Preview Session: Machine Learning Training

Deepak Narayanan
Stanford University (soon-to-be Microsoft Research)

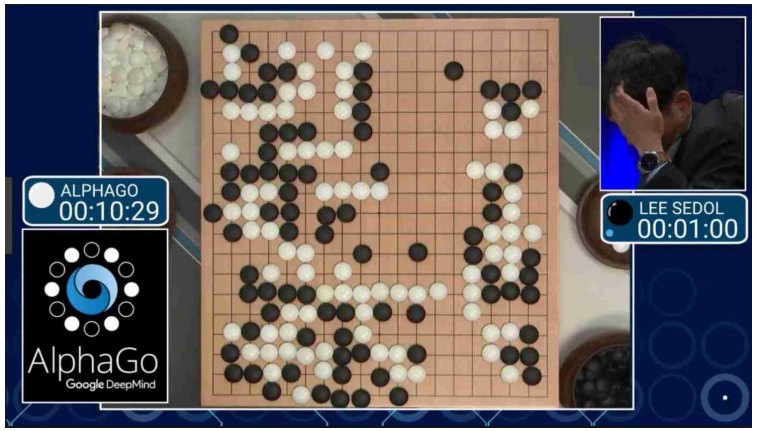
Deep Learning is powering new applications

வணக்கம் என் பெயர் தீபக்

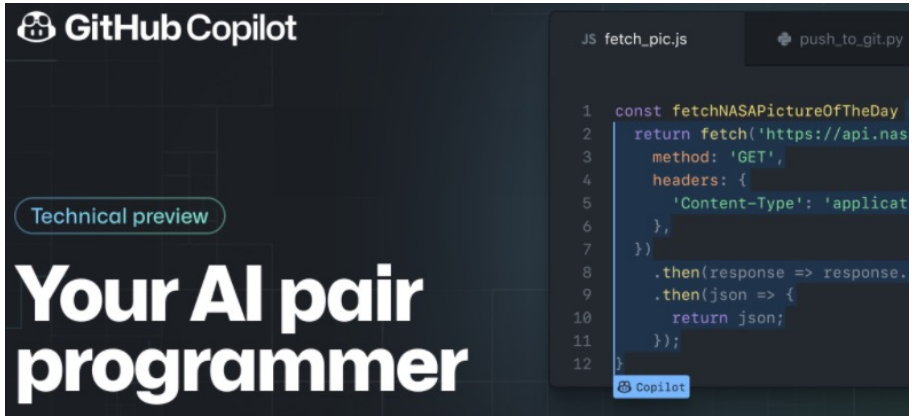
Hello, my name is Deepak

Machine Translation

Speech-to-Text



Game Playing



Code Autocompletion

…but extremely compute intensive!

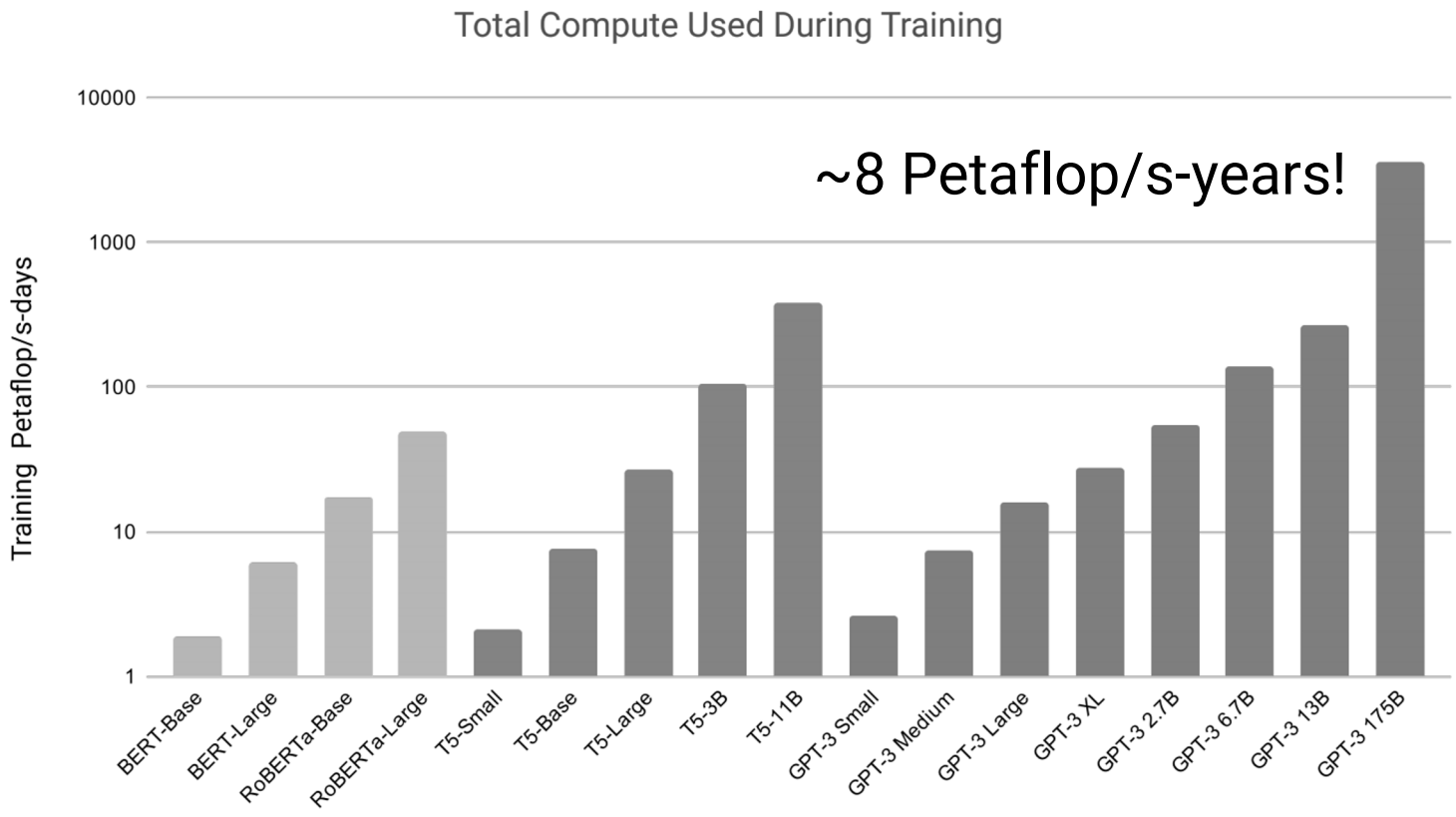
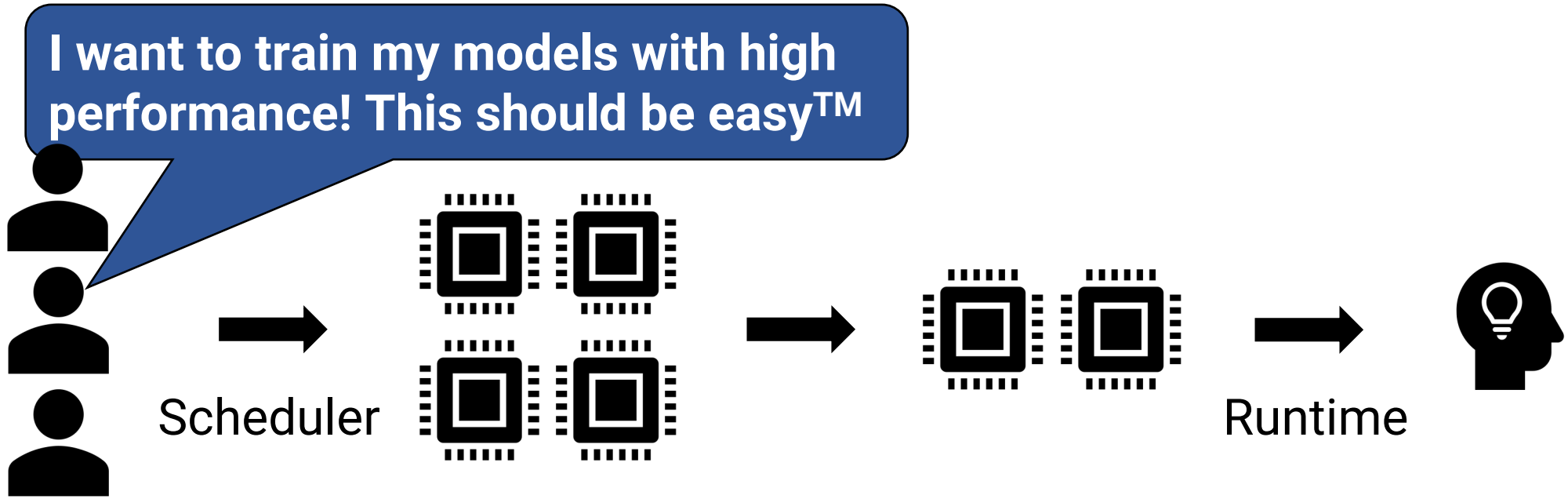


Figure from “Language Models are Few-Shot Learners”, Brown et al.

Research covered in this talk:

How can we train high-quality models fast using optimizations across the software stack?

Model training in datacenters



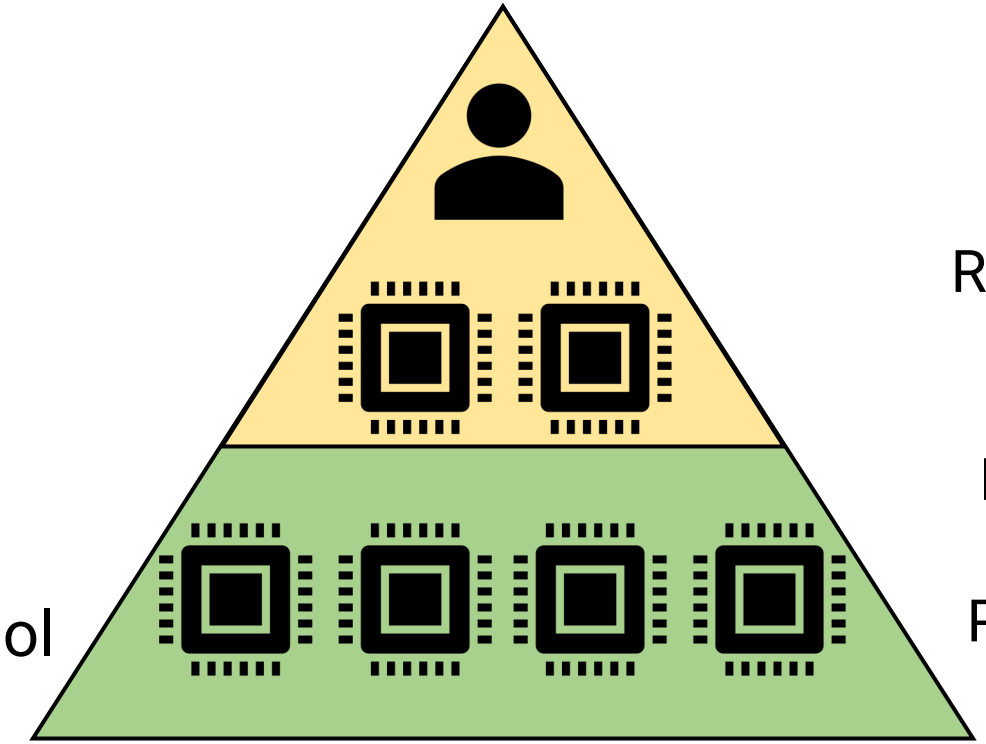
Iterative, long-running, and compute-intensive

Overview of work covered in this talk

User

Dedicated resources

Shared resource pool

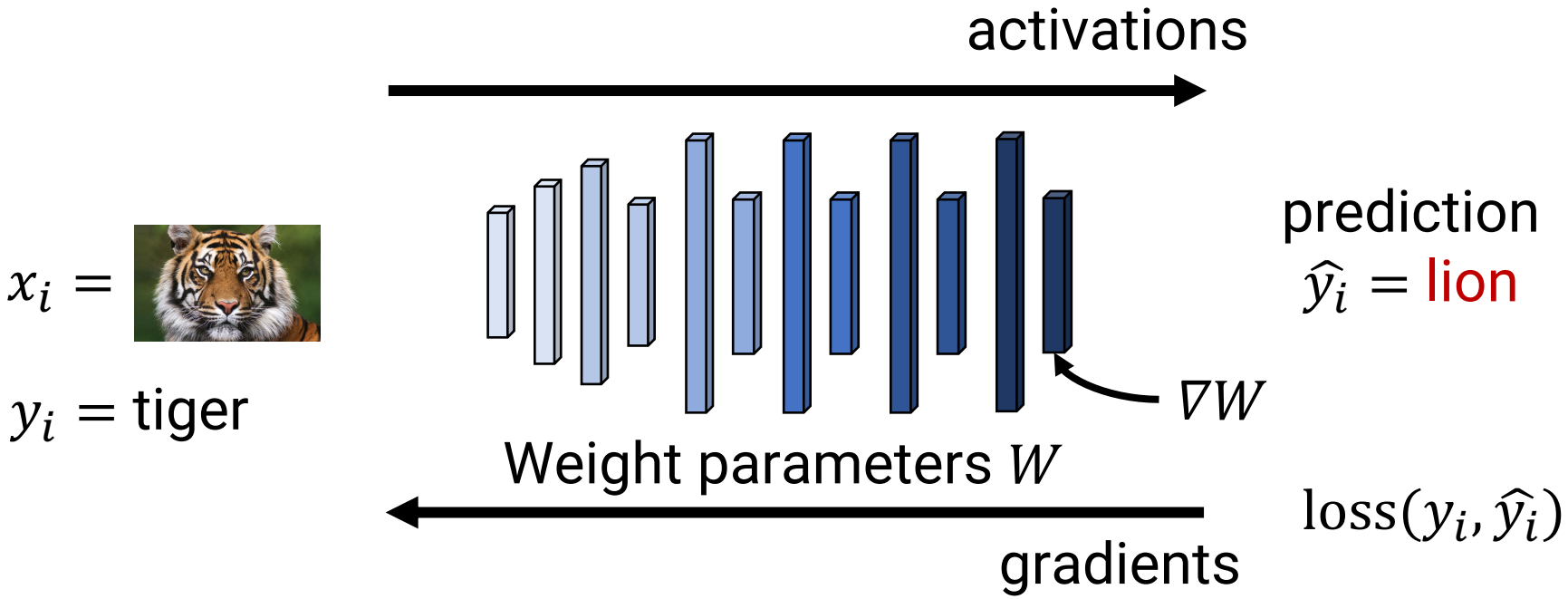


Efficient Training in the Datacenter
Zico (ATC 2021)
Refurbish your Training Data (ATC 2021)
ZeRO-Offload (ATC 2021)

Resource Allocation in the Datacenter
Pollux (OSDI 2021)
Privacy Budget Scheduling (OSDI 2021)
Habitat (ATC 2021)

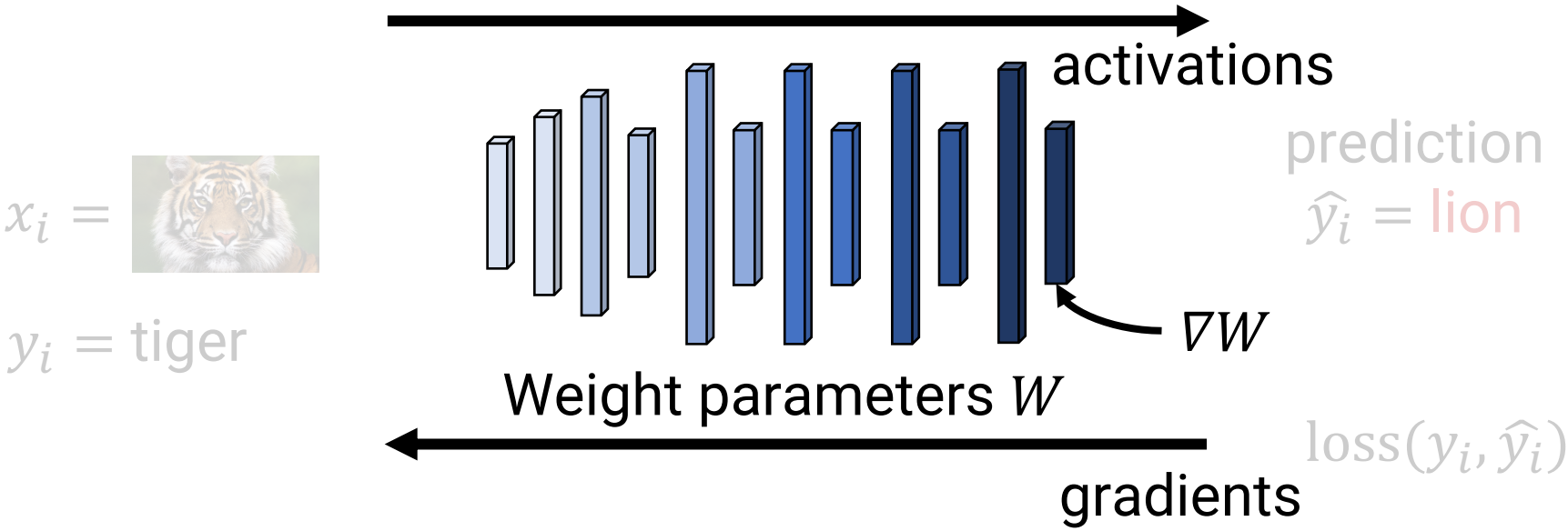
Efficient Training in the Edge
Oort (OSDI 2021)
Octo (ATC 2021)

Background: model training



Optimization performed in iterations; each iteration can be parallelized within an accelerator (GPU) and also across accelerators

Activations, gradients, and weights too large



- Activations, gradients, weights can be much larger than memory capacity of a single accelerator
- Need to either partition state across multiple accelerators or offload

Activations, gradients, and weights too large

ZeRO-Offload: Democratizing Billion-Scale Model Training

Jie Ren
UC Merced

Samyam Rajbhandari
Microsoft

Reza Yazdani Aminabadi
Microsoft

Olatunji Ruwase
Microsoft

Shuangyan Yang
UC Merced

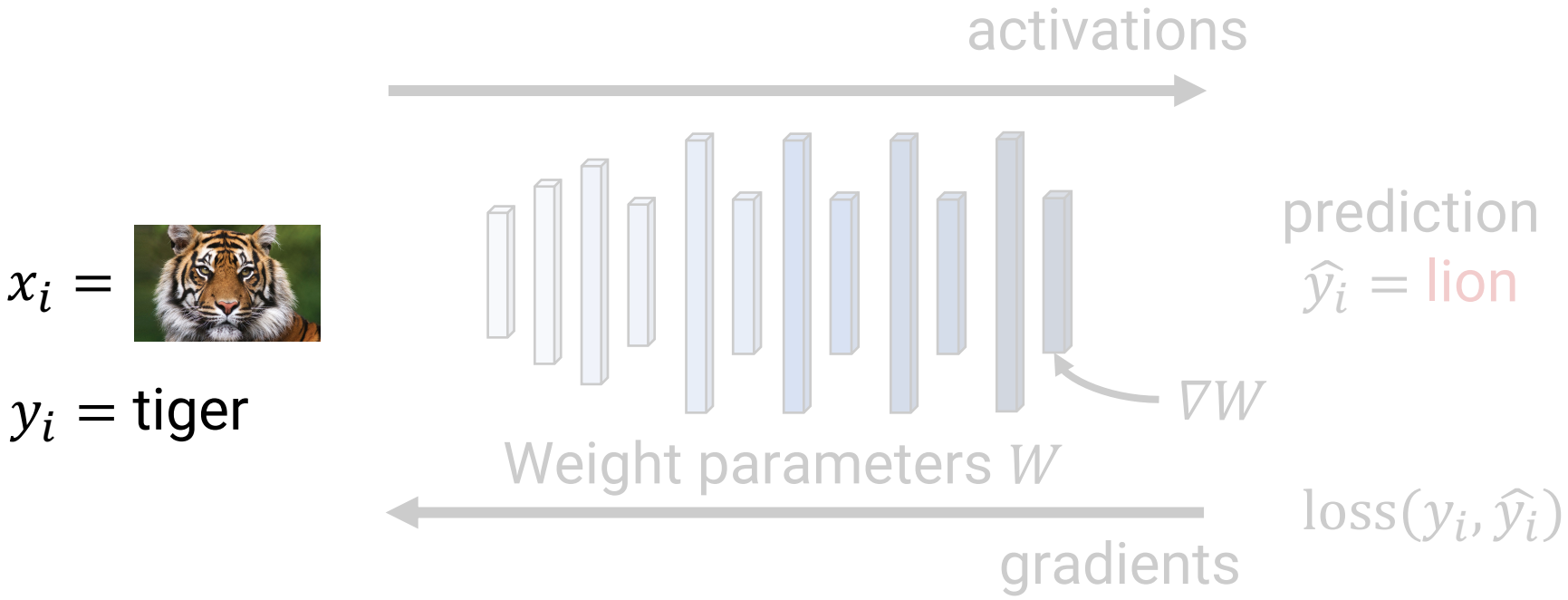
Minjia Zhang
Microsoft

Dong Li
UC Merced

Yuxiong He
Microsoft

ATC 2021
Sit, Fido!: Training Machine Learning Algorithms

Preprocessing can be a computational bottleneck



- Preprocessing performed on CPU usually (as opposed to model computation which is performed on accelerator)
- Can become a computational bottleneck (accelerator idle)

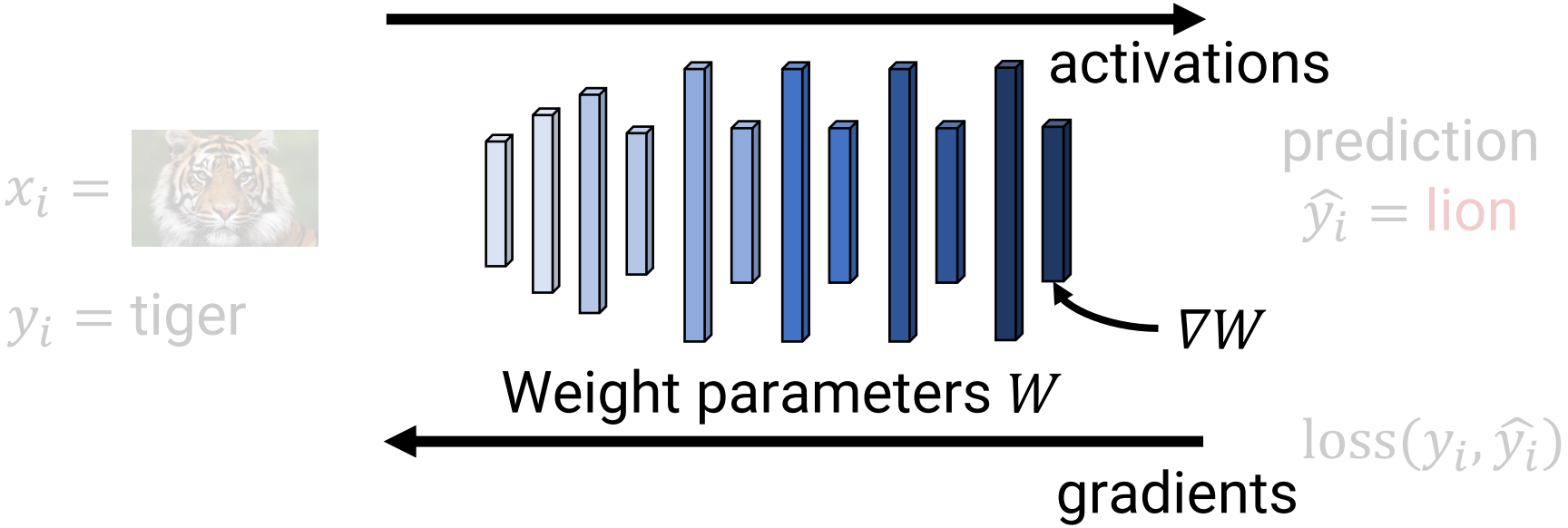
Preprocessing can be a computational bottleneck

Refurbish Your Training Data: Reusing Partially Augmented Samples for Faster Deep Neural Network Training

Gyewon Lee[1,3] Irene Lee[2] Hyeonmin Ha[1] Kyunggeun Lee[1]
Hwarim Hyun[1] Ahnjae Shin[1,3] Byung-Gon Chun[1,3]*
Seoul National University[1] *Georgia Institute of Technology*[2] *FriendliAI*[3]

ATC 2021
Sit, Fido!: Training Machine Learning Algorithms

How do multiple jobs share the same accelerator?



More intermediate state with concurrent jobs

- Footprint increases in forward pass, decreases in backward pass
- Peak memory footprint greatly increased if multiple training jobs are collocated on the same accelerator with phases aligned

How do multiple jobs share the same accelerator?

Zico: Efficient GPU Memory Sharing for Concurrent DNN Training

Gangmuk Lim
UNIST

Jeongseob Ahn
Ajou University

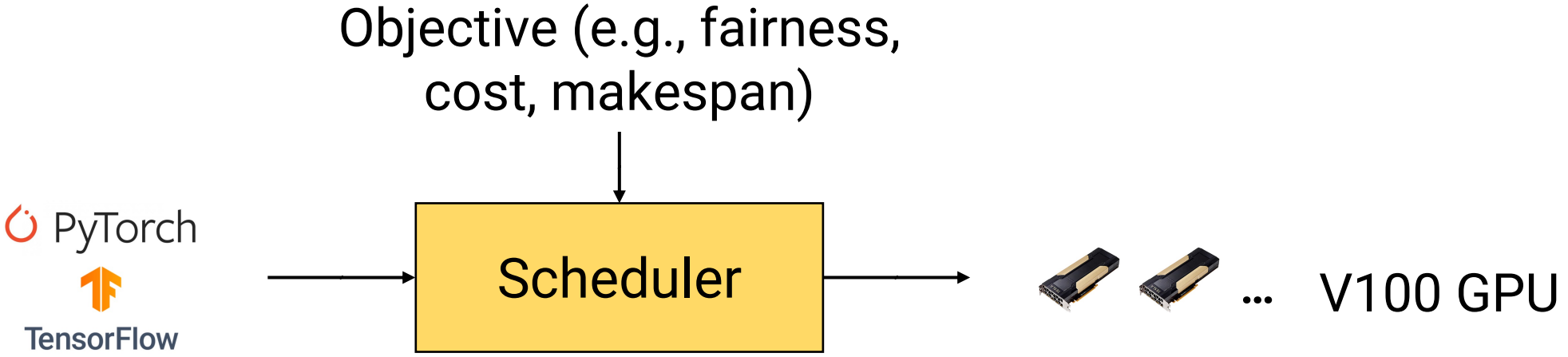
Wencong Xiao
Alibaba Group

Youngjin Kwon
KAIST

Myeongjae Jeon
UNIST

ATC 2021
Sit, Fido!: Training Machine Learning Algorithms

How should we allocate resources?

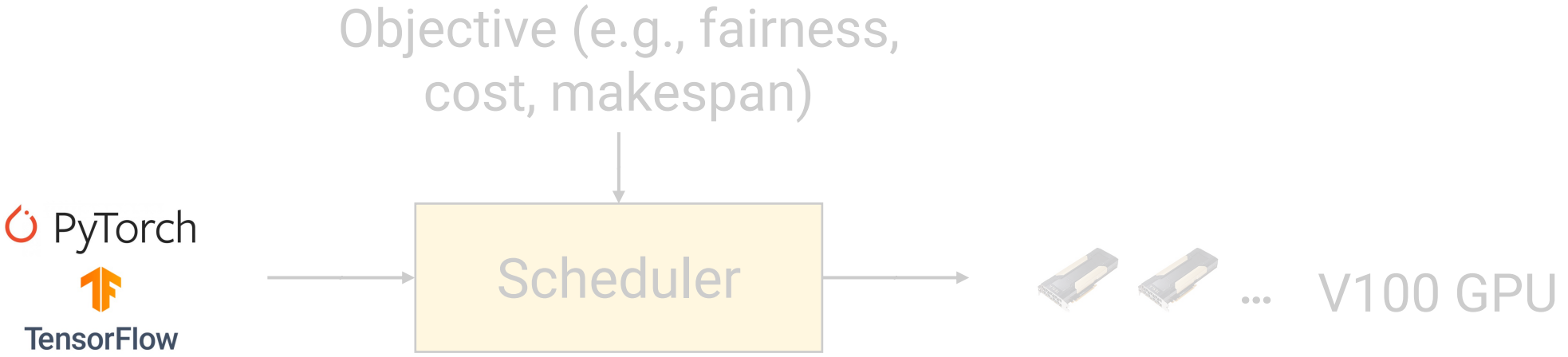


Training jobs in existing frameworks

Homogeneous cluster

Scheduling is a well-studied problem in computer systems in other contexts as well (e.g., big data clusters)

How do we incorporate elasticity into schedulers?



- Scheduler determining scale based on demand can allow resources to be better utilized
- Throughput and statistical efficiency affected by batch size and scale

How do we incorporate elasticity into schedulers?

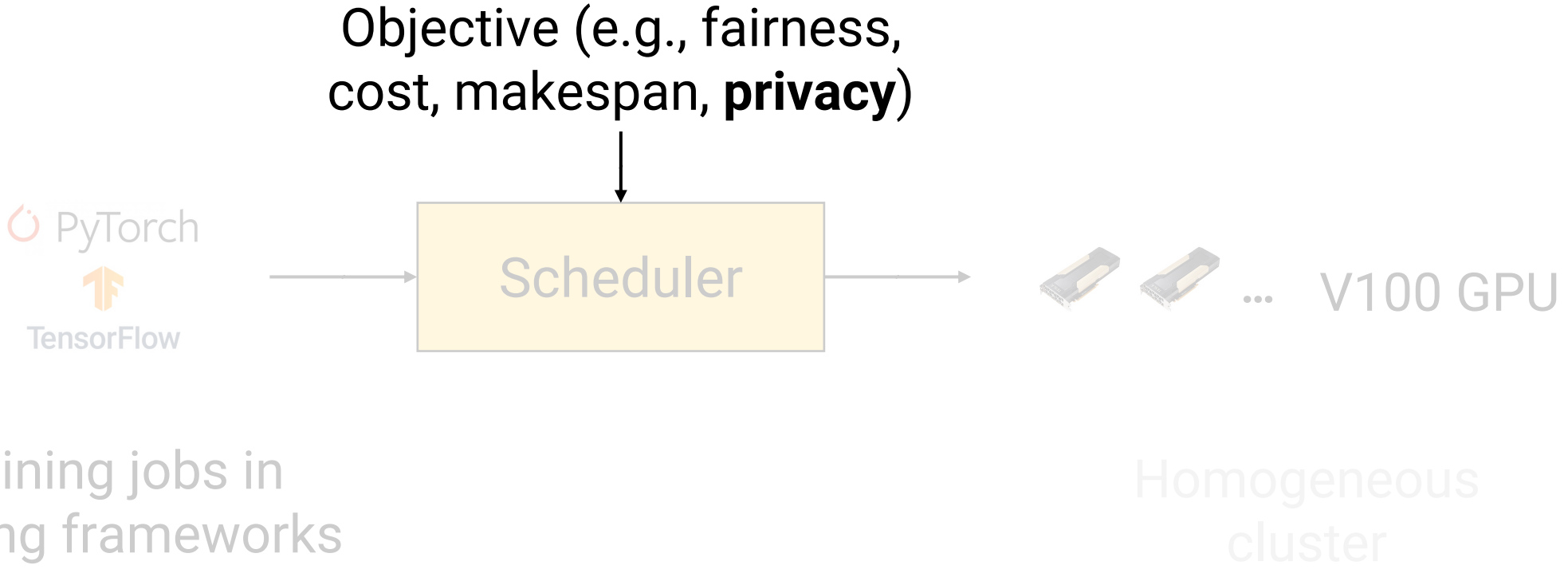
Pollux: Co-adaptive Cluster Scheduling for Goodput-Optimized Deep Learning

Aurick Qiao[1,2] Sang Keun Choe[2] Suhas Jayaram Subramanya[2] Willie Neiswanger[1,2]
Qirong Ho[1] Hao Zhang[1,3] Gregory R. Ganger[2] Eric P. Xing[4,1,2]

[1]*Petuum, Inc.* [2]*Carnegie Mellon University* [3]*UC Berkeley* [4]*MBZUAI*

OSDI 2021
Optimizations and Scheduling for Machine Learning

How about other kinds of objectives?



- Most objectives functions of throughput or cost (e.g., fairness)
- But what about privacy? (data leakage occurs every time a ML model is trained on a specific dataset)

How about other kinds of objectives?

Privacy Budget Scheduling

Tao Luo*
Columbia University

Mingen Pan*
Columbia University

Pierre Tholoniat*
Columbia University

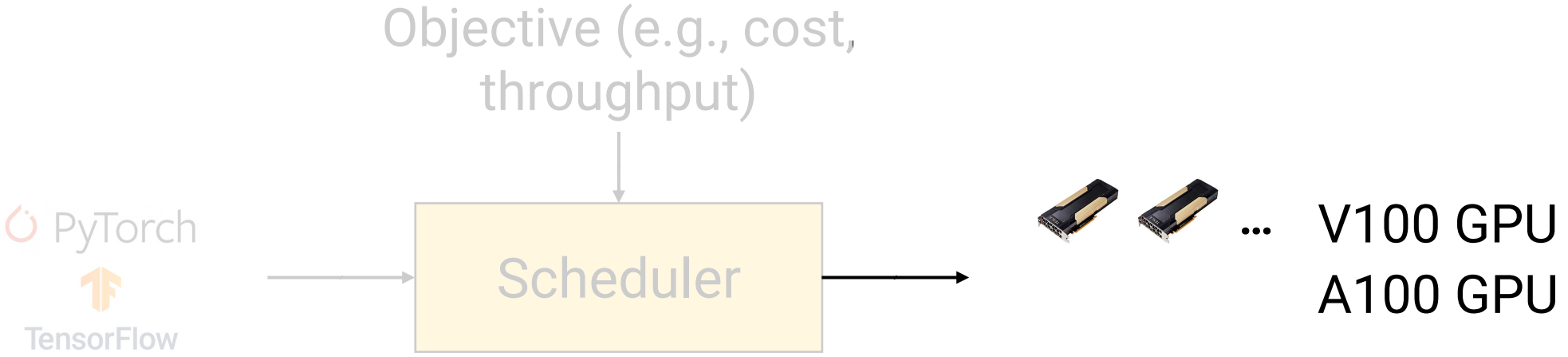
Asaf Cidon
Columbia University

Roxana Geambasu
Columbia University

Mathias Lécuyer
Microsoft Research

OSDI 2021
Optimizations and Scheduling for Machine Learning

How do we pick between accelerator types?



Training jobs in existing frameworks

Heterogeneous resources

- Models have different operators: some compute-bound, some memory-bound; optimal implementation hardware-specific
- Not easy to determine best accelerator type for a given objective

How do we pick between accelerator types?

Habitat: A Runtime-Based Computational Performance Predictor for Deep Neural Network Training

Geoffrey X. Yu
University of Toronto
Vector Institute

Yubo Gao
University of Toronto

Pavel Golikov
University of Toronto
Vector Institute

Gennady Pekhimenko
University of Toronto
Vector Institute

ATC 2021
Sit, Fido!: Training Machine Learning Algorithms

Training in the datacenter not always attractive!

- Privacy (don’t want to share sensitive data with the cloud)
- High latency (need to go to cloud and back)
- Hard to provide personalized models

Training on edge devices can be challenging!

- Limited computational capacity and memory
- Need to coordinate among multiple decentralized entities (to make sure model is not overfit to single user)

Can we train directly on edge devices?

Octo: INT8 Training with Loss-aware Compensation and Backward Quantization for Tiny On-device Learning

Qihua Zhou†, Song Guo†, Zhihao Qu‡, Jingcai Guo†, Zhenda Xu†,
Jiewei Zhang†, Tao Guo†, Boyuan Luo†, Jingren Zhou*
†*Hong Kong Polytechnic University,* ‡*Hohai University,* **Alibaba Group*

ATC 2021
I'm Old but I Learned a New Trick: Machine Learning

Can we train directly on edge devices?

Oort: Efficient Federated Learning via Guided Participant Selection

Fan Lai, Xiangfeng Zhu, Harsha V. Madhyastha, Mosharaf Chowdhury
University of Michigan

OSDI 2021
Optimizations and Scheduling for Machine Learning

ML Training @ OSDI and ATC 2021

- Zico
- Refurbish your Training Data
- ZeRO-Offload

Efficient training in the datacenter

- Pollux
- Privacy Budget Scheduling
- Habitat

Resource allocation in the datacenter

- Oort
- Octo

Efficient training in the edge

OSDI 2021

ATC 2021