DASYA

dasya.itu.dk

RAD

rad.itu.dk

IT UNIVERSITY OF COPENHAGEN

www.itu.dk

beauty & the beast:
deep learning & its resource needs

Pınar Tözün

Associate Professor
IT University of Copenhagen

EDBT
March 26, 2026

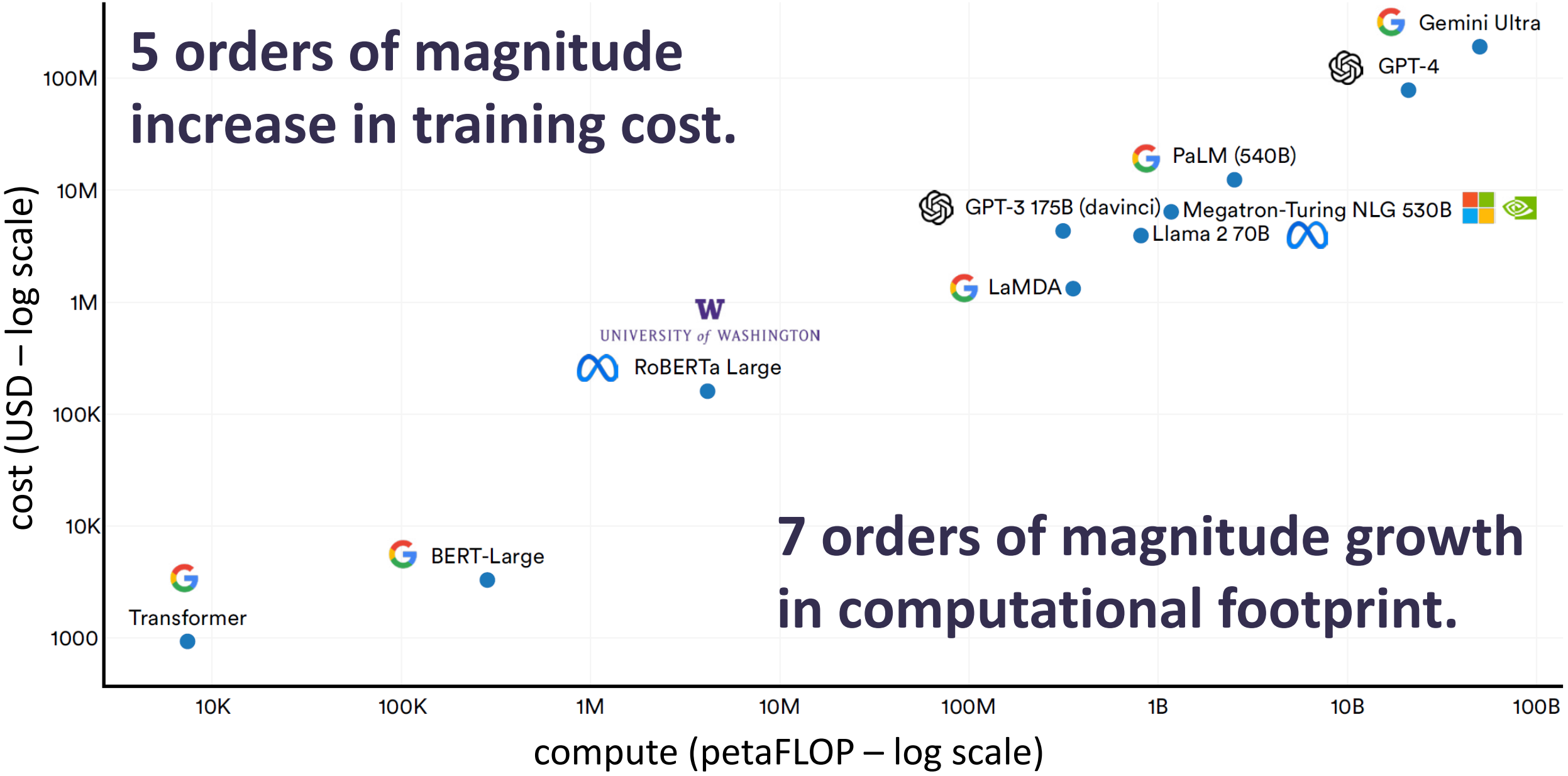
pito@itu.dk
pinartozun.com

Innovationsfonden

INDEPENDENT RESEARCH FUND DENMARK

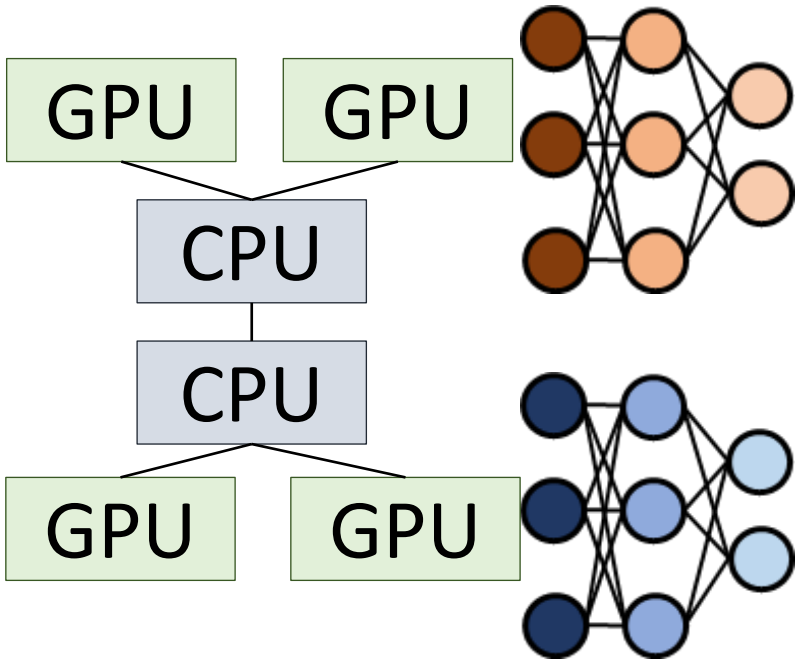
novo nordisk foundation

language model training (2017 – today)

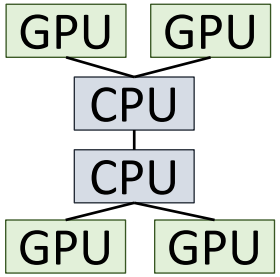


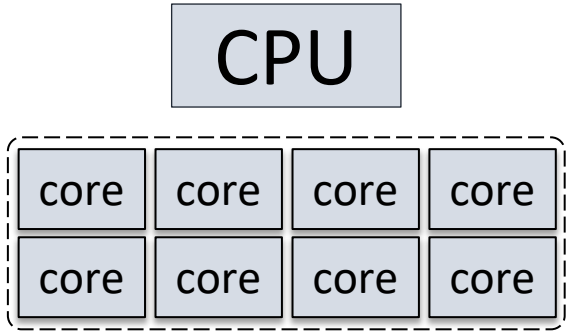
source: [Stanford AI Index Report 2024]

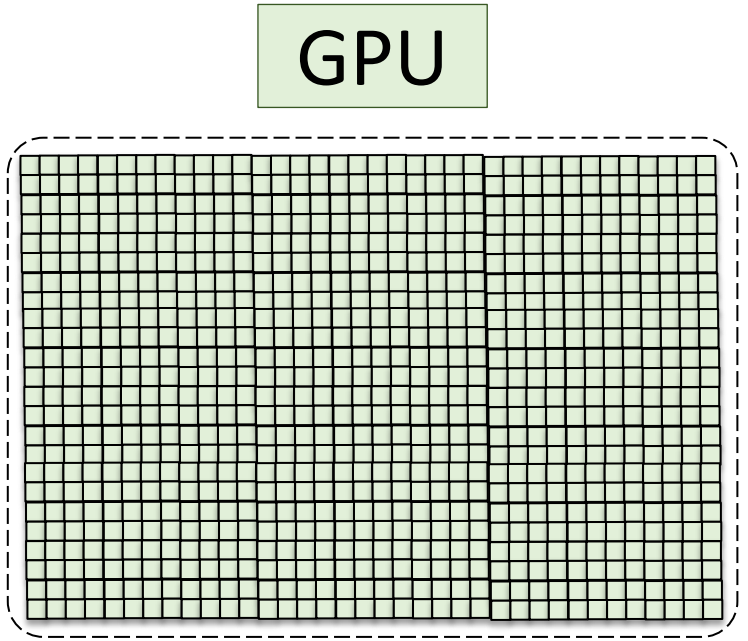
deep learning hardware



deep learning commodity hardware







central processing unit

➔ several (complex) cores

➔ good for latency-oriented tasks & single-core performance

- throughput- vs latency-oriented designs exist among CPUs as well

graphics processing unit

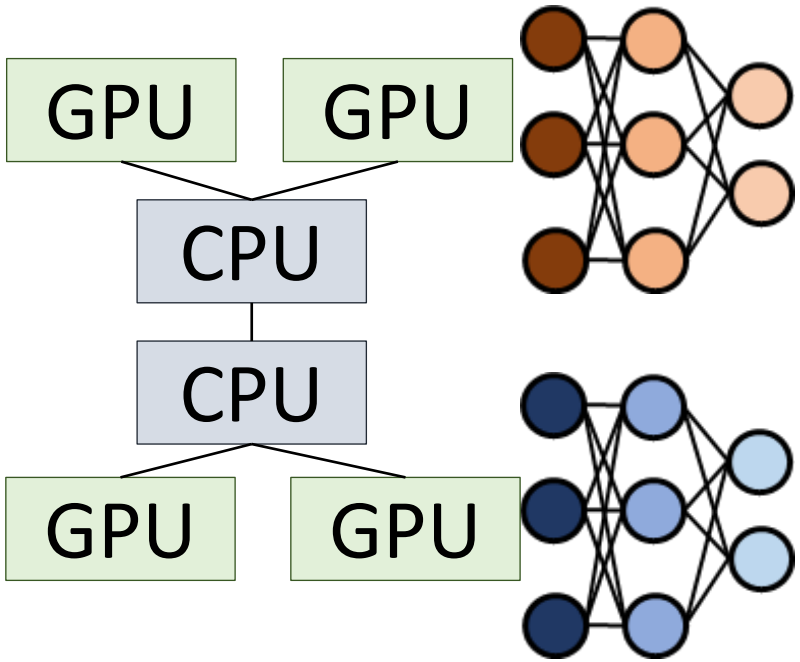
➔ many (simple) cores

➔ good for throughput-oriented & embarrassingly parallel tasks

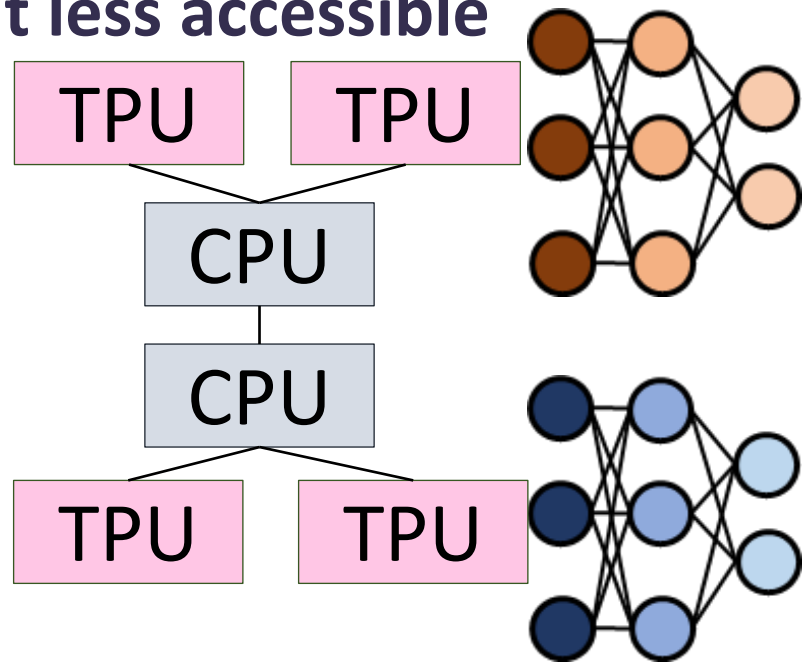
➔good for deep learning

- e.g., large matrix operations

deep learning hardware



- **tend to be more efficient**
- **but less accessible**

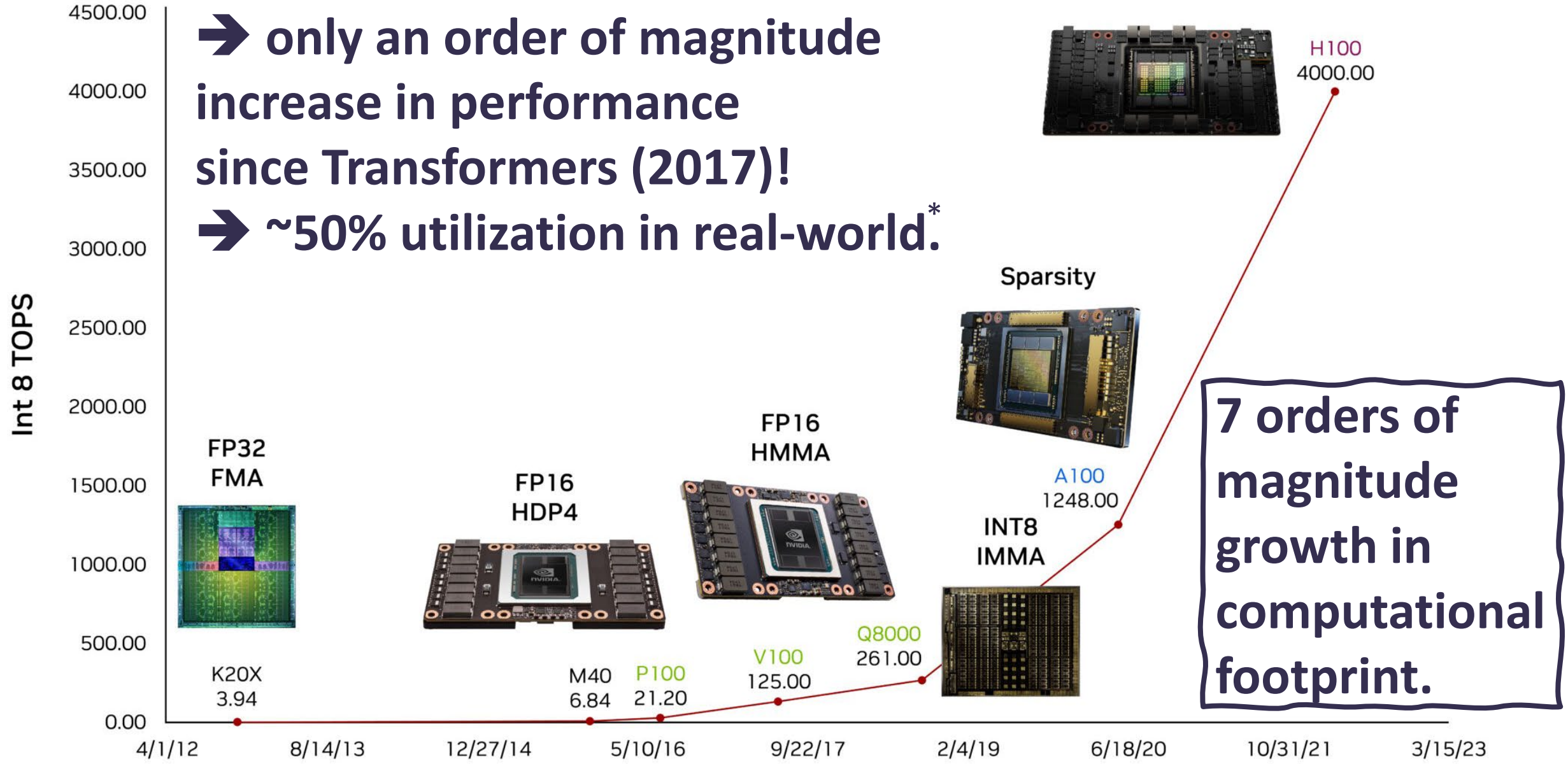


costs & progress depend on the performance & utilization of the available hardware

NVIDIA GPUs (2012 – 2023)

➔ only an order of magnitude increase in performance since Transformers (2017)!

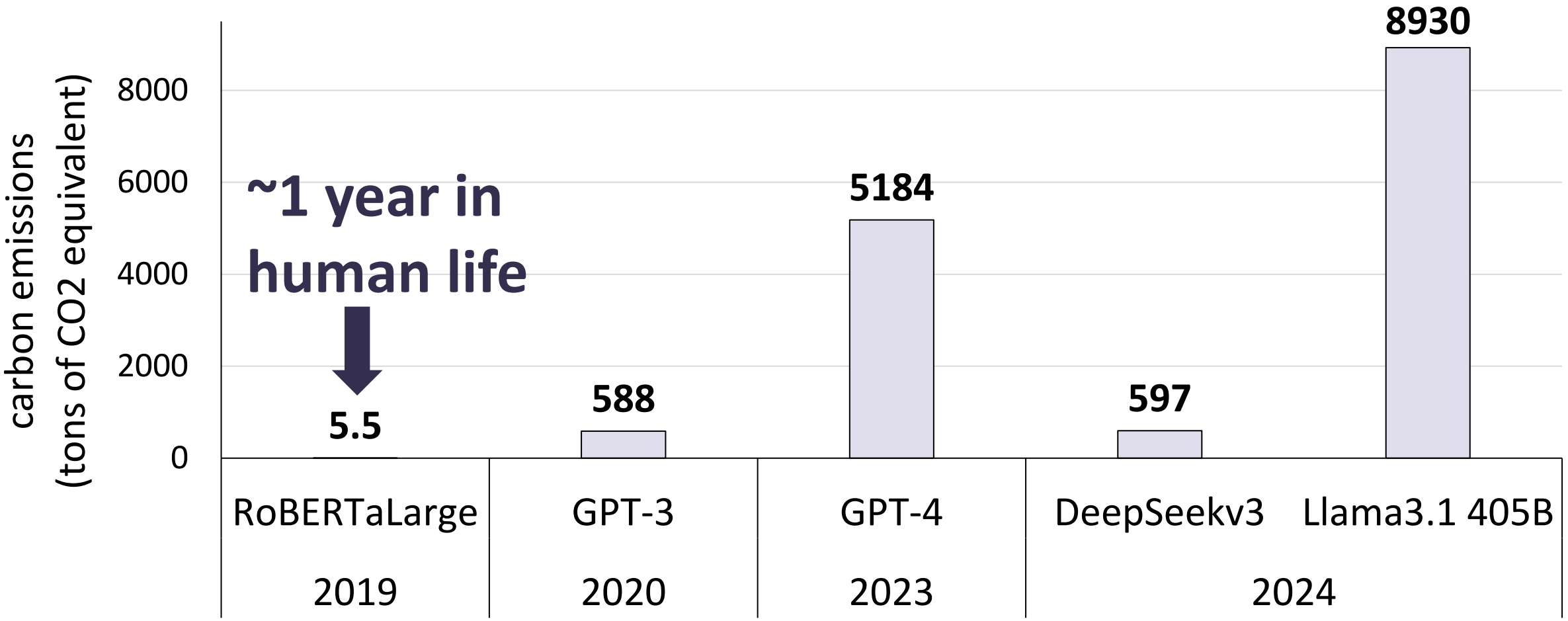
➔ ~50% utilization in real-world.*



7 orders of magnitude growth in computational footprint.

sources: figure by Bill Dally, NVIDIA & *Gao et al. "An Empirical Study on Low GPU Utilization of Deep Learning Jobs." ICSE 24

carbon footprint of language model training



can we do better while using fewer resources?

model accuracy cannot be the only metric to aim for!

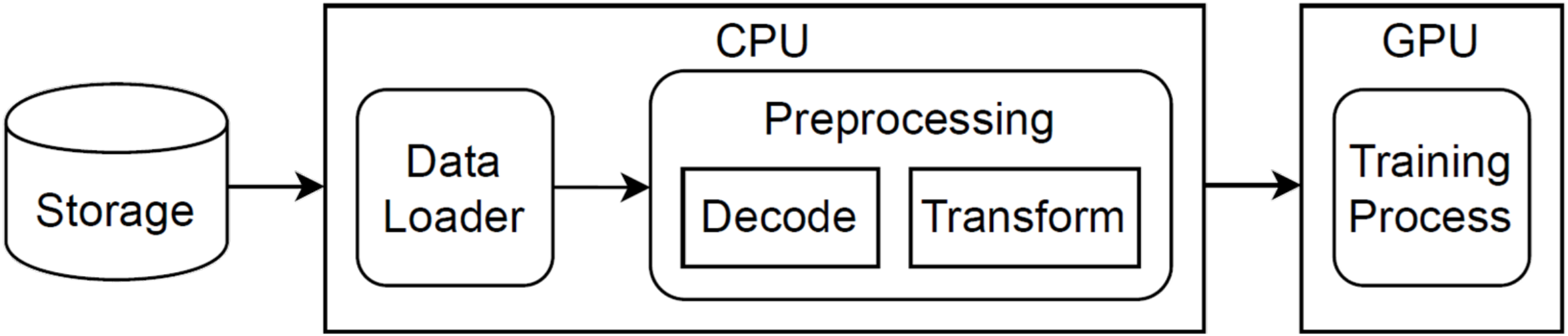
source: Stanford AI Index Report 2025

deep learning with fewer resources

- resource sharing
- data & work sharing

} for model training

deep learning training



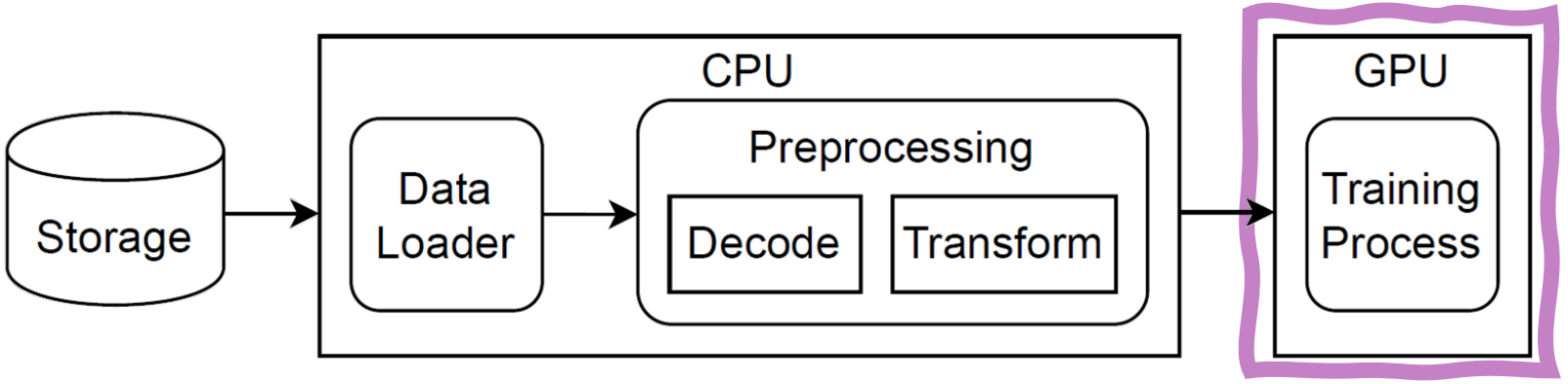
deep learning with fewer resources

An Analysis of Collocation on GPUs for Deep Learning Training
Ties Robroek, Ehsan Yousefzadeh-Asl-Miandoab, Pınar Tözün
EuroMLSys 2024

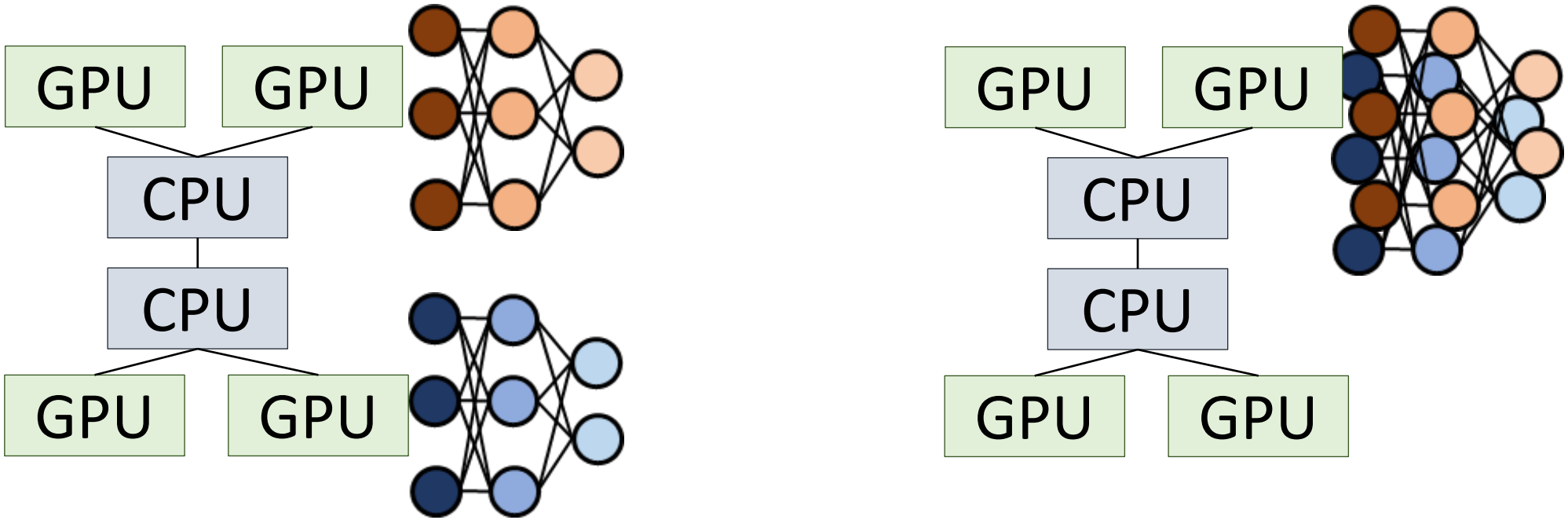
CARMA: Collocation-Aware Resource Manager
Ehsan Yousefzadeh-Asl-Miandoab, Florina M Ciorba, Pınar Tözün

- resource sharing

- data & work sharing

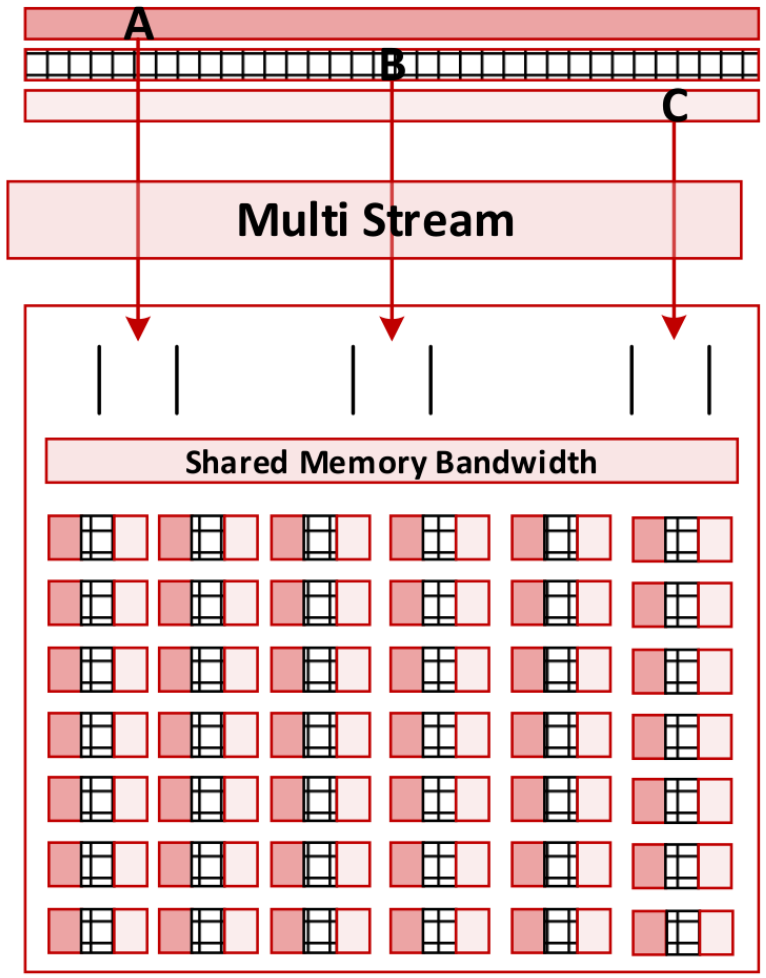


collocated training

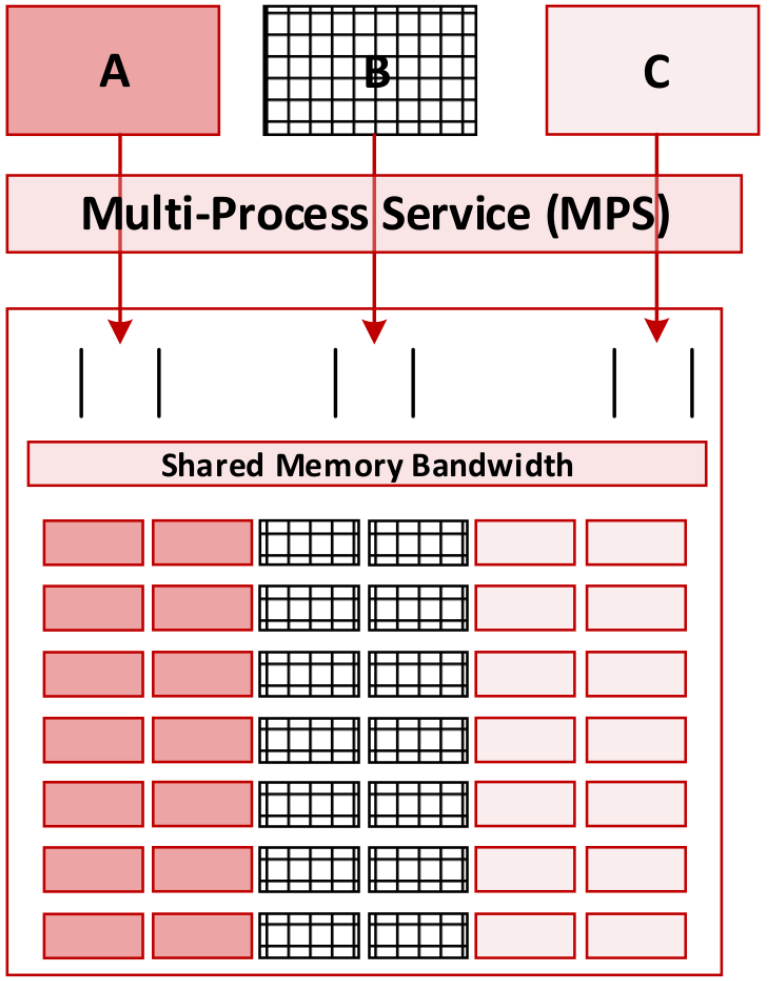


- **training more models with fewer hardware resources**
- **if done well ➔ better hardware utilization & reduces costs**
- **if not done well ➔ interference & performance degradation**

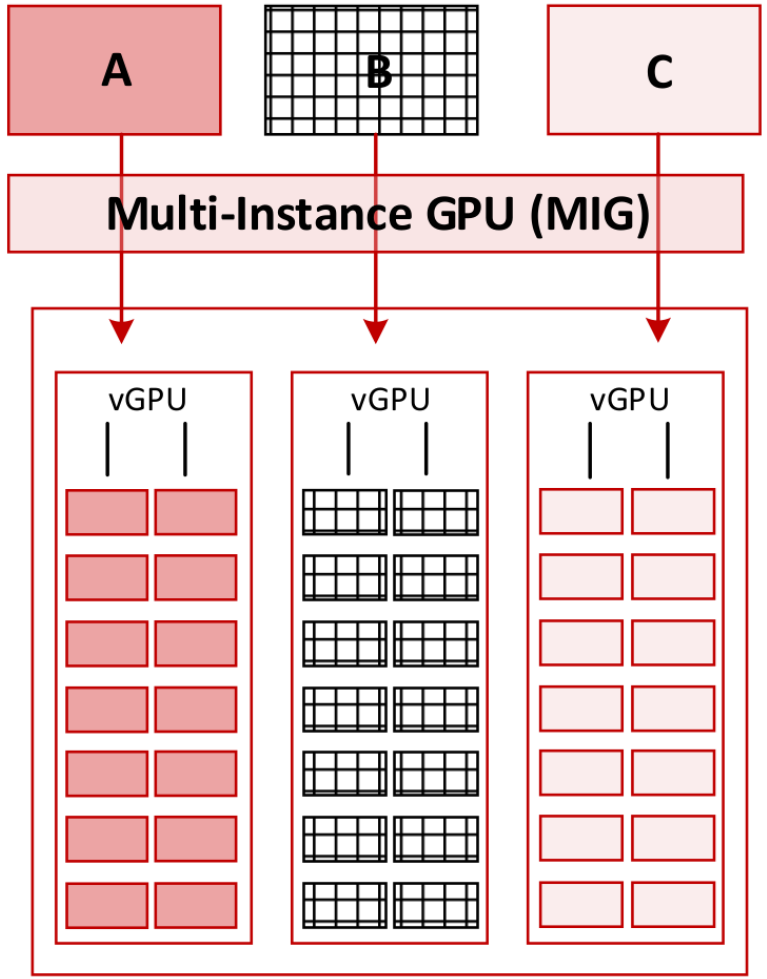
sharing resources on (NVIDIA) GPUs



- most straightforward
- time-multiplexing
- × limited parallelism



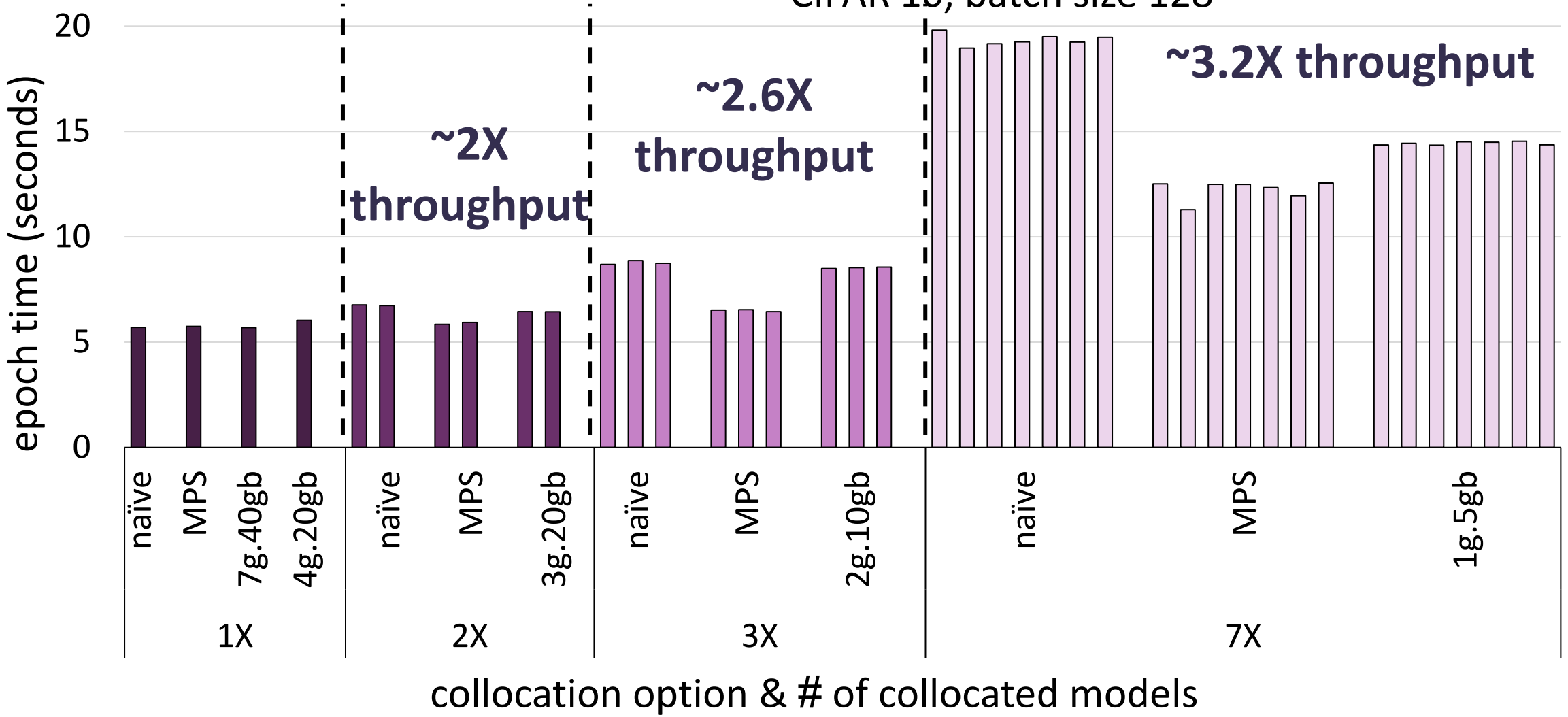
- finer-grained sharing
- × higher chances of interference



- hardware-support for resource split
- × rigid partitioning

small case – ResNet26

NVIDIA A100, PyTorch 2
CIFAR 10, batch size 128

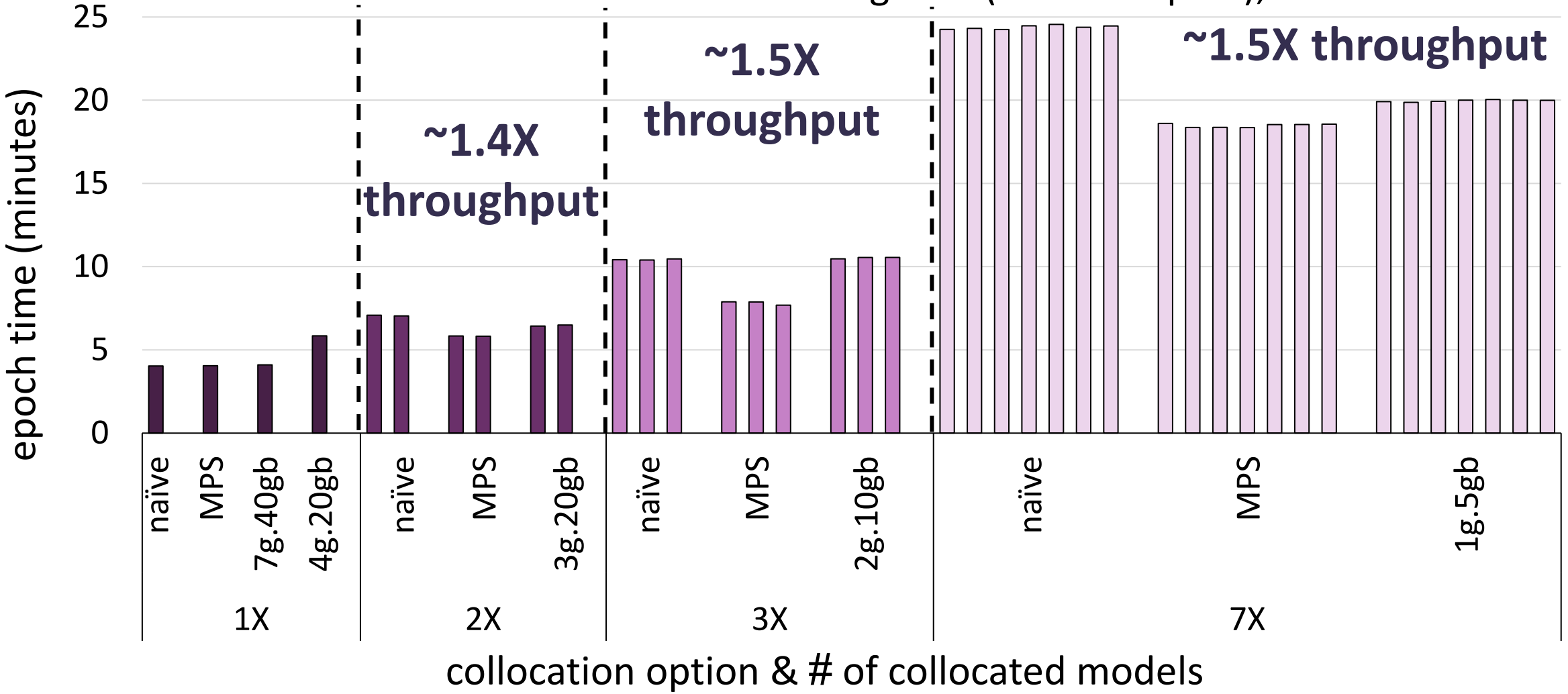


collocation benefits despite increased epoch time

MPS > MIG > naïve

medium case – ResNet50

NVIDIA A100, PyTorch 2
ImageNet (downsampled), batch size 128

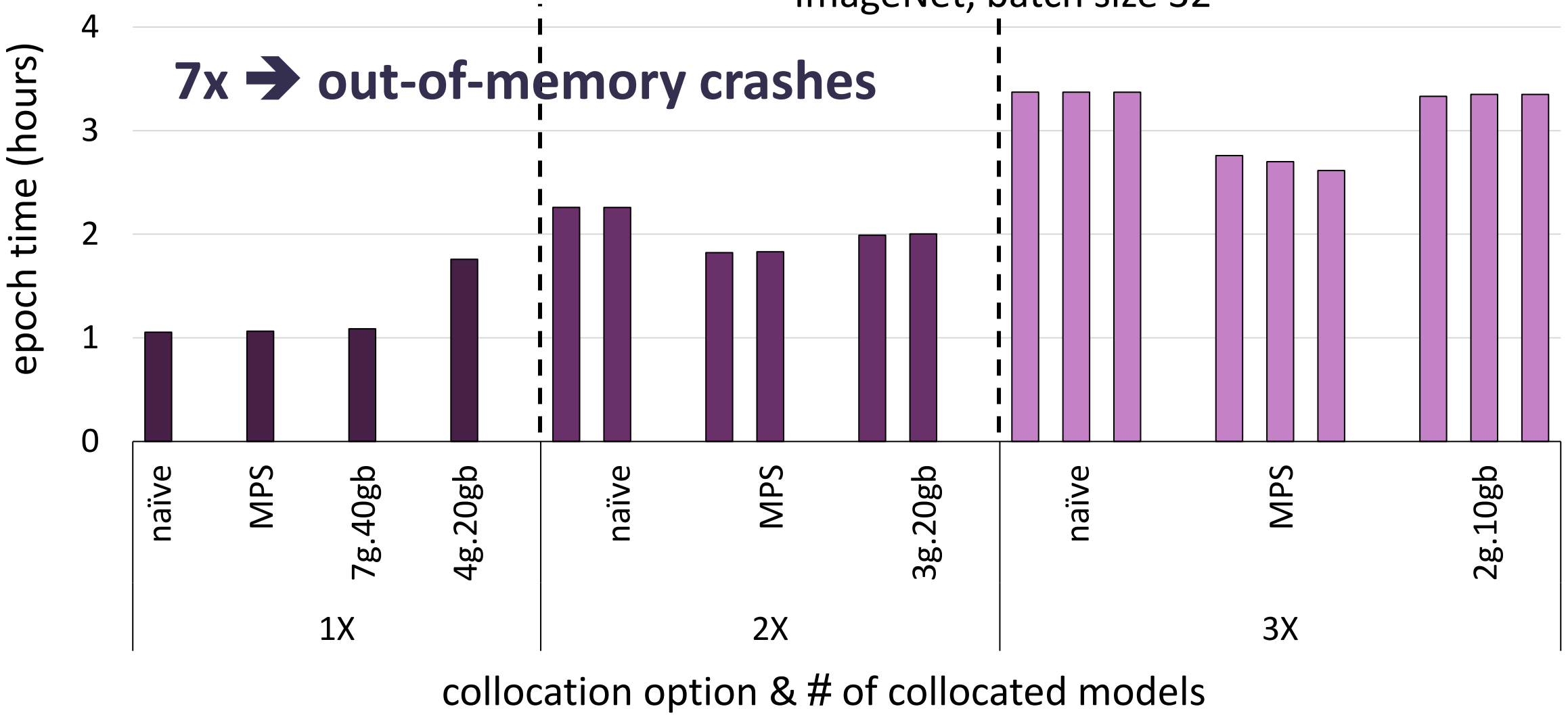


still some throughput benefits

but diminishing returns for increased collocation

large case – ResNet152

NVIDIA A100, PyTorch 2
ImageNet, batch size 32



no more throughput benefits – 80% utilization when training alone
better to collocate with smaller or less compute heavy tasks

mixed workloads: compute- & memory-heavy

	DLRM time per training block	ResNet152 time per epoch	sm activity	memory footprint
DLRM alone			5%	29.14 GB
ResNet152 alone			82%	8.47 GB

mixed workloads: compute- & memory-heavy

	DLRM time per training block	ResNet152 time per epoch	sm activity	memory footprint
DLRM alone			5%	29.14 GB
ResNet152 alone			82%	8.47 GB
MPS			81%	37.62 GB

mixed workloads: compute- & memory-heavy

	DLRM time per training block	ResNet152 time per epoch	sm activity	memory footprint
DLRM alone	5.36 h	-	5%	29.14 GB
ResNet152 alone	-	1.05 h	82%	8.47 GB
MPS			81%	37.62 GB

mixed workloads: compute- & memory-heavy

	DLRM time per training block	ResNet152 time per epoch	sm activity	memory footprint
DLRM alone	5.36 h	-	5%	29.14 GB
ResNet152 alone	-	1.05 h	82%	8.47 GB
MPS	5.57 h (+5%)	1.10 h (+4%)	81%	37.62 GB

collocation can lead to (almost) free lunch
when large models stress different hardware resources

collocation-aware resource management

requirements:

(1) don't overload the GPU compute
➔ degrades performance

(2) minimize & recover from out-of-memory crashes

CARMA: collocation-aware resource manager

submit training task

plug & play

monitoring

nvidia-smi

dcgmi

recovery queue

tasks' queue

parser

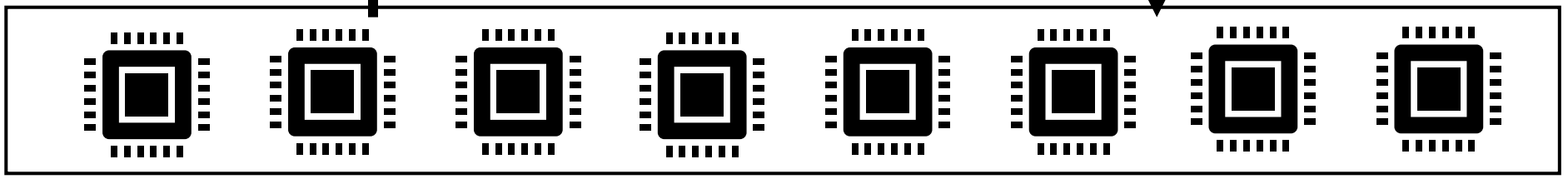
memory estimator

Horus,
Fake Tensors,
GPUMemNet …

most available memory,
least utilized GPU,
random …

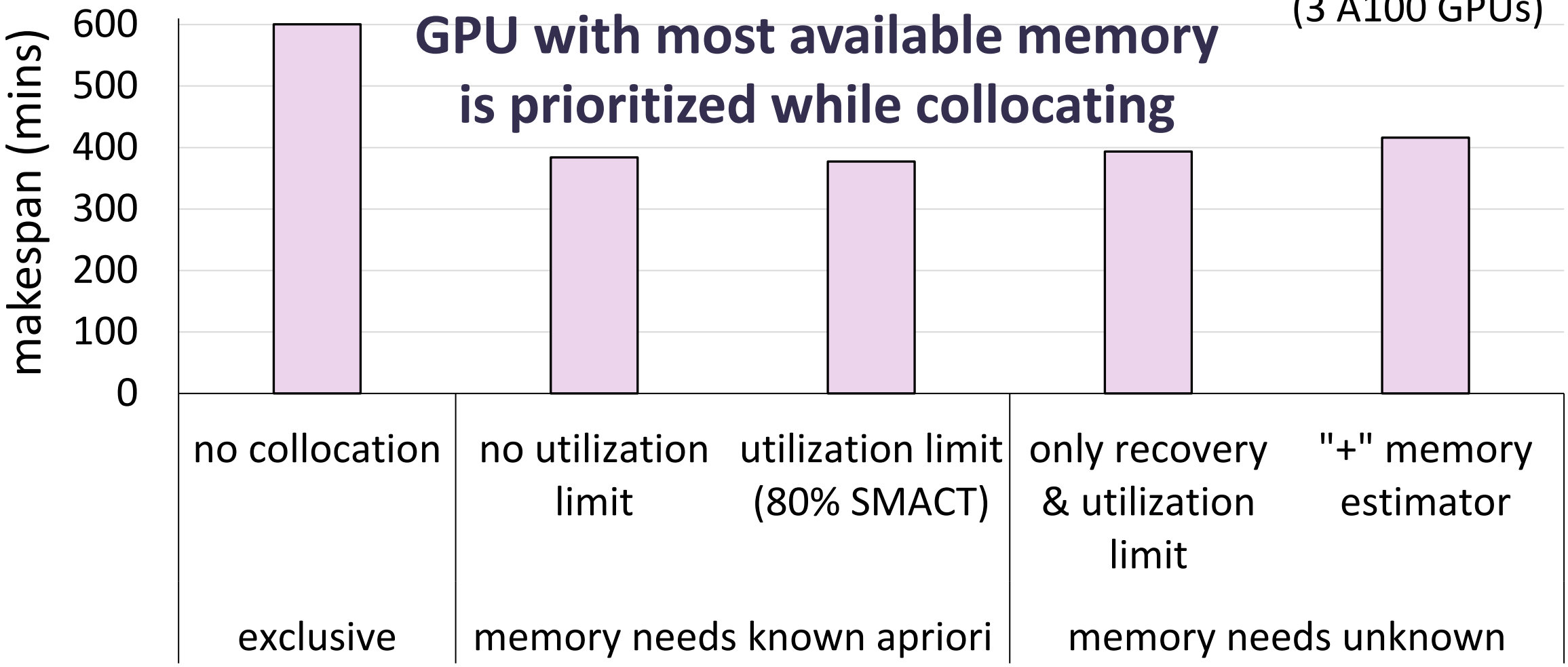
utilization information

mapping decisions



CARMA on a training workload trace

on NVIDIA DGX Station
(3 A100 GPUs)



30-37% reduction in end-to-end trace completion time leads to ~15% reduction in total energy need

workload collocation on GPUs

Orion: Interference-aware, Fine-grained GPU Sharing for ML Applications

Authors: Foteini Strati, Xianzhe Ma, Ana Klimovic | Authors Info & Claims

EuroSys '24: Proceedings of the Nineteenth European Conference on Computer Systems • Pages 1075 - 1092
https://doi.org/10.1145/3627703.3629578

training & inference

Data Movement-Aware GPU Sharing for Data-Intensive Systems

Yi Jiang
EPFL
Lausanne, Switzerland
yi.jiang@epfl.ch

Hamish Nicholson
EPFL
Lausanne, Switzerland
hamish.nicholson@epfl.ch

Viktor Sanca*
Oracle
Redwood City, United States
viktor.sanca@oracle.com

Anastasia Ailamaki
EPFL
Lausanne, Switzerland
anastasia.ailamaki@epfl.ch

CIDR 2026

inference & data analytics

Journals & Magazines > IEEE Transactions on Parallel... > Volume: 33 Issue: 1

Horus: Interference-Aware and Prediction-Based Scheduling in Deep Learning Systems

collocating based on memory prediction

Gandiva: Introspective Cluster Scheduling for Deep Learning

Wencong Xiao, *Beihang University & Microsoft Research;* Romil Bhardwaj, Ramachandran Ramjee, Muthian Sivathanu, and Nipun Kwatra, *Microsoft Research;* Zhenhua Han, *The University of Hong Kong & Microsoft Research;* Pratyush Patel, *Microsoft Research;* Xuan Peng, *Huazhong University of Science and Technology & Microsoft Research;* Hanyu Zhao, *Peking University & Microsoft Research;* Quanlu Zhang, Fan Yang, and Lidong Zhou, *Microsoft Research*

OSDI 2018

suspend & resume

...

workload collocation for model training

- not all training needs all the resources of a single GPU
- collocation on GPUs benefits when the aggregate compute & memory needs of the collocated training runs fit in the GPU
- a collocation-aware resource manager help reduce time & energy required for a training workload with various models

need to build collocation-aware resource managers for deep learning targeting both small & large scales!

deep learning with fewer resources

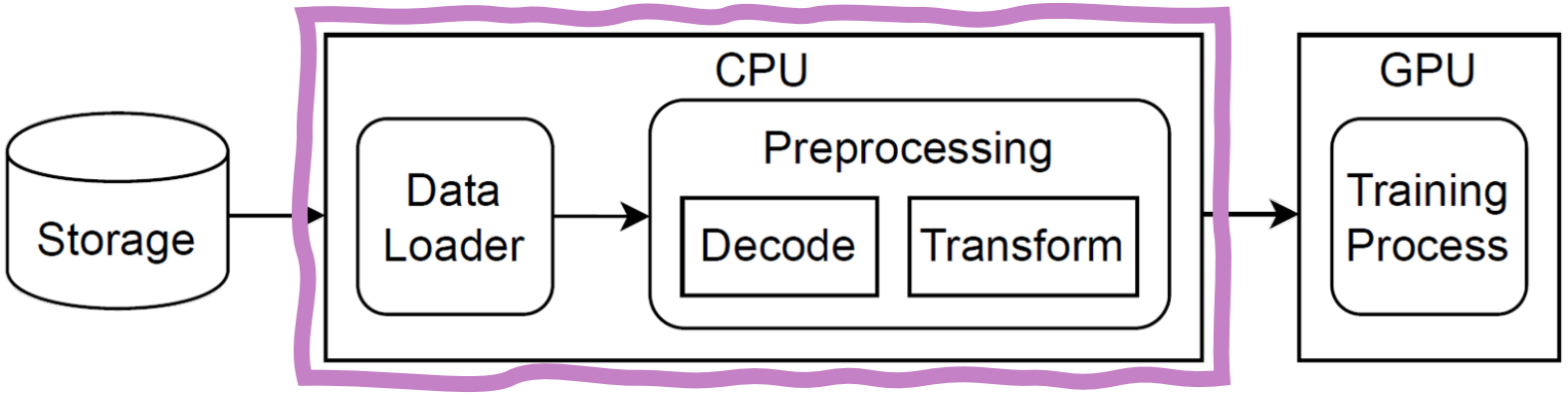
- resource sharing

TensorSocket: Shared Data Loading for Deep Learning Training

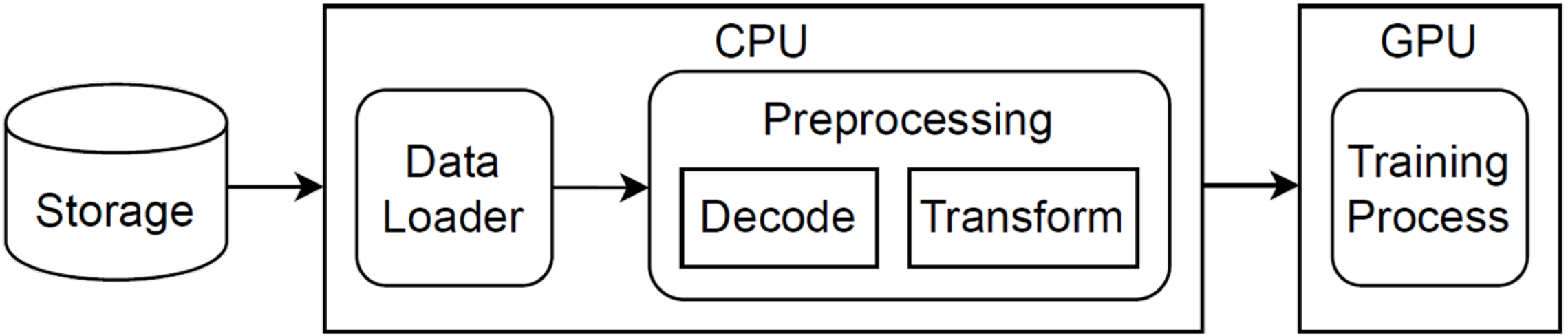
Ties Robroek, Neil Kim Nielsen, Pınar Tözün

SIGMOD 2026

- data & work sharing



journey of data in deep learning training



CPU feeds the GPU

- 16-64 CPU cores per GPU (recommended)
- 96 CPU cores per TPU*

➔ otherwise, GPU/TPU may be underutilized

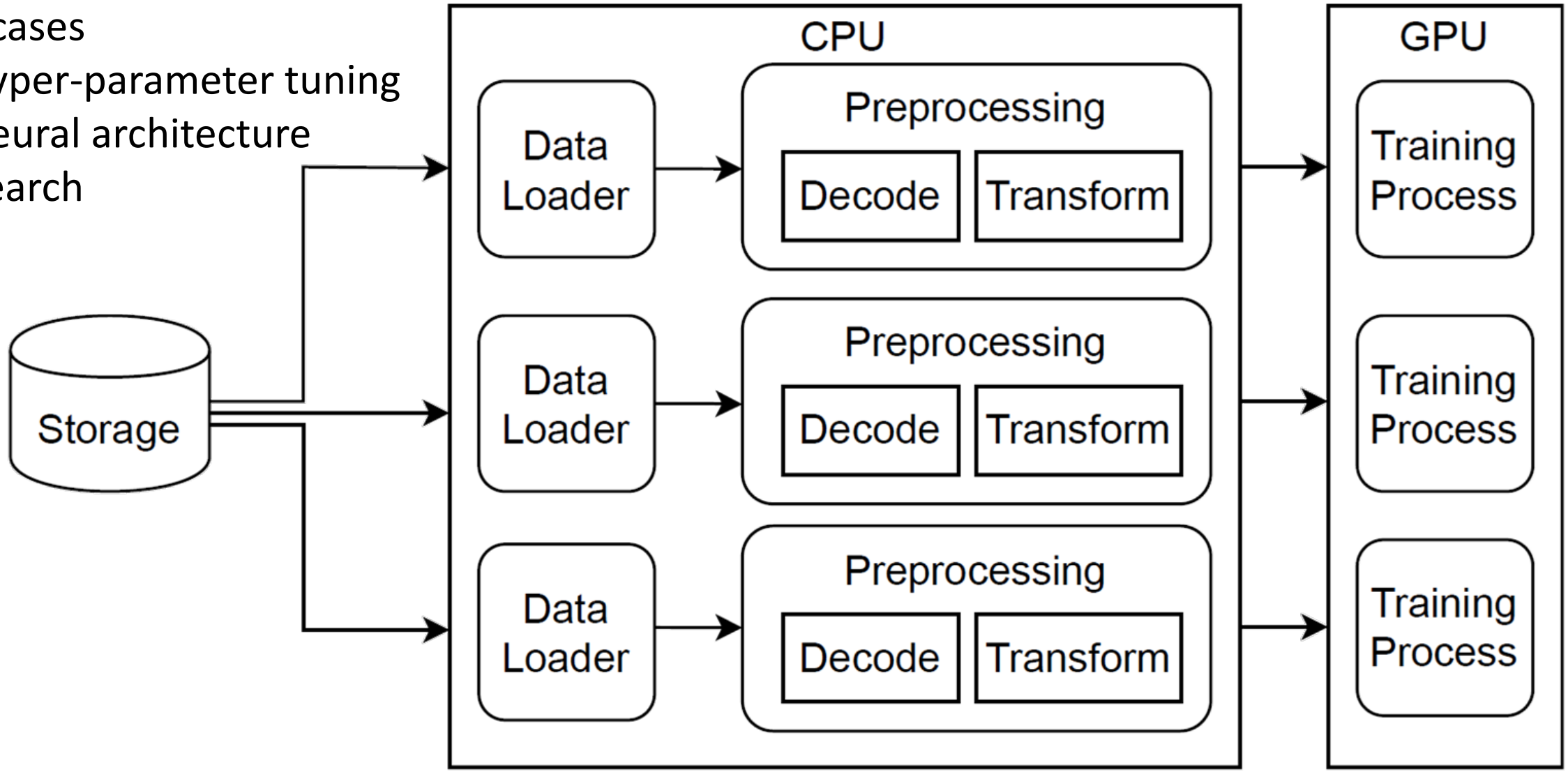
➔ can we do more with fewer CPU cores?

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

multiple model training on the same data

use cases

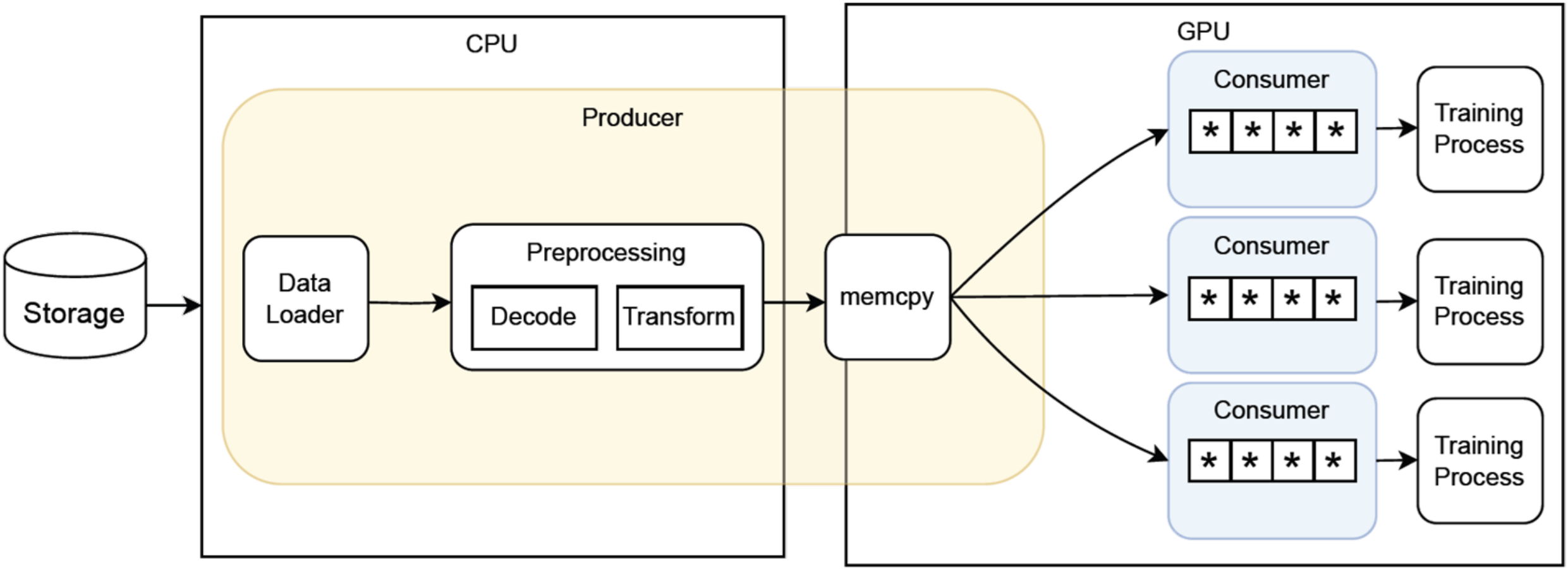
- hyper-parameter tuning
- neural architecture search
- …



redundant work & CPU use!

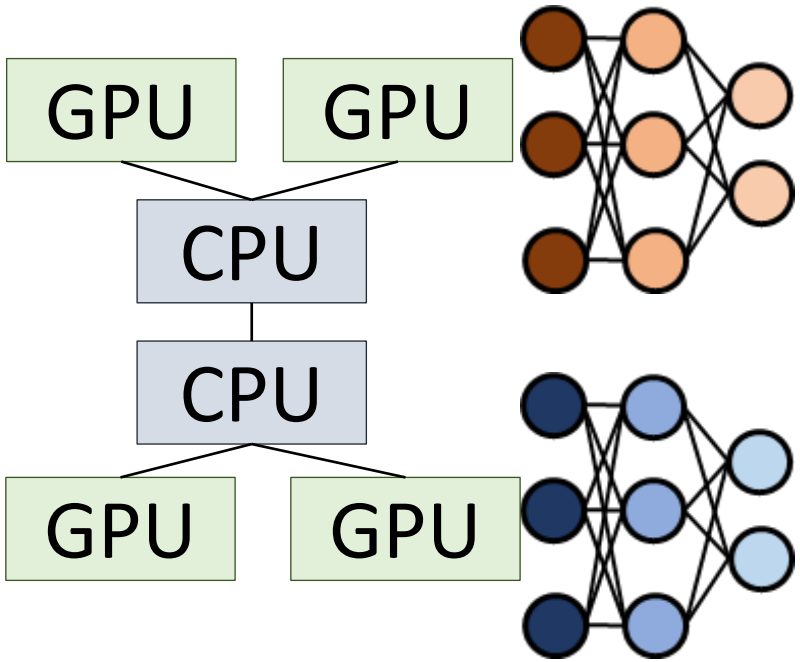
data sharing for collocated training

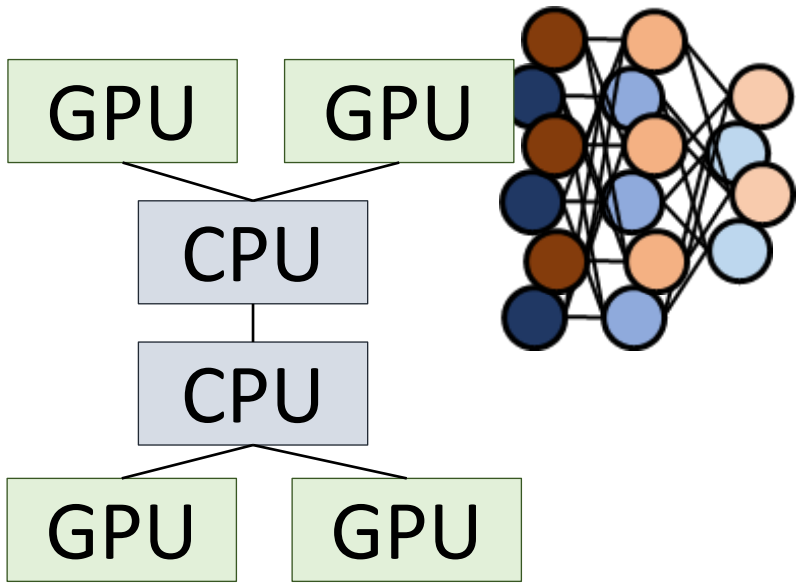
TensorSocket



minimize the redundancy!

collocated training





for *TensorSocket*, both are collocated training

TensorSocket requirements & limitations

➔ consumers go through the *same dataset* at the *same rate*

but consumers can …

- join at different epochs of training
- have different batch sizes
- be different models

➔ target is smaller scale

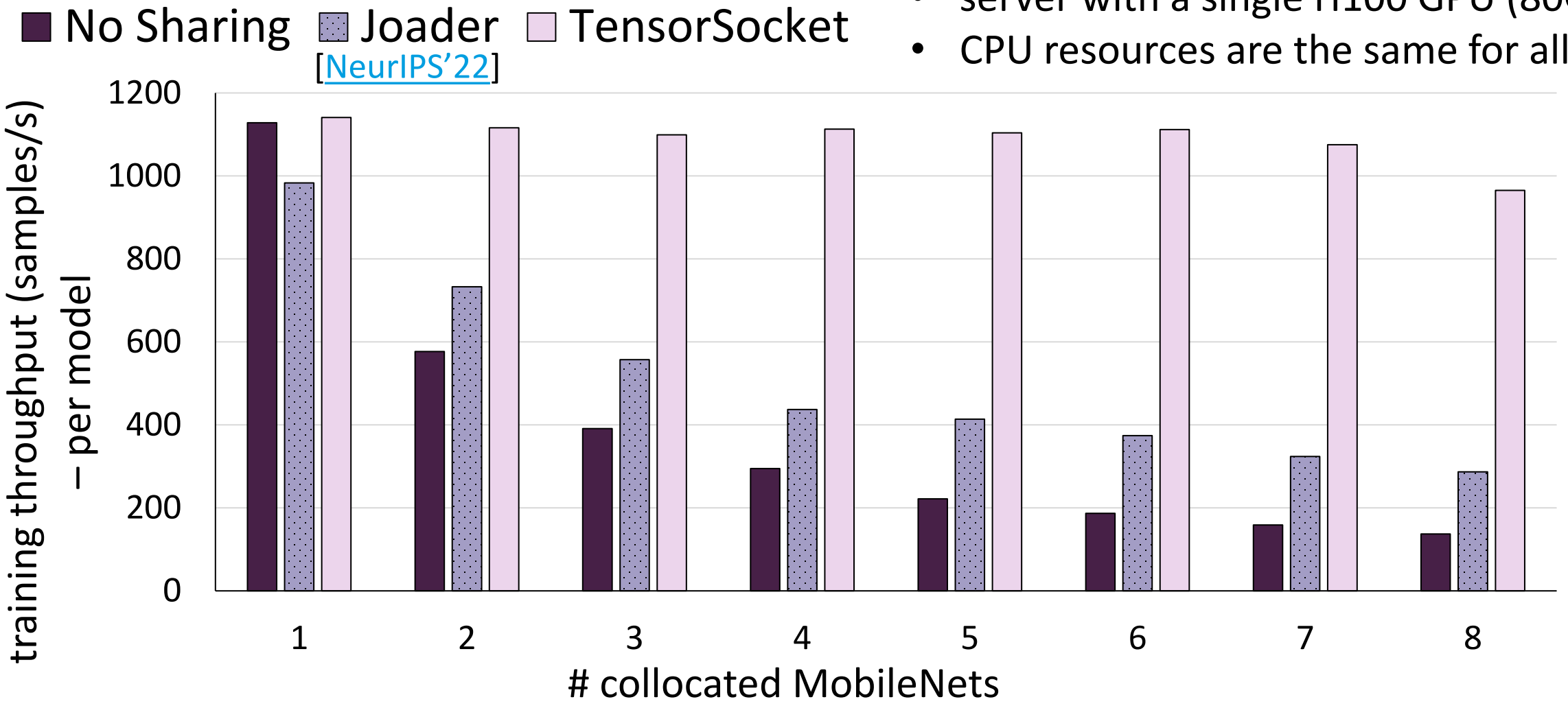
- collocation of model training on a single server
- models fit into the memory of a single GPU

not everyone needs “big” models & scale!

- Varoquaux et al. Hype, Sustainability, and the Price of the Bigger-is-Better Paradigm in AI
- Margot Seltzer, SIGMOD’25 keynote

resource savings

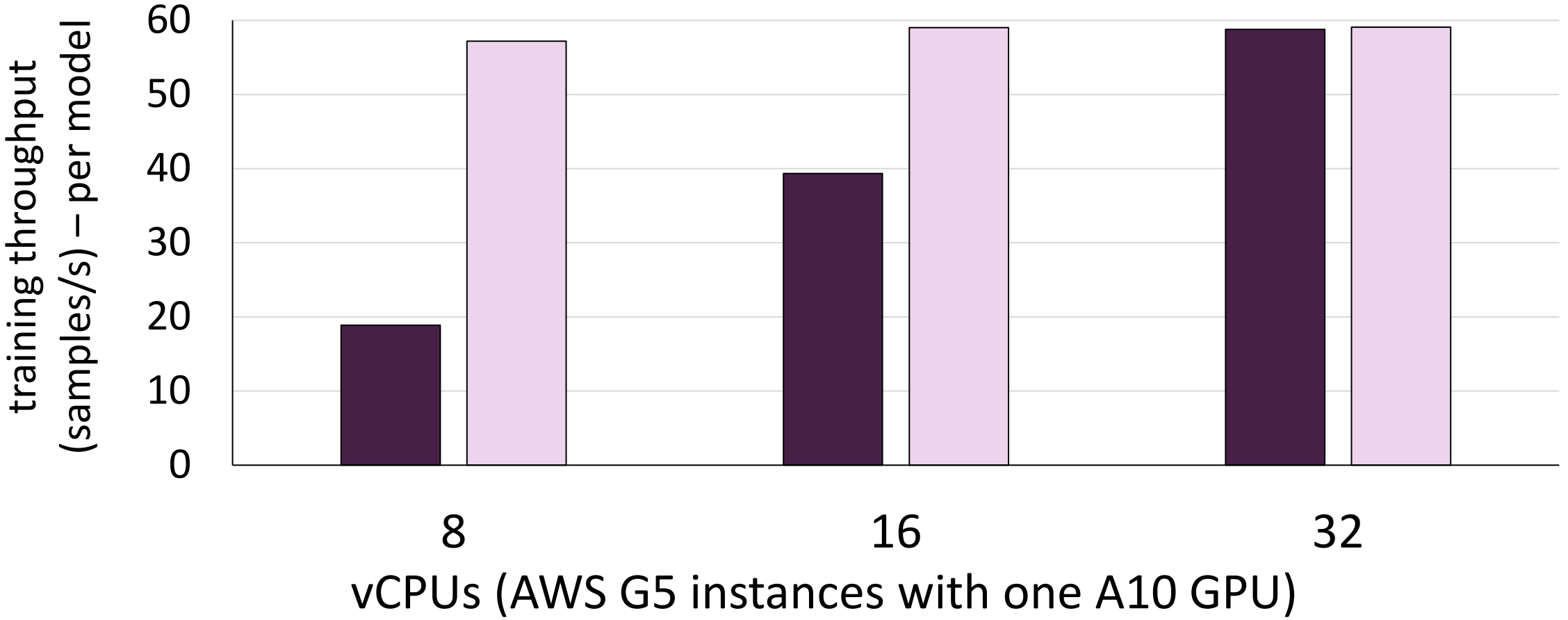
- server with a single H100 GPU (80GB)
- CPU resources are the same for all



TensorSocket sustains throughput even with GPU collocation & reduces both CPU and GPU needs for the whole workload.

cloud cost savings

- CLMR (audio classification model training)
- 4-way collocation



75% less vCPU need for the same training throughput

➔ 50% cost savings on AWS

data & work sharing in data systems

Analyzing and mitigating data stalls in DNN training

Authors: Jayashree Mohan, Amar Phanishayee, Ashish Raniwala, Vijay Chidambaram

Proceedings of the VLDB Endowment, Volume 14, Issue 5 • Pages 771 - 784 • https://doi.org/10.14778/3

tf.data service: A Case for Disaggregating ML Input Data Processing

Authors: Andrew Audibert, Yang Chen, Dan Graur, Ana Klimovic, Jiří Šimša, Chandramohan A. Thekkath | Authors Info & Claims

SoCC '23: Proceedings of the 2023 ACM Symposium on Cloud Computing • Pages 358 - 375 • https://doi.org/10.1145/3620

sharing for large scale ML

QPipe: a simultaneously pipelined relational query engine

Authors: Stavros Harizopoulos, Vladislav Shkapenyuk, Anastassia Ailamaki | Authors I

SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data
https://doi.org/10.1145/1066157.1066201

SharedDB: killing one thousand queries with one stone

Authors: Georgios Giannikis,  Gustavo Alonso, Donald Kossmann | Authors Inf

Proceedings of the VLDB Endowment, Volume 5, Issue 6 • Pages 526 - 537 • https://doi.org/1

sharing among concurrent queries

sharing for deep learning training

- workload collocation allows data & work sharing

- ***TensorSocket*** separates data loader as a producer of data for training jobs

➔ enables data & work sharing for collocated training jobs on the same dataset

reduces both the CPU & GPU needs (& costs) of training while increasing training throughput!

deep learning with fewer resources

An Analysis of Collocation on GPUs for Deep Learning Training

Ties Robroek, Ehsan Yousefzadeh-Asl-Miandoab, Pınar Tözün

EuroMLSys 2024

CARMA: Collocation-Aware Resource Manager

Ehsan Yousefzadeh-Asl-Miandoab, Florina M Ciorba, Pınar Tözün

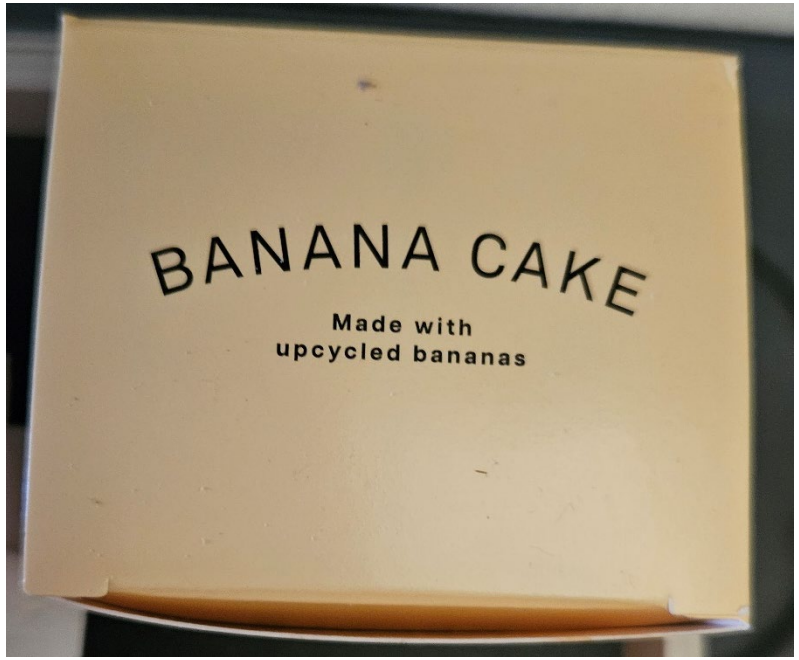
- resource sharing

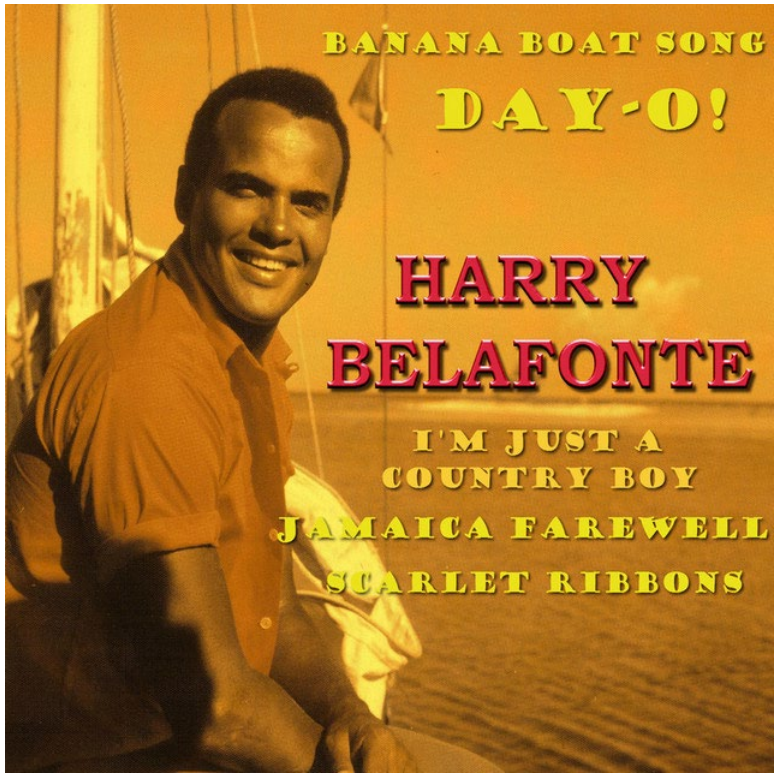
TensorSocket: Shared Data Loading for Deep Learning Training

Ties Robroek, Neil Kim Nielsen, Pınar Tözün

SIGMOD 2026

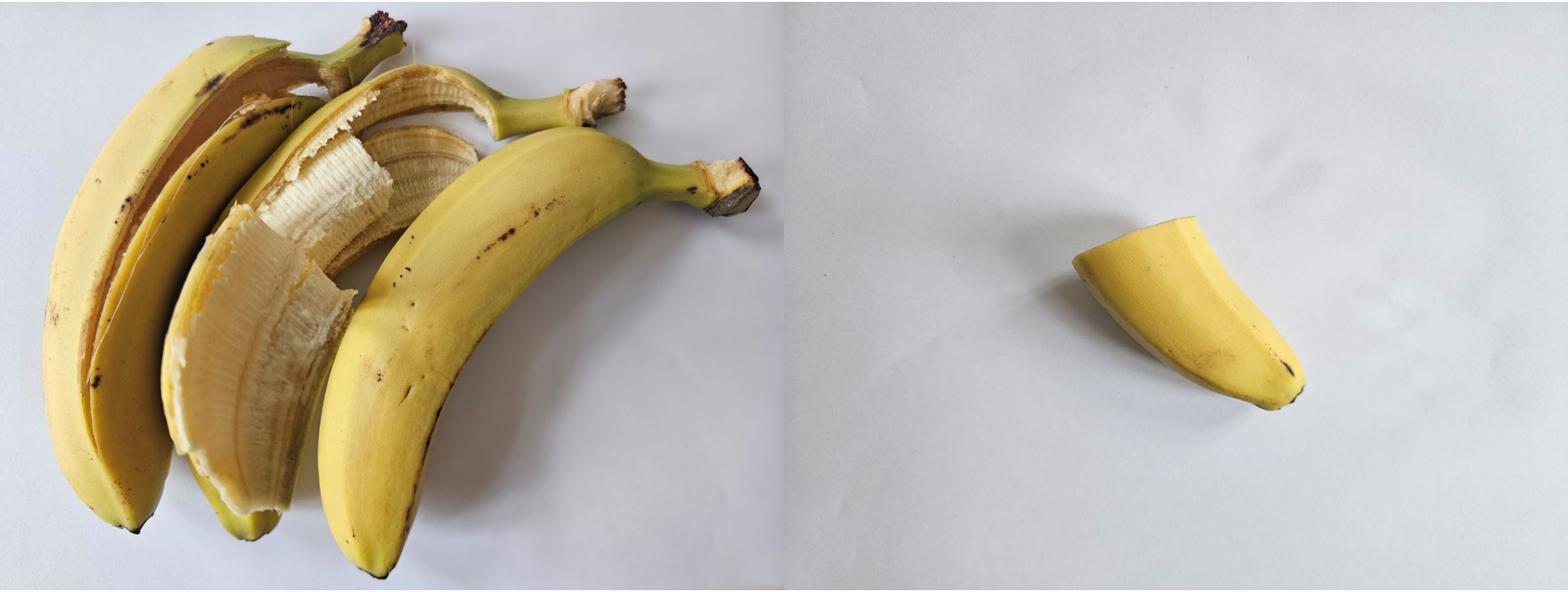
- data & work sharing


BANANA CAKE
Made with
upcycled bananas

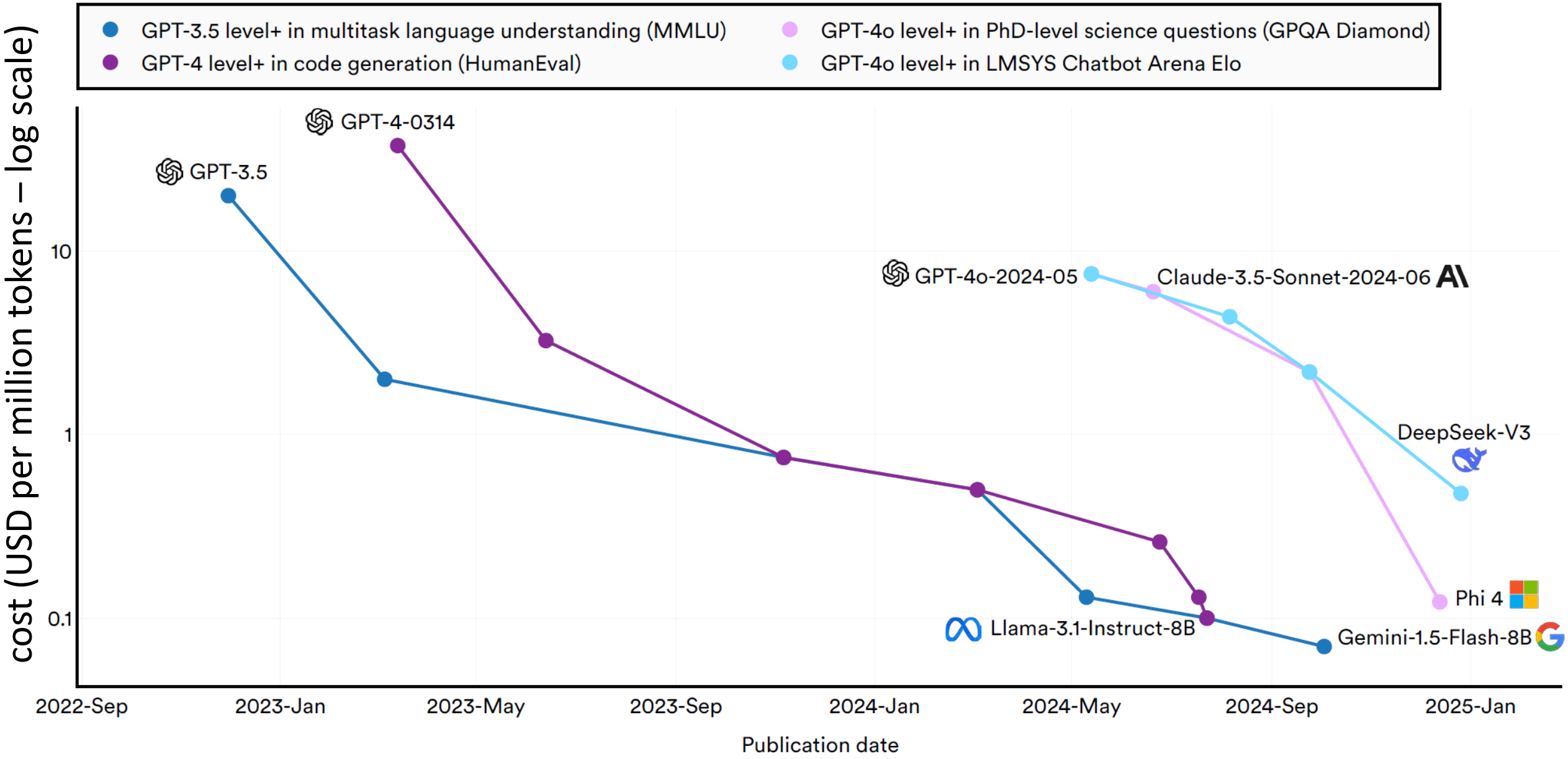

BANANA BOAT SONG
DAY-O!
HARRY
BELAFONTE
I'M JUST A
COUNTRY BOY
JAMAICA FAREWELL
SCARLET RIBBONS


PEEL SLOWLY AND SEE
Andy Warhol


BANANA



inference costs are going down



source: [Stanford AI Index Report 2025]

- a friend:
  I ask ChatGPT for advice when I cook
- a colleague:
  Use toolX for literature search, toolY for sparring for ideas, toolZ for writing, Claude for coding …
- a student:
  I asked ChatGPT/Gemini/… for this error, it gave me this
- Holger (yesterday):
  I am on the AI Agent train
- …

technology is a consumer product

cheaper product

➔ abundance / easy access

➔ higher consumption

➔ higher (carbon) footprint

jevon's paradox

there is a high cost of a cheap product

The Guardian

Eur

Robert Booth *UK technology editor*

Tue 4 Nov 2025 21.08 CET

Google plans to put datacentres in space to meet demand for AI

US technology company's engineers want to exploit solar power and the falling cost of rocket launches

THE OBSCENE ENERGY DEMANDS OF A.I.

How can the world reach net zero if it keeps inventing new ways to consume energy?

By Elizabeth Kolbert

March 9, 2024

Sam Altman gets defensive about AI's massive electricity usage: 'It also takes a lot of energy to train a human'

By Marco Quiroz-Gutierrez
Reporter

FORTUNE

February 24, 2026, 2:02 AM ET

The Atlantic

TECHNOLOGY

March 13, 2026

INSIDE THE DIRTY, DYSTOPIAN WORLD OF AI DATA CENTERS

The race to power AI is already remaking the physical world.

By Matteo Wong
Photographs by Landon Speers

path to sustainability

do we have to use a GenAI tool as often?

do we always need the biggest / latest GPU?

do we always need bigger scale?

how to decide? who decides?

how to incentivize lower use?

requires collaboration across disciplines!

ccit.itu.dk

CENTER FOR CLIMATE IT

Climate And Resource Awareness is Imperative to Achieving Sustainable AI (and Preventing a Global AI Arms Race)

Pedram Bakhtiarifard[⊥], Pınar Tözün[†], Christian Igel[⊥], Raghavendra Selvan[⊥]
[⊥]Department of Computer Science, University of Copenhagen, Denmark
[†]Data, Systems, & Robotics Section, IT University of Copenhagen, Denmark

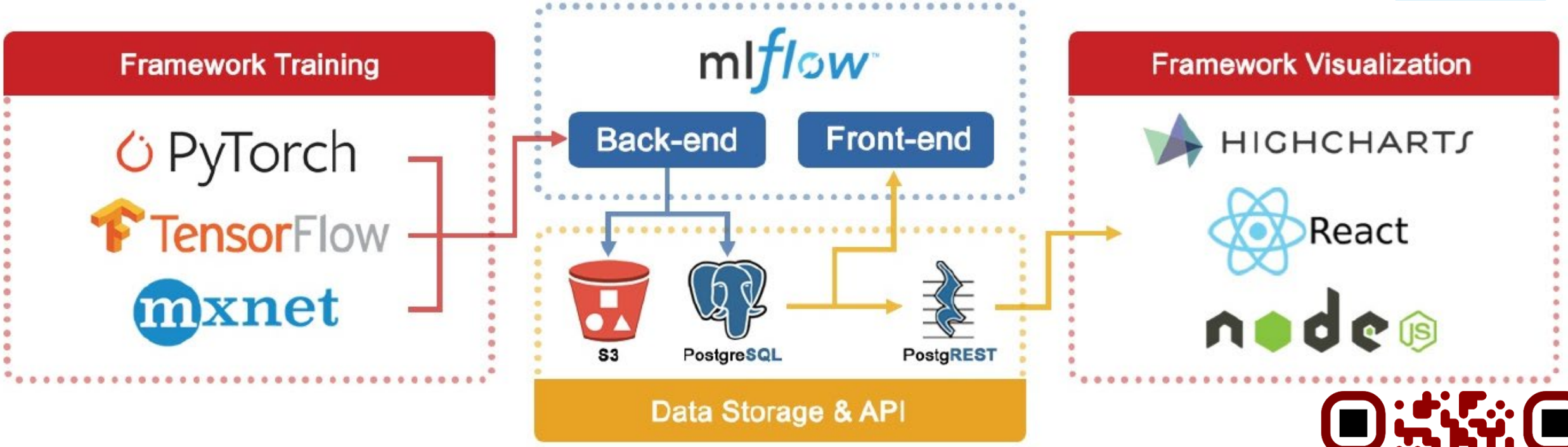
A Case for Ecological Efficiency in Database Server Lifecycles

Thomas Bodner, Martin Boissier, Tilmann Rabl, Ricardo Salazar-Diaz, Florian Schmeller, Nils Strassenburg, Ilin Tolovski, Marcel Weisgut, Wang Yue

DES

how to be more aware of hardware use?

radT [DEEM'23]



➔ easy, extensible, & scalable tracking of hardware metrics (GPU utilization, storage access, carbon footprint …)

➔ frontend for data exploration

used by our group & data scientists @ITU for systematic benchmarking of deep learning training & inference

RAD - resource-aware data systems

IT UNIVERSITY OF COPENHAGEN

postdocs

Ties Robroek

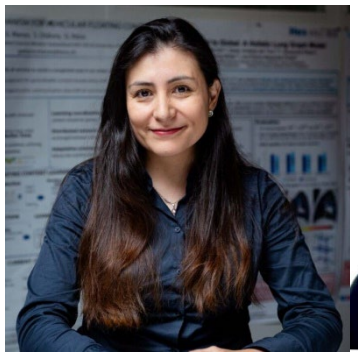
Ehsan Yousefzadeh-Asl-Miandoab

phd students

Robert Bayer

Jens Birk Andersen

ongoing collaborations



Pamela Delgado, Büşra Karatay
HEIG-VD

Tilmann Rabl, Marcel Weisgut
HPI

Julian Priest
ITU

Vivek Shah
Samsung

deep learning with fewer resources

thank you!

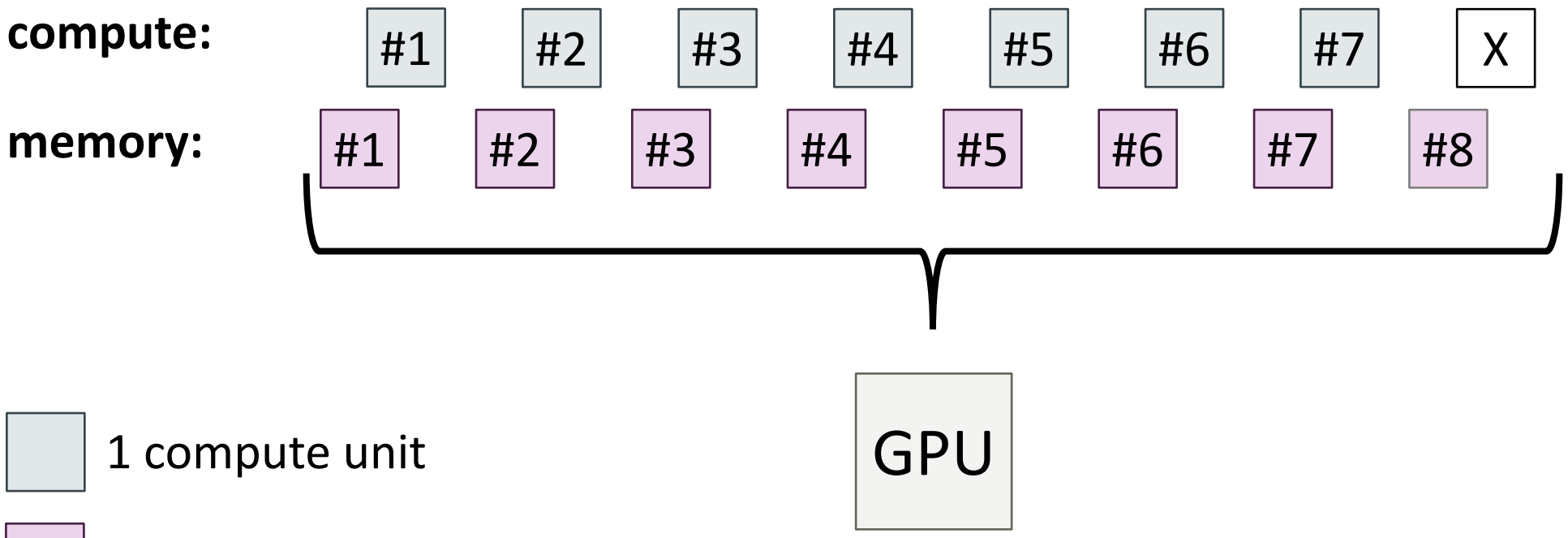
- not all training needs all the resources of a single GPU
- we need collocation-aware resource managers to reduce resource consumption of deep learning
- sharing data & work on preparing data can further reduce hardware resource needs / costs

pito@itu.dk
pinartozun.com

you don’t always need more!

backup

multi-instance GPU



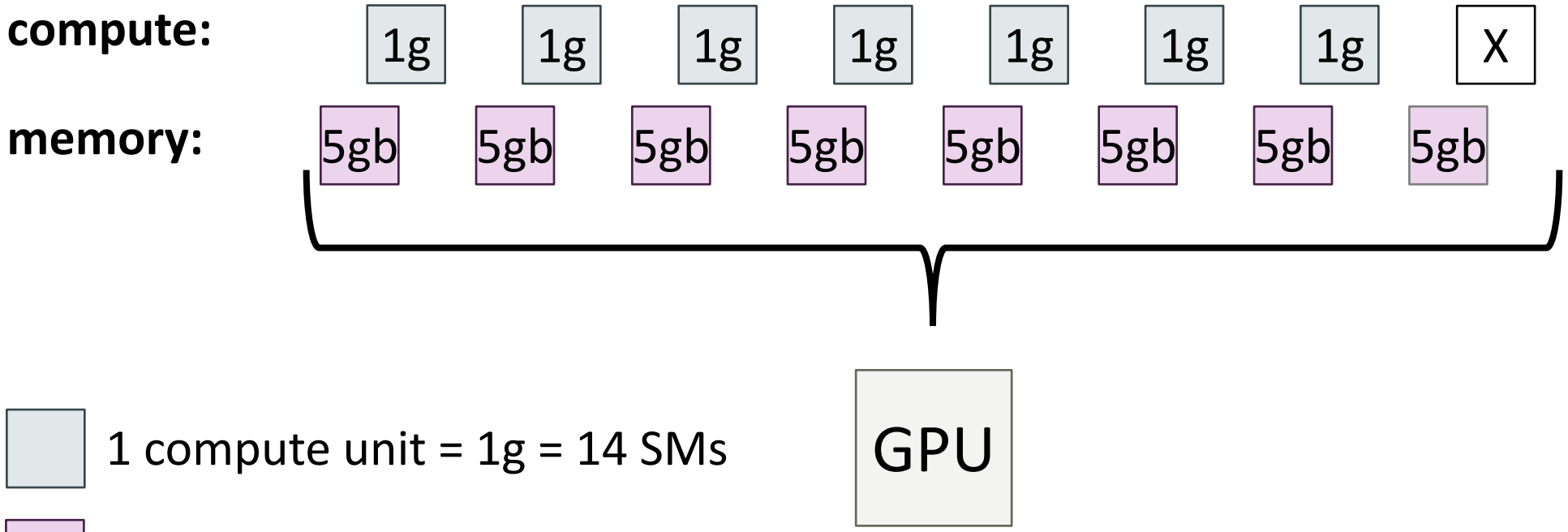
1 compute unit

1 memory unit

unused available (memory/compute) unit

X unavailable compute unit

multi-instance GPU on A100 (40GB)



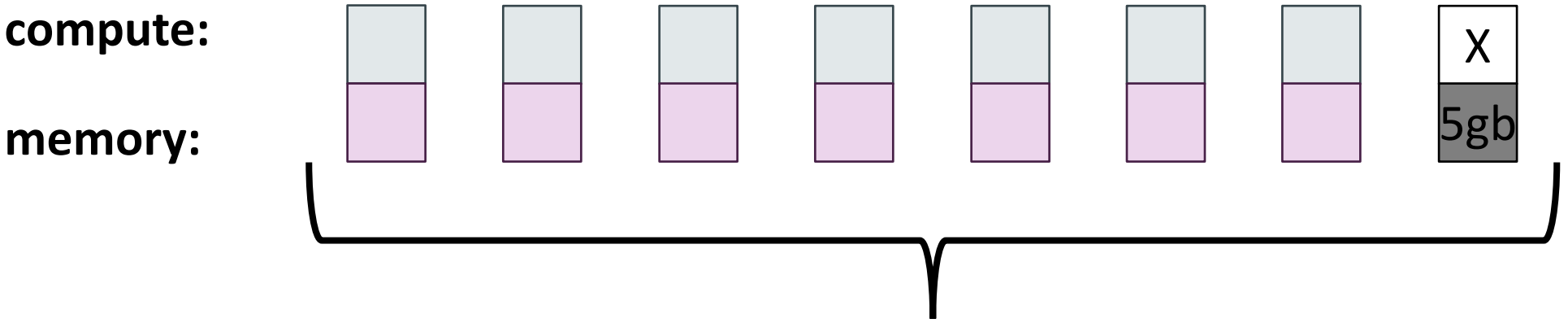
1 compute unit = 1g = 14 SMs

1 memory unit = 5GB

unused available (memory/compute) unit

X unavailable compute unit = 10 SMs (streaming multiprocessor)

multi-instance GPU on A100 (40GB)



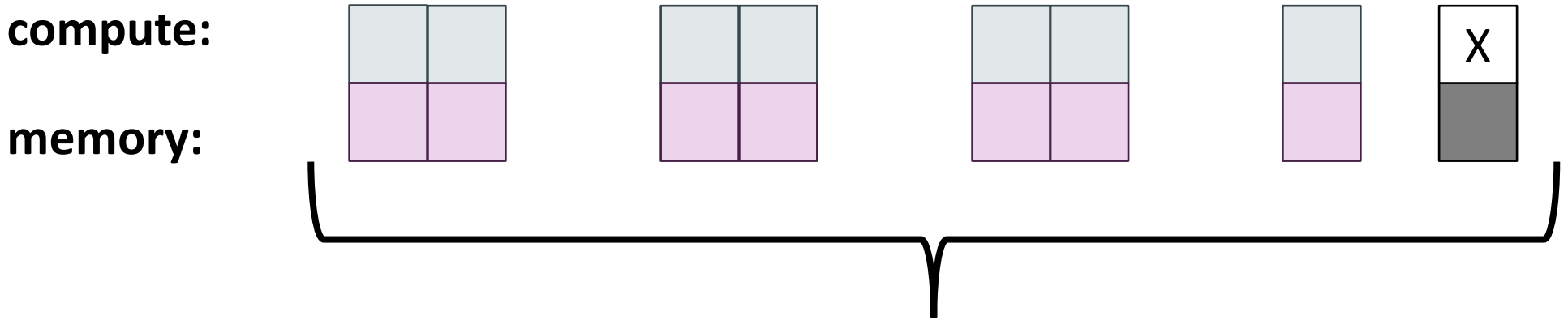
1 compute unit = 1g = 14 SMs

1 memory unit = 5GB

unused available (memory/compute) unit

X unavailable compute unit = 10 SMs (streaming multiprocessor)

multi-instance GPU on A100 (40GB)



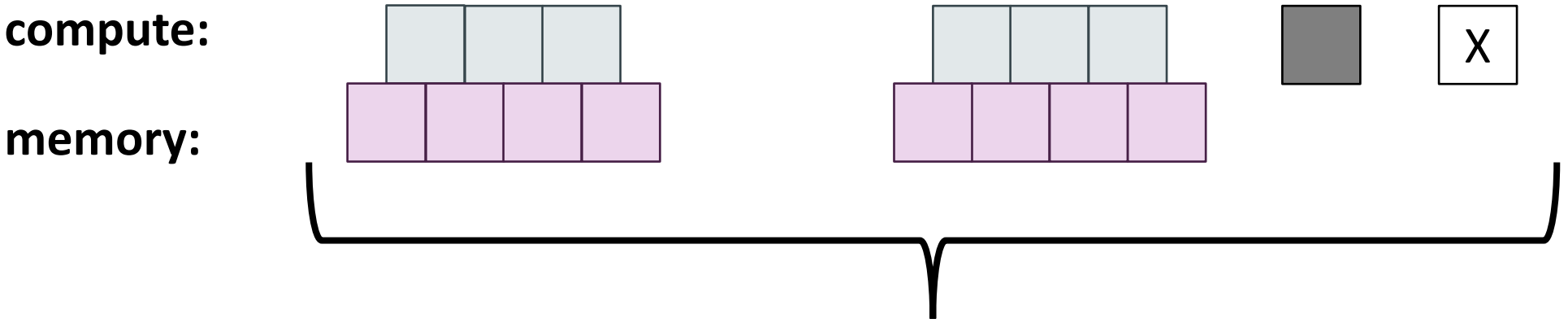
1 compute unit = 1g = 14 SMs

1 memory unit = 5GB

unused available (memory/compute) unit

X unavailable compute unit = 10 SMs (streaming multiprocessor)

multi-instance GPU on A100 (40GB)



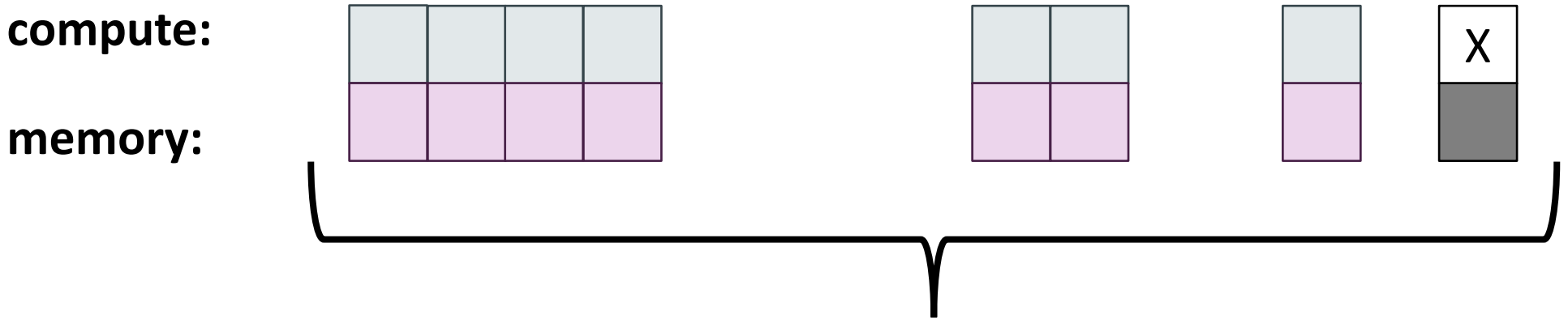
1 compute unit = 1g = 14 SMs

1 memory unit = 5GB

unused available (memory/compute) unit

X unavailable compute unit = 10 SMs (streaming multiprocessor)

multi-instance GPU on A100 (40GB)



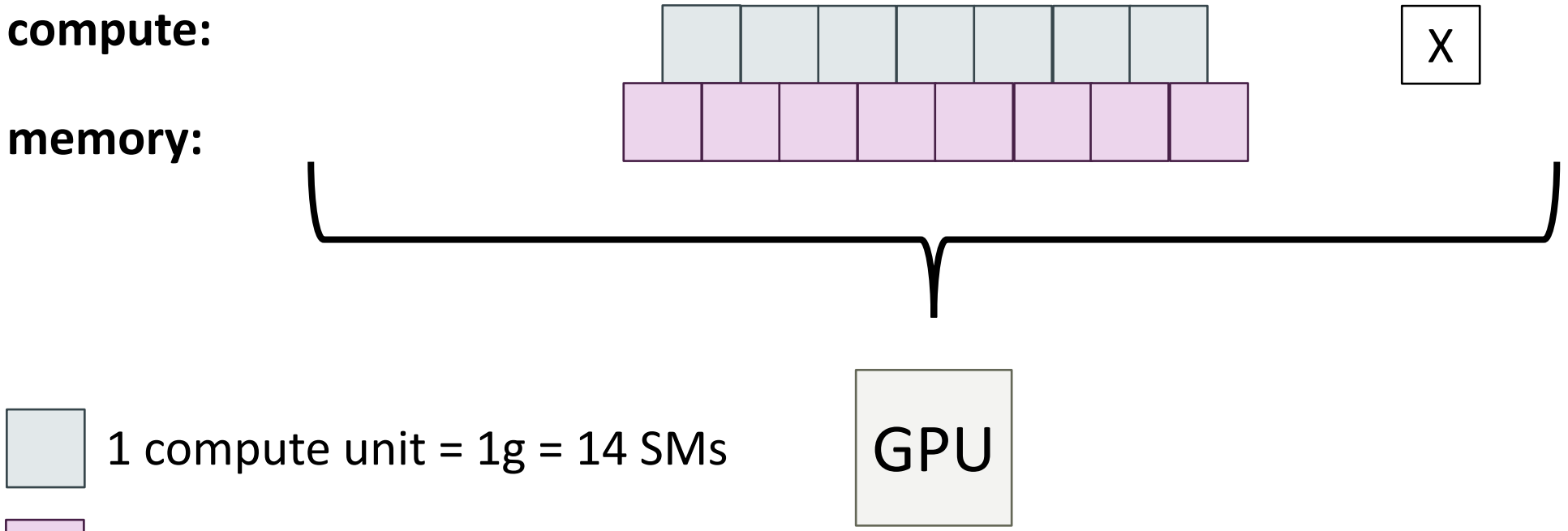
1 compute unit = 1g = 14 SMs

1 memory unit = 5GB

unused available (memory/compute) unit

X unavailable compute unit = 10 SMs (streaming multiprocessor)

multi-instance GPU on A100 (40GB)



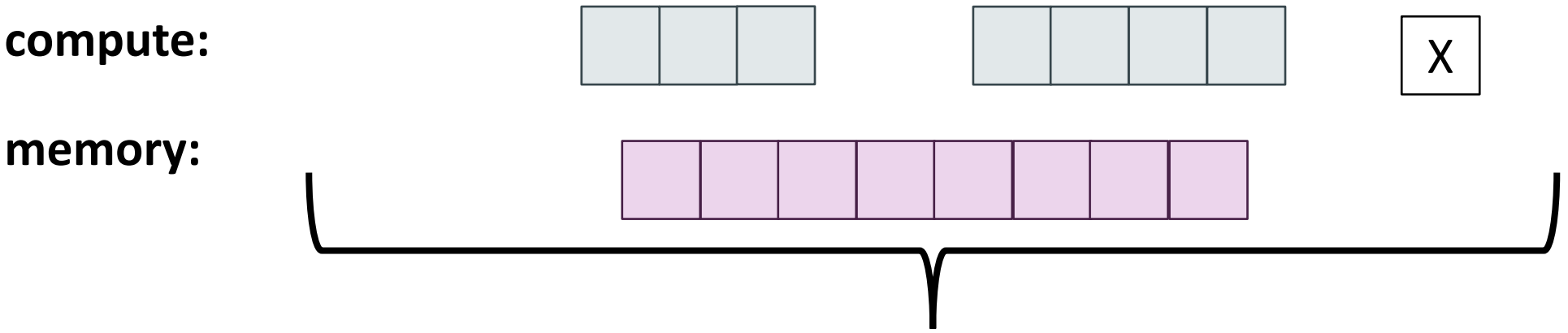
1 compute unit = 1g = 14 SMs

1 memory unit = 5GB

unused available (memory/compute) unit

X unavailable compute unit = 10 SMs (streaming multiprocessor)

multi-instance GPU on A100 (40GB)



1 compute unit = 1g = 14 SMs

1 memory unit = 5GB

unused available (memory/compute) unit

X unavailable compute unit = 10 SMs (streaming multiprocessor)

performance impact of collocation

NVIDIA DGX Station A100

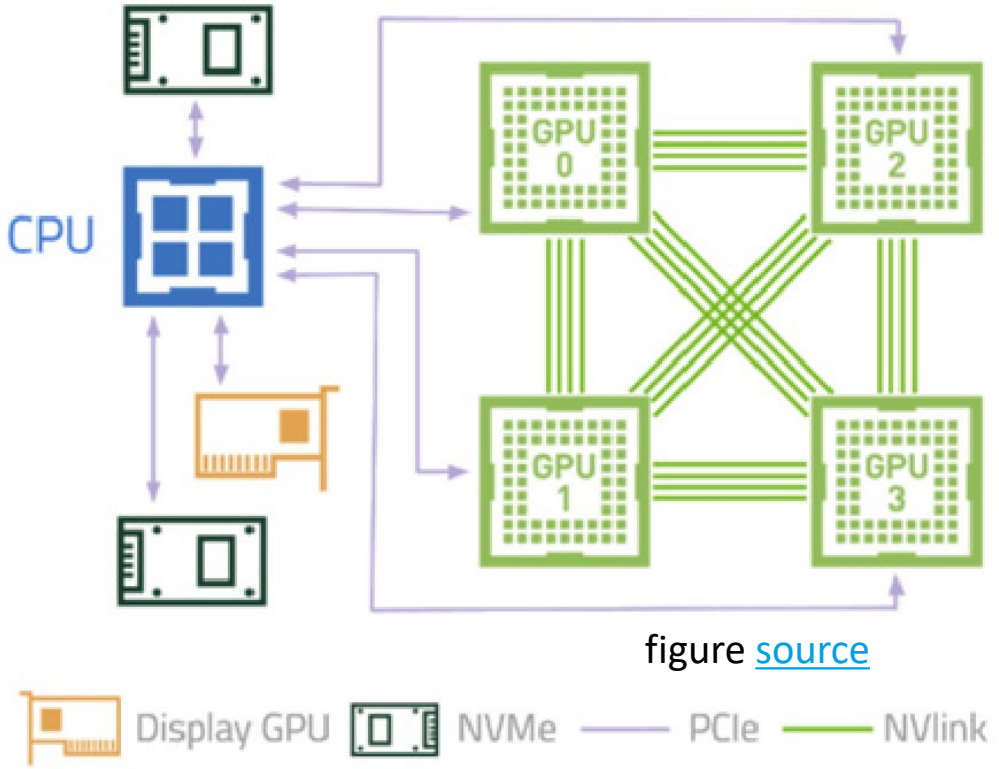


figure source

CPU = AMD 7742 – 512 GB RAM
64 physical cores
GPU = NVIDIA A100 – 40 GB RAM

workloads	model	batch size	dataset
small	**ResNet26** EfficientNet	128	CIFAR-10
medium	**ResNet50** EfficientNet	128	downsampled ImageNet*
large	**ResNet152** CaiT	32 128	ImageNet (2012)
xlarge	**DLRM**	1	Criteo Terabyte

- image models: CNN & transformers recommender model
- on single GPU with PyTorch v2.0
- results reported from 2nd epoch of training

CARMA evaluation

trace mix

- ***heavy:*** 1-2GPUs, epoch time: 7.5mins to >1hour, memory: 9GB to 30GB
  XLNet (base & large), BERT (base & large), GPT2 (large)
- ***medium:*** 1 GPU, epoch time: 1min to >1hour, memory: 1GB to 30GB
  EfficientNet, ResNet (50), MobileNet, VGG …
- ***light:*** 1 GPU, epoch time: up to 1min, memory: up to 1GB
  ResNet (18, 34), MobileNet (small) …

2 traces of 60 training tasks:

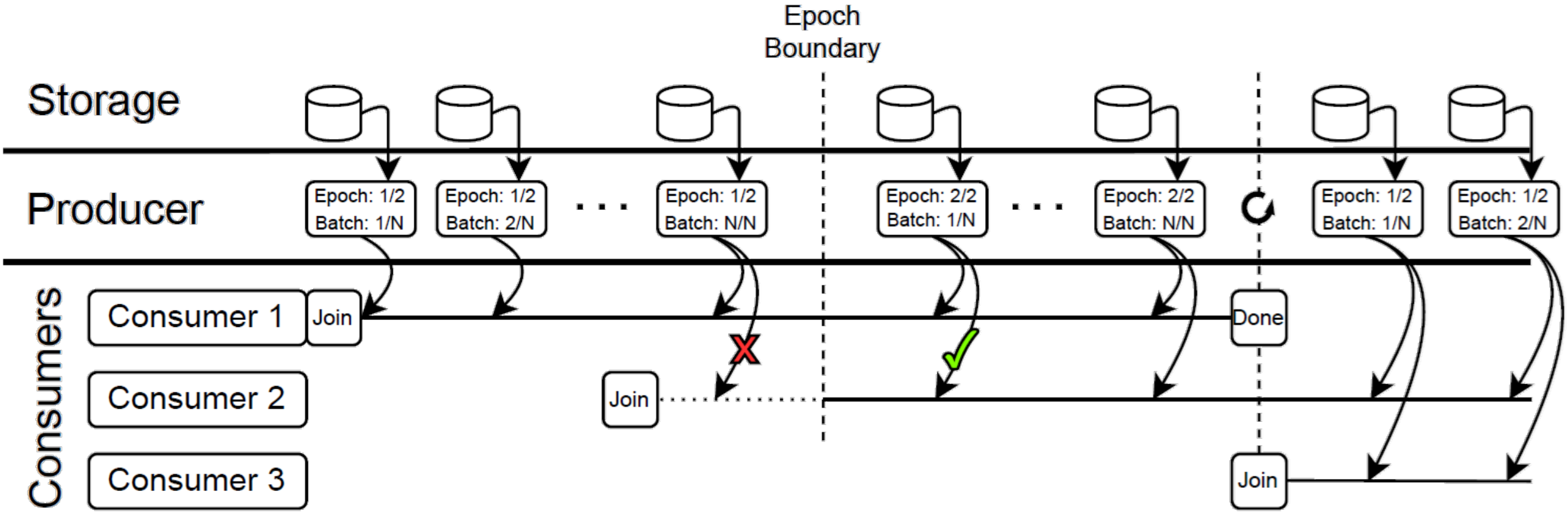
- 30% light, 60% medium/heavy, and 10% heavy 2 GPU models

based on real-world workload trace analysis

- task submission times: "Analysis of Large-Scale Multi-Tenant GPU Clusters for DNN Training Workloads." ATC 2019
- task distribution: "An Empirical Study on Low GPU Utilization of Deep Learning Jobs." TPDS 2022

consumers joining at different epochs

TensorSocket

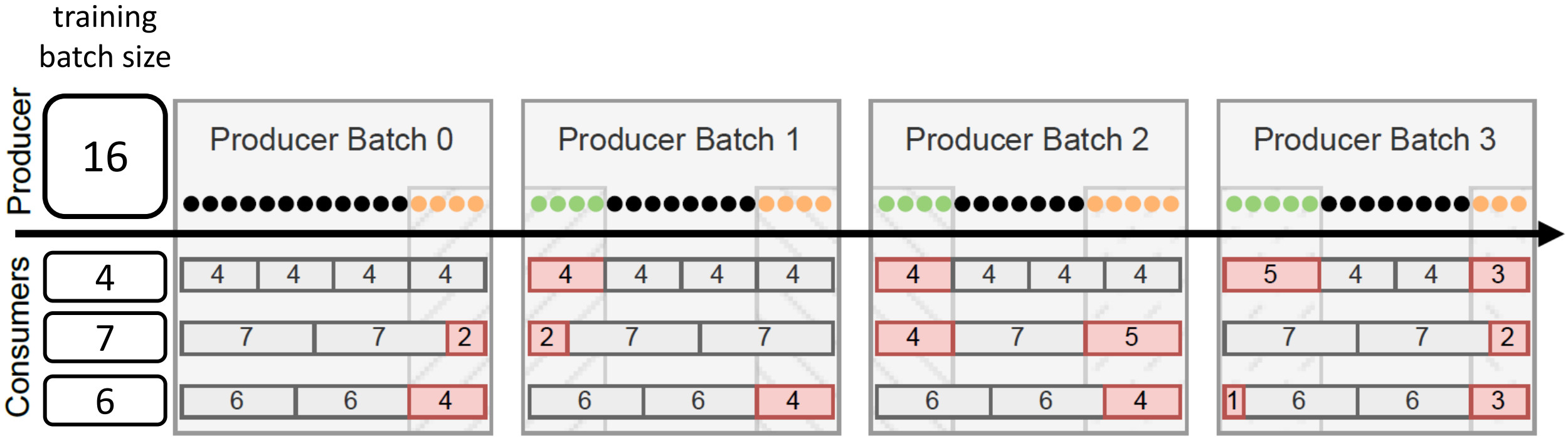


trade-off of training latency for throughput & resources

training is trail-&-error → latency is less critical

flexible batch sizing

TensorSocket



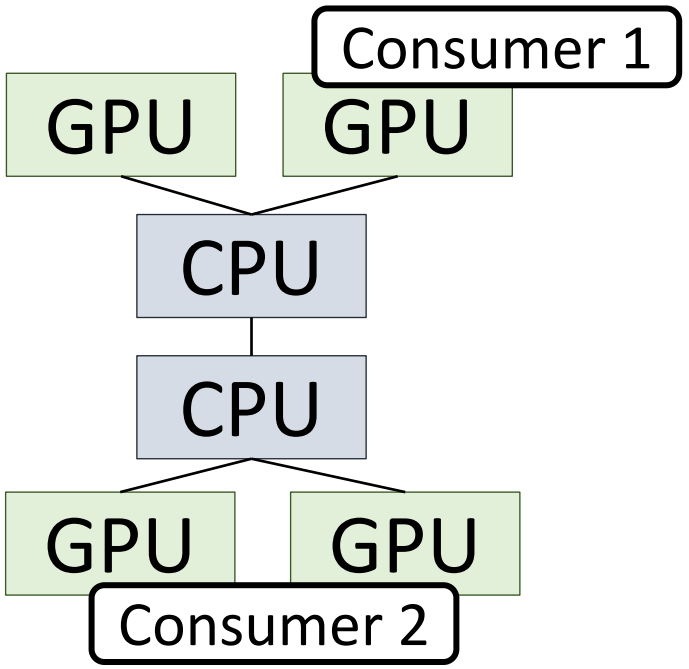
trade-off of repeated data to get flexibility

in practice, batch sizes tend to be multiples of 2

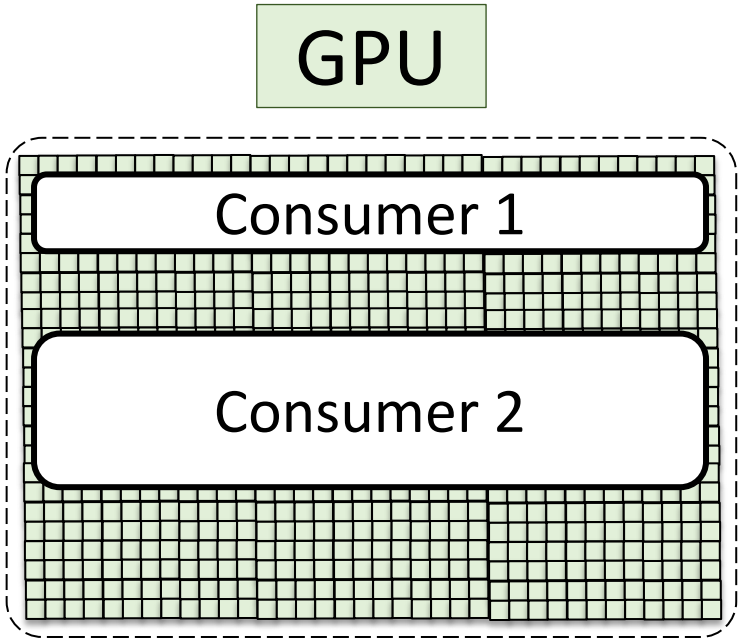
different consumers / models

Consumer 1
MobileNet Small

Consumer 2
MobileNet Large



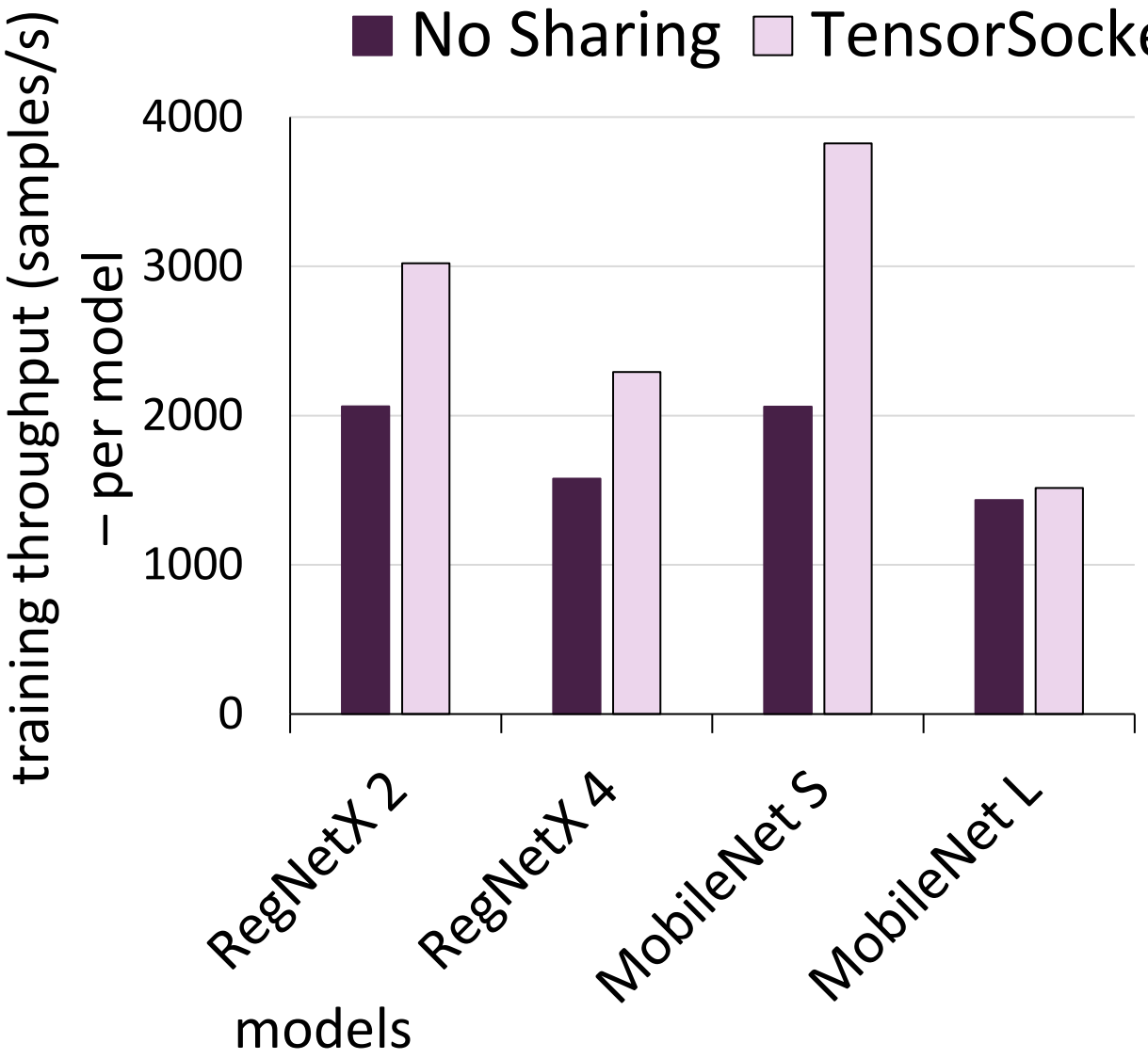
TensorSocket

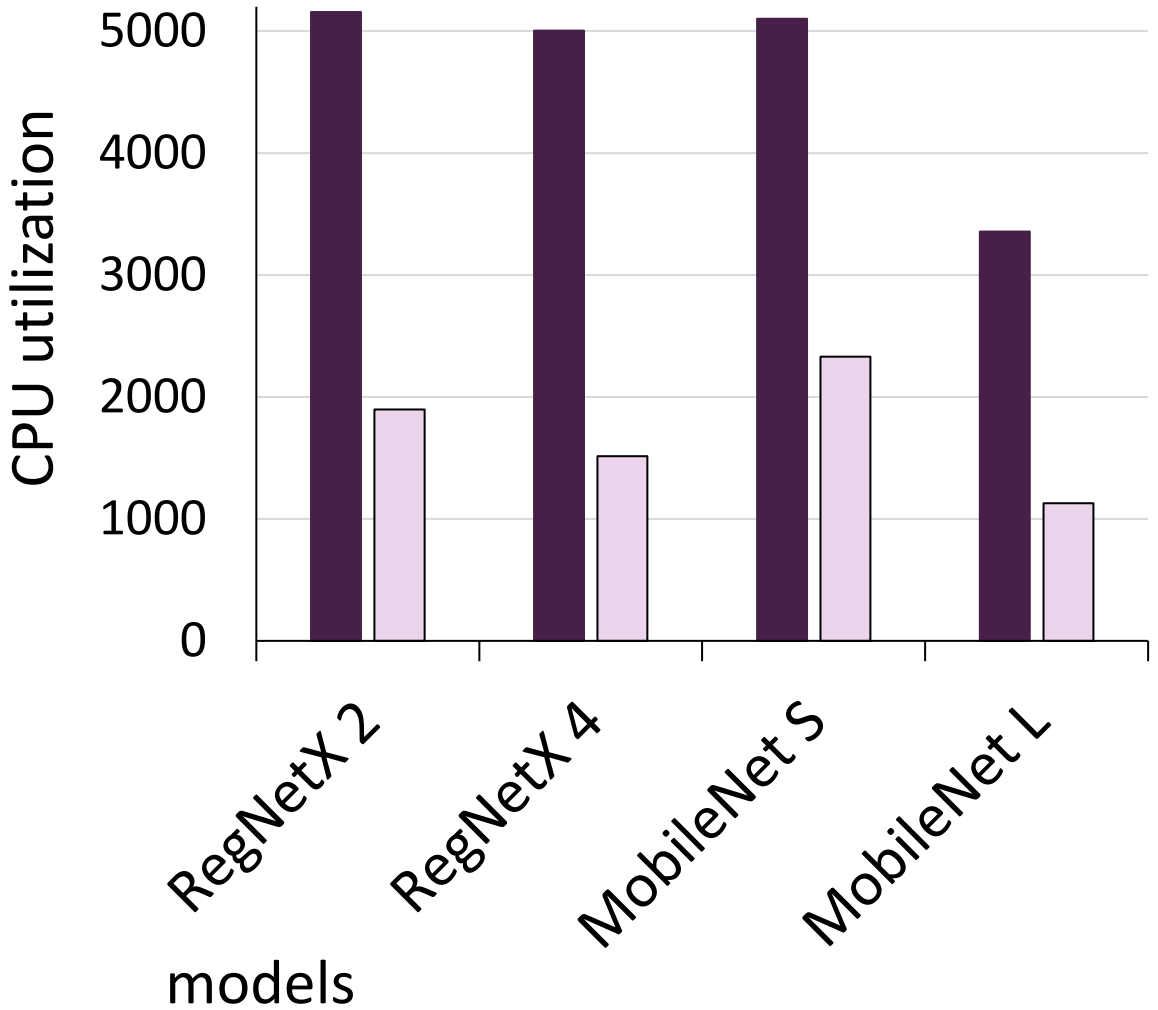


can adjust the hardware resources per consumer to ensure each goes over the data at the same rate

impact of data sharing

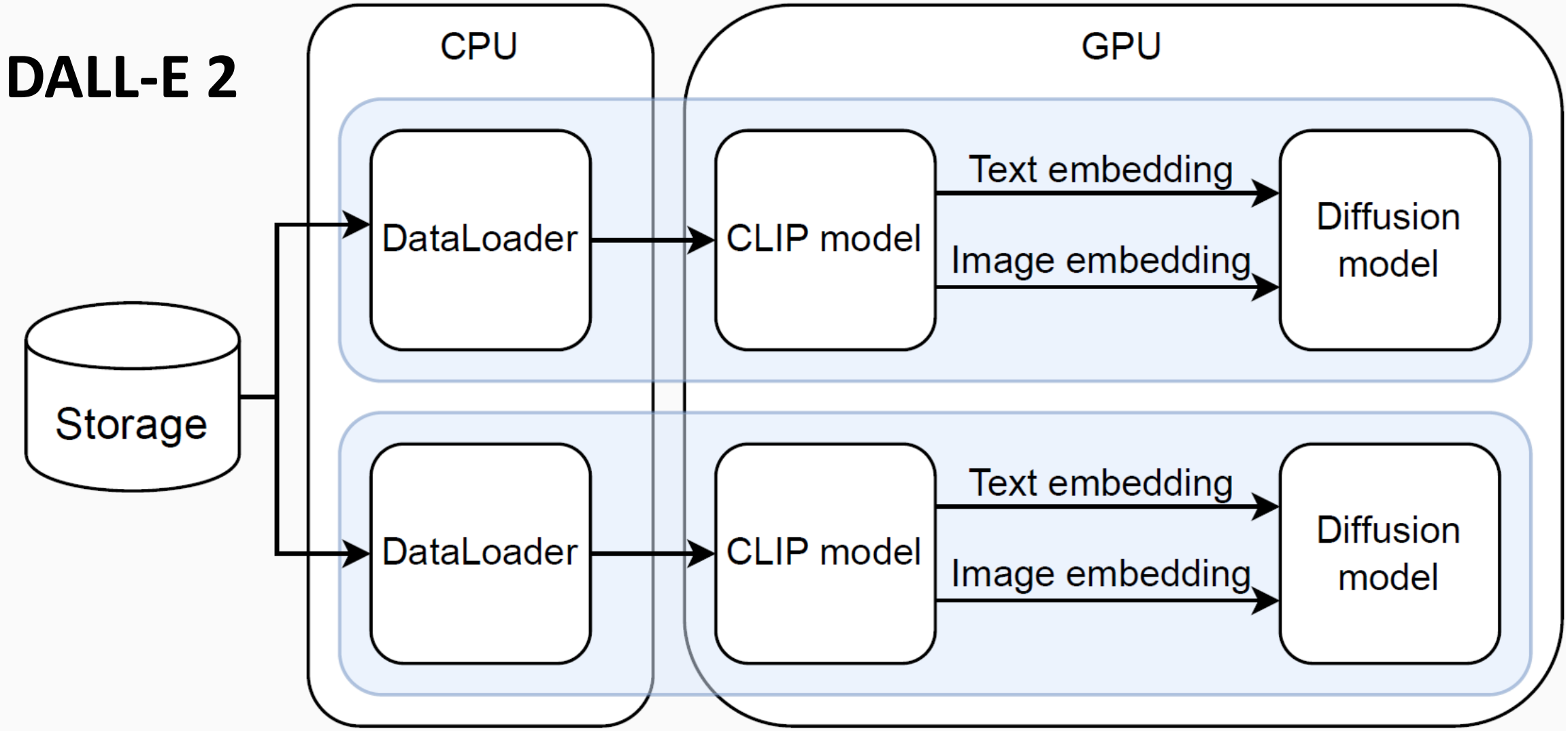
- on PyTorch
- a server with 4 A100 (40GB) GPUs
- one model training on each





higher overall throughput & reduced CPU need!

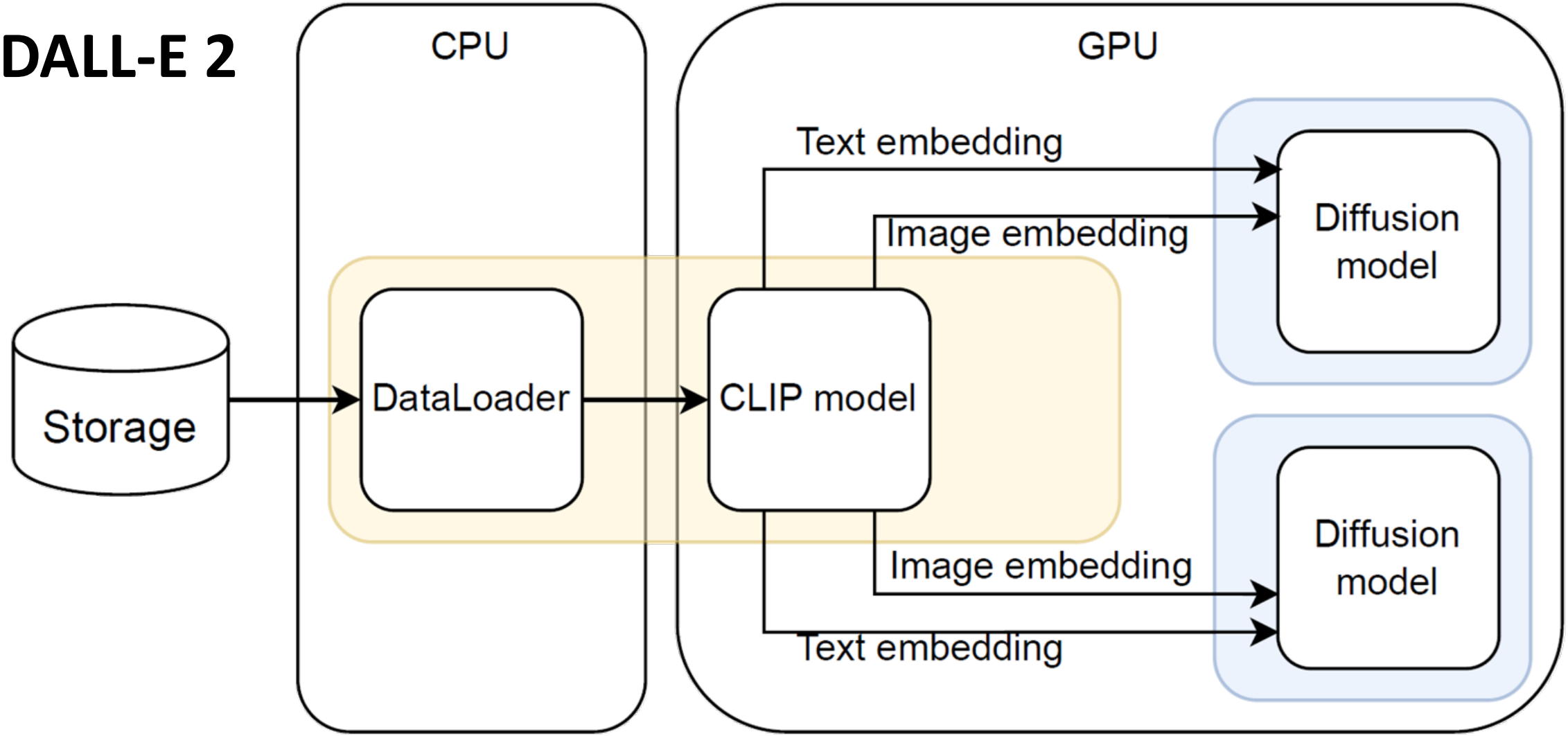
data sharing for collocated training



can also reduce work on GPUs!

data sharing for collocated training

DALL-E 2



can also reduce work on GPUs!