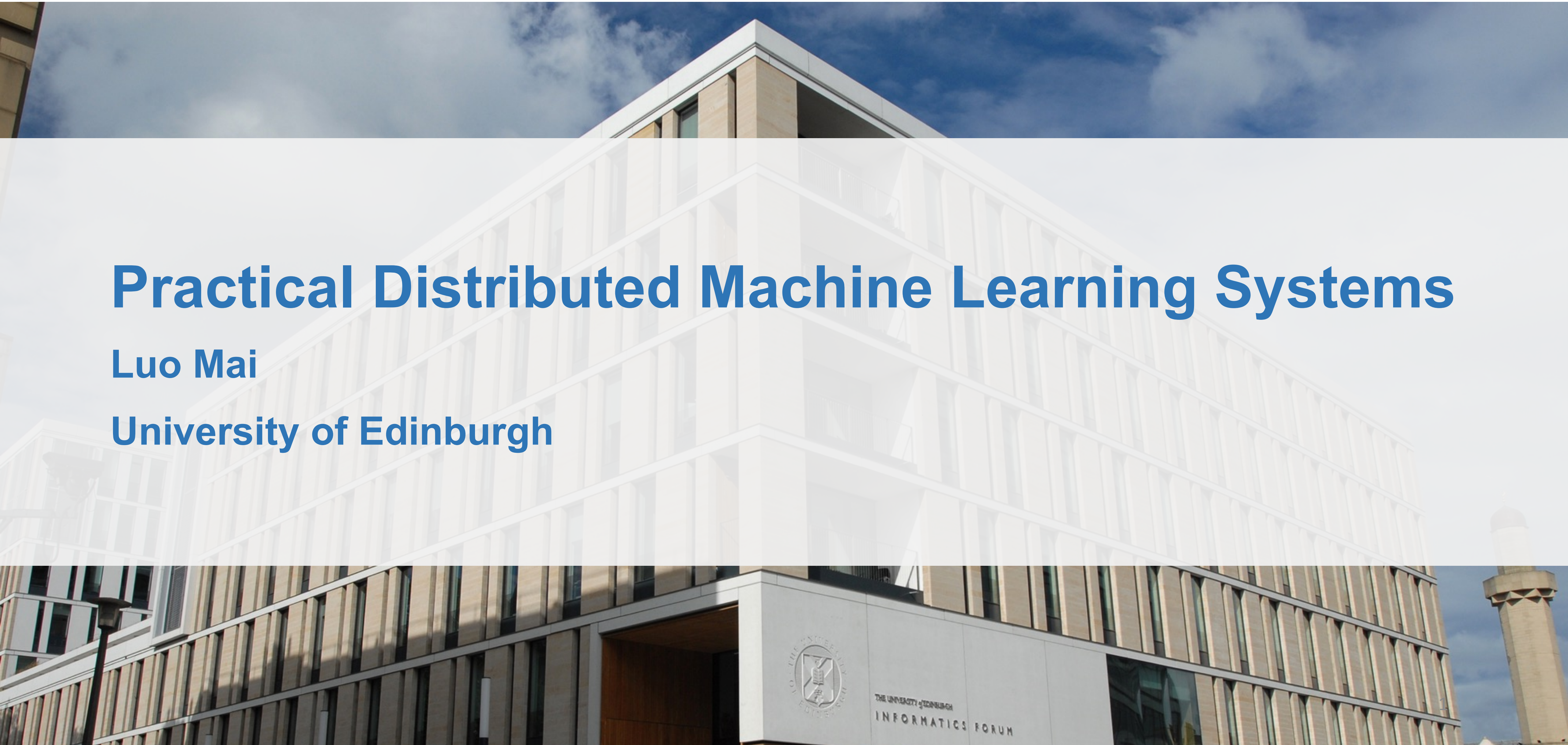
THE UNIVERSITY of EDINBURGH
informatics

Practical Distributed Machine Learning Systems

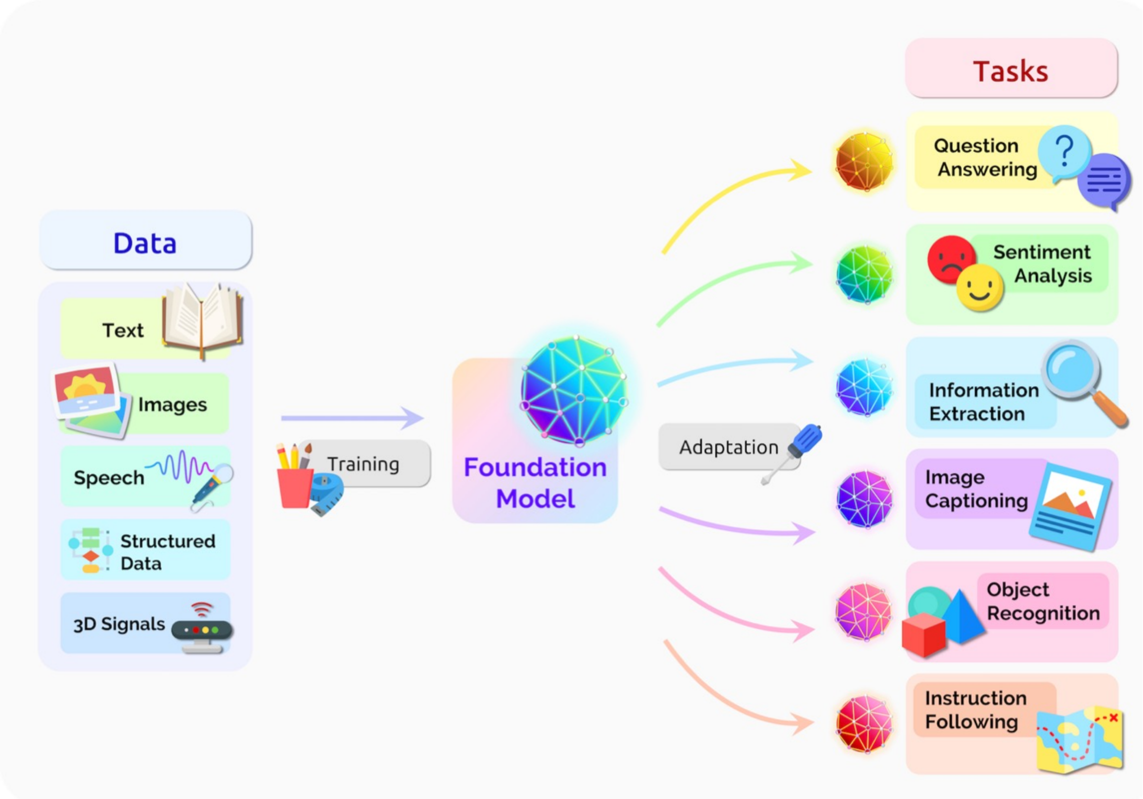
Luo Mai

University of Edinburgh



THE UNIVERSITY of EDINBURGH
informatics

Foundation Models



NLP Models:

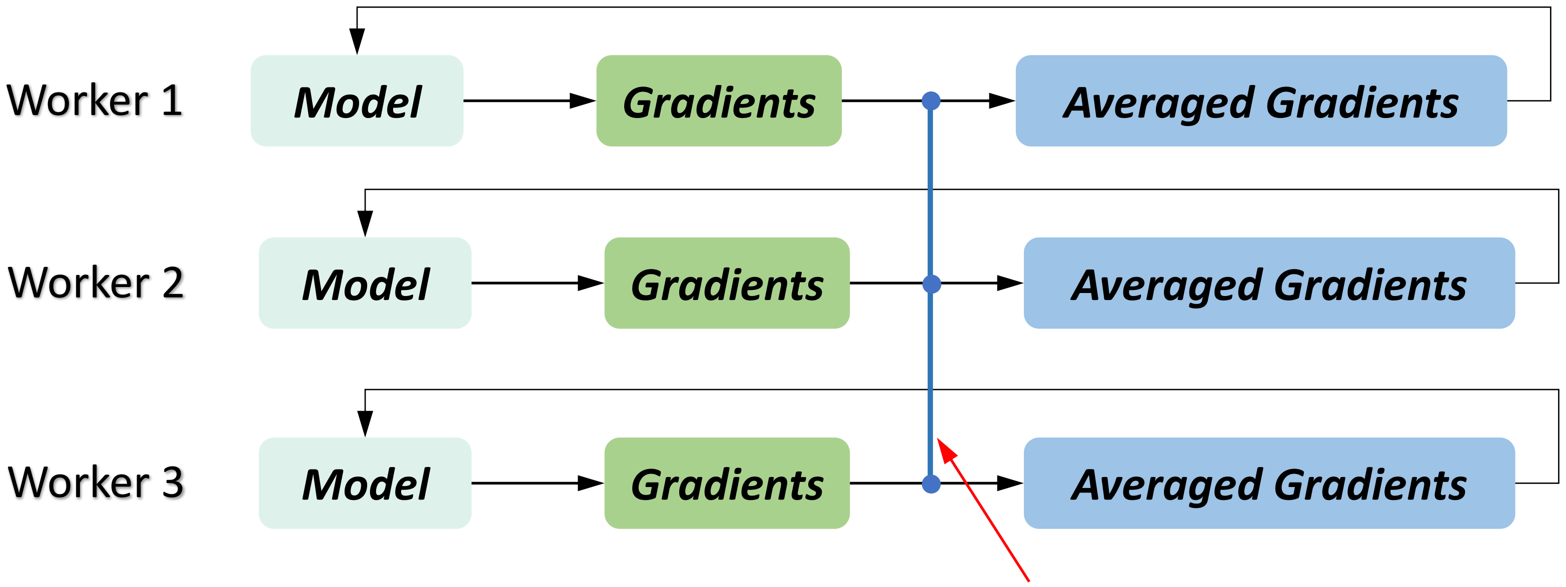
- GPT-3 (175 billion parameters)
- Switch Transformer (1 trillion parameters)

CV Models:

- Vision Transformers (10s billions parameters)

[1] On the Opportunities and Risks of Foundation Models, 2021

Synchronising Gradients is Network-Intensive

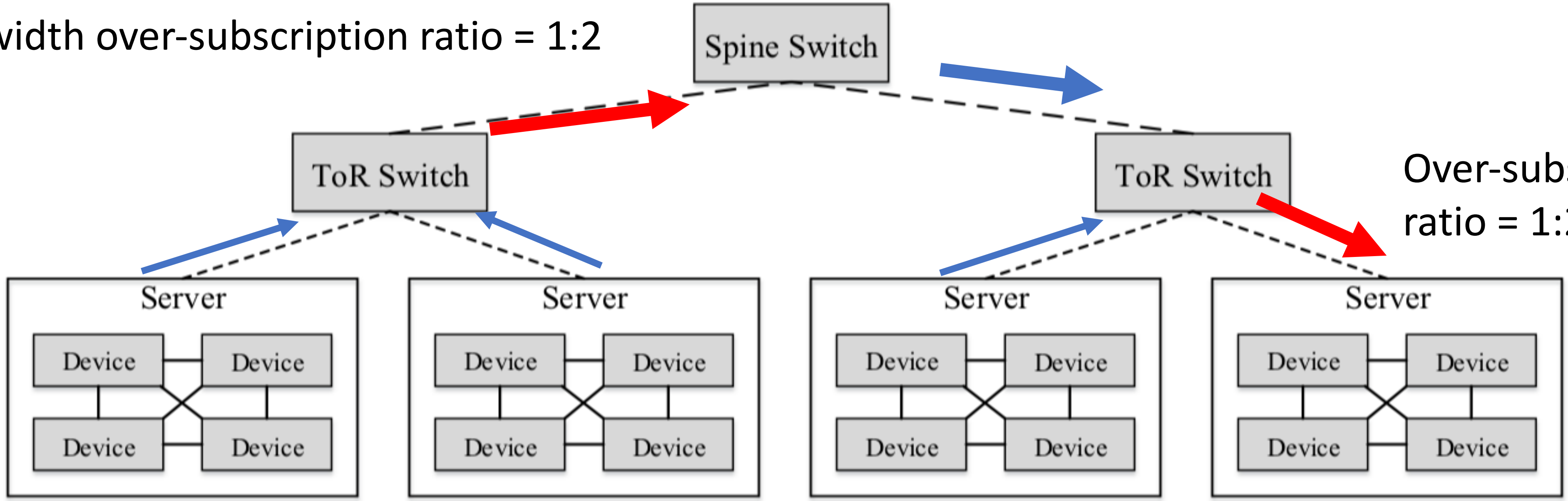


GPT-3 model: 175 billion parameters
175G x 4-byte float = 700 GB

Compute averaged gradients
1. Collect local gradients
2. Broadcast averaged gradients to all workers

Bandwidth is Limited in Data Centres

Bandwidth over-subscription ratio = 1:2

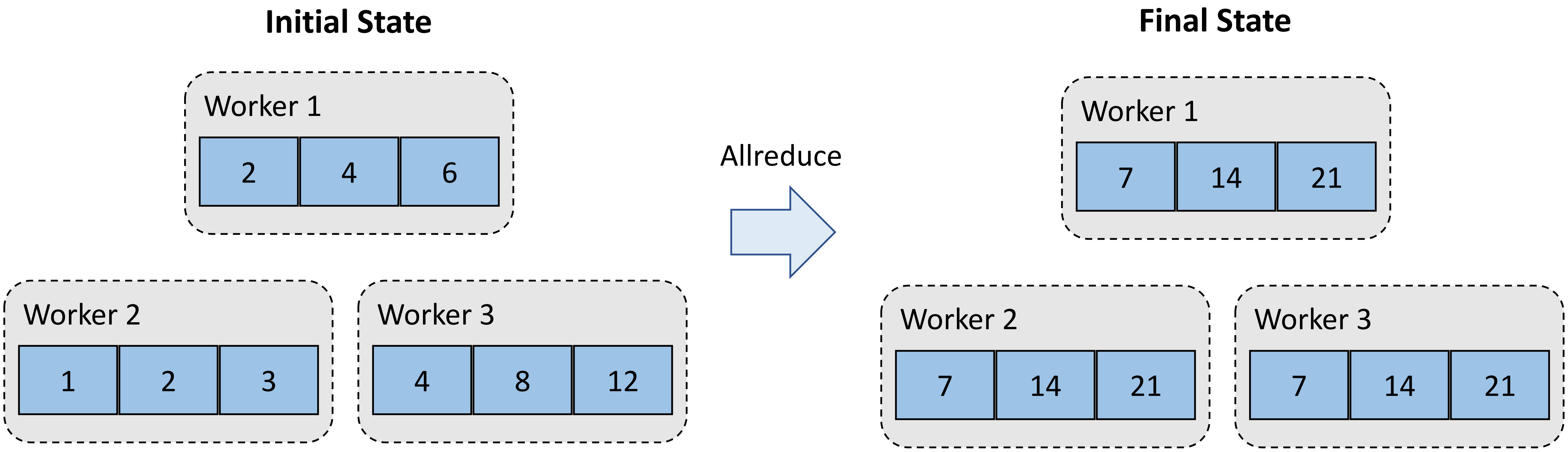


Over-subscription ratio = 1:2

Worker | Worker | Worker | Worker (Aggregator)

A real-world data centre

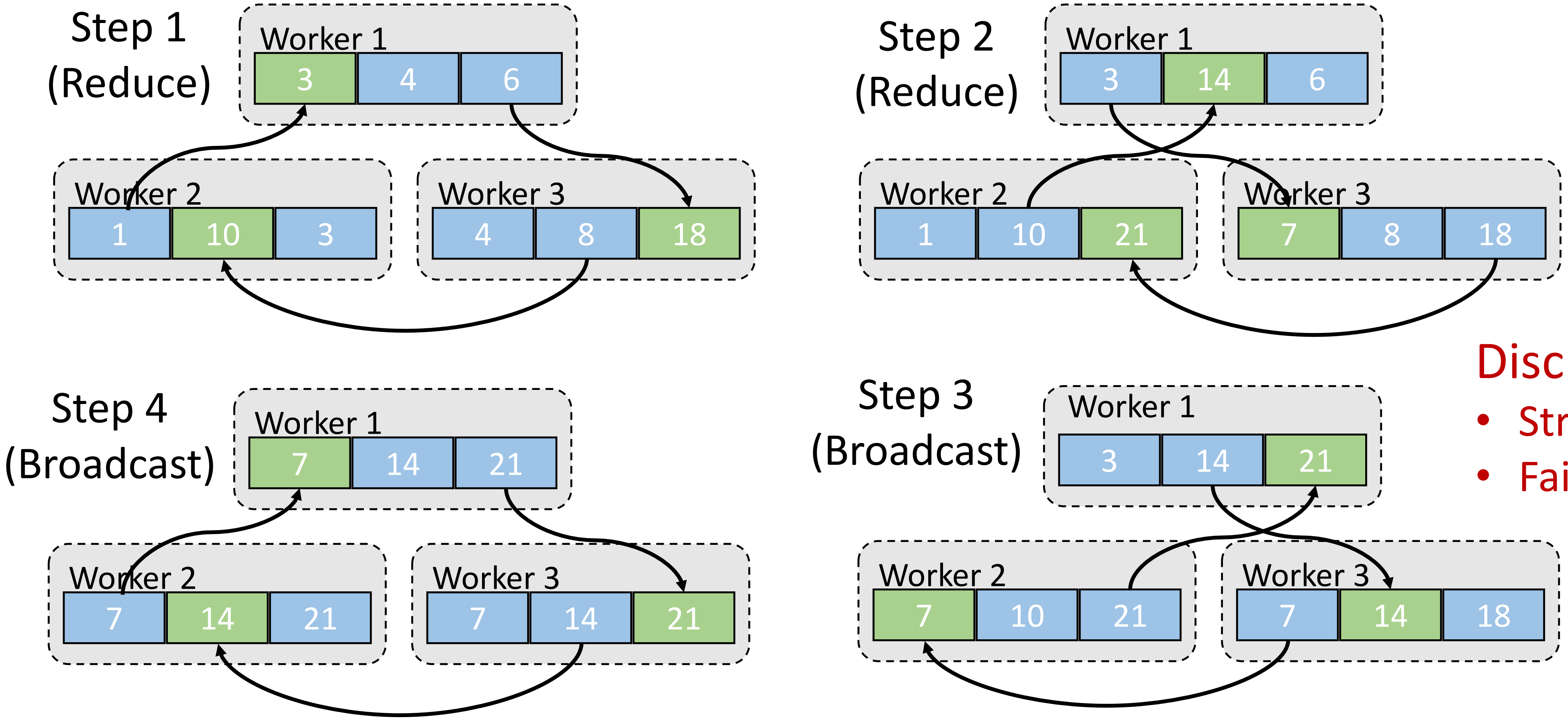
Bandwidth-efficient Allreduce



7 = 1 + 2 + 4 (first partition)

14 = 2 + 4 + 8 (second partition)

21 = 3 + 6 + 12 (third partition)



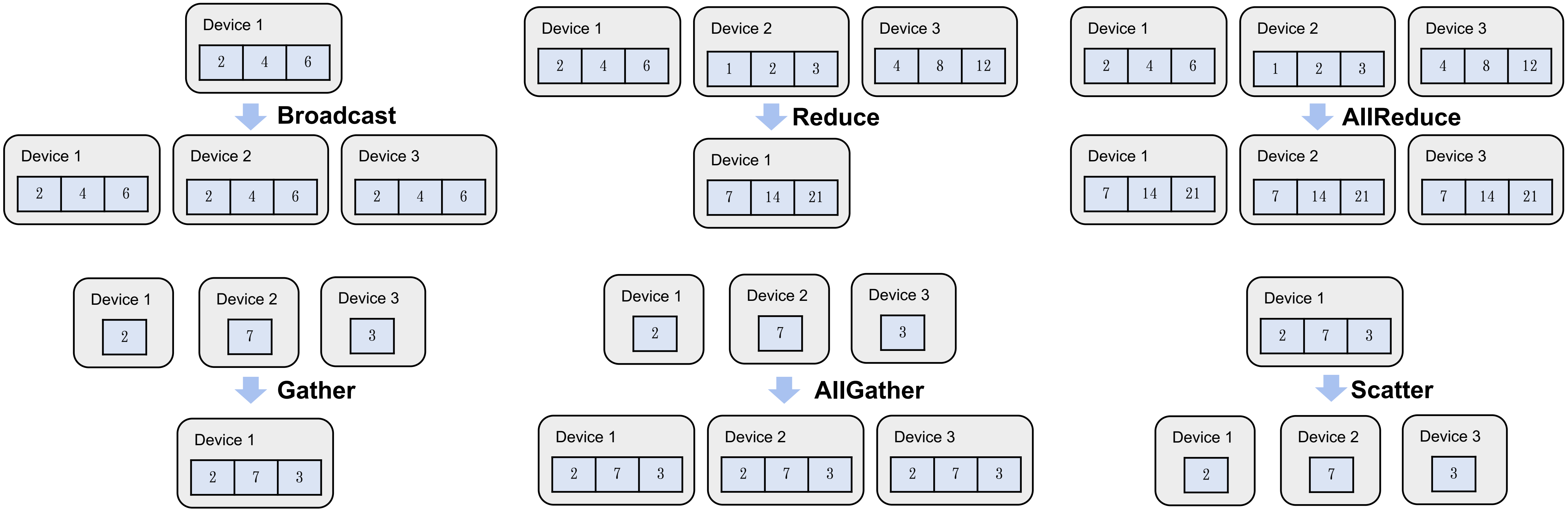
Discussion

- Stragglers
- Failures

Questions?

THE UNIVERSITY of EDINBURGH
informatics

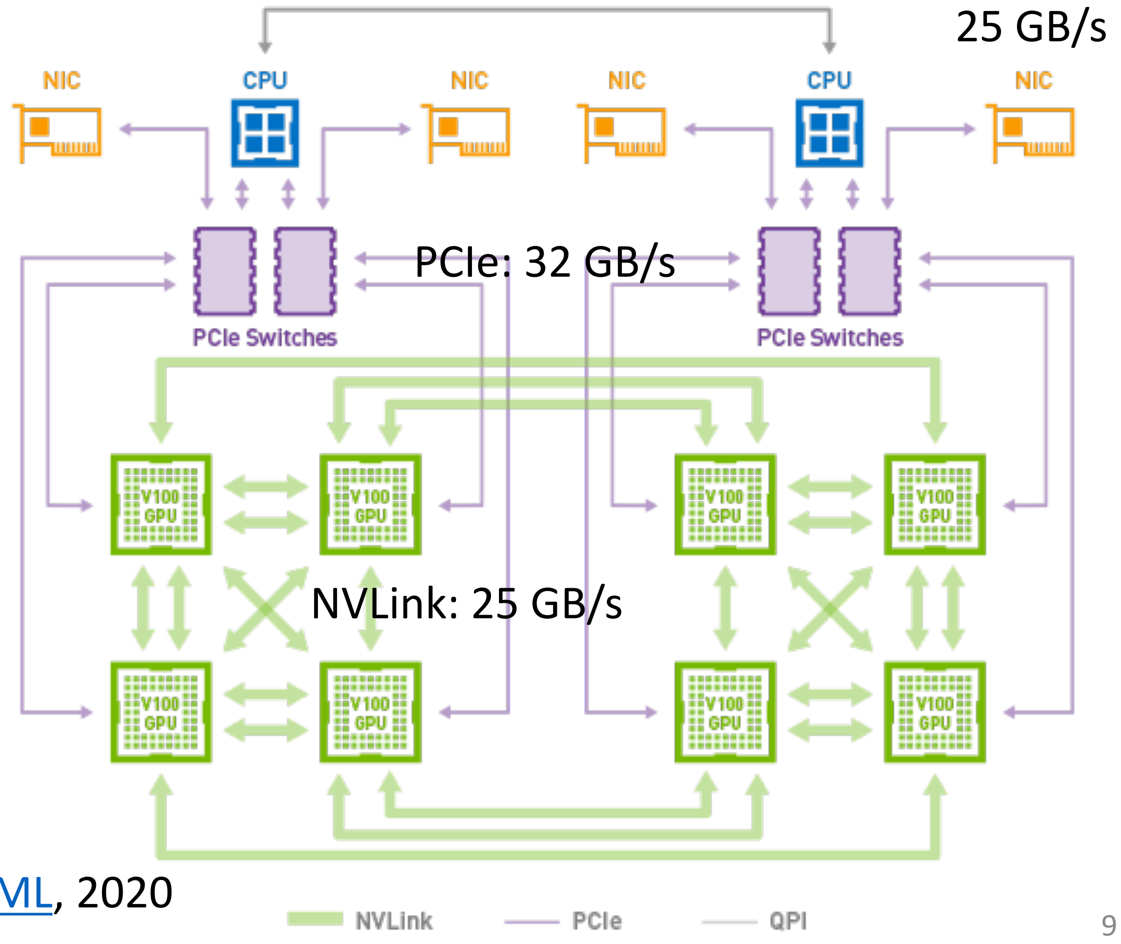
Common Collective communication operators



- Collective communication libraries: NVIDIA Collective Communications Library (NCCL), OpenMPI
- Integration with ML frameworks: PyTorch Distributed, Horovod, KungFu

ML Server Architecture

- Memory hierarchy
  - GPU memory bandwidth: ~2000 GB/s
  - System memory bandwidth: ~1600 GB/s
  - SSD: ~20 GB/s
- High-bandwidth networks
  - GPU-GPU direct: NVLink: 600 GB/s
  - CPU-GPU PCIe Switch: 64GB/s
  - Server-Server InfiniBand: 25 GB/s
- ML data placement
  - How to improve data locality?
  - How to reduce data access latency?



[1] Blink: Fast and Generic Collectives for Distributed ML, 2020

Future Systems for Foundation AI Models

- Automatic model parallelism
  - Parallelism cost model + parallelism strategy solver [1]
- Memory-efficient runtime
  - Memory swapping (CPU memory + GPU memory) [2]
- Efficient optimisers for giant AI models
  - Lookahead Optimizer [3]

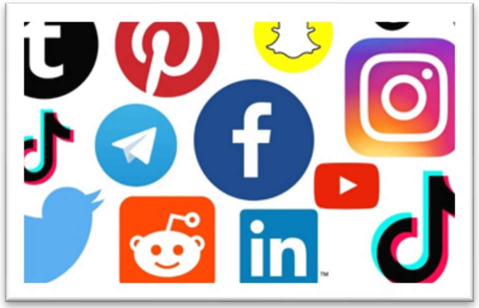
[1] GSPMD: General and Scalable Parallelization for ML Computation Graphs, 2021
[2] ZeRO-Offload: Democratizing Billion-Scale Model Training, 2021
[3] Lookahead Optimizer: k steps forward, 1 step back, 2020

Questions?

Recommender Systems

Digital Content
2.7 Billion
Monthly Active Users

E-Commerce
2 Billion
Digital Shoppers

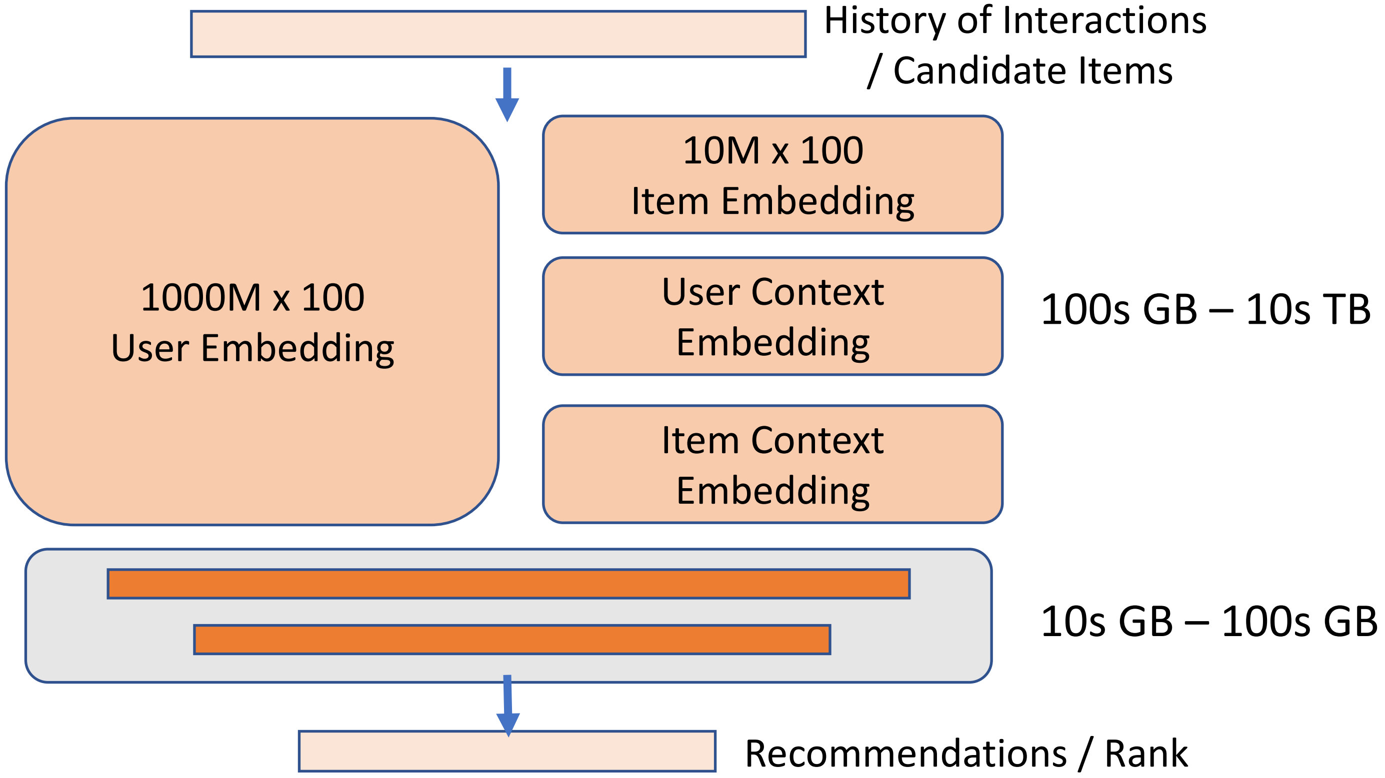


Social Media
3.8 Billion
Active Users

Digital Advertising
4.65 Billion
User Targeted

Deep Learning Recommender Models

Recommender system

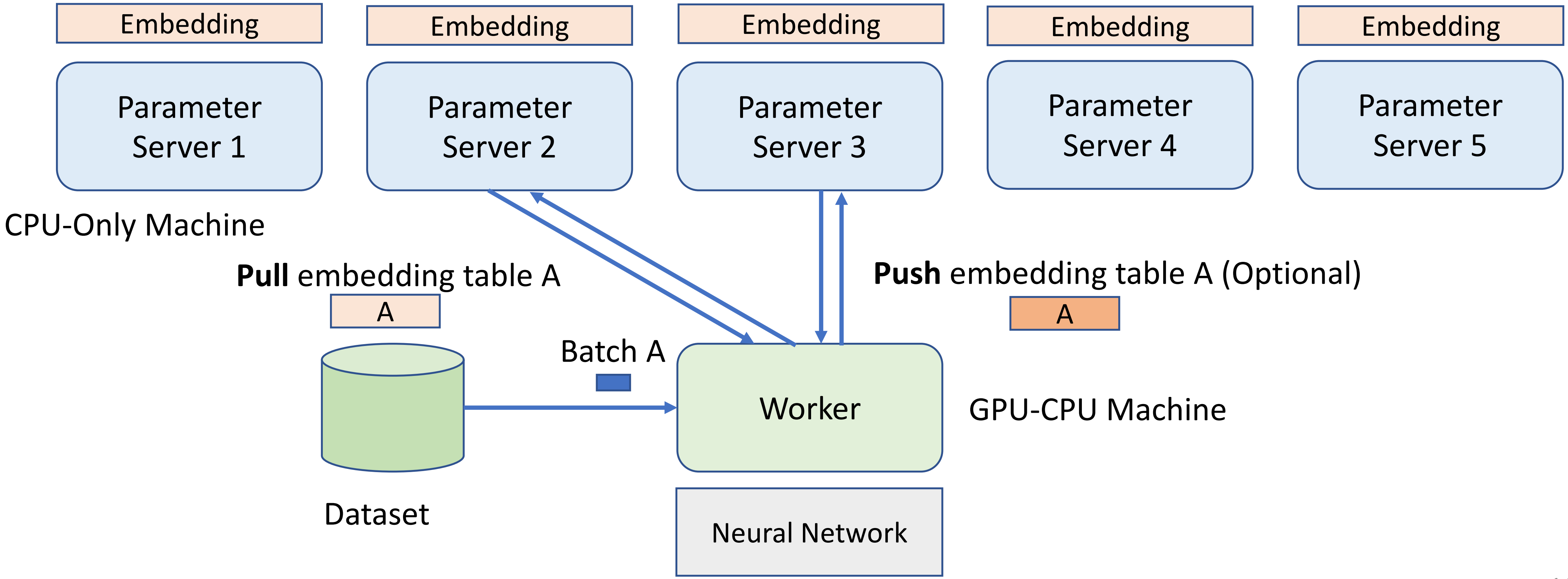


- Tremendous memory cost
- Small computation cost
- Embeddings are **sparsely updated**

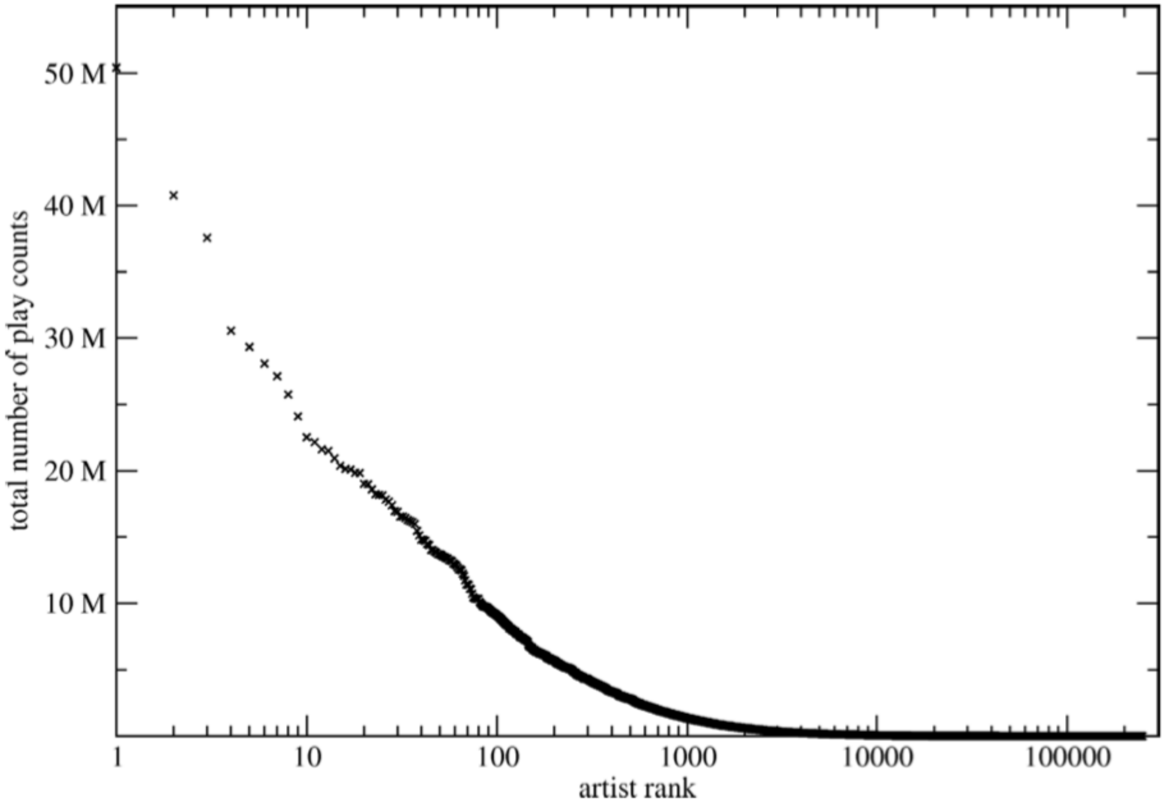
- Example: Deep learning recommendation models [1]

[1] DLRM: An advanced, open source deep learning recommendation model, 2020

Parameter Servers

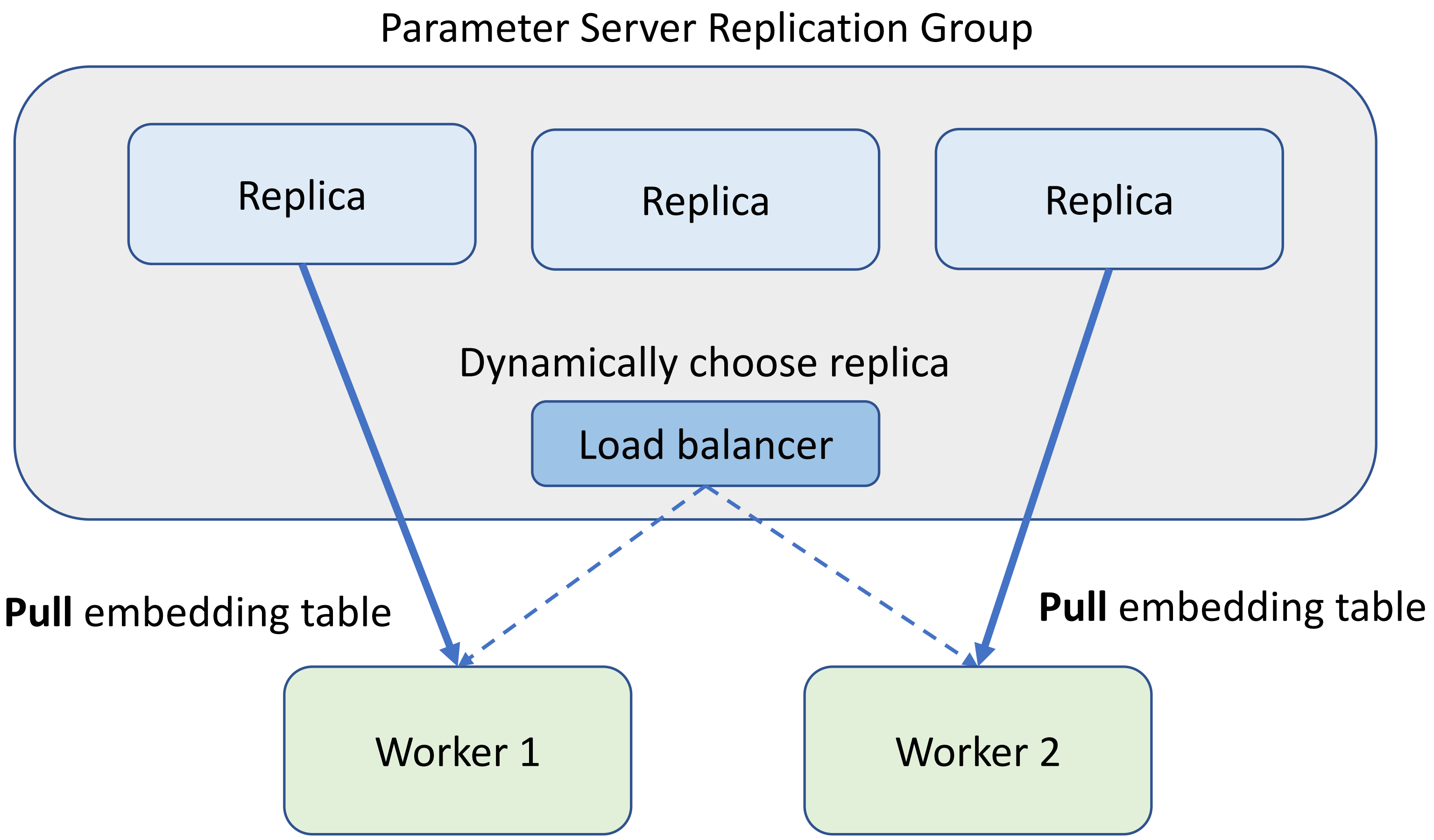


Users/Items Follow Power Law Distribution



Imbalanced workload is common in distributed computing systems

Handling Hot Spots via Data Replication



Future Recommender Systems

- Protecting user privacy
  - Federated Learning (FL)
    - Google Input Method with FL [1]
  - Trusted Execution Environments (TEEs)
    - x86 TEEs (AMD & Intel)
    - GPU TEE (Nvidia), coming soon

- Adaptive recommender engines
  - Updating models in real-time: Ekko [2]

[1] Towards Federated Learning at Scale: System Design, MLSys 2019
[2] Ekko: A Large-Scale Deep Learning Recommender System with Low-Latency Model Update, OSDI 2022

Summary

- Foundation AI models
- Collective communication systems
  - Bandwidth over-subscription, stragglers, failures, high-speed networks
- Deep learning recommendation systems
- Parameter servers
  - Data skews, replications

Reading

- Optional reading
  - [Dive into deep learning – computational performance]

THE UNIVERSITY of EDINBURGH
informatics

Questions?

Largs-Scale Computer Systems Group
https://luomai.github.io