FedScale

Benchmarking Model and System Performance of Federated Learning

Fan Lai, Yinwei Dai, Xiangfeng Zhu,

Harsha V. Madhyastha, Mosharaf Chowdhury

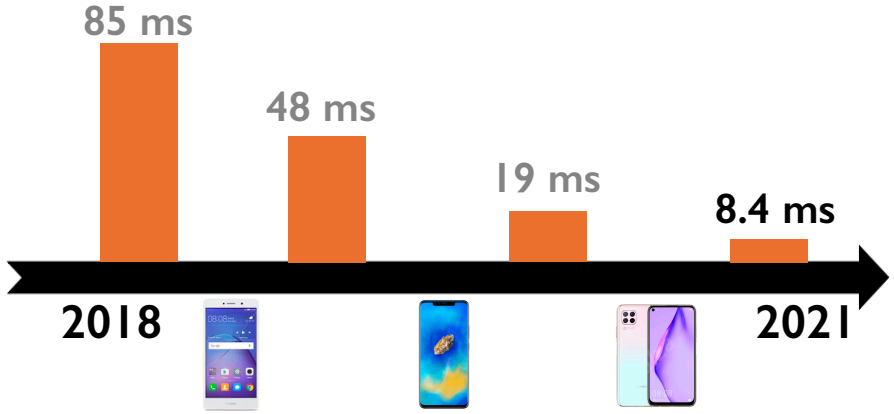
SymbioticLab

University of Michigan

Emerging Trend of Machine Learning

Edge devices generate massive data

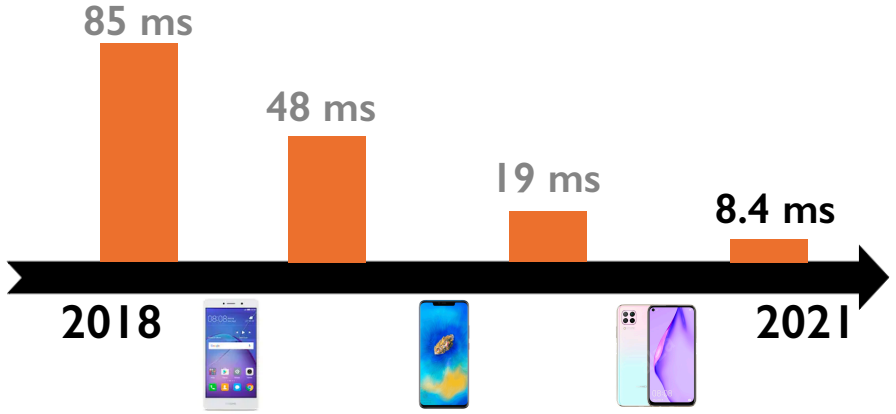
Model compute latency becomes smaller



Emerging Trend of Machine Learning

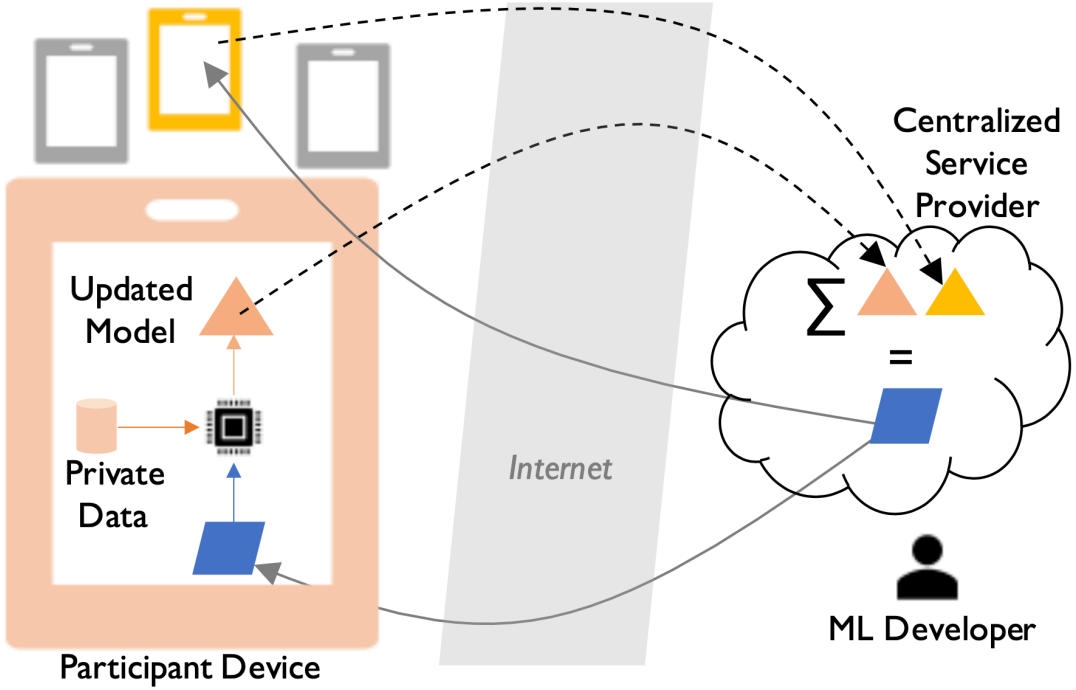
Edge devices generate massive data

Model compute latency becomes smaller



Federated learning helps

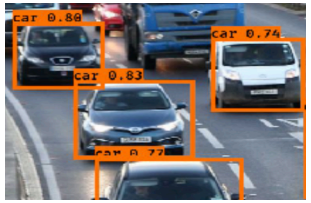
- Reduce data migration/privacy risk
- Learn on fresh real-world data
- …



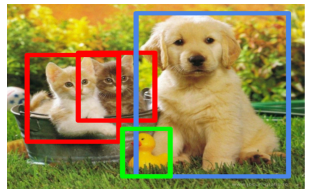
Challenges in Federated Learning (FL)

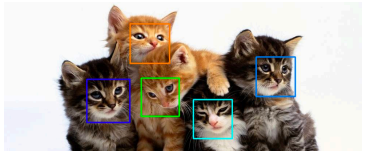
	FL	In-cluster ML
Data	Heter.	Homogeneous via shuffling

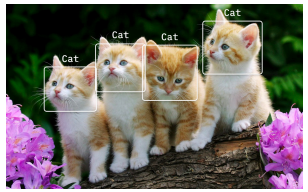
Client A



Client B







Heterogeneous data distribution

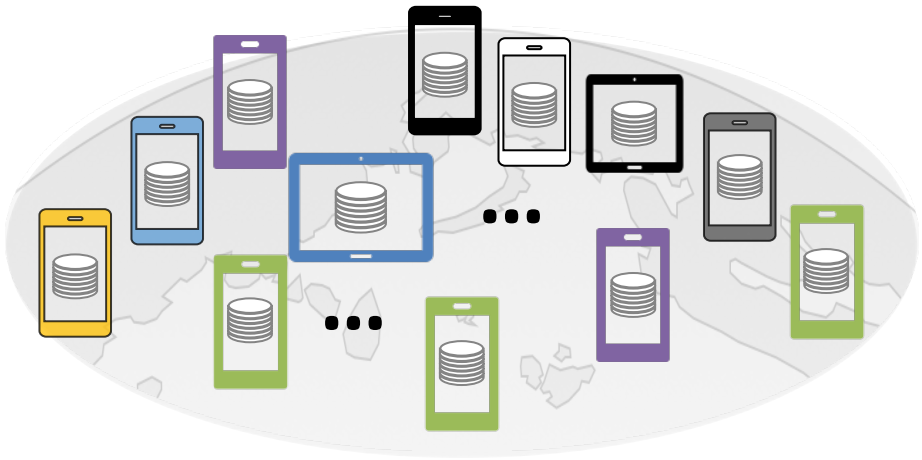
Challenges in Federated Learning (FL)

	FL	In-cluster ML
Data	Heter.	Homogeneous via shuffling
System	Heter.	Homogeneous

Heterogeneous system speed

Challenges in Federated Learning (FL)

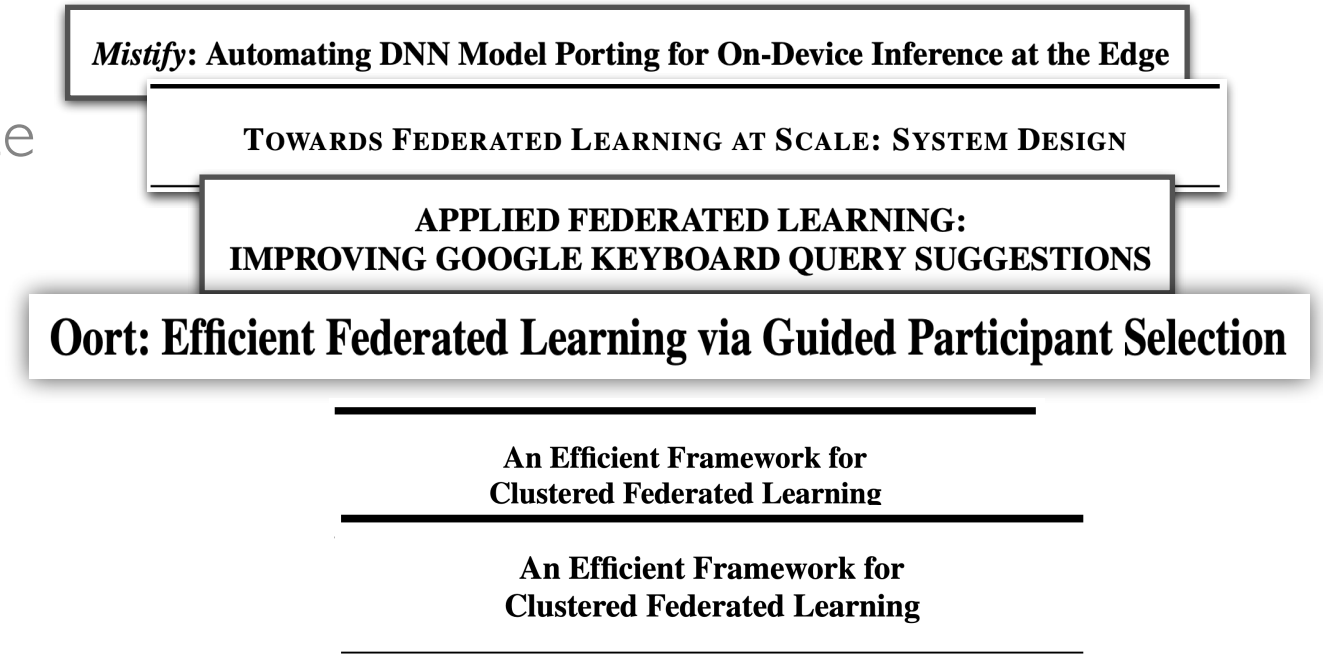
	FL	In-cluster ML
Data	Heter.	Homogeneous via shuffling
System	Heter.	Homogeneous
Scale	O(1M)	O(10)
Dynamics	Client can drop out/rejoin	Negligible
…	…	…



Large scale and dynamics

While Existing FL Optimizations Are Diverse …

- **To tackle challenges, we optimize**
  - Statistical efficiency
    - Round-to-accuracy convergence
    - Final model accuracy, …
  - System efficiency
    - Reduce network traffics
    - Discard system stragglers, …
  - Privacy and security
  - …



Diverse FL efforts in algorithm/system

Inefficiency of Existing FL Benchmarks

	LEAF	FedEval	FedML	Flower
Heter. Client Dataset	⭕	✕	⭕	⭕

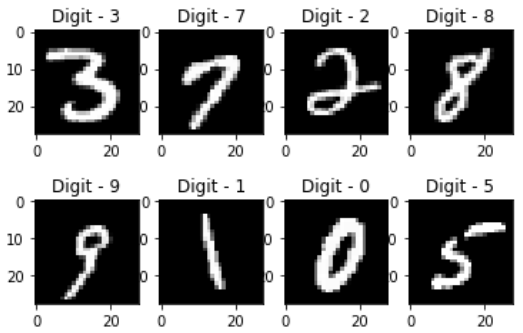
Few realistic datasets

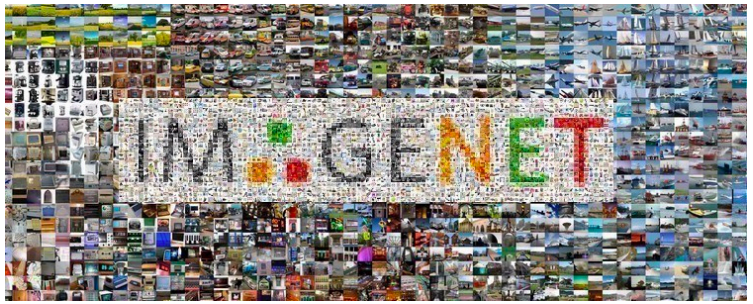
⭕: limited support

Inefficiency of Existing FL Benchmarks

	LEAF	FedEval	FedML	Flower
Heter. Client Dataset	⭕	×	⭕	⭕

Few realistic datasets





Traditional datasets are not representative

⭕: limited support

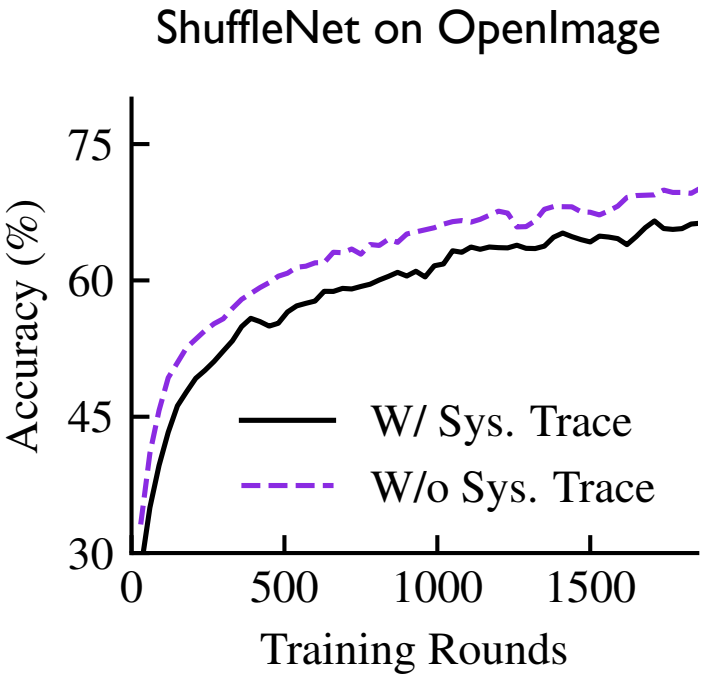
Inefficiency of Existing FL Benchmarks

	LEAF	FedEval	FedML	Flower
Heter. Client Dataset	⭕	✕	⭕	⭕
Heter. System Speed	✕	✕	✕	✕
Client Availability	✕	✕	✕	✕

⭕: limited support

Inefficiency of Existing FL Benchmarks

	LEAF	FedEval	FedML	Flower
Heter. Client Dataset	⭕	✕	⭕	⭕
Heter. System Speed	✕	✕	✕	✕
Client Availability	✕	✕	✕	✕



Overlooking system aspects leads to optimistic performance

⭕: limited support

Inefficiency of Existing FL Benchmarks

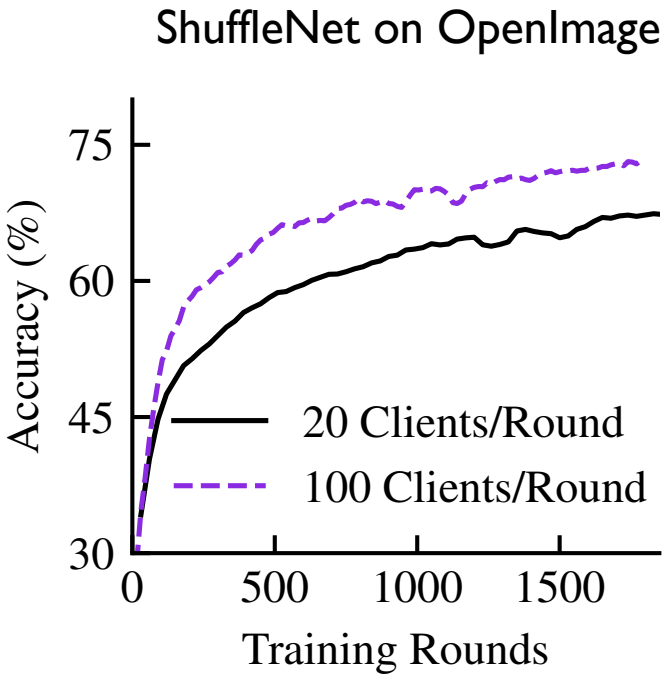
	LEAF	FedEval	FedML	Flower
Heter. Client Dataset	⭕	✕	⭕	⭕
Heter. System Speed	✕	✕	✕	✕
Client Availability	✕	✕	✕	✕
Scalable Platform	✕	⭕	⭕	✓

Suboptimal scalability in distributed setting

⭕: limited support

Inefficiency of Existing FL Benchmarks

	LEAF	FedEval	FedML	Flower
Heter. Client Dataset	○	×	○	○
Heter. System Speed	×	×	×	×
Client Availability	×	×	×	×
Scalable Platform	×	○	○	✓



Suboptimal scalability in distributed setting

Inability for large-scale evaluation under-reports FL performance

○: limited support

Inefficiency of Existing FL Benchmarks

	LEAF	FedEval	FedML	Flower
Heter. Client Dataset	○	×	○	○
Heter. System Speed	×	×	×	×
Client Availability	×	×	×	×
Scalable Platform	×	○	○	✓
Real FL Runtime	×	×	×	×
Flexible APIs	×	×	✓	✓

○: limited support

Missing FL runtime discourages system optimizations

Inefficiency of Existing FL Benchmarks

	LEAF	FedEval	FedML	Flower	FedScale
Heter. Client Dataset	○	×	○	○	✓
Heter. System Speed	×	×	×	×	✓
Client Availability	×	×	×	×	✓
Scalable Platform	×	○	○	✓	✓
Real FL Runtime	×	×	×	×	✓
Flexible APIs	×	×	✓	✓	✓

○: limited support

fedscale.ai

#1

Realistic FL Datasets

#2

Scalable Eval Platform

#3

Easy Development

Statistical Client Datasets

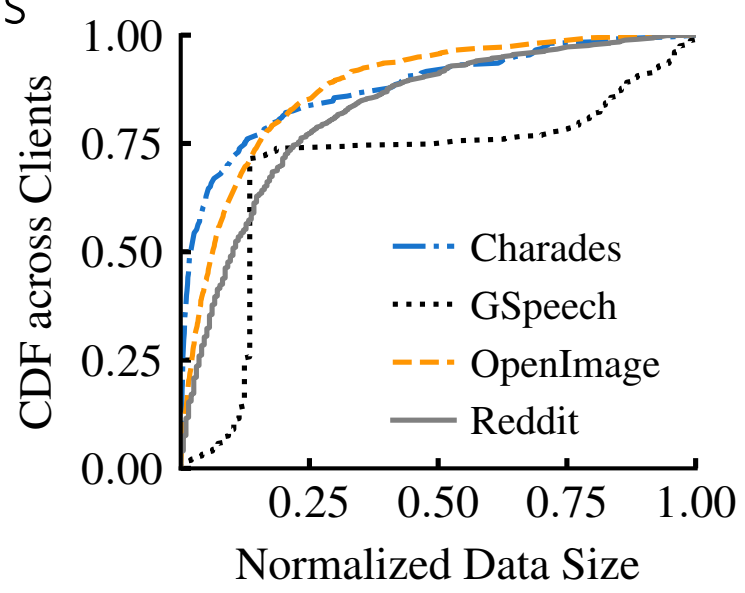
- **~20 diverse datasets w/ realistic partition**
  - CV, NLP, RL tasks
  - Small/Medium/Large scales
  - Standardized data loader

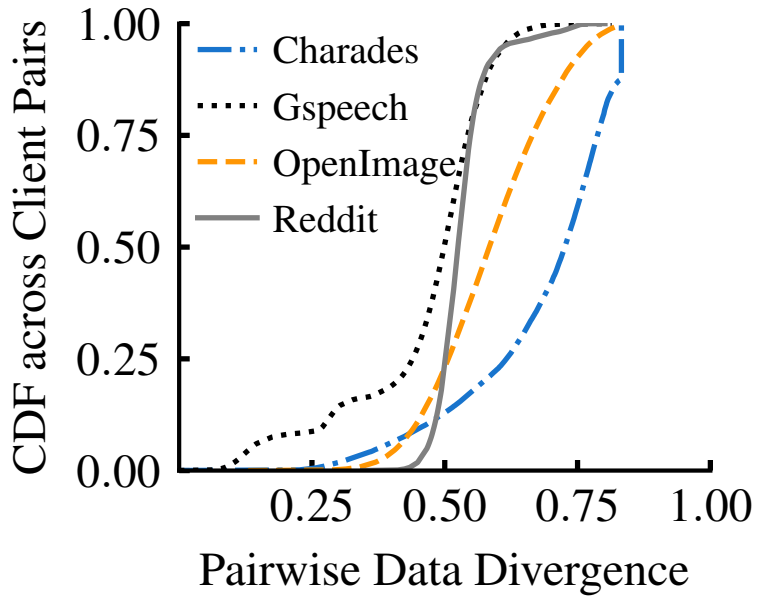
Category	Name	Data Type	#Clients	#Instances
CV	iNature	Image	2,295	193K
	FEMNIST	Image	3,400	640K
	OpenImage	Image	13,771	1.3M
	Google Landmark	Image	43,484	3.6M
	Charades	Video	266	10K
	VLOG	Video	4,900	9.6K
	Waymo Motion	Video	496,358	32.5M
NLP	Europarl	Text	27,835	1.2M
	Blog Corpus	Text	19,320	137M
	Stackoverflow	Text	342,477	135M
	Reddit	Text	1,660,820	351M
	Amazon Review	Text	1,822,925	166M
	CoQA	Text	7,189	114K
	LibriTTS	Text	2,456	37K
	Google Speech	Audio	2,618	105K
	Common Voice	Audio	12,976	1.1M
Misc ML	Taobao	Text	182,806	20.9M
	Fox Go	Text	150,333	4.9M

Some FedScale Datasets

Statistical Client Datasets

- **~20 diverse datasets w/ realistic partition**
  - CV, NLP, RL tasks
  - Small/Medium/Large scales
  - Standardized data loader





Realistic client data is heterogeneous in size and distribution

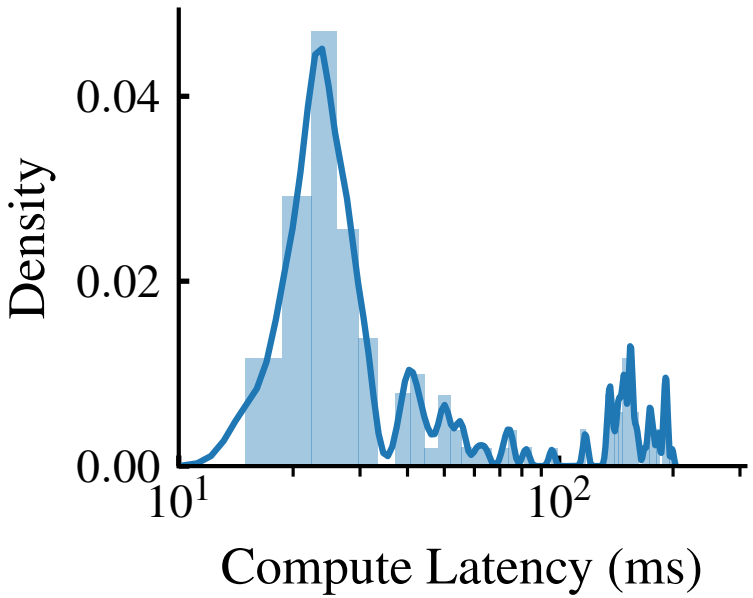
Statistical Client Datasets

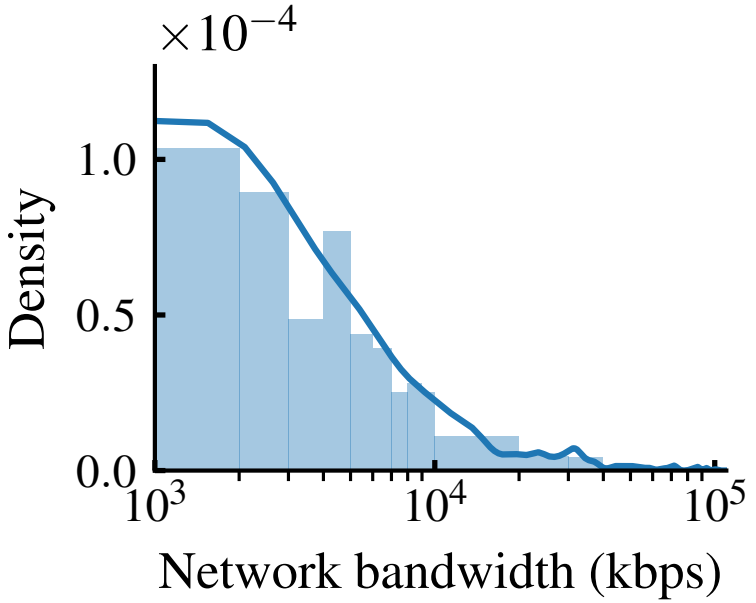
- **~20 diverse datasets w/ realistic partition**
  - CV, NLP, RL tasks
  - Small/Medium/Large scales
  - Standardized data loader
- **Use cases**
  - Investigate convergence under realistic dist.
  - Pinpoint accuracy fairness across clients
  - Personalize models for different clients
  - …

Category	Name	Data Type	#Clients	#Instances
CV	iNature	Image	2,295	193K
	FEMNIST	Image	3,400	640K
	OpenImage	Image	13,771	1.3M
	Google Landmark	Image	43,484	3.6M
	Charades	Video	266	10K
	VLOG	Video	4,900	9.6K
	Waymo Motion	Video	496,358	32.5M
NLP	Europarl	Text	27,835	1.2M
	Blog Corpus	Text	19,320	137M
	Stackoverflow	Text	342,477	135M
	Reddit	Text	1,660,820	351M
	Amazon Review	Text	1,822,925	166M
	CoQA	Text	7,189	114K
	LibriTTS	Text	2,456	37K
	Google Speech	Audio	2,618	105K
	Common Voice	Audio	12,976	1.1M
Misc ML	Taobao	Text	182,806	20.9M
	Fox Go	Text	150,333	4.9M

FedScale Datasets

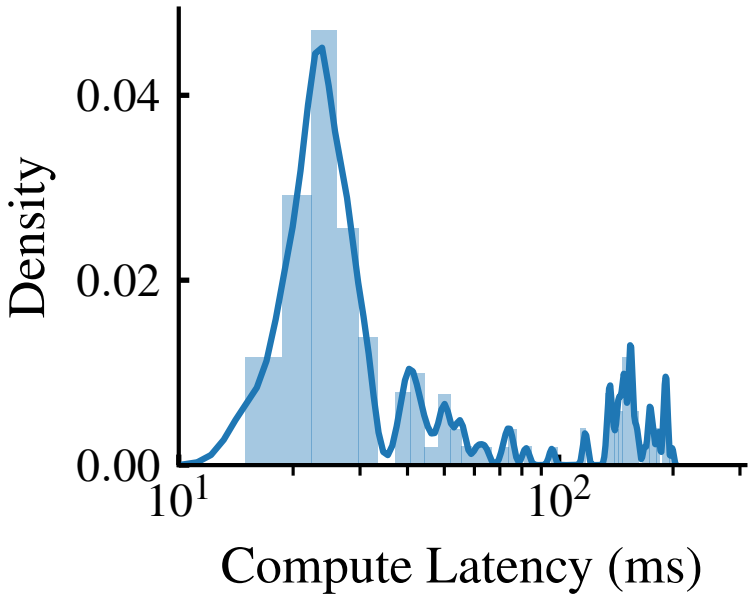
Millions of Client System Traces

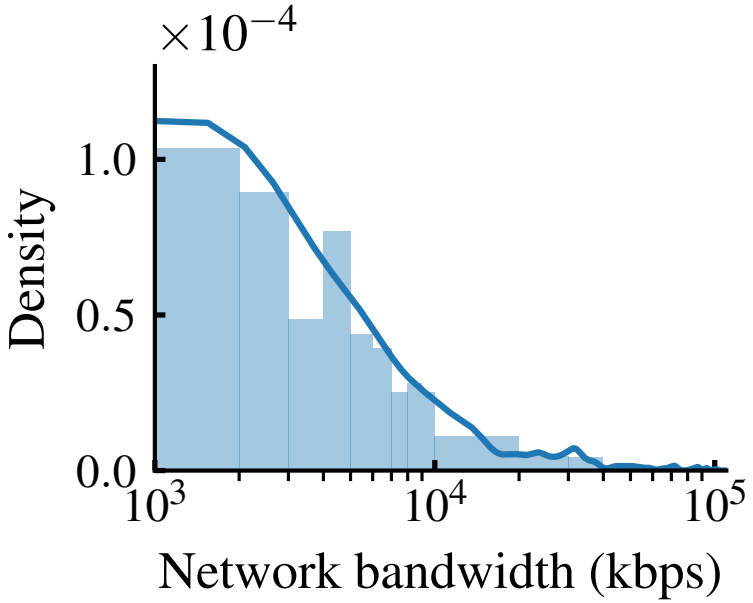




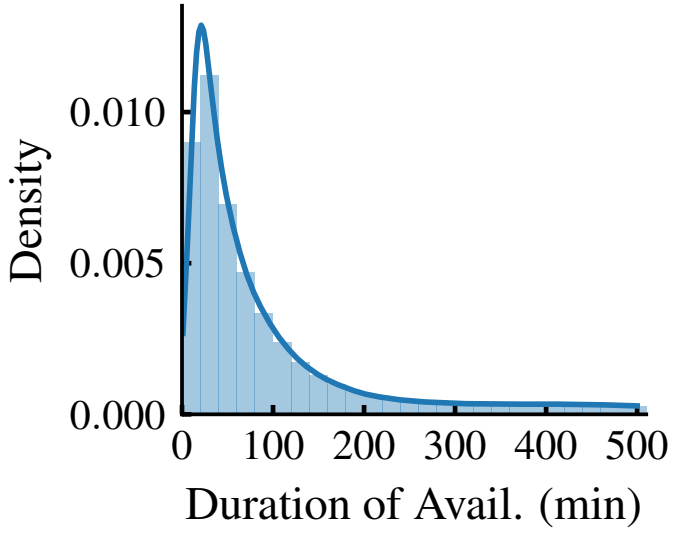
Heterogeneous computation/ communication speed

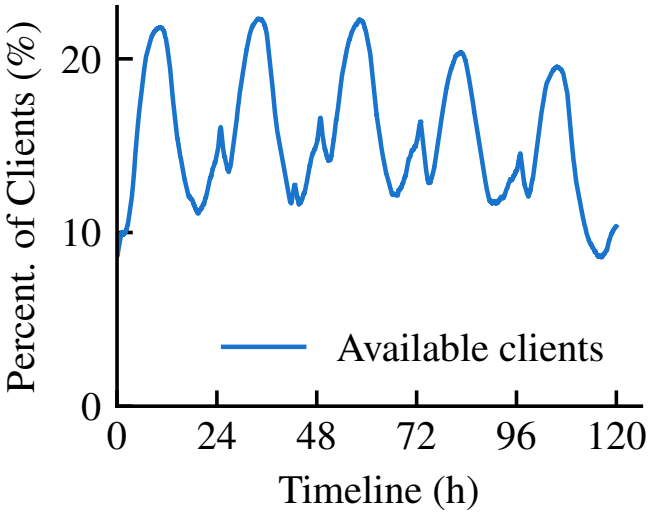
Millions of Client System Traces





Heterogeneous computation/ communication speed





Dynamics of client availability in the wild

#1

Realistic FL Datasets

Statistical client datasets

- ~20 datasets w/ real partition
- Various FL tasks/scales

Client system traces

- Client device speed
- Device availability over time

#2

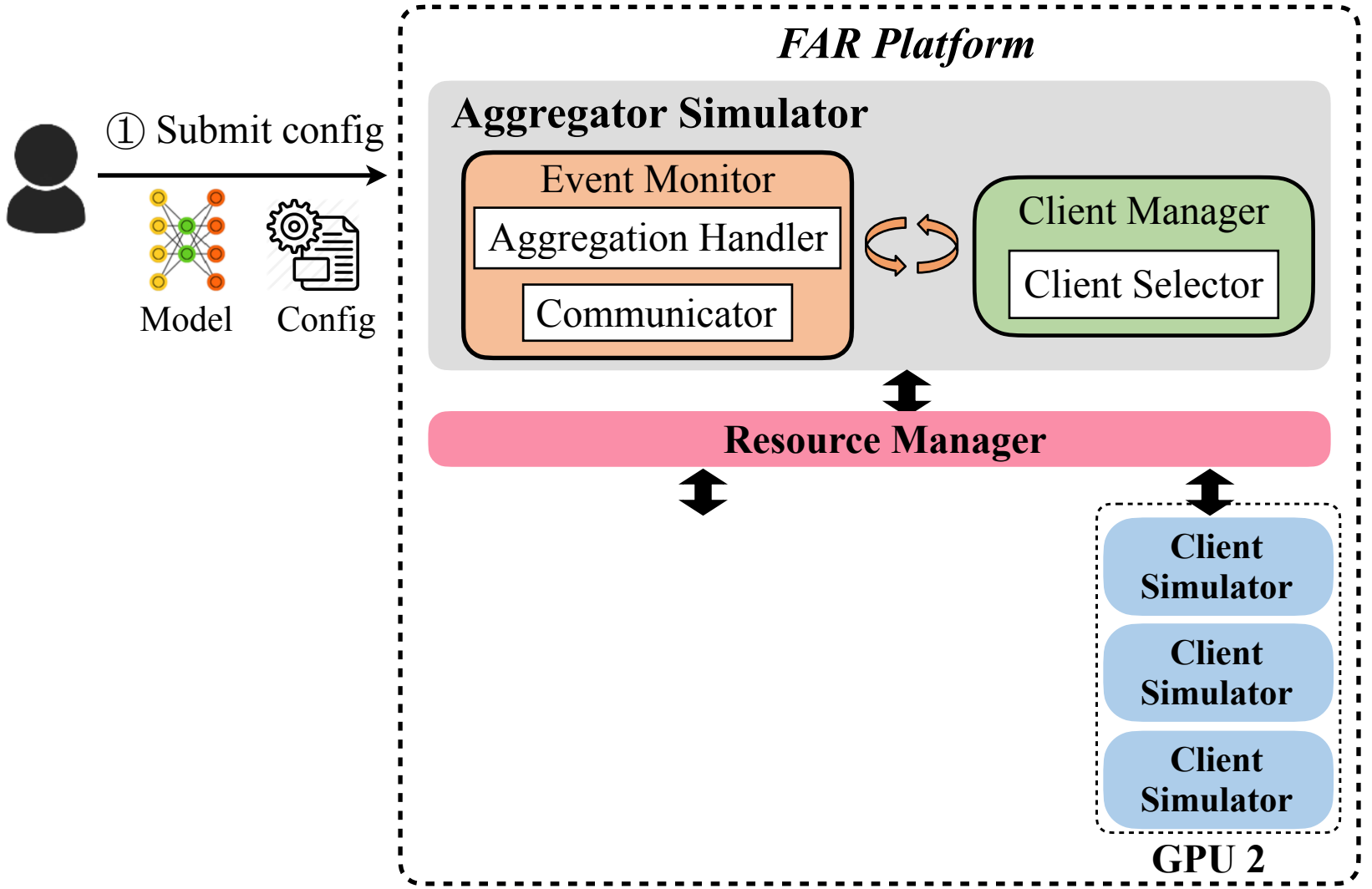
Scalable Eval Platform

#3

Easy Development

FAR: FedScale Automated Runtime

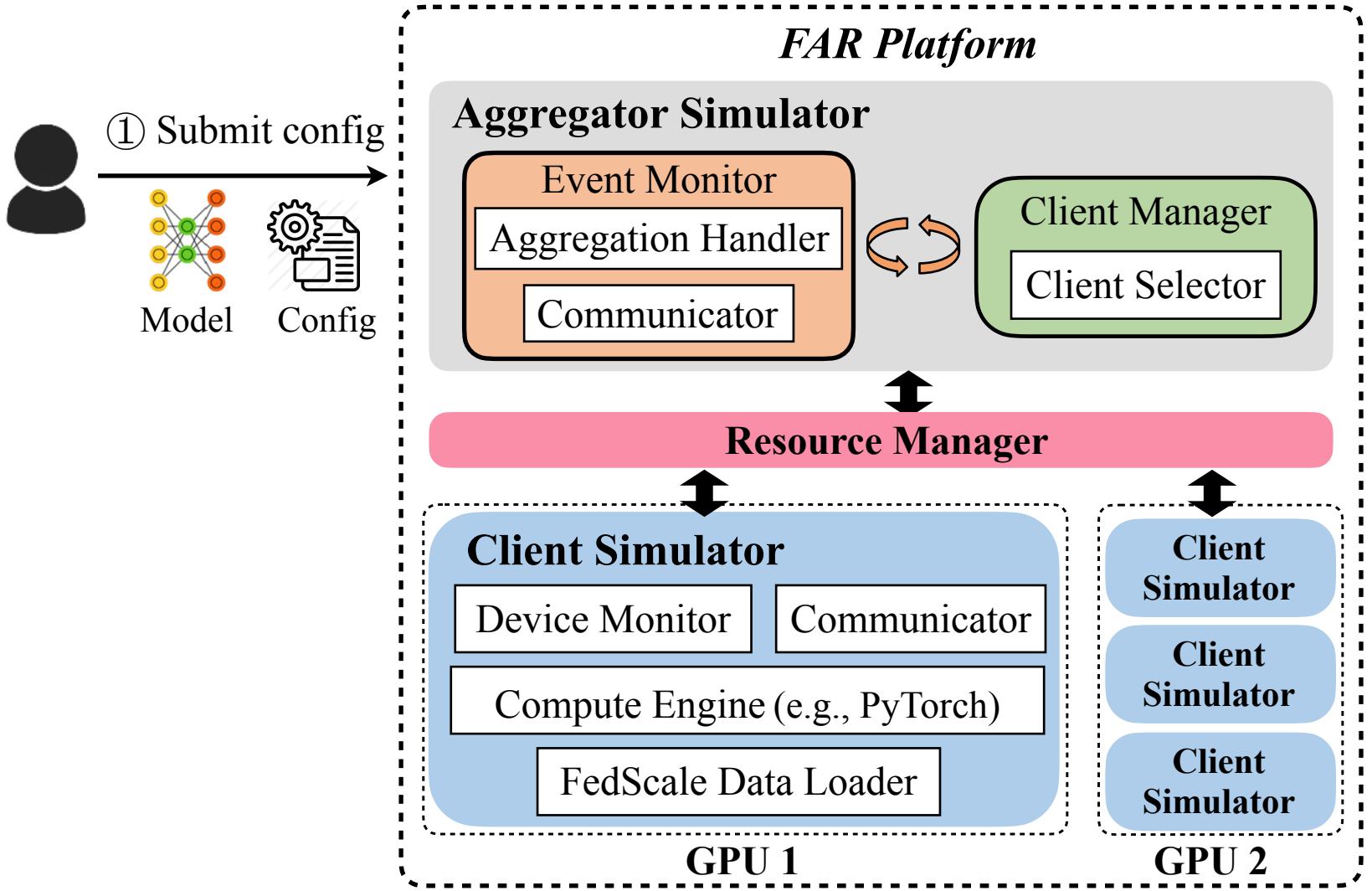
- **Scalable eval platform**
  - GPUs/CPUs
  - High resource util.



② Simulation of Practical FL

FAR: FedScale Automated Runtime

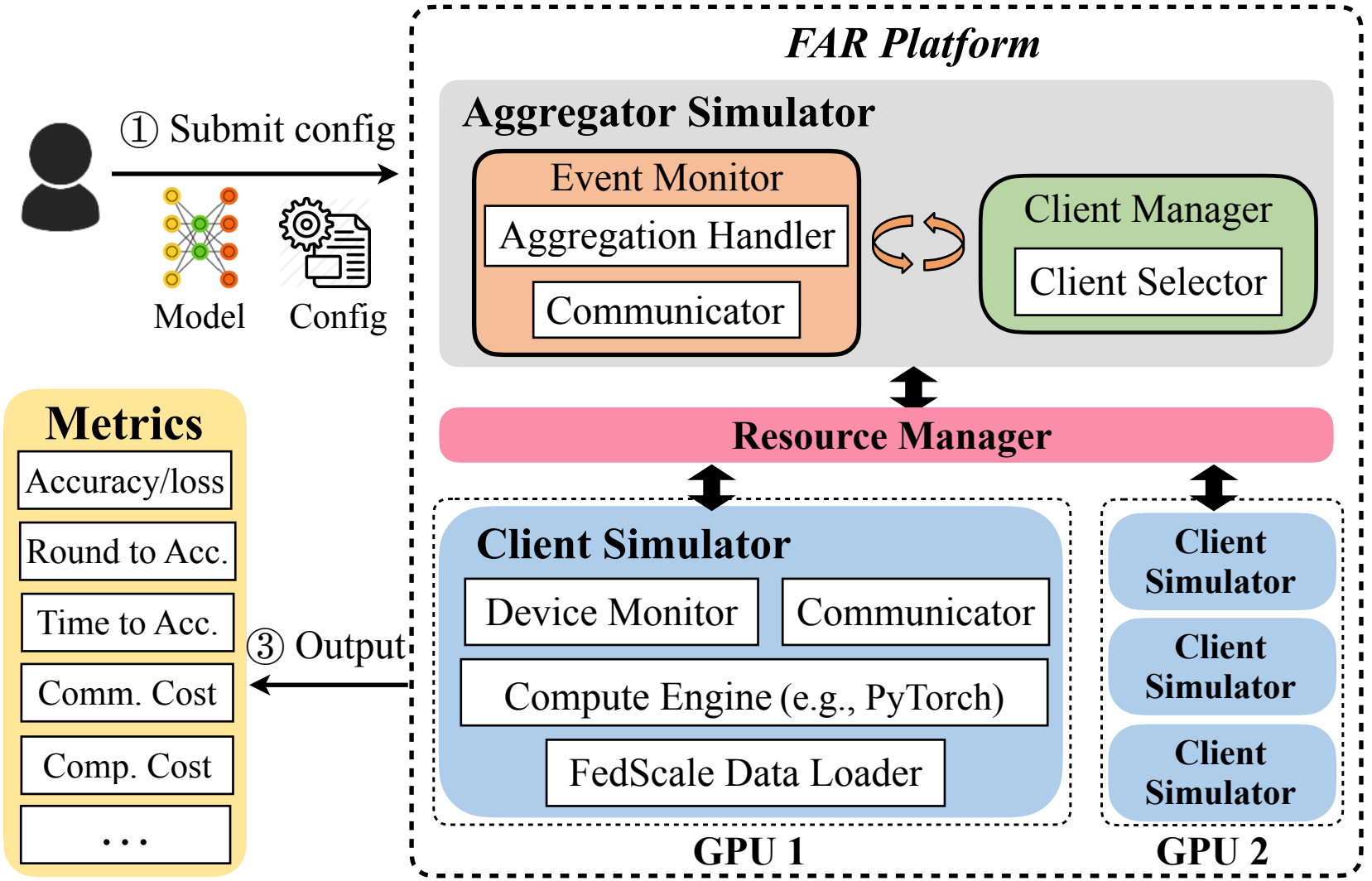
- **Scalable eval platform**
  - GPUs/CPUs
  - High resource util.



② Simulation of Practical FL

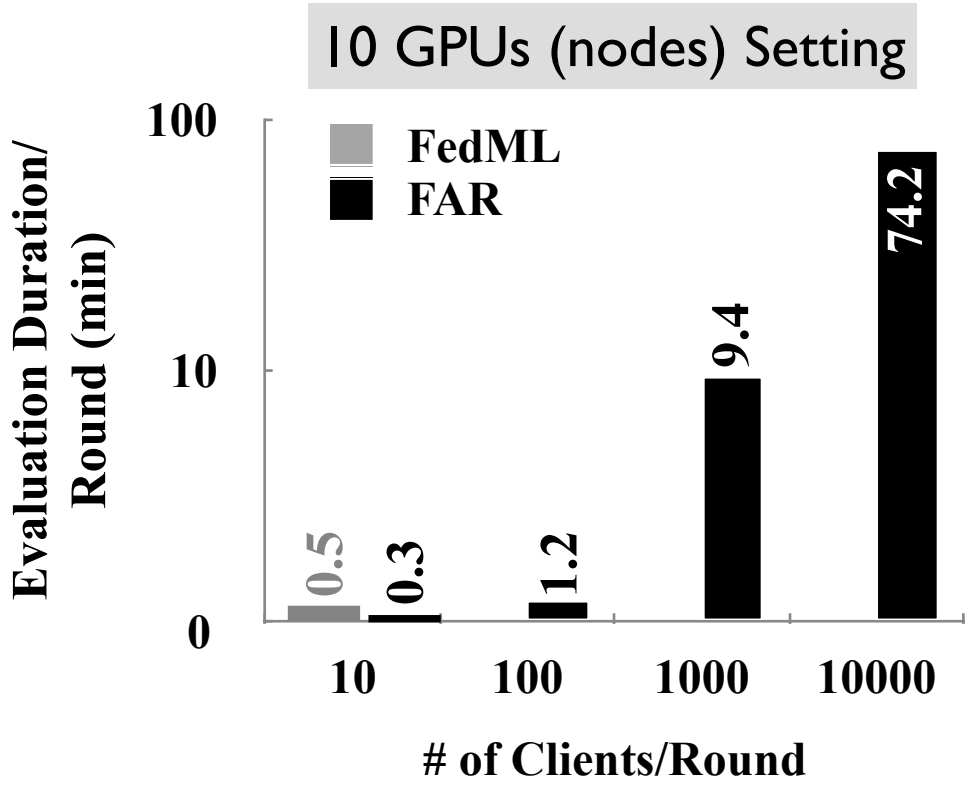
FAR: FedScale Automated Runtime

- **Scalable eval platform**
  - GPUs/CPUs
  - High resource util.
- **Practical FL runtime**
  - Convergence
  - System duration



FAR: FedScale Automated Runtime

- **Scalable eval platform**
  - GPUs/CPUs
  - High resource util.
- **Practical FL runtime**
  - Convergence
  - System duration



FAR can evaluate the practical scale of federated learning O(1k) participants/round

FAR: Easily-Deployable Benchmarking

- **Flexible APIs to automatically integrate new plugins**
  - Little effort to customize/benchmark new designs

Module	API Name	Example Use Case
Aggregator	`round_completion_handler`(*args)	Adaptive/secure model aggregation
Simulator	`client_completion_handler`(client_id, msg)	Straggler mitigation
	`push_msg_to_client`(client_id, msg)	Model compression
Client	`select_clients`(*args)	Client selection
Manager	`select_model_for_client`(client_id)	Adaptive model selection
Client	`train`(client_data, model, config)	Local SGD/malicious attack
Simulator	`push_msg_to_aggregator`(msg)	Model compression

Oort [OSDI'21]

Some Example APIs

FAR: Easily-Deployable Benchmarking

- **Flexible APIs to automatically integrate new plugins**
  - Little effort to customize/benchmark new designs

```
from fedscale.core.client import Client

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

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

A few lines are enough for benchmarking

FAR: Easily-Deployable Benchmarking

- **Flexible APIs to automatically integrate new plugins**
  - Little effort to customize/benchmark new designs

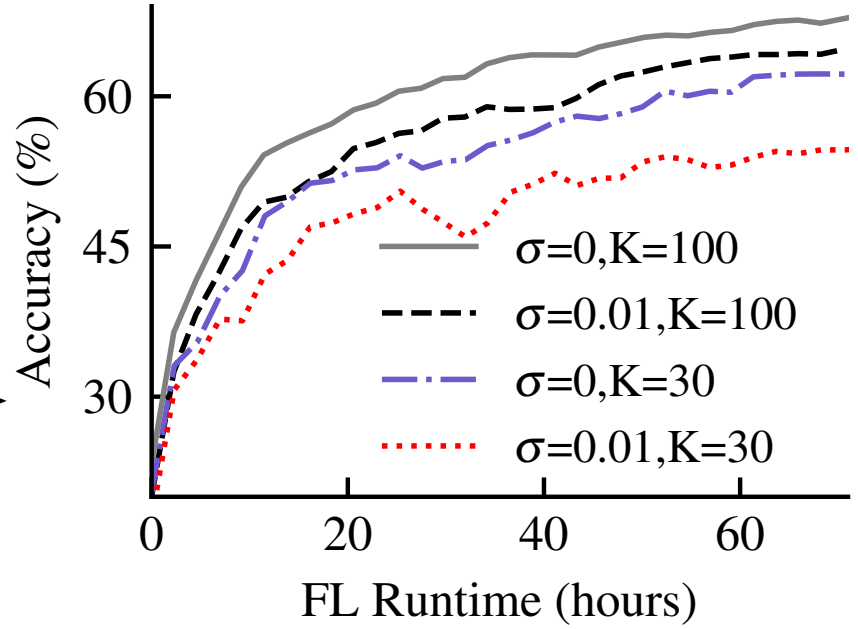
```
from fedscale.core.client import Client

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

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

A few lines are enough for benchmarking

Differential Private-SGD

σ (privacy target)

K (# participants/round)



FedScale can benchmark more realistic statistical/system performance

#1

Realistic FL Datasets

Statistical client datasets

- ~20 datasets w/ real partition
- Various FL tasks/scales

Client system traces

- Client device speed
- Device availability over time

#2

Scalable Eval Platform

Standardized benchmarking

- GPUs/CPUs
- Standalone/Distributed
- Practical FL runtime

#3

Easy Development

Easily-deployable to plugins

- Flexible APIs
- Extensible to new traces
- Automated integration

#1

Realistic FL Datasets

Statistical client datasets

- ~20 datasets w/ real partition
- Various FL tasks/scales

Client system traces

- Client device speed
- Device availability over time

#2

Scalable Eval Platform

Standardized benchmarking

- GPUs/CPUs
- Standalone/Distributed
- Practical FL runtime

#N

FedScale is open-source:

fedscale.ai

We welcome
your contributions!