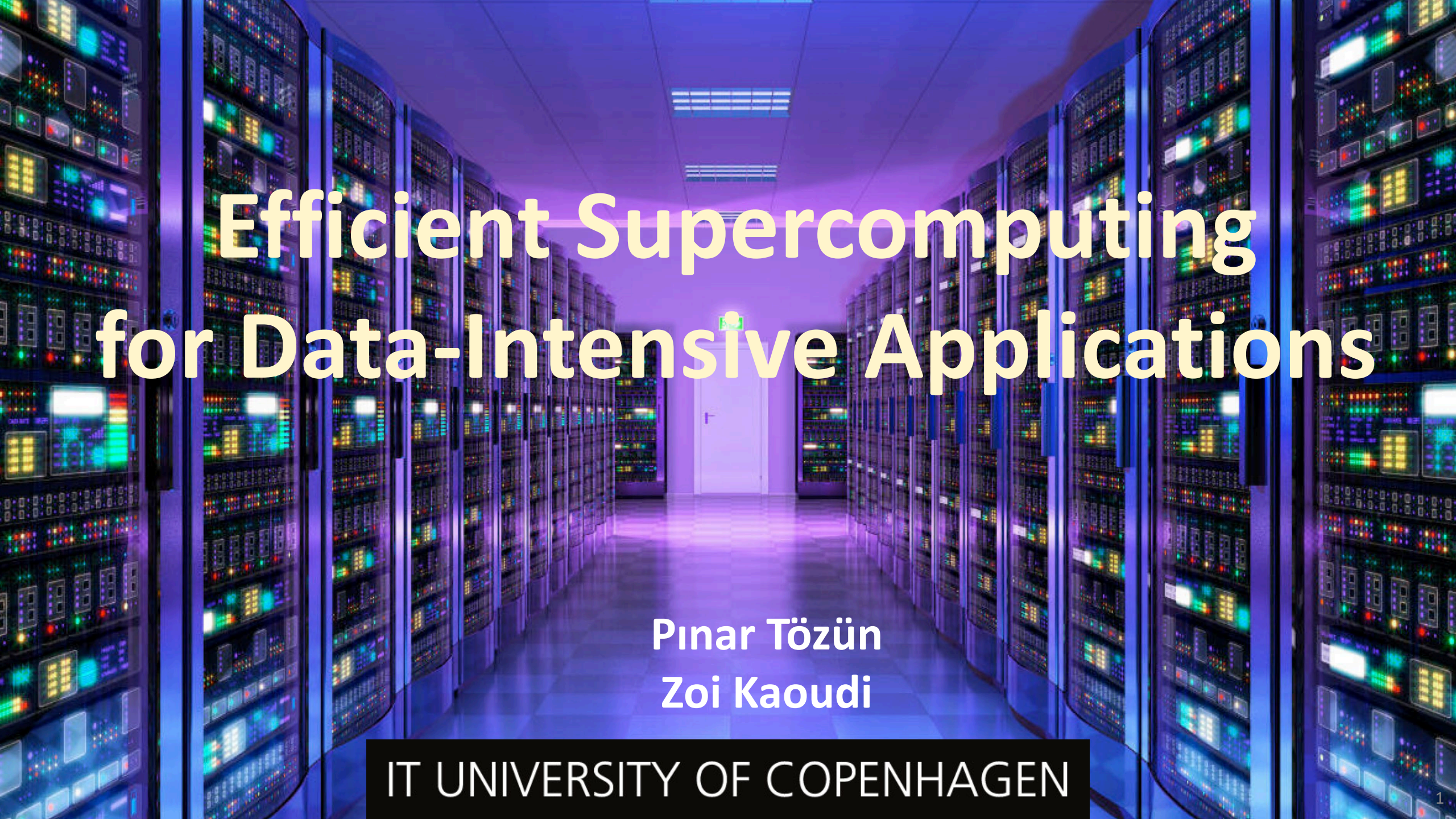


Efficient Supercomputing for Data-Intensive Applications

Pınar Tözün

Zoi Kaoudi

IT UNIVERSITY OF COPENHAGEN

Efficient Supercomputing for Data-Intensive Applications

Pınar Tözün
Associate Professor
IT University of Copenhagen

Zoi Kaoudi
Associate Professor
IT University of Copenhagen

Resource-aware Deep Learning

Cost-aware Data Platform Integration

DASYA

resource-aware deep learning

Pınar Tözün

Associate Professor, IT University of Copenhagen

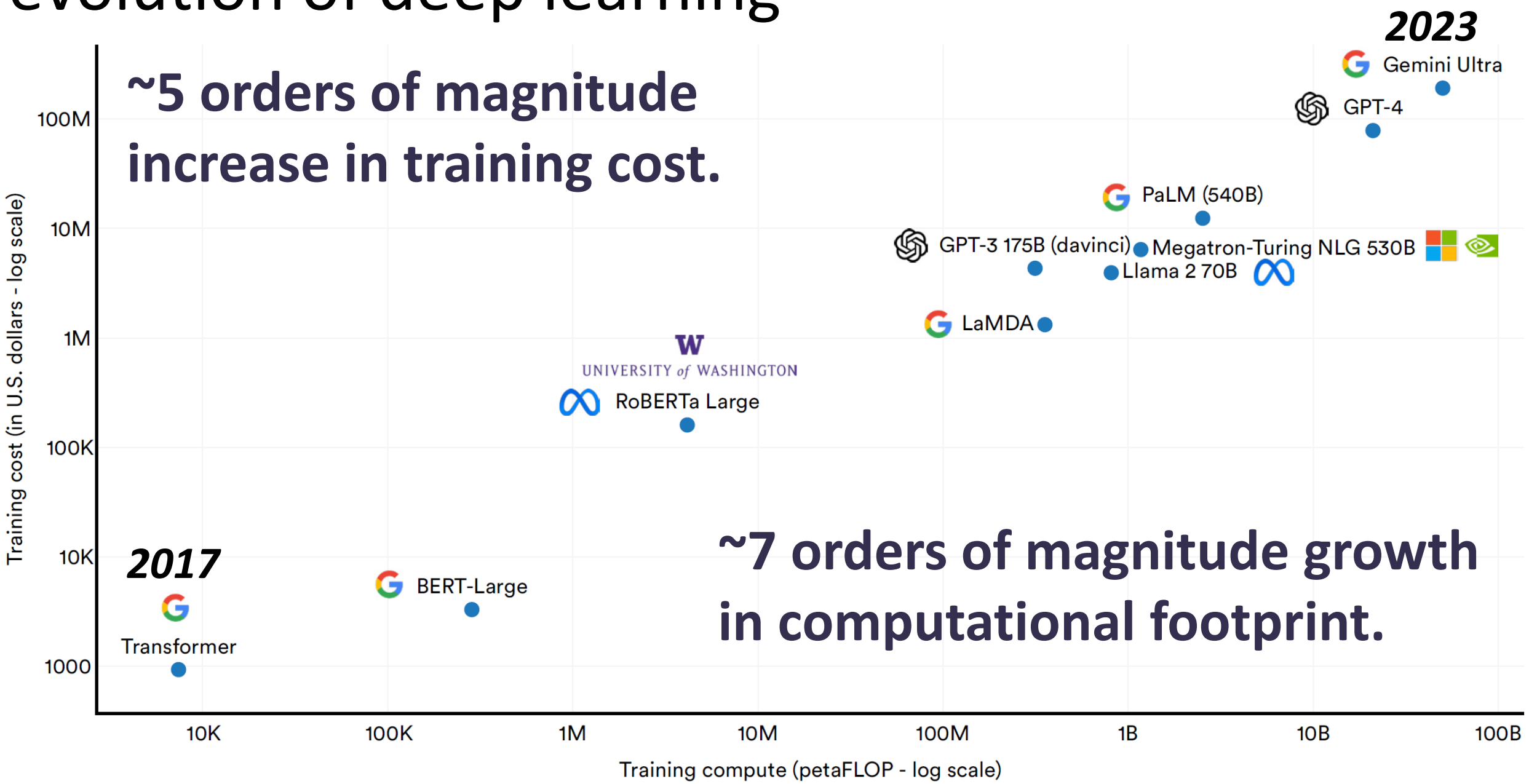
pito@itu.dk, www.pinartozun.com, @pinartozun

INDEPENDENT RESEARCH FUND DENMARK

novo nordisk foundation

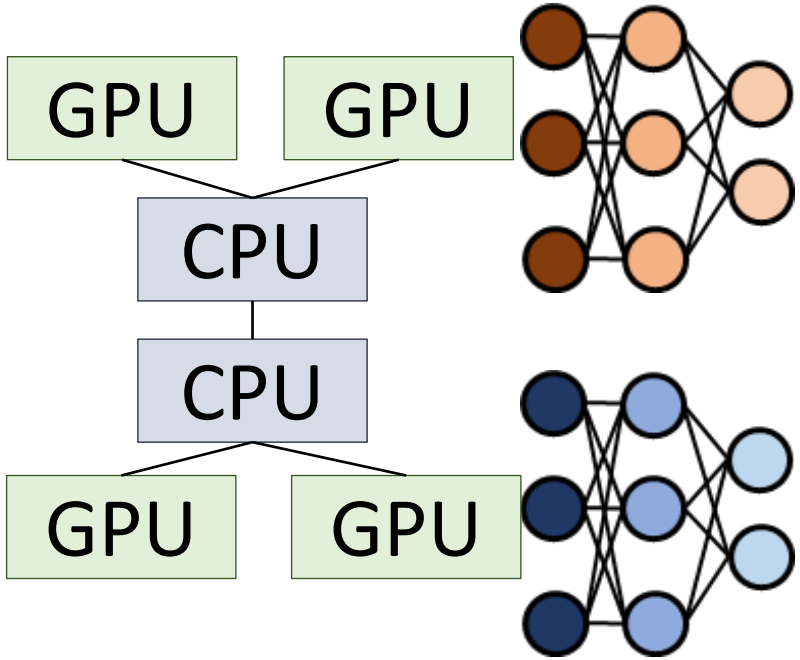
Digital Tech Summit
31/10/2024

evolution of deep learning

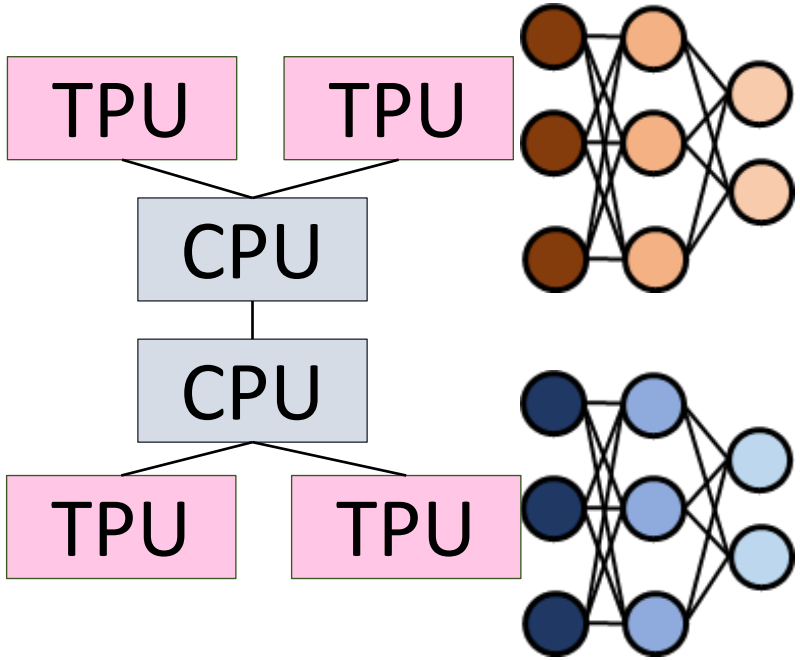


source: Stanford AI Index Report 2024

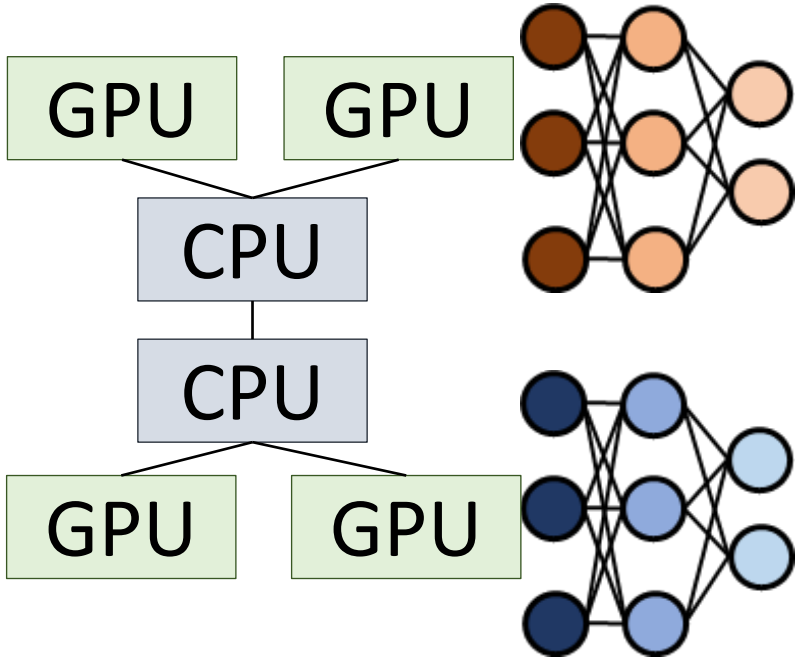
deep learning hardware



deep learning hardware



deep learning hardware



in real-world*, 52% GPU utilization

on average for 100,000 jobs

*Jeon et al. "Analysis of Large-Scale Multi-Tenant GPU Clusters for DNN Training Workloads." USENIX ATC 2019

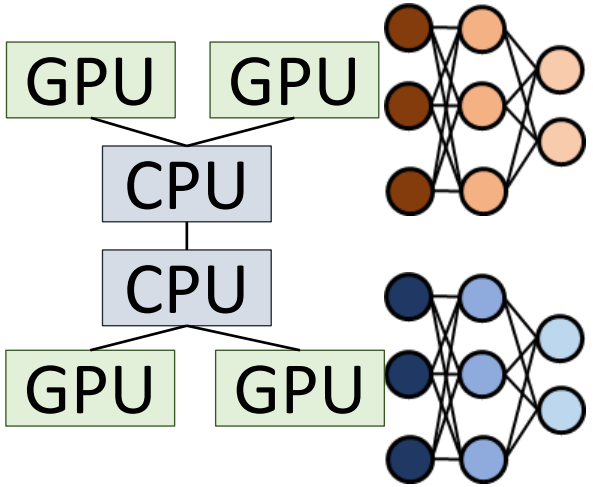
can we do better while using fewer hardware resources?

hardware resource management

conventional wisdom

➔ exclusive GPU access per job

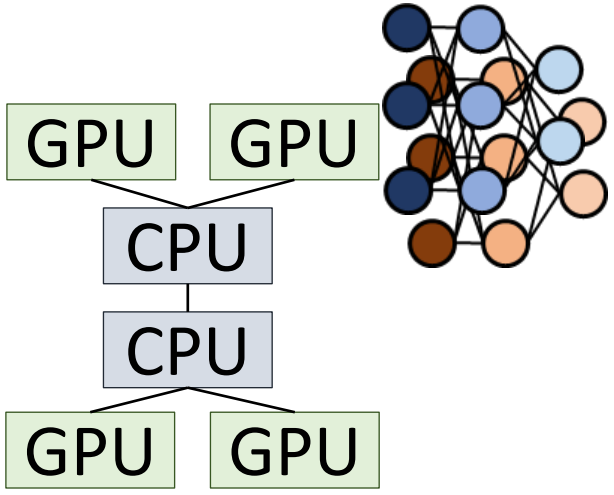
➔ pessimistic, but easy to manage



workload collocation on GPUs

➔ leads to better GPU utilization

➔ reduces costs



need for resource managers that incorporate GPU collocation!

data path of deep learning training

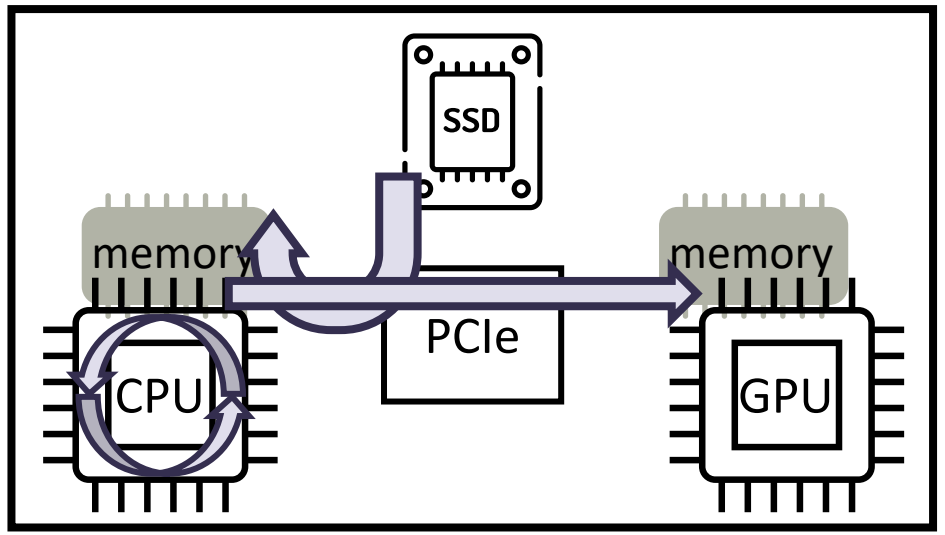
CPU feeds the accelerators with data.

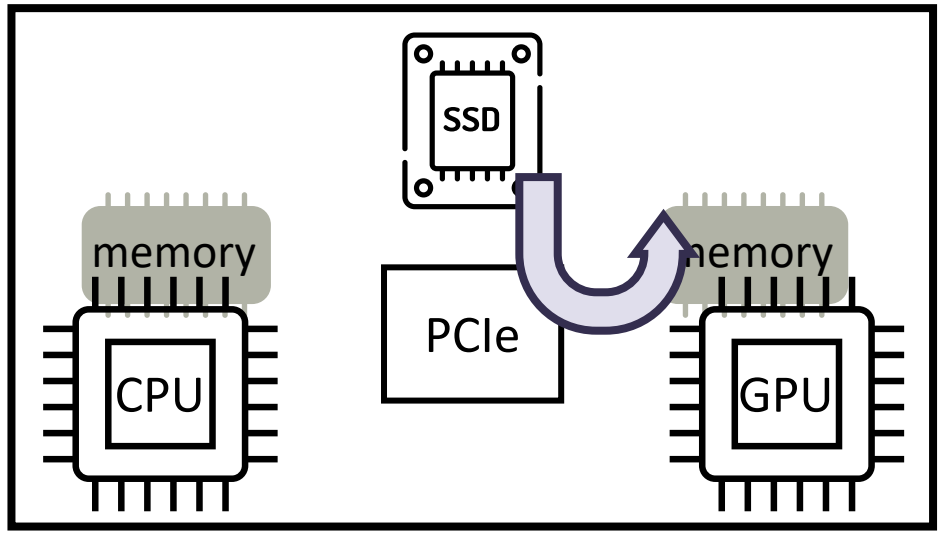
- 16-64 cores per GPU
- 96 cores per TPU*

more direct data paths exist!

need to make such paths accessible to deep learning practitioners!

conventional

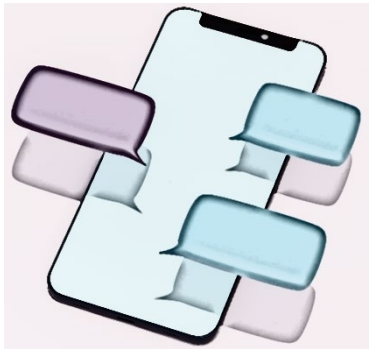




direct

*Audibert et al., "tf.data service: A Case for Disaggregating ML Input Data Processing" ACM SoCC 2023

data processing @ the edge



conventional-approach

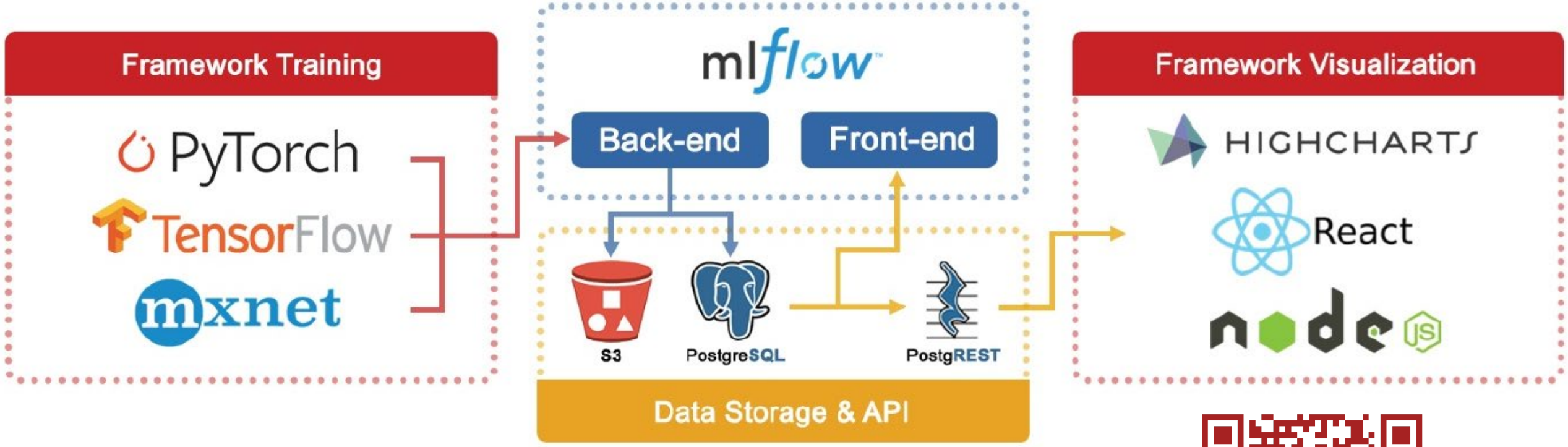
- do (most) data processing in the cloud

cannot satisfy

- low-latency & real-time applications
- poor / non-existing connectivity
- legal restrictions & privacy

need for efficient & complex data processing closer to data sources; *at the edge*!

how to monitor hardware?



➔ easy, extensible, and scalable tracking of hardware metrics

➔ frontend for data exploration

used by several members of our group, including data scientists, for systematic benchmarking of deep learning training

can we do better with fewer resources?

yes, but no free lunch!

- must have more effective workload collocation on accelerators
- data path requires optimizations to reduce data movement
- different scales of hardware devices need different tools
- higher awareness on hardware utilization

thank you!