FedScale

Benchmarking Model and System Performance of Federated Learning At Scale

fedscale.ai

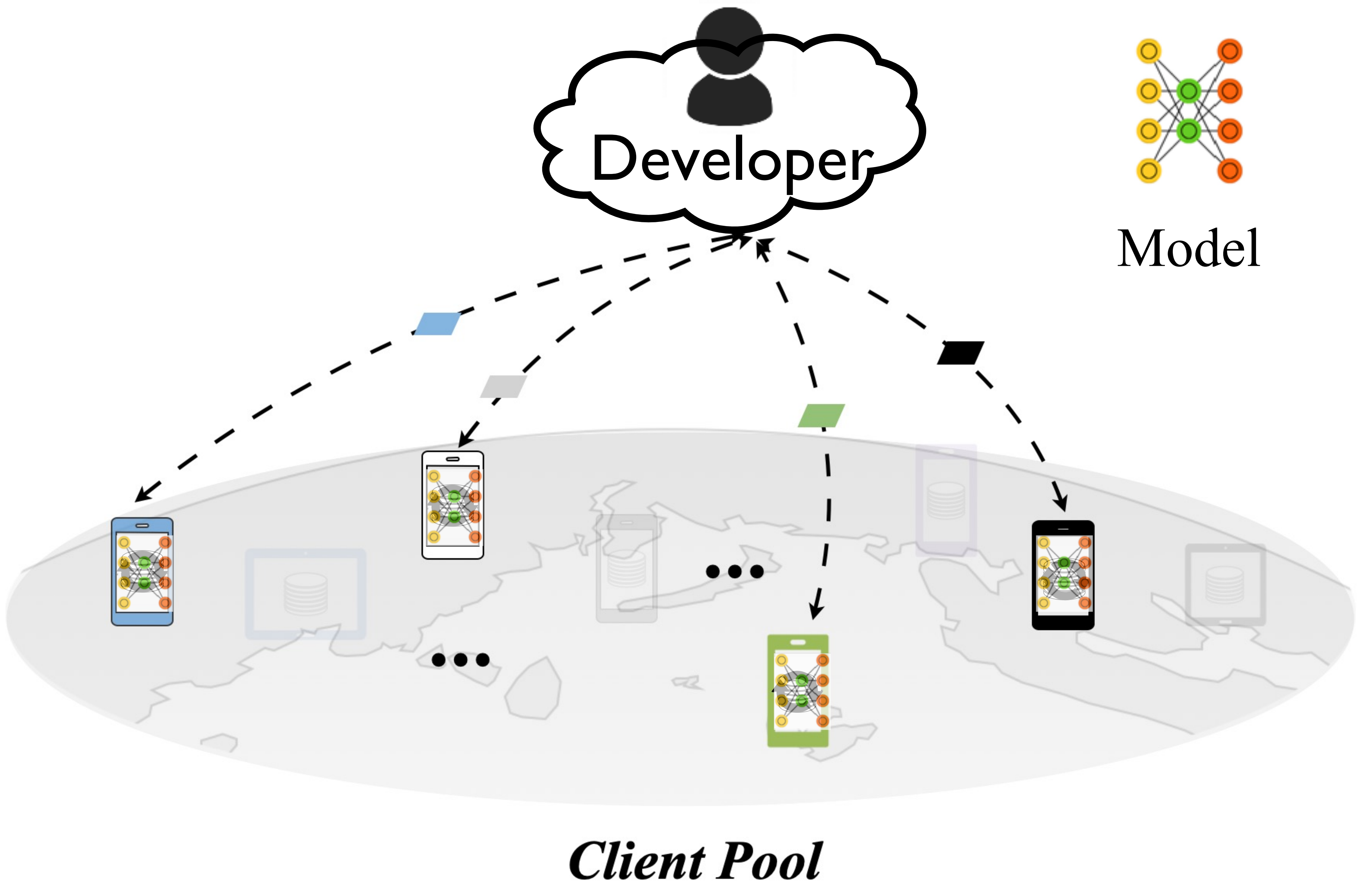
Fan Lai, Yinwei Dai, Sanjay Singapuram, Jiachen Liu, Xiangfeng Zhu,

Harsha V. Madhyastha, Mosharaf Chowdhury

University of Michigan

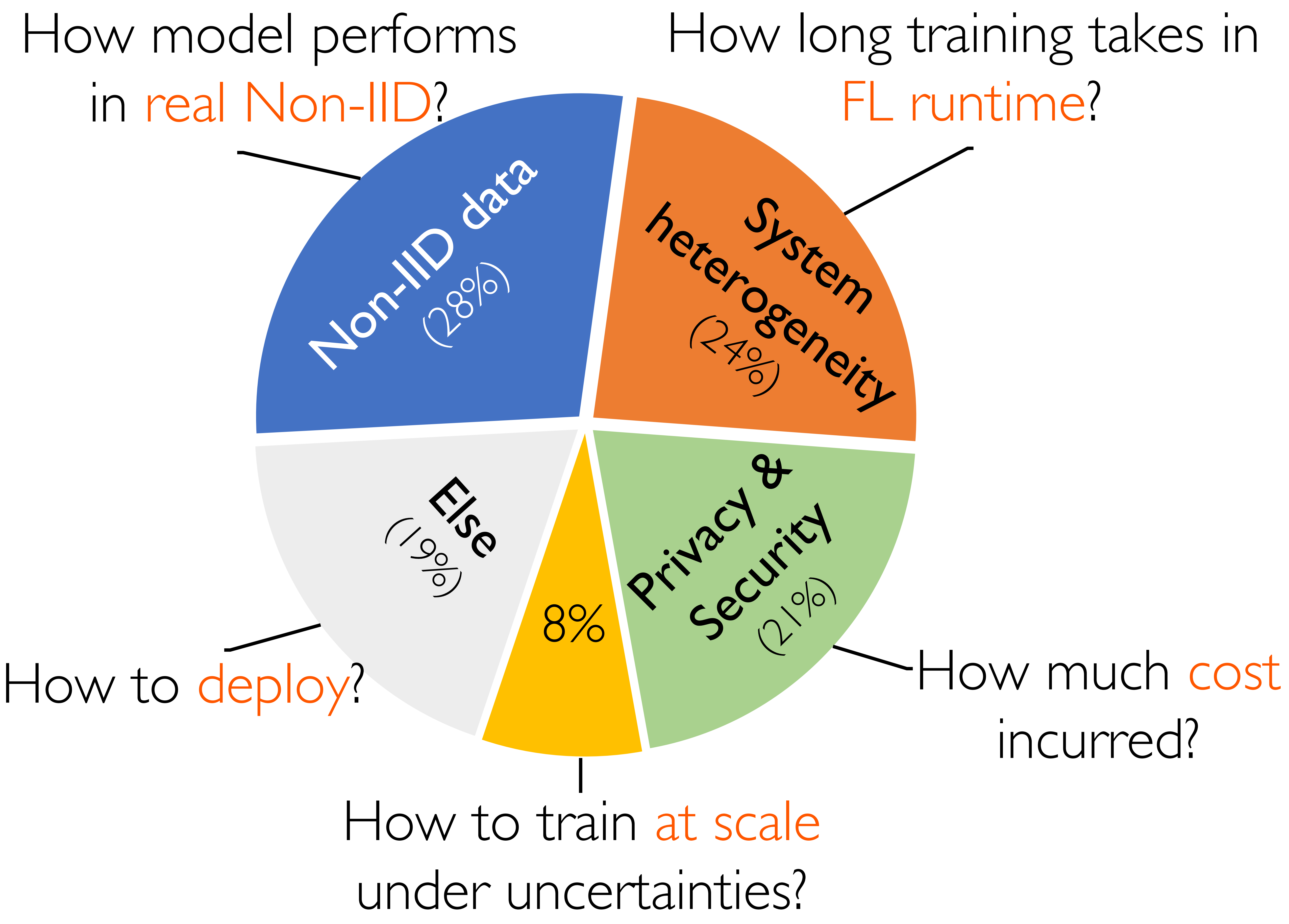
SymbioticLab

Federated Learning (FL) in Practice



- @ Google, Meta, Apple, Nvidia, …
- Training across millions of clients

Efforts for FL Challenges[1]



How model performs in real Non-IID?

How long training takes in FL runtime?

How to deploy?

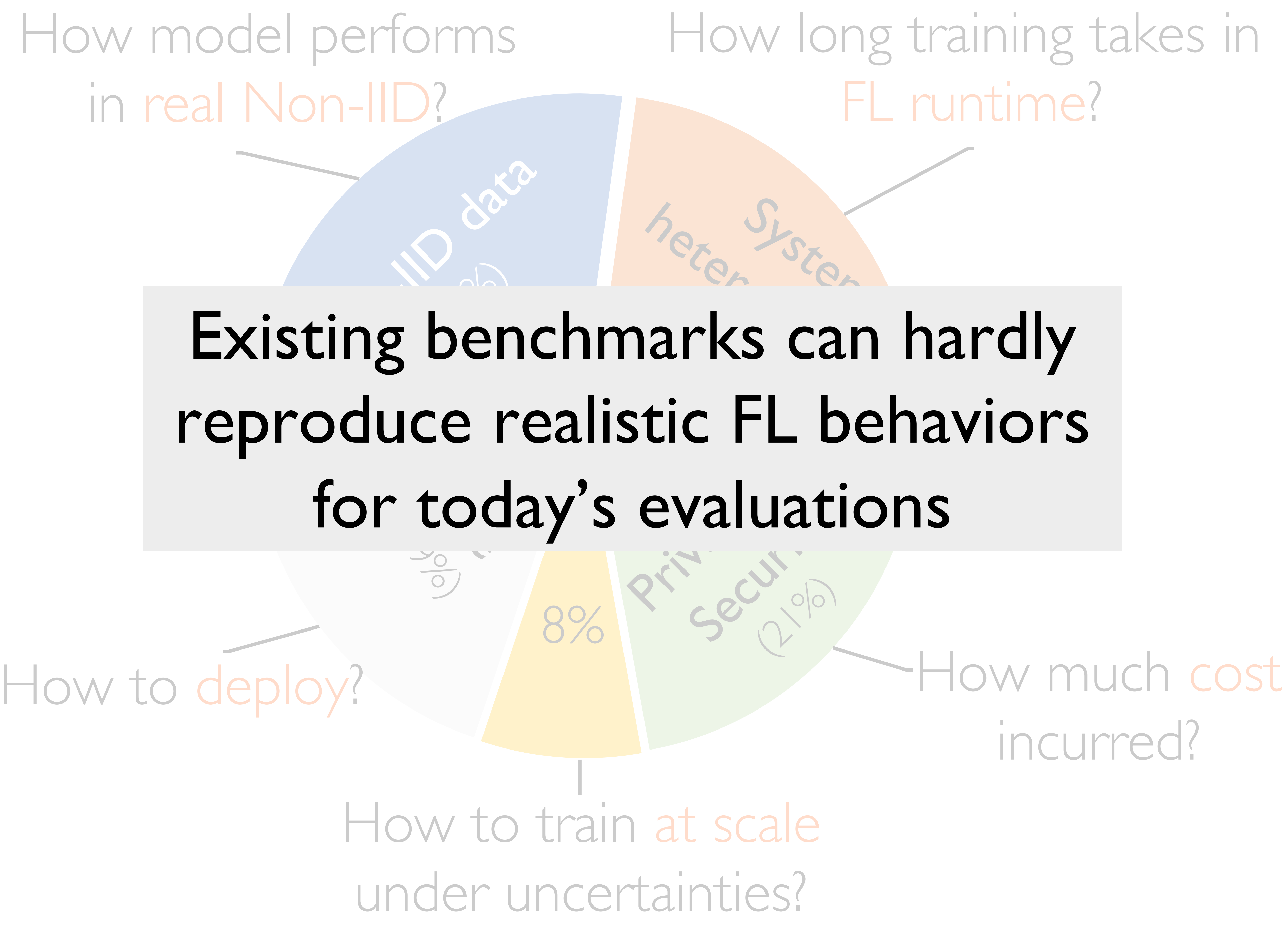
How much cost incurred?

How to train at scale under uncertainties?

[1] A Systematic Literature Review on Federated Machine Learning: From A Software Engineering Perspective. ACM Computing Surveys, 2022.

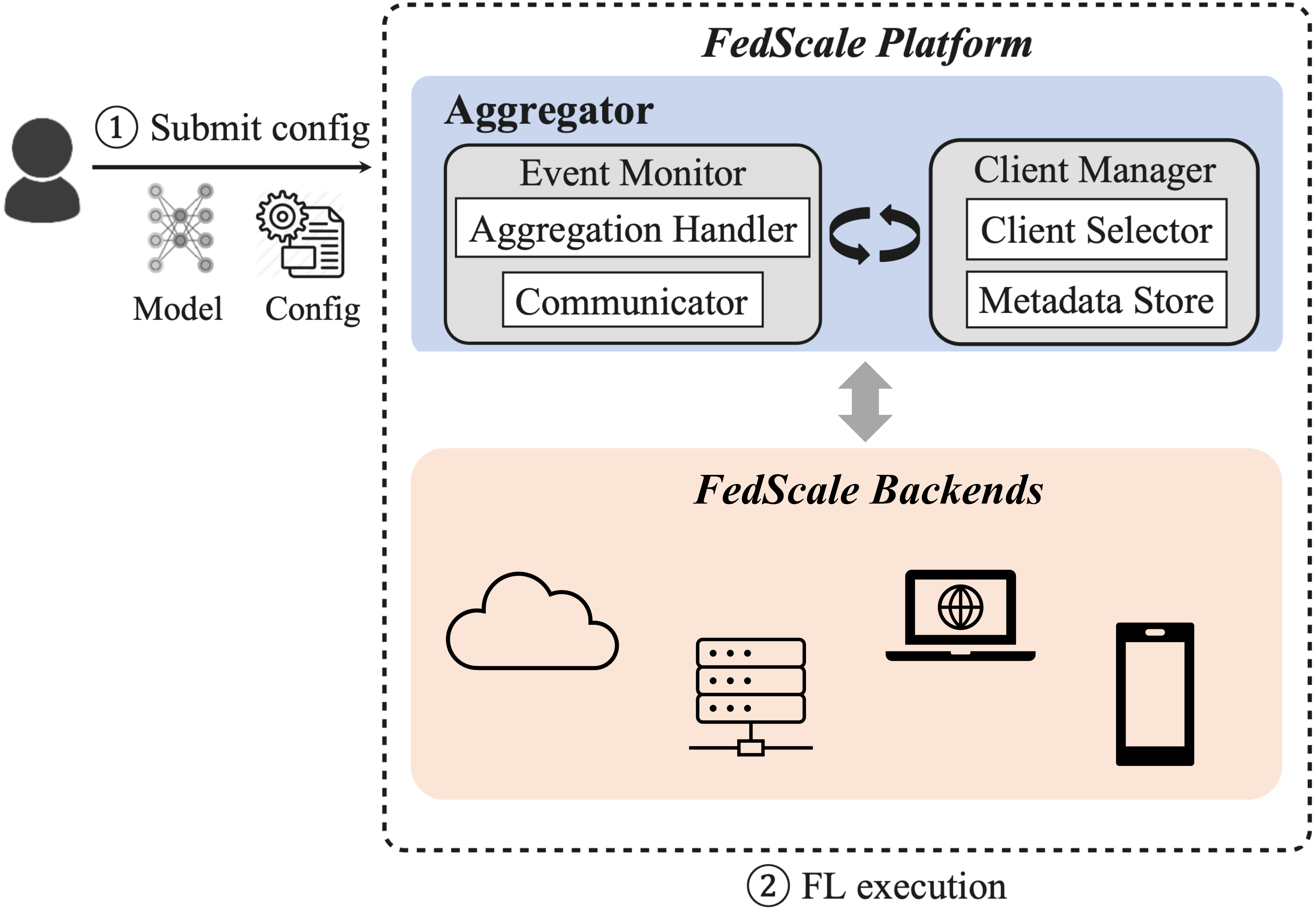
Federated Learning in Existing Benchmarks

- **Few realistic datasets**
  - Synthesized data (e.g., CIFAR)
- **Missing system details**
  - Hard to evaluate MLSys optimization
- **Suboptimal scalability**
  - Hard to reproduce practical FL scale

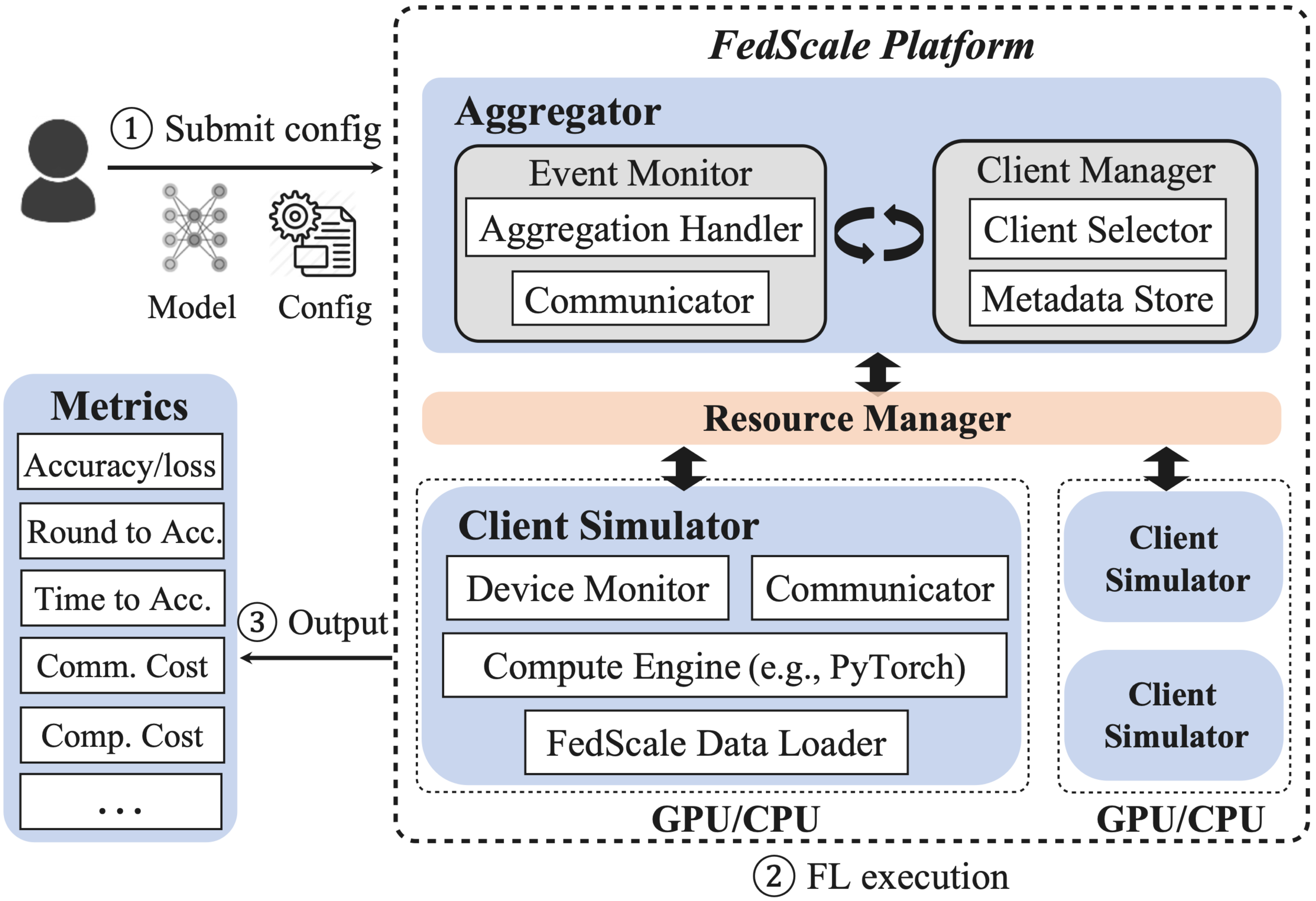


Existing benchmarks can hardly reproduce realistic FL behaviors for today's evaluations

FedScale as a Comprehensive Benchmark



FedScale as a Comprehensive Benchmark



① Submit config

Model

Config

FedScale Platform

Aggregator

Event Monitor

Aggregation Handler

Communicator

Client Manager

Client Selector

Metadata Store

Resource Manager

Client Simulator

Device Monitor

Communicator

Compute Engine (e.g., PyTorch)

FedScale Data Loader

GPU/CPU

Client Simulator

Client Simulator

GPU/CPU

② FL execution

③ Output

Metrics

Accuracy/loss

Round to Acc.

Time to Acc.

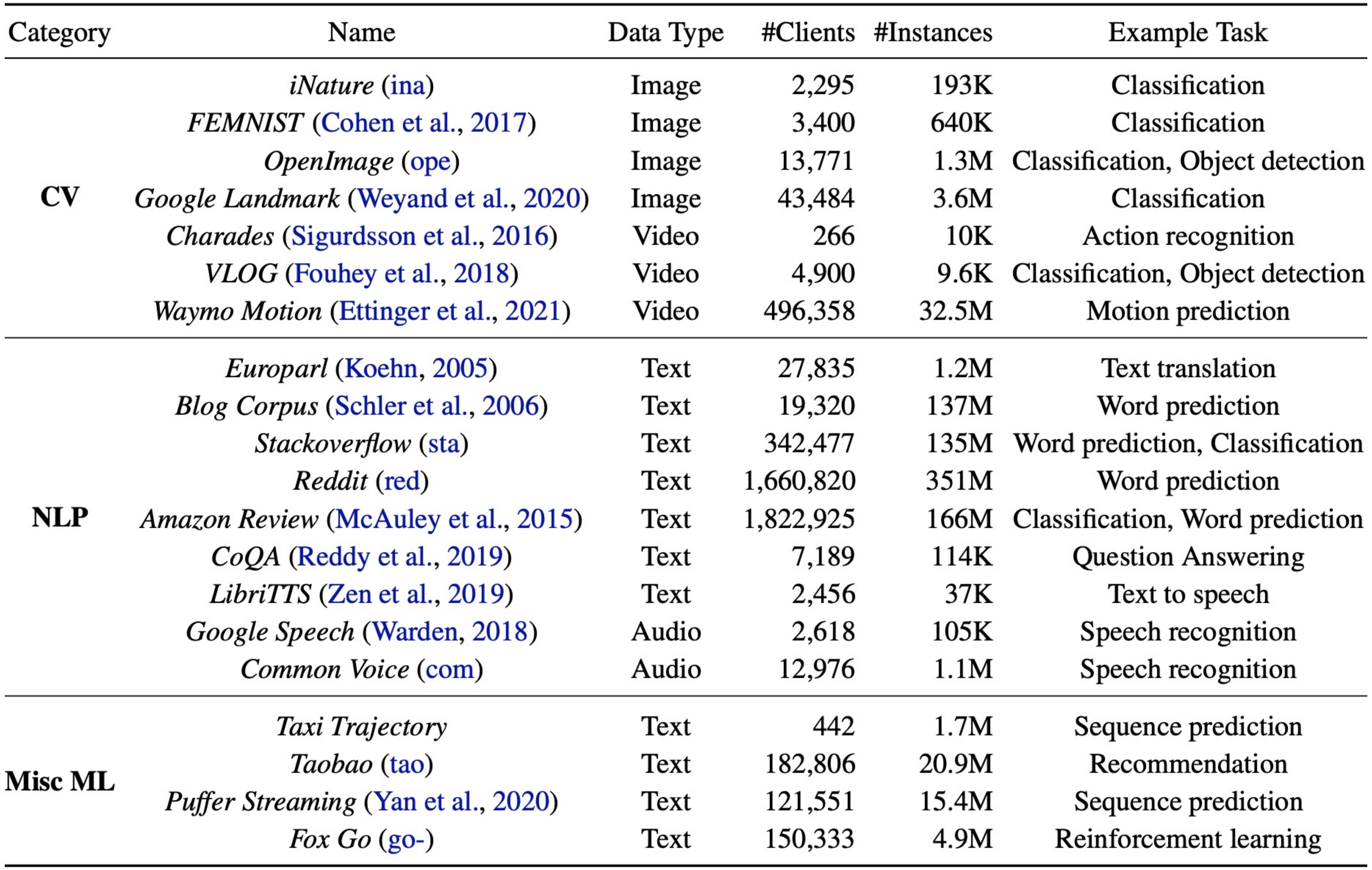
Comm. Cost

Comp. Cost

. . .

Realistic Client Datasets

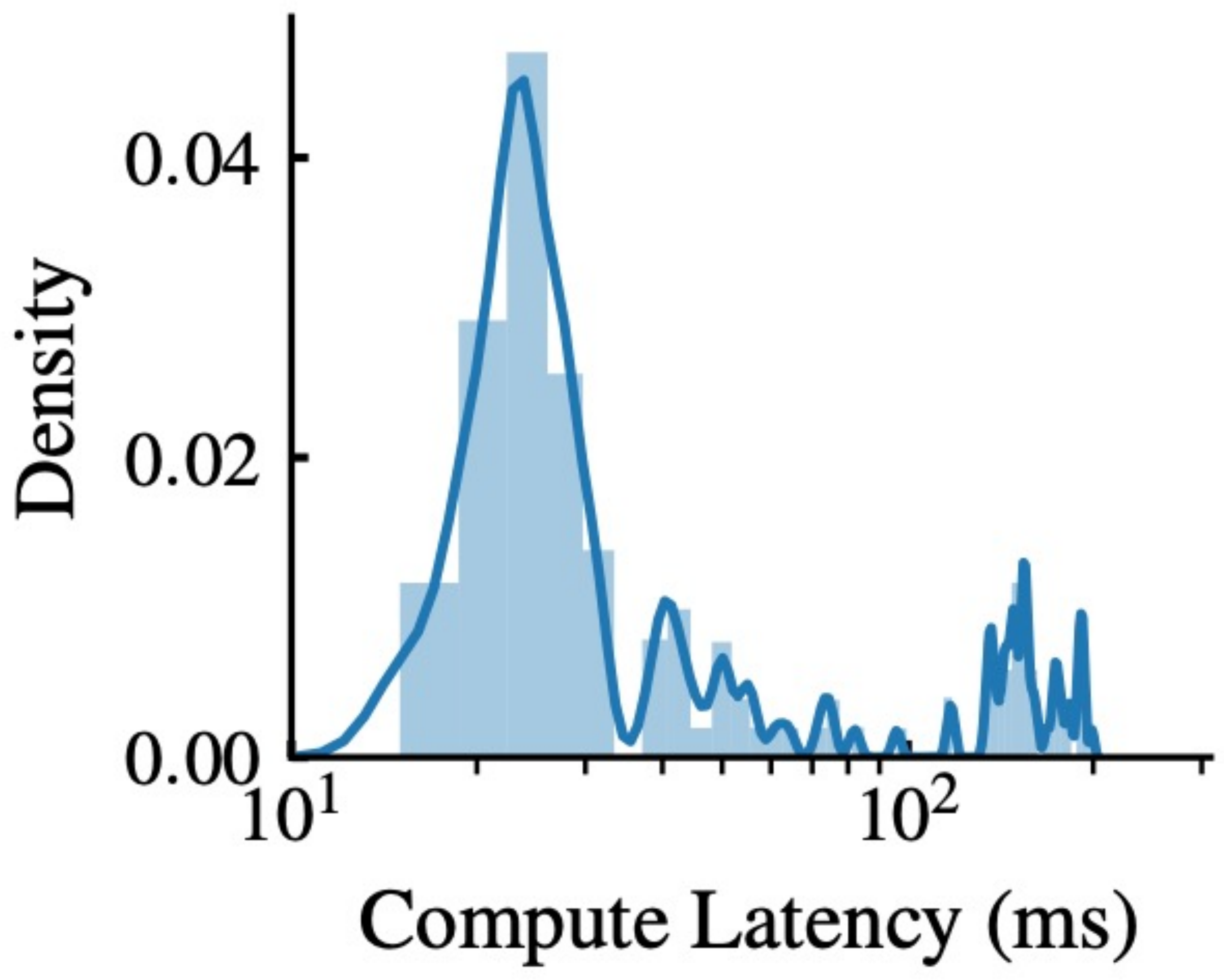
- **> 20 realistic datasets**
  - For CV, NLP, …
  - Small/Medium/Large scales

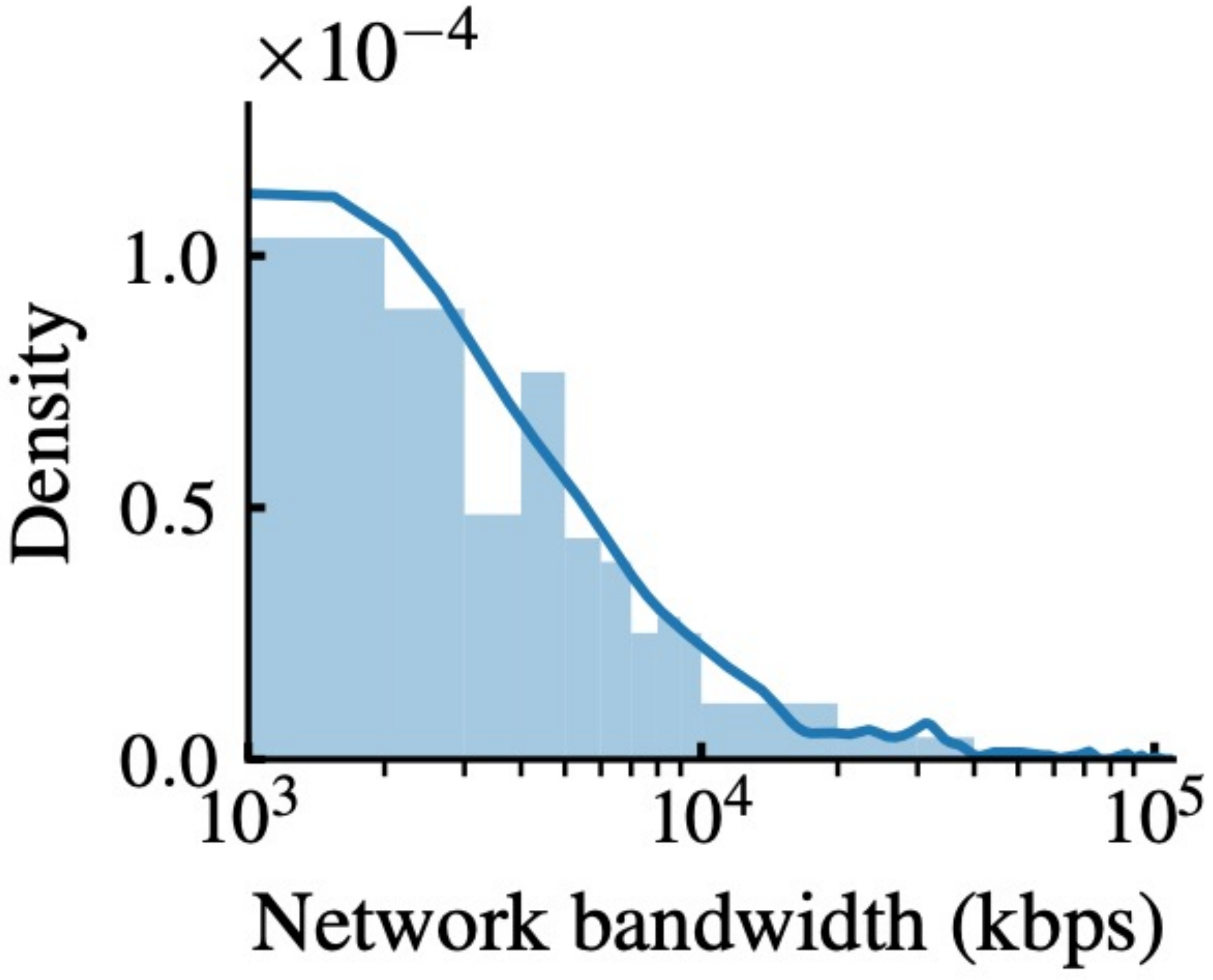


Category	Name	Data Type	#Clients	#Instances	Example Task
CV	*iNature* (ina)	Image	2,295	193K	Classification
	FEMNIST (Cohen et al., 2017)	Image	3,400	640K	Classification
	OpenImage (ope)	Image	13,771	1.3M	Classification, Object detection
	Google Landmark (Weyand et al., 2020)	Image	43,484	3.6M	Classification
	Charades (Sigurdsson et al., 2016)	Video	266	10K	Action recognition
	VLOG (Fouhey et al., 2018)	Video	4,900	9.6K	Classification, Object detection
	Waymo Motion (Ettinger et al., 2021)	Video	496,358	32.5M	Motion prediction
NLP	*Europarl* (Koehn, 2005)	Text	27,835	1.2M	Text translation
	Blog Corpus (Schler et al., 2006)	Text	19,320	137M	Word prediction
	Stackoverflow (sta)	Text	342,477	135M	Word prediction, Classification
	Reddit (red)	Text	1,660,820	351M	Word prediction
	Amazon Review (McAuley et al., 2015)	Text	1,822,925	166M	Classification, Word prediction
	CoQA (Reddy et al., 2019)	Text	7,189	114K	Question Answering
	LibriTTS (Zen et al., 2019)	Text	2,456	37K	Text to speech
	Google Speech (Warden, 2018)	Audio	2,618	105K	Speech recognition
	Common Voice (com)	Audio	12,976	1.1M	Speech recognition
Misc ML	*Taxi Trajectory*	Text	442	1.7M	Sequence prediction
	Taobao (tao)	Text	182,806	20.9M	Recommendation
	Puffer Streaming (Yan et al., 2020)	Text	121,551	15.4M	Sequence prediction
	Fox Go (go-)	Text	150,333	4.9M	Reinforcement learning

Some FedScale Datasets

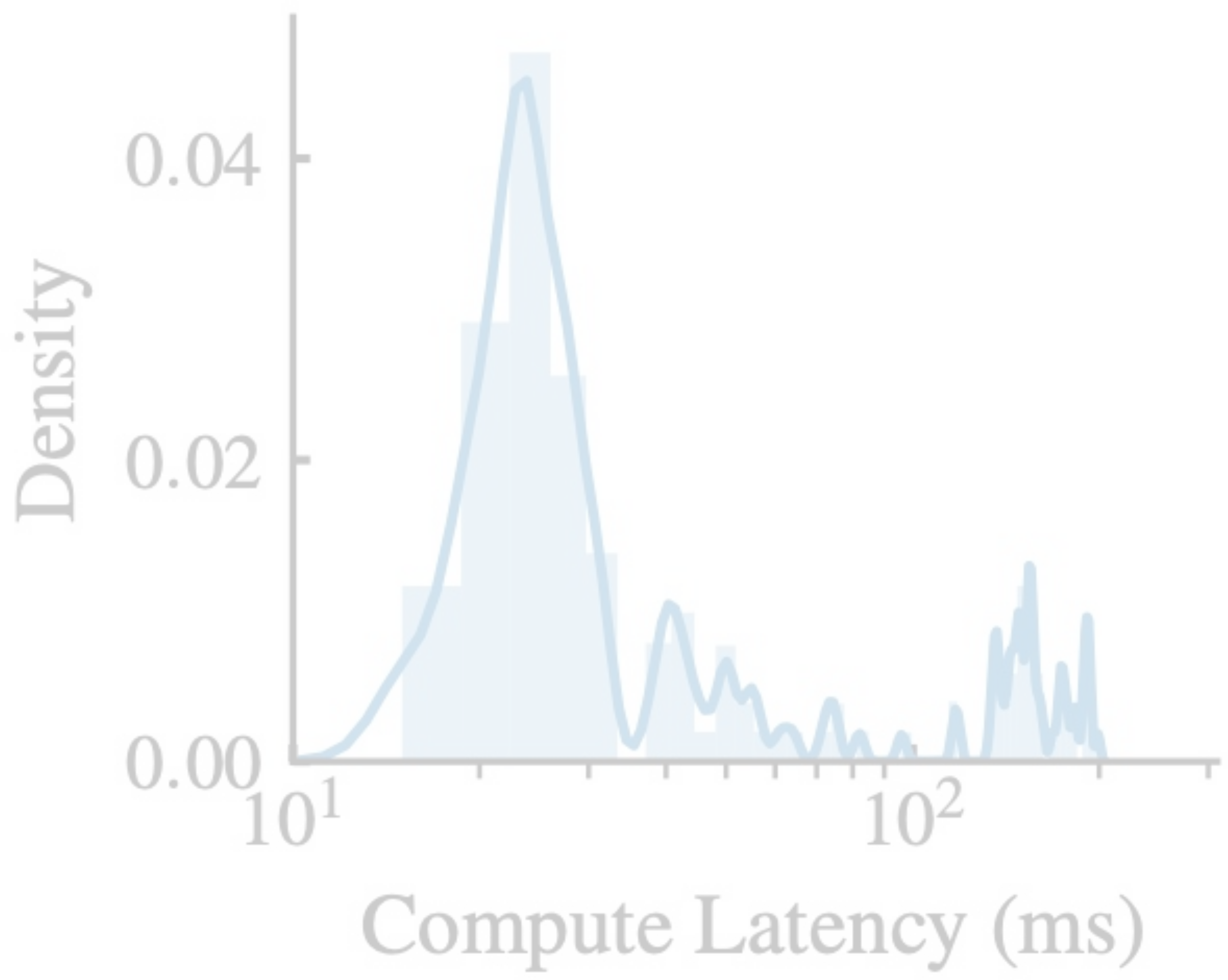
Millions of Client System Traces

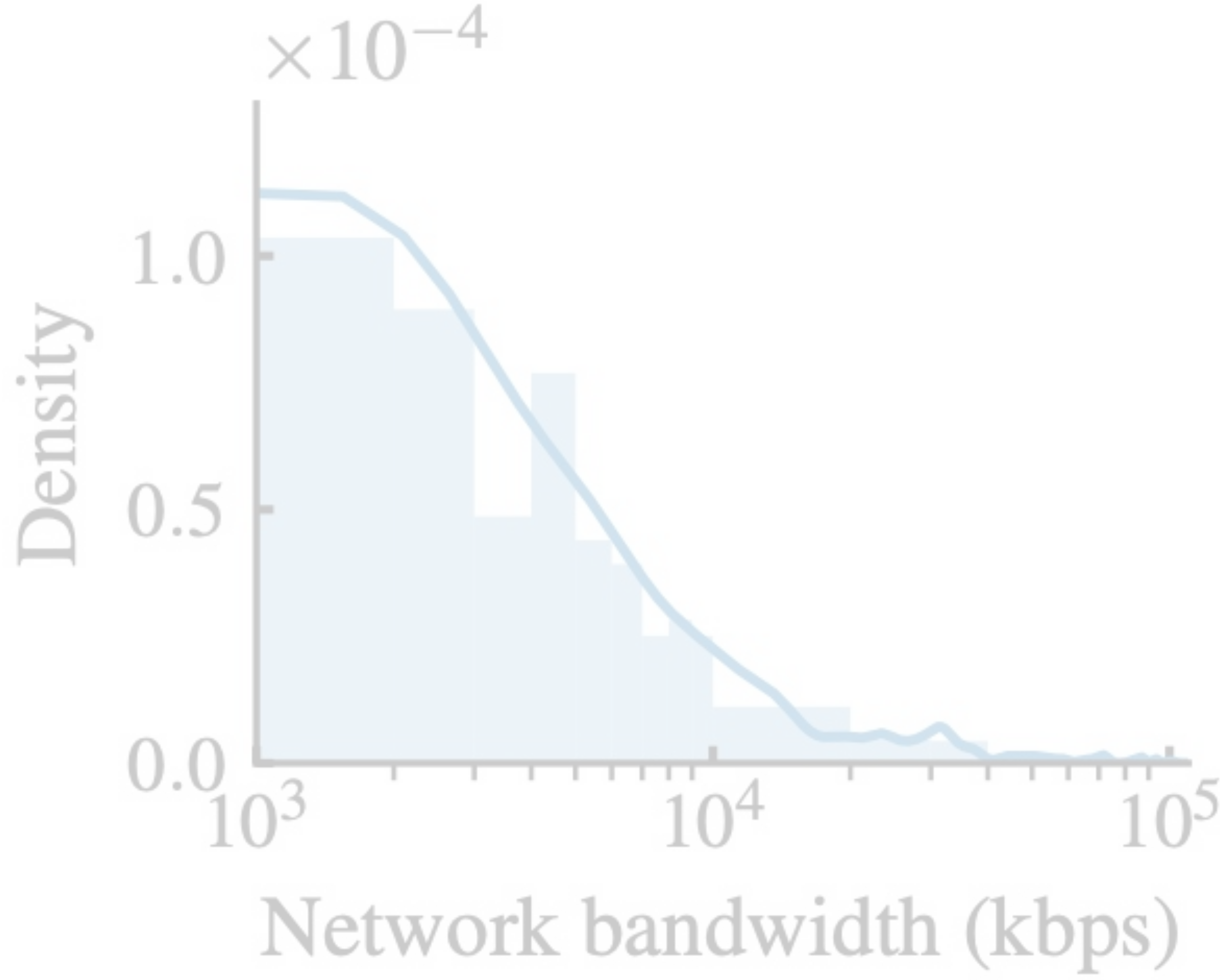




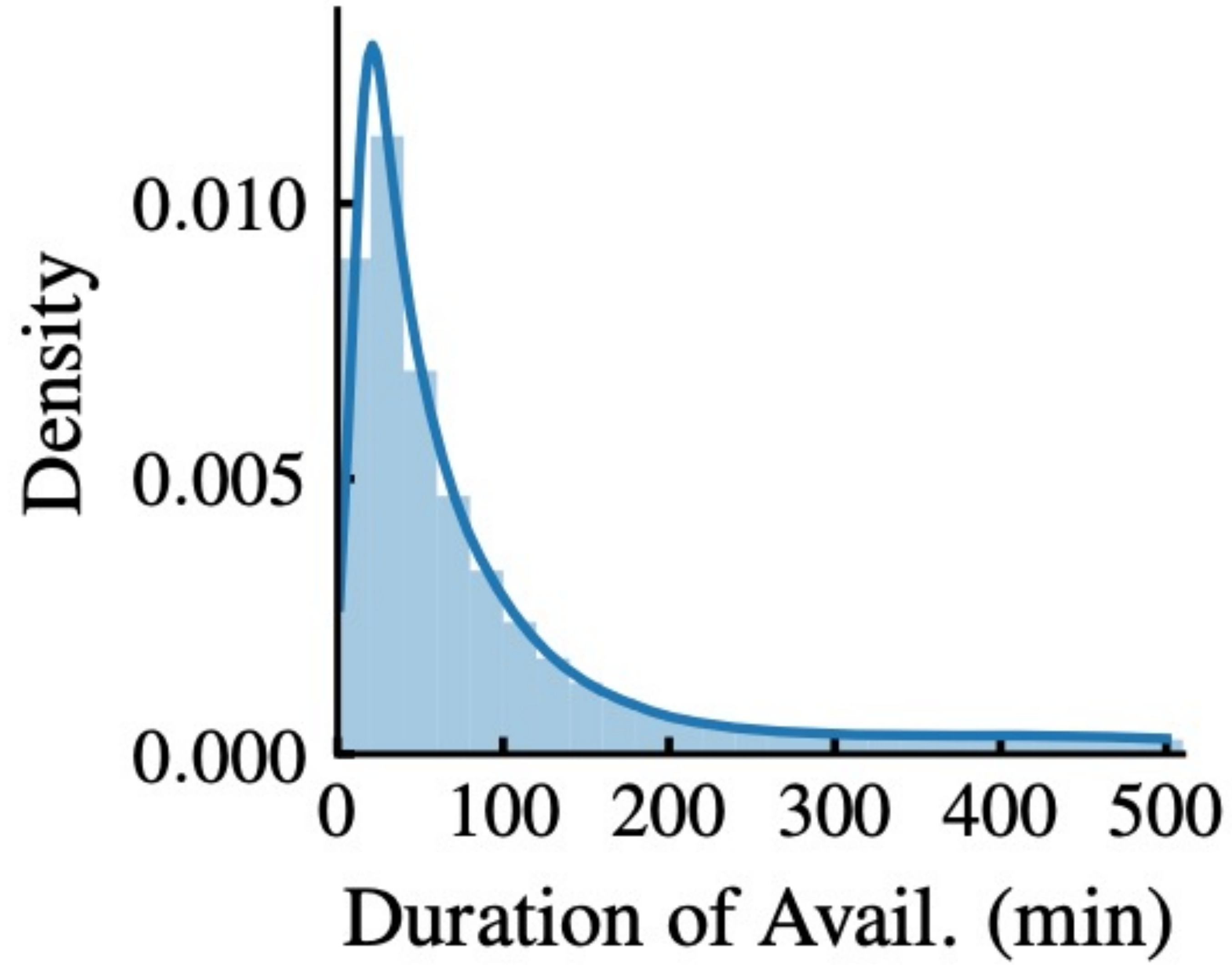
Heterogeneous computation/ communication speed

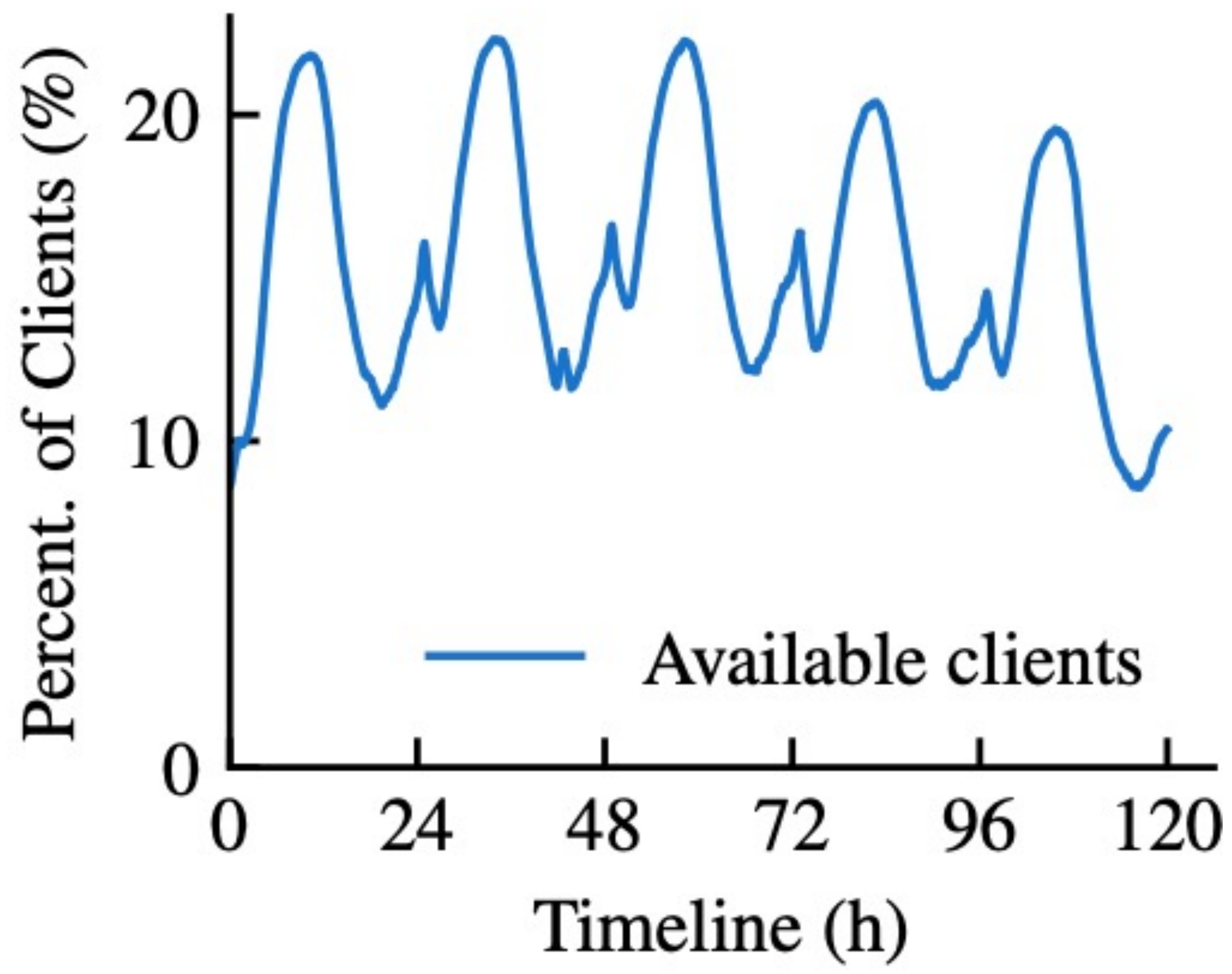
Millions of Client System Traces





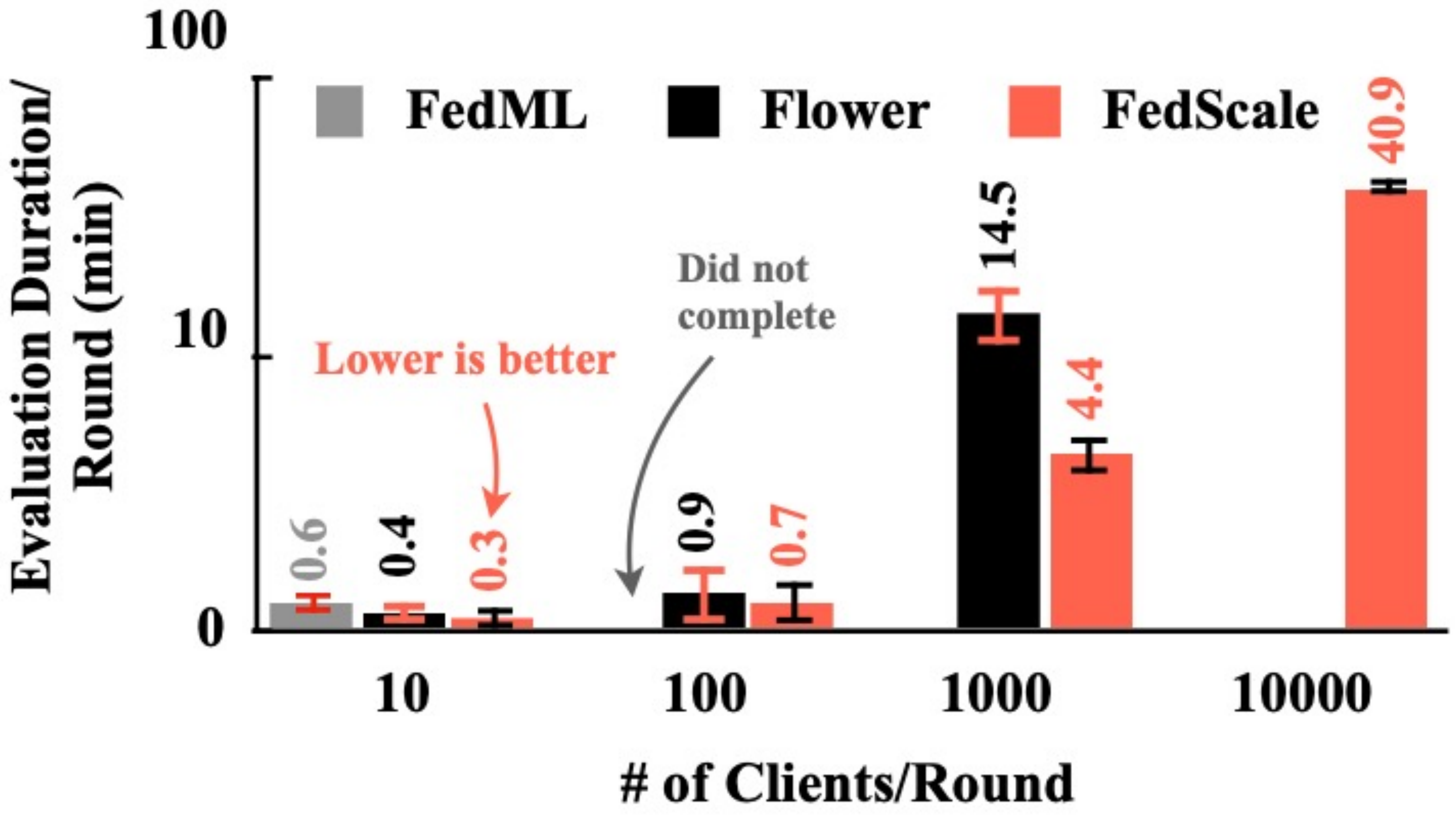
Heterogeneous computation/ communication speed





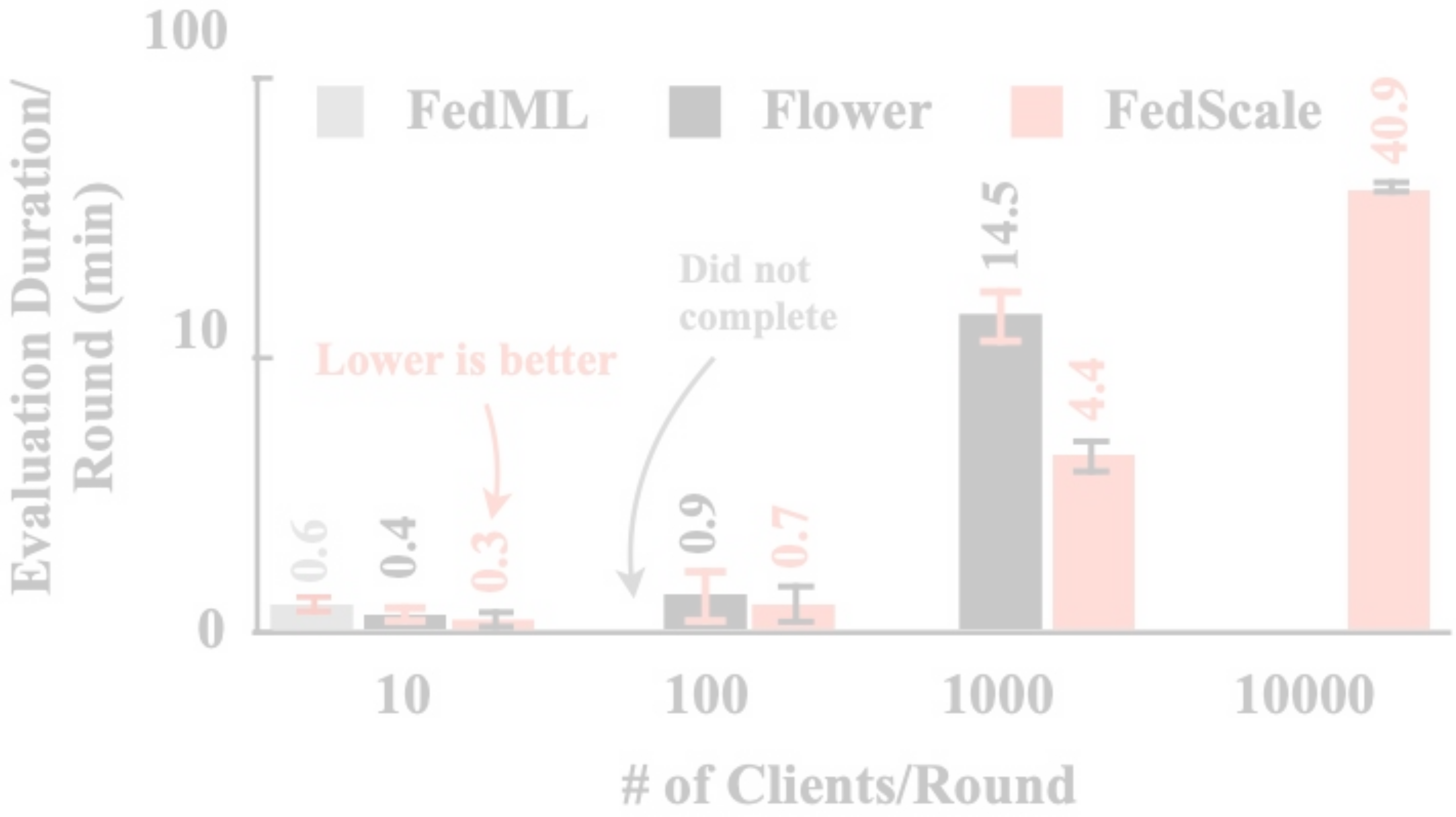
Dynamics of client availability in the wild

Scalable & Extensible Runtime

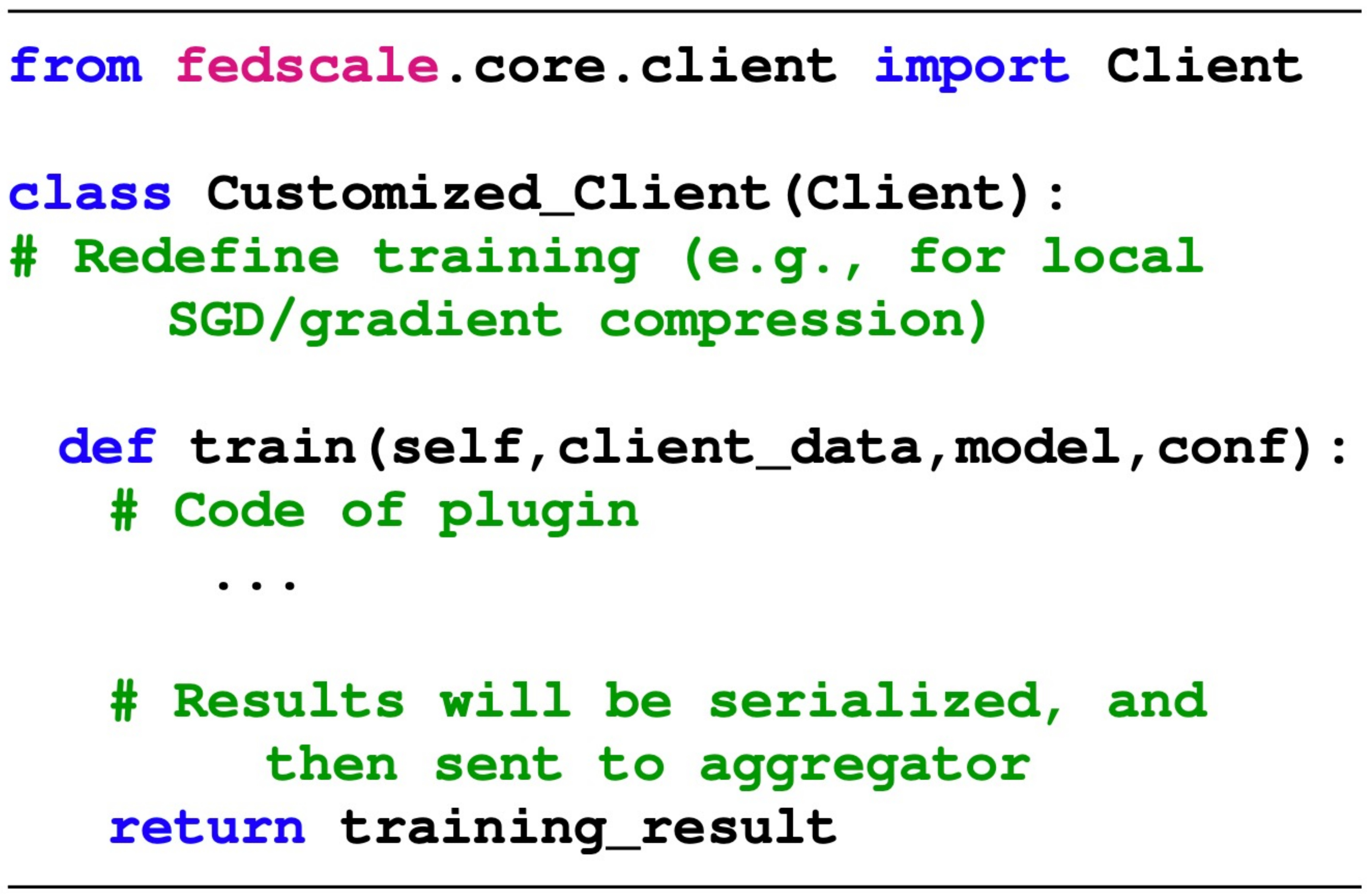


FedScale supports orders-of-magnitude more clients on the same underlying cluster

Scalable & Extensible Runtime



FedScale supports orders-of-magnitude more clients on the same underlying cluster



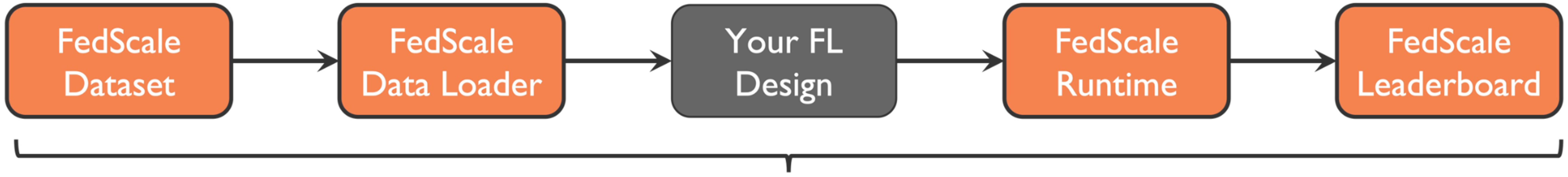
```
from fedscale.core.client import Client

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
```

Implementing new FL designs only takes a few lines of code

FedScale.ai

FL Benchmark & Platform:

- 20+ realistic datasets
- 70+ models
- Up to O(10k) clients/round
- Sync/Async training mode
- Easy extension
- On-device deployment
- Practical FL runtime



FedScale automates FL benchmarking