FedScale: Benchmarking Model and System Performance of Federated Learning at Scale

Fan Lai 1 Yinwei Dai 1 Sanjay S. Singapuram 1 Jiachen Liu 1 Xiangfeng Zhu 1,2 Harsha V. Madhyastha 1 Mosharaf Chowdhury 1

Abstract

We present FedScale, a federated learning (FL) benchmarking suite with realistic datasets and a scalable runtime to enable reproducible FL research. FedScale datasets encompass a wide range of critical FL tasks, ranging from image classification and object detection to language modeling and speech recognition. Each dataset comes with a unified evaluation protocol using real-world data splits and evaluation metrics. To reproduce realistic FL behavior, FedScale contains a scalable and extensible runtime. It provides high-level APIs to implement FL algorithms, deploy them at scale across diverse hardware and software backends, and evaluate them at scale, all with minimal developer efforts. We combine the two to perform systematic benchmarking experiments and highlight potential opportunities for heterogeneity-aware co-optimizations in FL. FedScale is open-source and actively maintained by contributors from different institutions at http://fedscale.ai. We welcome feedback and contributions from the community.

1. Introduction

Federated learning (FL) is an emerging machine learning (ML) setting where a logically centralized coordinator orchestrates many distributed clients (e.g., smartphones or laptops) to collaboratively train or evaluate a model (Bonawitz et al., 2019; Kairouz et al., 2021b) (Figure 1). It enables model training and evaluation on end-user data, while circumventing high cost and privacy risks in gathering the raw data from clients, with applications across diverse ML tasks.

In the presence of heterogeneous execution speeds of client devices as well as non-IID data distributions, existing efforts have focused on optimizing different aspects of FL: (1) System efficiency: reducing computation load (e.g., using smaller models (Sandler et al., 2018)) or communication traffic (e.g., local SGD (McMahan et al., 2017)) to achieve shorter round duration; (2) Statistical efficiency: designing data heterogeneity-aware algorithms (e.g., client clustering (Ghosh et al., 2020b)) to obtain better training accuracy with fewer training rounds; (3) Privacy and security: developing reliable strategies (e.g., differentially private training (Kairouz et al., 2021a)) to make FL more privacy-preserving and robust to potential attacks.

A comprehensive benchmark to evaluate an FL solution must investigate its behavior under the practical FL setting with (1) data heterogeneity and (2) device heterogeneity under (3) heterogeneous connectivity and (4) availability conditions at (5) multiple scales on a (6) broad variety of ML tasks. While the first two aspects are oft-mentioned in the literature (Li et al., 2020), realistic network connectivity and the availability of client devices can affect both types of heterogeneity (e.g., distribution drift (Eichner et al., 2019)), impairing model convergence. Similarly, evaluation at a large scale can expose an algorithm’s robustness, as practical FL deployment often runs across thousands of concurrent participants out of millions of clients (Yang et al., 2018). Overlooking any one aspect can mislead FL evaluation ($\S$2).

Unfortunately, existing FL benchmarks often fall short across multiple dimensions (Table 1). First, they are limited in the versatility of data for various real-world FL applications. Indeed, even though they may have quite a few
datasets and FL training tasks (e.g., LEAF (Caldas et al., 2019)), their datasets often synthetically generated partitions derived from conventional datasets (e.g., CIFAR) and do not represent realistic characteristics. This is because these benchmarks are mostly borrowed from traditional ML benchmarks (e.g., MLPerf (Mattson et al., 2020)) or designed for simulated FL environments like TensorFlow Federated (TFF) (tff) or PySyft (pys). Second, existing benchmarks often overlook system speed, connectivity, and availability of the clients (e.g., FedML (He et al., 2020) and Flower (Beutel et al., 2021)). This discourages FL efforts from considering system efficiency and leads to overly optimistic statistical performance (§2). Third, their datasets are primarily small-scale, because their experimental environments are unable to emulate large-scale FL deployments. While real FL often involves thousands of participants in each training round (Kairouz et al., 2021b; Yang et al., 2018), most existing benchmarking platforms can merely support the training of tens of participants per round. Finally, most of them lack user-friendly APIs for automated integration, resulting in great engineering efforts for benchmarking at scale. We attach a detailed comparison of existing benchmarks against FedScale in Appendix C.

**Contributions** We introduce an FL benchmark and accompanying runtime, FedScale, to enable comprehensive and standardized FL evaluations:

- To the best of our knowledge, FedScale presents the most comprehensive collection of FL datasets for evaluating different aspects of real FL deployments. It currently has 20 realistic FL datasets with small, medium, and large scales for a wide variety of task categories, such as image classification, object detection, word prediction, speech recognition, and reinforcement learning. To account for practical client behaviors, we include real-world measurements of mobile devices and associate each client with their computation and communication speeds, as well as the availability status over time.
- We present an automated evaluation platform, FedScale Runtime, to simplify and standardize FL evaluation in more realistic settings. FedScale Runtime provides a mobile backend to enable on-device FL evaluation and a cluster backend to benchmark various practical FL metrics (e.g., real client round duration) on GPUs/CPUs using real FL statistical and system datasets. The cluster backend can efficiently train thousand of clients in each round on a handful of GPUs. FedScale Runtime is also extensible, allowing easy deployment of new algorithms and ideas with flexible APIs.
- We perform systematic experiments to show how FedScale facilitates comprehensive FL benchmarking and highlight the pressing need for co-optimizing system and statistical efficiency, especially in tackling system stragglers, accuracy bias, and device energy trade-offs.

2. Background

**Existing efforts toward practical FL** To tackle heterogeneous client data, FedProx (Li et al., 2020), FedYoGi (Reddi et al., 2020) and Scaffold (Karimireddy et al., 2020) introduce adaptive client/server optimizations that use control variates to account for the ‘drift’ in model updates. Instead of training a single global model, some efforts enforce guided client selection (Lai et al., 2021), train a mixture of models (Shi et al., 2021; Fallah et al., 2020), or cluster clients over training (Ghosh et al., 2020a); To tackle the scarce and heterogeneous device resource (Lai et al., 2020), FedAvg (McMahan et al., 2017) reduces communication cost by performing multiple local SGD steps, while some works compress the model update by filtering out or quantizing unimportant parameters (Rothchild et al., 2020; Karimireddy et al., 2019); After realizing the privacy risk in FL (Geiping et al., 2020; Wang et al., 2020), DP-SGD (Geyer et al., 2017) enhances the privacy by employing differential privacy, and DP-FTRL (Kairouz et al., 2021a) applies the tree aggregation to add noise to the sum of mini-batch gradients. These FL efforts often navigate accuracy-computation-privacy trade-offs. As such, a realistic FL setting is crucial for comprehensive evaluations.

### Table 1. Comparing FedScale with existing FL benchmarks and libraries.

<table>
<thead>
<tr>
<th>Features</th>
<th>LEAF</th>
<th>TFF</th>
<th>FedML</th>
<th>Flower</th>
<th>FedScale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heter. Client Dataset</td>
<td>✔️</td>
<td>X</td>
<td>✔️</td>
<td>X</td>
<td>✔️</td>
</tr>
<tr>
<td>Heter. System Speed</td>
<td>X</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Client Availability</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Scalable Platform</td>
<td>X</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Real FL Runtime</td>
<td>X</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Flexible APIs</td>
<td>X</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>

*Figure 2. Existing benchmarks can be misleading. We train ShuffleNet on OpenImage classification ( Detailed setup in Section 5).*

---

**Figure 2.** Existing benchmarks can be misleading. We train ShuffleNet on OpenImage classification (Detailed setup in Section 5).
We next introduce how we curate realistic datasets in FedScale to fulfill the desired properties of FL datasets.

### 3. FedScale Dataset: Realistic FL Workloads

We next introduce how we curate realistic datasets in FedScale to fulfill the desired properties of FL datasets.

#### 3.1. Client Statistical Dataset

FedScale currently has 20 realistic FL datasets (Table 2) across diverse practical FL scenarios and disciplines. For example, Puffer dataset (Yan et al., 2020) is from FL video streaming deployed to edge users over the Internet. The raw data of FedScale datasets are collected from different sources and stored in various formats. We clean up the raw data, partition them into new FL datasets, streamline new datasets into consistent formats, and categorize them into different FL use cases. Moreover, FedScale provides standardized APIs, a Python package, for the user to easily leverage these datasets (e.g., using different distributions of the same data or new datasets) in other frameworks.

#### Realistic data and partitions

We target realistic datasets with client information, and partition the raw dataset using the unique client identification. For example, OpenImage is a vision dataset collected by Flickr, wherein different mobile users upload their images to the cloud for public use. We use the AuthorProfileUrl attribute of the OpenImage data to map data instances to each client, whereby we extract the realistic distribution of the raw data. Following the practical FL deployments (Yang et al., 2018), we assign the clients of each dataset into the training, validation and testing set. Here, we pick four real-world datasets – video (Charades), audio (Google Speech), image (OpenImage), and text (Reddit) – to illustrate practical FL characteristics. Each dataset consists of hundreds or up to millions of clients and millions of data points. Figure 3 reports the Probability Density Function (PDF) of the data distribution, wherein we see a high statistical deviation (e.g., wide distribution of the density) across clients not only in the quantity of samples (Figure 3(a)) but also in the data distribution (Figure 3(b)).

#### Existing FL benchmarks can be misleading

Existing benchmarks often lack realistic client statistical and system behavior datasets and/or fail to reproduce large-scale FL deployments. As a result, they are not only insufficient for benchmarking diverse FL optimizations but can even mislead performance evaluations. For example, (1) As shown in Figure 2(a), statistical performance becomes worse when encountering realistic client behavior (e.g., training failures and availability dynamics), which indicates existing benchmarks that do not have systems traces can produce overly optimistic statistical performance; (2) FL training with hundreds of participants each round performs better than that with tens of participants (Figure 2(b)). As such, existing benchmark platforms can under-report FL optimizations as they cannot support the practical FL scale with a large number of participants.

### FedScale: Benchmarking Model and System Performance of Federated Learning at Scale

#### Table 2. Statistics of partial FedScale datasets (the full list with more details is available in Appendix B). Currently, FedScale has 20 real-world federated datasets: each dataset is partitioned by its real client-data mapping, and we have removed sensitive information in these datasets.

<table>
<thead>
<tr>
<th>Category</th>
<th>Name</th>
<th>Data Type</th>
<th>#Clients</th>
<th>#Instances</th>
<th>Example Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV</td>
<td>OpenImage</td>
<td>Image</td>
<td>13,771</td>
<td>1.3M</td>
<td>Classification, Object detection</td>
</tr>
<tr>
<td></td>
<td>Google Landmark</td>
<td>Image</td>
<td>43,484</td>
<td>3.6M</td>
<td>Classification</td>
</tr>
<tr>
<td></td>
<td>Charades</td>
<td>Video</td>
<td>266</td>
<td>10K</td>
<td>Action recognition</td>
</tr>
<tr>
<td></td>
<td>VLOG</td>
<td>Video</td>
<td>4,900</td>
<td>9.6K</td>
<td>Classification, Object detection</td>
</tr>
<tr>
<td></td>
<td>Waymo Motion</td>
<td>Video</td>
<td>496,358</td>
<td>32.5M</td>
<td>Motion prediction</td>
</tr>
<tr>
<td>NLP</td>
<td>Europarl</td>
<td>Text</td>
<td>27,835</td>
<td>1.2M</td>
<td>Text translation</td>
</tr>
<tr>
<td></td>
<td>Reddit</td>
<td>Text</td>
<td>1,660,820</td>
<td>351M</td>
<td>Word prediction</td>
</tr>
<tr>
<td></td>
<td>LibriTTS</td>
<td>Text</td>
<td>2,456</td>
<td>37K</td>
<td>Text to speech</td>
</tr>
<tr>
<td></td>
<td>Google Speech</td>
<td>Audio</td>
<td>2,618</td>
<td>105K</td>
<td>Speech recognition</td>
</tr>
<tr>
<td></td>
<td>Common Voice</td>
<td>Audio</td>
<td>12,976</td>
<td>1.1M</td>
<td>Speech recognition</td>
</tr>
<tr>
<td>Misc ML</td>
<td>Taobao</td>
<td>Text</td>
<td>182,806</td>
<td>20.9M</td>
<td>Recommendation</td>
</tr>
<tr>
<td></td>
<td>Puffer Streaming</td>
<td>Text</td>
<td>121,551</td>
<td>15.4M</td>
<td>Sequence prediction</td>
</tr>
<tr>
<td></td>
<td>Fox Go</td>
<td>Text</td>
<td>150,333</td>
<td>4.9M</td>
<td>Reinforcement learning</td>
</tr>
</tbody>
</table>

We report the pairwise Jensen–Shannon distance of the label distribution between two clients.

---

1 We report the pairwise Jensen–Shannon distance of the label distribution between two clients.
stragglers are paramount.

**Client device availability is dynamic** We incorporate a large-scale user behavior dataset spanning across 136k users (Yang et al., 2021) to emulate the behaviors of clients. It includes 180 million trace items of client devices (e.g., battery charge or screen lock) over a week. We follow the real FL setting, which considers the device in charging to be available (Bonawitz et al., 2019) and observe great dynamics in their availability: (i) the number of available clients reports diurnal variation (Figure 5(a)). This confirms the cyclic patterns in the client data, which can deteriorate the statistical performance of FL (Eichner et al., 2019). (ii) the duration of each available slot is not long-lasting (Figure 5(b)). This highlights the need of handling failures (e.g., clients become offline) during training, as the round duration (also a few minutes) is comparable to that of each available slot. This, however, is largely overlooked in the literature.

4. **FedScale Runtime: Evaluation Platform**

Existing FL evaluation platforms are poor at reproducing practical, large-scale FL deployment scenarios. Worse, they often lack user-friendly APIs and require significant developer efforts to introduce new plugins. We introduce, FedScale Runtime, an automated, extensible, and easily-deployable evaluation platform equipped with mobile and cluster backends, to simplify and standardize FL evaluation under realistic settings.

4.1. **FedScale Runtime: Mobile Backend**

FedScale Runtime deploys a mobile backend to enable on-device FL evaluation on smartphones. The first principle in building our mobile backend is to minimize any engineering effort for the developer (e.g., without reinventing their Python code) to benchmark FL on mobiles. To this end, FedScale mobile backend (Singapuram et al., 2022) is built atop the Termux app (ter), an Android terminal that supports Linux environment.

```python
from fedscale.core.client import Client
class Mobile_Client(Client):
    def train(self,client_data,model,conf):
        for local_step in range(conf.local_steps):
            optimizer.zero_grad()
            loss.backward()
            optimizer.step()
        # Results will be sent to cloud aggregator via gRPC
        return gradient_update
```

**Figure 6.** Training on mobile client.

Figure 6 shows a snippet of code running on FedScale mobile backend. By integrating with Termux, FedScale Runtime allows the developer to run an unmodified version...
of Python script (e.g., PyTorch) built from source on the mobile device; the full-operator set (e.g., PyTorch modules) is available too. This speeds up the deployment cycle: FL models and algorithms that were prototyped on server GPUs/CPUs can also be deployed using FedScale Runtime. We are currently implementing the Google Remote Procedure Call (gRPC) for distributed mobile devices to interface with FedScale Runtime cloud server.

**Benchmarking the mobile backend** FedScale mobile backend enables developers to benchmark realistic FL training/testing performance on mobile phones. For example, Figure 7 reports the performance metrics of training ShuffleNet and ResNet34 on one mini-batch (batch size 32), drawn from the ImageNet and CIFAR-10 datasets, on Xiaomi Mi10 and Samsung S10e Android devices. We benchmark the average training time. We notice that ResNet34 runs at higher instantaneous power than ShuffleNet on both devices, but it requires less total energy to train since it takes shorter latency. ImageNet takes longer than CIFAR-10 per mini-batch, as the larger training image sizes lead to longer execution. The heterogeneity in computational capacity is evident as the Xiaomi Mi10 device outperforms the Samsung S10e device due to a more capable processor. As such, we believe that FedScale mobile backend can facilitate future on-device FL optimizations (e.g., hardware-aware neural architecture search (He et al., 2018)).

4.2. FedScale Runtime: Cluster Backend

FedScale Runtime provides an automated cluster backend that can support FL evaluations in real deployments and in-cluster simulations. In the deployment mode, FedScale Runtime acts as the cloud aggregator and orchestrates FL executions across real devices (e.g., laptops, mobiles, or even cloud servers). To enable cost-efficient FL benchmarking, FedScale Runtime also includes a simulation mode that performs FL training/testing on GPUs/CPUs, while providing various practical FL metrics by emulating realistic FL behaviors, such as computation/communication cost, latency and wall clock time. To the best of our knowledge, FedScale Runtime is the first platform that enables FL benchmarking with practical FL runtime on GPUs/CPUs. More importantly, FedScale Runtime can run the same code with little changes in both modes to minimize the migration overhead.

Throughout the rest of the paper, we focus on the simulation mode as benchmarking is the primary focus of this paper.

**Overview of the simulation mode** FedScale Runtime consists of three primary components (Figure 8):

- **Aggregator Simulator**: It acts as the aggregator in practical FL, which selects participants, distributes execution profiles (e.g., model weight), and handles result (e.g., model updates) aggregation. In each round, its client manager uses the client behavior trace to monitor whether a client is available; then it selects the specified number of clients to participate that round. Once receiving new events, the event monitor activates the handler (e.g., aggregation handler to perform model aggregation), or the gRPC communicator to send/receive messages. The communicator records the size (cost) of every network traffic, and its runtime latency in FL wall-clock time ($t_{client\_bandwidth}$).

- **Client Simulator**: It works as the FL client. FedScale data loader loads the federated dataset of that client and feeds this data to the compute engine to run real training/testing. The computation latency is determined by ($\#_{\text{processed\_sample}} \times \text{latency\_per\_sample}$), and the communicator handles network traffic and records the communication latency ($t_{\text{traffic\_size}} \times t_{\text{traffic\_size}}$). The device monitor will terminate the simulation of a client if the current FL runtime exceeds his available slot (e.g.,...
FedScale: Benchmarking Model and System Performance of Federated Learning at Scale

FedScale Runtime enables automated FL simulation
FedScale Runtime incorporates realistic FL traces, using the aforementioned trace by default or the developer-specified profile from the mobile backend, to automatically emulate the practical FL workflow: ① Task submission: FL developers specify their configurations (e.g., model and dataset), which can be federated training or testing, and the resource manager will initiate the aggregator and client simulator on available resource (GPU, CPU, other accelerators, or even smartphones); ② FL simulation: Following the standardized FL lifecycle (Figure 1), in each training round, the aggregator inquires the client manager to select participants, whereby the resource manager distributes the client configuration to the available client simulators. After the completion of each client, the client simulator pushes the model update to the aggregator, which then performs the model aggregation. ③ Metrics output: During training, the developer can query the practical evaluation metrics on the fly. Figure 8 lists some popular metrics in FedScale.

FedScale Runtime is easily deployable and extensible with plugins
FedScale Runtime provides flexible APIs, which can accommodate with different execution backends (e.g., PyTorch) by design, for the developer to quickly benchmark new plugins. Table 3 illustrates some example APIs that can facilitate diverse FL efforts, and Figure 9 dictates an example showing how these APIs help to benchmark a new design of local client training with a few lines of code by inheriting the base Client module. Moreover, FedScale Runtime can embrace new realistic (statistical client or system behavior) datasets with the built-in APIs. For example, the developer can import his own dataset of the client availability trace (e.g., PyTorch) by design, for the developer to quickly benchmark new plugins. Table 3 illustrates some example APIs that can facilitate diverse FL efforts, and Figure 9 dictates an example showing how these APIs help to benchmark a new design of local client training with a few lines of code by inheriting the base Client module. Moreover, FedScale Runtime can embrace new realistic (statistical client or system behavior) datasets with the built-in APIs. For example, the developer can import his own dataset of the client availability trace (Yu & Chowdhury, 2020) to divide GPU among tasks so that multiple client simulators can co-locate

Table 3. Some example APIs. FedScale provides APIs to deploy new plugins for various designs. We omit input arguments for brevity here.

<table>
<thead>
<tr>
<th>Module</th>
<th>API Name</th>
<th>Example Use Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregator</td>
<td>round_completion_handler()</td>
<td>Adaptive/secure aggregation</td>
</tr>
<tr>
<td>Simulator</td>
<td>client_completion_handler()</td>
<td>Straggler mitigation</td>
</tr>
<tr>
<td>Client Manager</td>
<td>select_participants()</td>
<td>Client selection</td>
</tr>
<tr>
<td></td>
<td>select_model_for_client()</td>
<td>Adaptive model selection</td>
</tr>
<tr>
<td>Client</td>
<td>train()</td>
<td>Local SGD/malicious attack</td>
</tr>
<tr>
<td>Simulator</td>
<td>serialize_results()</td>
<td>Model compression</td>
</tr>
</tbody>
</table>

Note that capturing runtime performance (e.g., wall clock time) is rather slow and expensive in practical FL – each mobile device takes several minutes to train a round – but our simulator enables fast-forward simulation, as training on CPUs/GPUs takes only a few seconds per round, while providing simulated runtime using realistic traces.

FedScale Runtime is scalable and efficient
In the simulation mode, FedScale Runtime can perform large-scale simulations (thousands of clients per round) in both standalone (single CPU/GPU) and distributed (multiple machines) settings. This is because: (1) FedScale Runtime uses GPU sharing techniques (Yu & Chowdhury, 2020) to divide GPU among tasks so that multiple client simulators can co-locate

from fedscale.core.client import Client

class Customized_Client(Client):
  # Redefine training (e.g., for local SGD/gradient compression)
  def train(self,client_data,model,conf):
    # Code of plugin
    ...  
    # Results will be serialized, and then sent to aggregator
    return training_result

Figure 9. Add plugins by inheritance.

Figure 10. FedScale Runtime can run thousands of clients/round on 10 GPUs where others fail. More results are in Appendix C.

# Results will be serialized, and then sent to aggregator
return training_result

Figure 10. FedScale Runtime can run thousands of clients/round on 10 GPUs where others fail. More results are in Appendix C.
FedScale: Benchmarking Model and System Performance of Federated Learning at Scale

<table>
<thead>
<tr>
<th>Task</th>
<th>Dataset</th>
<th>Model</th>
<th>IID</th>
<th>FedAvg</th>
<th>FedProx</th>
<th>FedYoGi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Classification</td>
<td>FEMNIST</td>
<td>ResNet-18</td>
<td>86.40%</td>
<td>78.50%</td>
<td>78.40%</td>
<td>76.30%</td>
</tr>
<tr>
<td></td>
<td>OpenImage</td>
<td>ShuffleNet-V2</td>
<td>81.37%</td>
<td>70.27%</td>
<td>69.54%</td>
<td>74.04%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MobileNet-V2</td>
<td>80.83%</td>
<td>70.09%</td>
<td>70.34%</td>
<td>75.25%</td>
</tr>
<tr>
<td>Text Classification</td>
<td>Amazon Review</td>
<td>Logistic Regression</td>
<td>66.10%</td>
<td>65.80%</td>
<td>65.10%</td>
<td>65.30%</td>
</tr>
<tr>
<td>Language Modeling</td>
<td>Reddit</td>
<td>Albert</td>
<td>73.5 perplexity (ppl)</td>
<td>77.3 ppl</td>
<td>76.6 ppl</td>
<td>81.6 ppl</td>
</tr>
<tr>
<td>Speech Recognition</td>
<td>Google Speech</td>
<td>ResNet-34</td>
<td>72.58%</td>
<td>63.37%</td>
<td>63.25%</td>
<td>62.67%</td>
</tr>
</tbody>
</table>

Table 4. Benchmarking of different FL algorithms across realistic FL datasets. We report the mean test accuracy over 5 runs.

(a) Convergence on Google speech.

(b) Convergence on OpenImage.

(c) Final model performance.

Figure 11. FedScale can benchmark the statistical FL performance. (c) shows existing benchmarks can under-report the FedYoGi performance as they cannot support a large number of participants.

5. Experiments

In this section, we show how FedScale can facilitate better benchmarking of FL efforts over its counterparts.

Experimental setup We use 10 NVIDIA Tesla P100 GPUs in our evaluations. Following the real FL deployments (Bonawitz et al., 2019; Yang et al., 2018), the aggregator collects updates from the first $N$ completed participants out of $1.3N$ participants to mitigate system stragglers in each round, and $N = 100$ by default. We experiment with representative FedScale datasets in different scales and tasks (detailed experiment setup in Appendix A).

5.1. How Does FedScale Help FL Benchmarking?

Existing benchmarks are insufficient to evaluate the various metrics needed in today’s FL. We note that the performance of existing benchmarks and FedScale are quite close in the same settings if we turn off the optional system traces in FedScale. Because the underlying training and FL protocols in evaluations are the same. However, the limited scalability can mislead the practical FL performance. Next, we show the effectiveness of FedScale in benchmarking different FL aspects over its counterparts.

Benchmarking FL statistical efficiency. FedScale provides various realistic client datasets to benchmark the FL statistical efficiency. Here, we experiment with state-of-the-art optimizations (FedAvg, FedProx and FedYoGi) – each aims to mitigate the data heterogeneity – and the traditional IID data setting. Figure 11 and Table 4 report that: (1) the round-to-accuracy performance and final model accuracy of the non-IID setting is worse than that of the IID setting, which is consistent with existing findings (Kairouz et al., 2021b); (2) different tasks can have different preferences.
on the optimizations. For example, FedYoGi performs the best on OpenImage, but it is inferior to FedAvg on Google Speech. With much more FL datasets, FedScale enables extensive studies of the sweet spot of different optimizations; and (3) existing benchmarks can under-report the FL performance due to their inability to reproduce the FL setting. Figure 11(c) reports the final model accuracy using FedML and FedScale, where we attempt to reproduce the scale of practical FL with 100 participants per round in both frameworks, but FedML can only support 30 participants because of its suboptimal scalability, which under-reports the FL performance that the algorithm can indeed achieve.

Benchmarking FL system efficiency. Existing system optimizations for FL focus on the practical runtime (e.g., wall-clock time in real FL training) and the FL execution cost. Unfortunately, existing benchmarks can hardly evaluate the FL runtime due to the lack of realistic system traces, but we now show how FedScale can help such benchmarking: (1) FedScale Runtime enables fast-forward evaluations of the practical FL wall-clock time with fewer evaluation hours. Taking different number of local steps \( K \) in local SGD as an example (McMahan et al., 2017), Figure 12(a) and Table 12(b) illustrate that FedScale can evaluate this impact of \( K \) on practical FL runtime in a few hours. This allows the developer to evaluate large-scale system optimizations efficiently; and (2) FedScale Runtime can dictate the FL execution cost by using realistic system traces. For example, Figure 12(c) reports the practical FL communication cost in achieving the performance of Figure 11, while Figure 15 reports the system duration of individual clients. These system metrics can facilitate developers to navigate the accuracy-cost trade-off.

Benchmarking FL privacy and security. FedScale can evaluate the statistical and system efficiency for privacy and security optimizations more realistically. Here, we give an example of benchmarking DP-SGD (Geyer et al., 2017; Kairouz et al., 2021a), which applies differential privacy to improve the client privacy. We experiment with different privacy targets \( \sigma \) (\( \sigma=0 \) indicates no privacy enhancement) and different number of participants per round \( N \). Figure 13 shows that the scale of participants (e.g., \( N=30 \)) that today’s benchmarks can support can mislead the privacy evaluations too: for \( \sigma=0.01 \), while we notice great performance degradation (12.8%) in the final model accuracy when \( N=30 \), this enhancement is viable in practical FL (\( N=100 \)) with decent accuracy drop (4.6%). Moreover, FedScale is able to benchmark more practical FL metrics, such as wall-clock time, communication cost added in privacy optimizations, and the number of rounds needed to leak the client privacy under realistic individual client data and Non-IID distributions.

As for benchmarking FL security, we follow the example setting of recent backdoor attacks (Sun et al., 2019; Wang et al., 2020) on the OpenImage, where corrupted clients flip their ground-truth labels to poison the training. We benchmarked two settings: one without security enhancement, while the other clips the model updates as (Sun et al., 2019). As shown in Figure 14, while state-of-the-art optimizations report this can mitigate the attacks without hurting the overall performance on their synthesized datasets, large accuracy drops can occur in more practical FL settings.

5.2. Opportunities for Future FL Optimizations

Next, we show FedScale can shine light on the need for yet unexplored optimizations owing to its realistic FL settings.

Heterogeneity-aware co-optimizations of communication and computation. Existing optimizations for the system efficiency often apply the same strategy on all clients (e.g., using the same number of local steps (McMahan et al., 2017) or compression threshold (Rothchild et al., 2020)), while ignoring the heterogeneous client system speed. When we outline the timeline of 5 randomly picked participants in training of the ShuffleNet (Figure 15), we find: (1) system
stragglers can greatly slow down the round aggregation in practical FL; and (2) simply optimizing the communication or computation efficiency may not lead to faster rounds, as the last participant can be bottlenecked by the other resource. Here, optimizing the communication can greatly benefit Client 4, but it achieves marginal improvement on the round duration as Client 5 is computation-bound. This implies an urgent need of heterogeneity-aware co-optimizations of communication and computation efficiency.

Co-optimizations of statistical and system efficiency
Most of today’s FL efforts focus on either optimizing the statistical or the system efficiency, whereas we observe a great opportunity to jointly optimize both efficiencies: (1) As the system behavior determines the availability of client data, predictable system performance can benefit statistical efficiency. For example, in alleviating the biased model accuracy (Figure 16), we may prioritize the use of upcoming offline clients to curb the upcoming distribution drift of client data; (2) Statistical optimizations should be aware of the heterogeneous client system speed. For example, instead of applying one-fit-all strategies (e.g., local steps or gradient compression) for all clients, faster workers can trade more system latency for more statistical benefits (e.g., transferring more traffic with less intensive compression).

FL design-decisions considering mobile environment
Existing efforts have largely overlooked the interplay of client devices and training speed (e.g., using a large local steps to save communication (McMahan et al., 2017)), however, as shown in Figure 7, running intensive on-device computation for a long time can quickly drain the battery, or even burn the device, leading to the unavailability of clients. Therefore, we believe that a power and temperature-aware training algorithm (e.g., different local steps across clients or device-aware NAS) can be an important open problem.

6. Conclusion
To enable scalable and reproducible FL research, we introduce FedScale, a diverse set of realistic FL datasets in terms of scales, task categories, and client system behaviors, along with a scalable and extensible evaluation platform, FedScale Runtime. FedScale Runtime performs fast-forward evaluation of the practical FL runtime metrics needed in today’s works. More importantly, FedScale Runtime provides ready-to-use realistic datasets and flexible APIs to allow more FL applications, such as benchmarking NAS, model inference, and a broader view of federated computation (e.g., multi-party computation). FedScale is open-source at http://fedscale.ai, and we hereby invite the community to develop and contribute state-of-the-art FL efforts.

Acknowledgments
We would like to thank the anonymous reviewers and SybionticLab members for their insightful feedback. We also thank FedScale contributors and users from many different academic institutions and industry for their valuable inputs. This work was supported in part by NSF grants CNS-1900665, CNS-1909067, and CNS-2106184, and a grant from Cisco.

References


MobiPerf. https://www.measurementlab.net/tests/mobiperf/.


Reddit Comment Data. https://files.pushshift.io/reddit/comments/.
FedScale: Benchmarking Model and System Performance of Federated Learning at Scale


FedScale: Benchmarking Model and System Performance of Federated Learning at Scale


A. Experiment Setup

Scalability Evaluations We evaluate the scalability of FedScale Runtime, FedML (GitHub commit@2ee0517) and Flower (v0.17.0 atop Ray v1.9.2) using a cluster with 10 GPU nodes. Each GPU node has a P100 GPU with 12GB GPU memory and 192GB RAM. We train the ShuffleNet-V2 model on the OpenImage dataset. We set the minibatch size of each participant to 32, and the number of local steps $K$ to 20, which takes around 2800MB GPU memory for each model training. As such, we allow each GPU node to run 4 processes in benchmarking these three frameworks.

Evaluation Setup Applications and models used in our evaluations are widely used on mobile devices. We set the minibatch size of each participant to 32, and the number of local steps $K$ to 20. We cherry-pick the hyper-parameters with grid search, ending up with an initial learning rate 0.04 for CV tasks and 5e-5 for NLP tasks. The learning rate decays by 0.98 every 10 training rounds. These settings are consistent with the literature. More details about the input dataset are available in Appendix B.

B. Introduction of FedScale Datasets

FedScale currently has 20 realistic federated datasets across a wide range of scales and task categories (Table 5). Here, we provide the description of some representative datasets.

Google Speech Commands. A speech recognition dataset (Warden, 2018) with over ten thousand clips of one-second-long duration. Each clip contains one of the 35 common words (e.g., digits zero to nine, "Yes", "No", "Up", "Down") spoken by thousands of different people.

OpenImage. OpenImage (ope) is a vision dataset collected from Flickr, an image and video hosting service. It contains a total of 16M bounding boxes for 600 object classes (e.g., Microwave oven). We clean up the dataset according to the provided indices of clients.

Reddit and StackOverflow. Reddit (red) (StackOverflow (sta)) consists of comments from the Reddit (StackOverflow) website. It has been widely used for language modeling tasks, and we consider each user as a client. In this dataset, we restrict to the 30k most frequently used words, and represent each sentence as a sequence of indices corresponding to these 30k frequently used words.

VLOG. VLOG (Fouhey et al., 2018) is a video dataset collected from YouTube. It contains more than 10k Lifestyle Vlogs, videos that people purportedly record to show their lives, from more than 4k actors. This dataset is aimed at understanding everyday human interaction and contains labels for scene classification, hand-state prediction, and hand detection tasks.

LibriTTS. LibriTTS (Zen et al., 2019) is a large-scale text-to-speech dataset. It is derived from audiobooks in LibriVox project. There are 585 hours of read English speech from 2456 speakers at a 24kHz sampling rate.

Taobao. Taobao Dataset (tao) is a dataset of click rate prediction about display Ad, which is displayed on the website of Taobao. It is composed of 1,140,000 users ad display/click logs for 8 days, which are randomly sampled from the website of Taobao.

Waymo Motion. Waymo Motion (Ettinger et al., 2021) is composed of 103,354 segments each containing 20 seconds of object tracks at 10Hz and map data for the area covered by the segment. These segments are further broken into 9 second scenarios, and we consider each scenario as a client.

Puffer Streaming. Puffer is a Stanford University research study about using machine learning (e.g., reinforcement learning (Mao et al., 2017)) to improve video-streaming algorithms: the kind of algorithms used by services such as YouTube, Netflix, and Twitch. Puffer dataset (Yan et al., 2020) consists of 15.4M sequences of network throughput on edge clients over time.

C. Comparison with Existing FL Benchmarks

In this section, we compare FedScale with existing FL benchmarks in more details.

Data Heterogeneity Existing benchmarks for FL are mostly limited in the variety of realistic datasets for real-world FL applications. Even they have various datasets (e.g., LEAF (Caldas et al., 2019)) and FedEval (Chai et al., 2020)), their datasets are mostly synthetically derived from conventional datasets (e.g., CIFAR) and limited to quite a few FL tasks. These statistical client datasets can not represent realistic characteristics and are inefficient to benchmark various real FL applications. Instead, FedScale provides 20 comprehensive realistic datasets for a wide variety of tasks and across small, medium, and large scales, which can also be used in data analysis to motivate more FL designs.

System Heterogeneity The practical FL statistical performance also depends on the system heterogeneity (e.g., internet bandwidth (Lai et al., 2018) and client availability), which has inspired lots of optimizations for FL system efficiency. However, existing FL benchmarks have largely overlooked the system behaviors of FL clients, which can produce misleading evaluations, and discourage the benchmarking of system efforts. To emulate the heterogeneous
FedScale: Benchmarking Model and System Performance of Federated Learning at Scale

<table>
<thead>
<tr>
<th>Category</th>
<th>Name</th>
<th>Data Type</th>
<th>#Clients</th>
<th>#Instances</th>
<th>Example Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV</td>
<td>iNature (ina)</td>
<td>Image</td>
<td>2,295</td>
<td>193K</td>
<td>Classification</td>
</tr>
<tr>
<td></td>
<td>FEMNIST (Cohen et al., 2017)</td>
<td>Image</td>
<td>3,400</td>
<td>640K</td>
<td>Classification</td>
</tr>
<tr>
<td></td>
<td>OpenImage (ope)</td>
<td>Image</td>
<td>13,771</td>
<td>1.3M</td>
<td>Classification, Object detection</td>
</tr>
<tr>
<td></td>
<td>Google Landmark (Weyand et al., 2020)</td>
<td>Image</td>
<td>43,484</td>
<td>3.6M</td>
<td>Classification</td>
</tr>
<tr>
<td></td>
<td>Charades (Sigurdsson et al., 2016)</td>
<td>Video</td>
<td>266</td>
<td>10K</td>
<td>Action recognition</td>
</tr>
<tr>
<td></td>
<td>VLOG (Fouhey et al., 2018)</td>
<td>Video</td>
<td>4,900</td>
<td>9.6K</td>
<td>Classification, Object detection</td>
</tr>
<tr>
<td></td>
<td>Waymo Motion (Ettinger et al., 2021)</td>
<td>Video</td>
<td>496,358</td>
<td>32.5M</td>
<td>Motion prediction</td>
</tr>
<tr>
<td>NLP</td>
<td>Europarl (Koehn, 2005)</td>
<td>Text</td>
<td>27,835</td>
<td>1.2M</td>
<td>Text translation</td>
</tr>
<tr>
<td></td>
<td>Blog Corpus (Schler et al., 2006)</td>
<td>Text</td>
<td>19,320</td>
<td>137M</td>
<td>Word prediction</td>
</tr>
<tr>
<td></td>
<td>Stackoverflow (sta)</td>
<td>Text</td>
<td>342,477</td>
<td>135M</td>
<td>Word prediction, Classification</td>
</tr>
<tr>
<td></td>
<td>Reddit (red)</td>
<td>Text</td>
<td>1,660,820</td>
<td>351M</td>
<td>Word prediction</td>
</tr>
<tr>
<td></td>
<td>Amazon Review (McAuley et al., 2015)</td>
<td>Text</td>
<td>1,822,925</td>
<td>166M</td>
<td>Classification, Word prediction</td>
</tr>
<tr>
<td></td>
<td>CoQA (Reddy et al., 2019)</td>
<td>Text</td>
<td>7,685</td>
<td>116K</td>
<td>Question Answering</td>
</tr>
<tr>
<td></td>
<td>LibriTTS (Zen et al., 2019)</td>
<td>Text</td>
<td>2,456</td>
<td>37K</td>
<td>Text to speech</td>
</tr>
<tr>
<td></td>
<td>Google Speech (Warden, 2018)</td>
<td>Audio</td>
<td>2,618</td>
<td>105K</td>
<td>Speech recognition</td>
</tr>
<tr>
<td>Misc ML</td>
<td>Common Voice (com)</td>
<td>Audio</td>
<td>12,976</td>
<td>1.1M</td>
<td>Speech recognition</td>
</tr>
<tr>
<td></td>
<td>Taxi Trajectory</td>
<td>Text</td>
<td>442</td>
<td>1.7M</td>
<td>Sequence prediction</td>
</tr>
<tr>
<td></td>
<td>Taobao (tao)</td>
<td>Text</td>
<td>182,806</td>
<td>20.9M</td>
<td>Recommendation</td>
</tr>
<tr>
<td></td>
<td>Puffer Streaming (Yan et al., 2020)</td>
<td>Text</td>
<td>121,551</td>
<td>15.4M</td>
<td>Sequence prediction</td>
</tr>
<tr>
<td></td>
<td>Fox Go (go-)</td>
<td>Text</td>
<td>150,333</td>
<td>4.9M</td>
<td>Reinforcement learning</td>
</tr>
</tbody>
</table>

Table 5. Statistics of FedScale datasets. FedScale has 20 realistic client datasets, which are from the real-world collection, and we partitioned each dataset using its real client-data mapping.

system behaviors in practical FL, FedScale incorporates real-world traces of mobile devices, associates each client with his system speeds, as well as the availability. Moreover, we develop a more efficient evaluation platform, FedScale Runtime, to automate FL benchmarking.

Scalability Existing frameworks, perhaps due to the heavy burden of building complicated system support, largely rely on the traditional ML architectures (e.g., the primitive parameter-server architecture of Pytorch). These architectures are primarily designed for the traditional large-batch training on a number of workers, and each worker often trains a single batch at a time. However, this is ill-suited to the simulation of thousands of clients concurrently: (1) they lack tailored system implementations to orchestrate the synchronization and resource scheduling, for which they can easily run into synchronization/memory issues and crash down; (2) their resource can be under-utilized, as FL evaluations often use a much smaller batch size.

Tackling all these inefficiencies requires domain-specific system designs. Specifically, we built a cluster resource scheduler that monitors the fine-grained resource utilization of machines, queues the overcommitted simulation requests, adaptively dispatches simulation requests of the client across machines to achieve load balance, and then orchestrates the simulation based on the client virtual clock. Moreover, given a much smaller batch size in FL, we maximize the resource utilization by overlapping the communication and computation phrases of different client simulations. The former and the latter make FedScale more scalable across machines and on single machines, respectively.

Empirically, we have run the 20-GPU set up on different datasets and models in Figure 10 and Table 6, and are aware of at least one group who ran FedScale Runtime with more than 60 GPUs (Lai et al., 2021).

<table>
<thead>
<tr>
<th>Eval. Duration/Round vs. # of Clients/Round</th>
<th>10</th>
<th>100</th>
<th>1K</th>
<th>10K</th>
</tr>
</thead>
<tbody>
<tr>
<td>FedScale</td>
<td>0.03 min</td>
<td>0.16 min</td>
<td>1.14 min</td>
<td>10.9 min</td>
</tr>
<tr>
<td>FedML</td>
<td>0.58 min</td>
<td>4.4 min</td>
<td>fail to run</td>
<td>fail to run</td>
</tr>
<tr>
<td>Flower</td>
<td>0.05 min</td>
<td>0.23 min</td>
<td>2.4 min</td>
<td>fail to run</td>
</tr>
</tbody>
</table>

Table 6. FedScale is more scalable and faster. Image classification on iNature dataset using MobileNet-V2 on 20-GPU setting.

Modularity As shown in Table 1, some existing frameworks (e.g., LEAF and FedEval) do not provide user-friendly modularity, which requires great engineering efforts to benchmark different components, and we recognize that FedML and Flower provide the API modularity too.

On the other hand, FedScale Runtime’s modularity for easy deployments and broader use cases is not limited to APIs (Figure 8): (1) FedScale Runtime Data Loader: it simplifies...
FedScale: Benchmarking Model and System Performance of Federated Learning at Scale

```python
from fedscale.core.client_manager import ClientManager
import Oort

class Customized_ClientManager(ClientManager):
    def __init__(self, *args):
        super().__init__(*args)
        self.oort_selector = Oort.create_training_selector(*args)

    # Replace default client selection algorithm w/ Oort
    def select_participants(self, numOfClients, cur_time, feedbacks):
        # Feed Oort w/ execution feedbacks
        oort_selector.update_client_info(feedbacks)
        selected_clients = oort_selector.select_participants(numOfClients, cur_time)
        return selected_clients
```

Figure 17. Evaluate client selection algorithm (Lai et al., 2021).

```python
from fedscale.core.client import Client

class Customized_Client(Client):
    # Customize the training on each client
    def train(self, client_data, model, conf):
        training_result = super().train(client_data, model, conf)

        # Implementation of compression
        compressed_result = compress_impl(training_result)
        return compressed_result
```

Figure 18. Evaluate model compression (Rothchild et al., 2020).

```python
from fedscale.core.client import Client

class Customized_Client(Client):
    # Customize the training on each client
    def train(self, client_data, model, conf):
        training_result = super().train(client_data, model, conf)

        # Clip updates and add noise
        secure_result = secure_impl(training_result)
        return secure_result
```

Figure 19. Evaluate security enhancement (Sun et al., 2019).

and expands the use of realistic datasets. e.g., developers can load and analyze the realistic FL data to motivate new algorithm designs, or imports new datasets/customize data distributions in FedScale evaluations; (2) Client simulator: it emulates the system behaviors of FL clients, and developers can customize their system traces in evaluating the FL system efficiency too; (3) Resource manager: it hides the system complexity in training large-scale participants.

D. Examples of New Plugins

In this section, we demonstrate examples to show the ease of integrating today’s FL efforts in FedScale evaluations.

At its core, FedScale Runtime provides flexible APIs on each module so that the developer can access and customize methods of the base class. Table 3 illustrates some example APIs that can facilitate diverse FL efforts. Note that FedScale Runtime will automatically integrate new plugins into evaluations, and then produces practical FL metrics. Figure 17 demonstrates that we can easily evaluate new client selection algorithms, Oort (Lai et al., 2021), by modifying a few lines of the `clientManager` module. Similarly, Figure 18 and Figure 19 show that we can extend the basic `Client` module to apply new gradient compression (Rothchild et al., 2020) and enhancement for malicious attack (Sun et al., 2019), respectively.

Comparison with other work Figure 20 shows the same evaluation of gradient compression (Rothchild et al., 2020) as that in Figure 18 using flower (Beutel et al., 2021), which requires much more human efforts than Figure 18. The gray components in figure 20 requires implementation. Flower falls short in providing APIs for passing metadata between client and server, for example client id, which makes server and running workers client-agnostic. To customized anything for clients during FL training, developers have to go through the source code and override many components to share client meta data between server and workers.