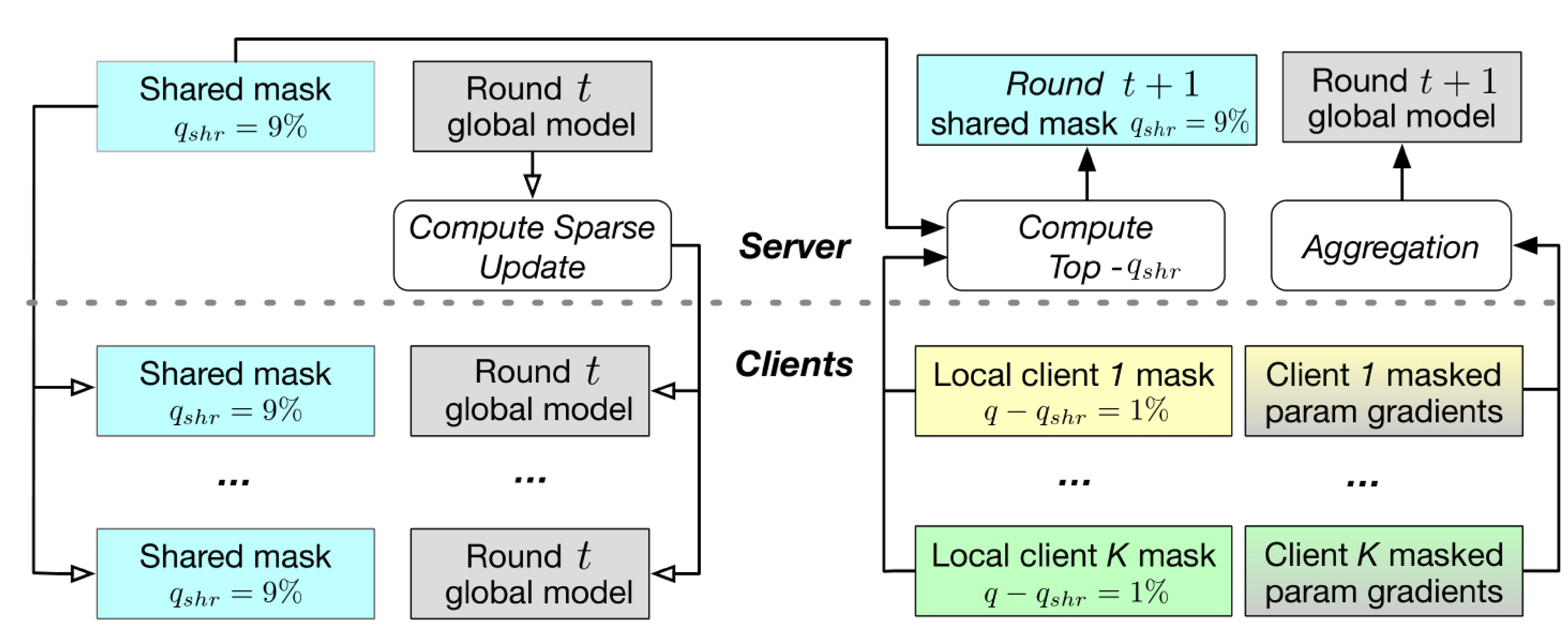
Sticky Sampling:

- Prioritize sampling "sticky" clients with the most up-to-date local state because they can download less
- Updates are appropriately reweighted for unbiasedness



Mask Shifting:

- Increase overlap of server updates across successive rounds with shared masks
- Further allow clients to download less if sampled repeatedly

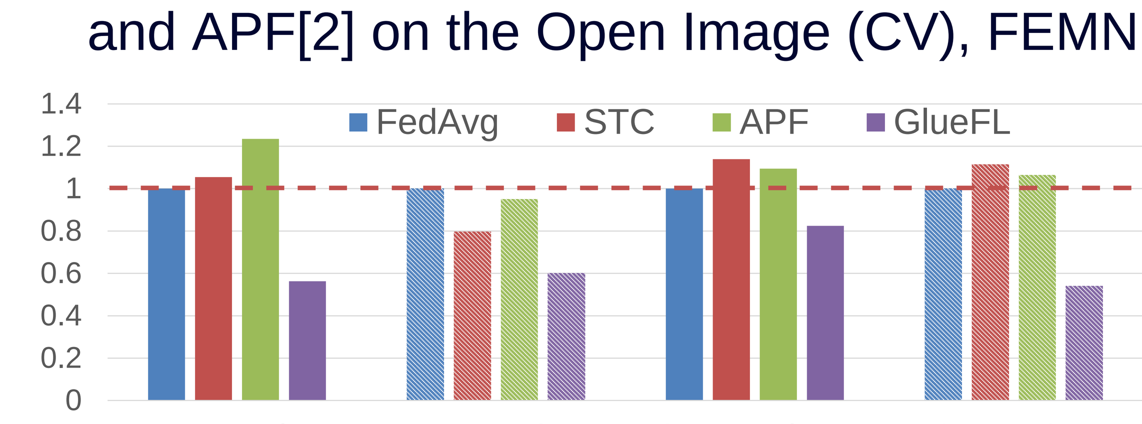


(a) Server sends shared mask and latest model state to clients

(b) Clients send their local masks and masked model updates to server

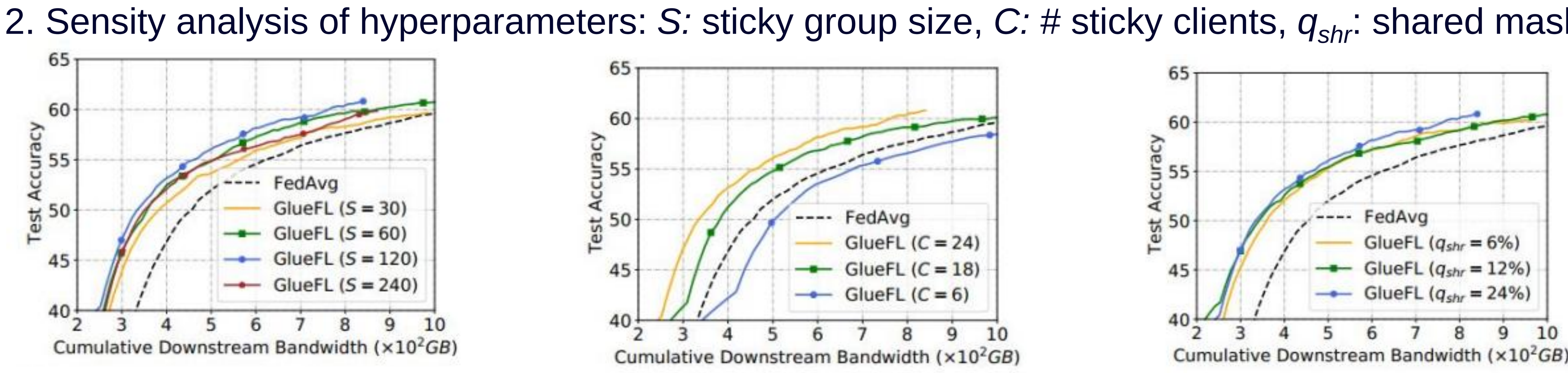
Evaluation results

1. A comparison of the downstream time and bandwidth usage between GlueFL, FedAvg [3], STC[1], and APF[2] on the Open Image (CV), FEMNIST (CV), and Google Speech (NLP) datasets



To reach the same target test accuracy, GlueFL needs significantly less downstream bandwidth and time for CV and NLP tasks on average

2. Sensity analysis of hyperparameters: *S:* sticky group size, *C:* # sticky clients, q_{shr}: shared mask size



With most hyperparameter choices, GlueFL outperforms FedAvg, showing its robustness

Future work

Prefetching strategies for further reducing the staleness of client local model states

1. Significantly shortens download time at minimal extra downstream transmission overhead
2. Generally applicable to a broad variety of masking and quantization techniques with minimal setup
3. Encourages slower clients to participate to decrease biasedness in settings where only fastest clients run

References

1. Sattler et al. "*Robust and communication-efficient federated learning from non-iid data.*", TNNLS '19
2. Chen et al. "*Communication-efficient federated learning with adaptive parameter freezing.*", ICDCS' 21
3. McMahan et al. "*Communication-Efficient Learning of Deep Networks from Decentralized Data.*", AISTATS '17

THE UNIVERSITY OF BRITISH COLUMBIA

SYSTOPIA

Partners

NSERC CRSNG

HUAWEI