Pollen: High-throughput Simulation of Federated Learning via Resource-Aware Client Placement

Lorenzo Sani [1*], Pedro P.B. Gusmao[1*], Alex Iacob[1*],
Wanru Zhao[1], Xinchi Qiu[1], Yan Gao[1], Javier Fernandez-Marques[2], Nicholas Donald Lane[1]
[1] Department of Computer Science and Technology, University of Cambridge
[2]Samsung AI

Federated Learning (FL) is the privacy-preserving machine learning paradigm which collaboratively trains a model across millions of devices. FL brings unique challenges relating to clients' diverse hardware and data distributions which become more arduous as the size of the federation increases. Simulated environments are fundamental to large-scale FL research, allowing researchers to quickly test new ideas to solve such system and statistical heterogeneity issues. However, FL simulation is not straightforward as FL frameworks face challenges when distributing many small workloads across heterogeneous resources, i.e. FL simulation needs to efficiently allocate the training of many small datasets on diverse hardware, including GPUs, CPU-only machines, TPUs, and others.

Our analysis of this problem led us to focus on three aspects: (1) the simulated clients cannot singularly take full advantage of a single GPU; (2) clients having very different datasets sizes take a different time to train when placed on the same GPU; (3) different GPUs, or more generally different machines, take different time in training the same client.

Our contribution is threefold. (1) We experimentally investigated and characterised the relationship between the size of clients' dataset and their training time across diverse workloads and GPU types. (2) We propose *Pollen*, a novel *push-based resource-aware* system capable of speeding up FL simulations by efficiently placing clients across distributed and heterogeneous hardware. (3) Efficient client placement strategies are also proposed based on the inherent trade-offs of FL client placement on heterogeneous GPUs. We compared the proposed strategies against each other across different homogeneous and heterogeneous hardware configurations identifying the situations where using *Pollen* is needed. Finally, based on our evaluation, we make the argument that our *push-based* method for scheduling tasks performs better than *pull-based* methods seen in existing FL frameworks research [1–3] and should be implemented as a core feature.

We explored these trade-offs through relevant baselines on three popular FL tasks: image classification, speech recognition and text generation. Compared to existing ad-hoc FL frameworks, such as Flower, Flute and FedScale, and show that *Pollen* provides gains of 50% to 400% in speed.

References

[1] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, T. Parcollet, and N. D. Lane, "Flower: A friendly federated learning research framework," *CoRR*, vol. abs/2007.14390, 2020.

[2] D. Dimitriadis, M. Hipolito Garcia, D. Madrigal, A. Manoel, and R. Sim, "Flute: A scalable, extensible framework for high-performance federated learning simulations," March 2022.

[3] F. Lai, Y. Dai, S. S. V. Singapuram, J. Liu, X. Zhu, H. V. Madhyastha, and M. Chowdhury, "Fedscale: Benchmarking model and system performance of federated learning at scale," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 11 814–11 827.