Democratize Large-scale ML Training with Full-stack Systems

Yifan Qiao
Advisors: Harry Xu and Miryung Kim
Large-scale ML Model: A Double-edged Sword

Large-scale ML model powers many applications.

- Recommendation
- Game Playing
- Code Generation
Large-scale ML Model: A Double-edged Sword

Resource intensive (data, compute power, memory, etc.) and expensive.

Training GCN on large graphs:
- TBs of memory and
tens or hundreds of GPUs

Training GPT-3:
- ~355 Nvidia V100 GPU-years and
- ~$12 million

My Research: Full-stack Systems with Co-design

- Algorithms / Models
- Frameworks / Runtimes
- Infrastructures
- Hardware
My Research: Full-stack Systems with Co-design

- Algorithms / Models
- Frameworks / Runtimes
- Infrastructures
- Hardware

- ✔ Cross-layer optimizations
- ✔ Cheaper resources
- ✔ Performance isolation
Dorylus: GNN Training with Serverless Threads

Algorithms / Models

Frameworks / Runtimes

Infrastructure: serverless

Hardware

Graph Neural Network (GNN)

CPU servers + thousands of serverless threads offer **massive parallelism** with a **much lower price** than GPU servers!

**CPU**

$0.432/h

**Serverless**

Pay-as-you-go pricing

**GPU**

$3.06/h
Dorylus: GNN Training with Serverless Threads

- Async. pipelined training algorithm
- Frameworks / Runtimes
- Infrastructure: serverless
- Hardware

Thousands of + AWS Lambda

pipeline
Dorylus: GNN Training with Serverless Threads

Async. pipelined training algorithm
Framework: computation separation
Infrastructure: serverless
Hardware

Thousands of
pipeline

graph task

(tensor task)

AWS Lambda
Dorylus: GNN Training with Serverless Threads

Async. pipelined training algorithm

Framework: computation separation

Infrastructure: serverless

Hardware

pipeline

Thousands of

AWS Lambda

graph task

tensor task

✔

✔

✔
Dorylus: GNN Training with Serverless Threads

- Async. pipelined training algorithm
- Framework: computation separation
- Infrastructure: serverless
- Hardware

1.2× faster and 4.8× cheaper than GPUs with convergence guarantee!
Canvas: Isolated & Semantics-aware Data Plane

- Algorithms / Models
- Frameworks / Runtimes
- Infrastructures
- Hardware

Improve the throughput by up to $6.1 \times$
Reduce the performance variation by $31 \times$

- Runtime
  - OS: isolation & semantics-aware
  - Network: scheduler
My Research: Full-stack Systems with Co-design

- Algorithms / Models
- Frameworks / Runtimes
- Infrastructures
- Hardware

- ✔ Higher accuracy
- ✔ Better resiliency
- ✔ Higher throughput
- ✔ Lower cost

UCLA
Samueli
School of Engineering