DASYA

RAD

IT UNIVERSITY OF COPENHAGEN

www.dasya.dk
@dasyaITU

rad.itu.dk

www.itu.dk

different scales of resource-aware deep learning & how to tackle them

Pınar Tözün

Associate Professor, IT University of Copenhagen

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

INDEPENDENT RESEARCH FUND DENMARK

novo nordisk foundation

Imperial College London
June 6, 2024

unsustainable growth of deep learning

2012 → **present**

- powerful hardware
- larger datasets
- deep learning frameworks

APIs (python, R, …)

deep learning software — PyTorch, TensorFlow

CPU | GPU | TPU | FPGA | …

$\underbrace{\text{CPU} \quad \text{GPU}}_{\text{commodity hardware}}$

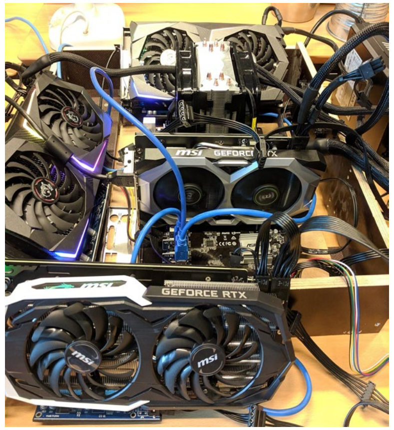
- ***several orders of magnitude increase*** in the ***computational need*** for models.
- estimated carbon footprint for large language model training = ***average yearly energy of several US homes***

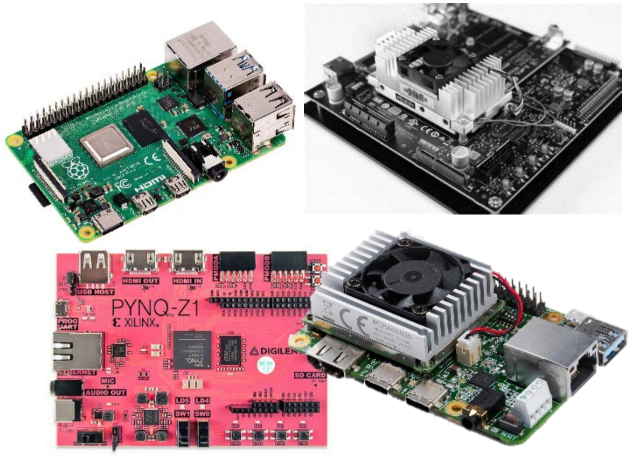
need for higher computational efficiency!

sources: https://openai.com/blog/ai-and-compute/
Dodge et al. "Measuring the Carbon Intensity of AI in Cloud Instances." FAccT 2022

hardware scales for deep learning

large ⟷ **tiny**





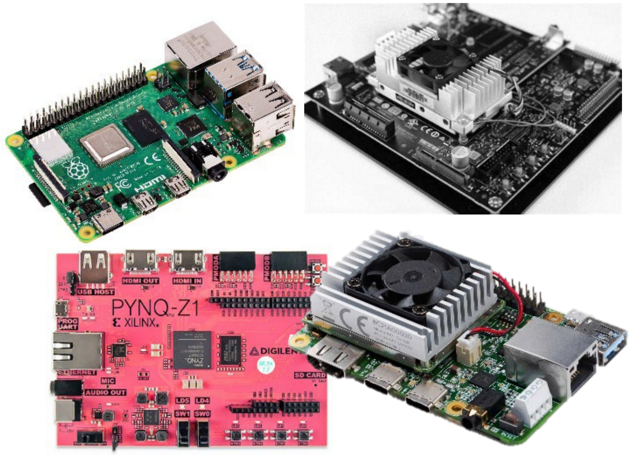
➔can we utilize these hardware well?

➔can we do more with less?

hardware scales for deep learning

large ⟷ **tiny**

“An Analysis of Collocation on GPUs for Deep Learning Training”, EuroMLSys 2024



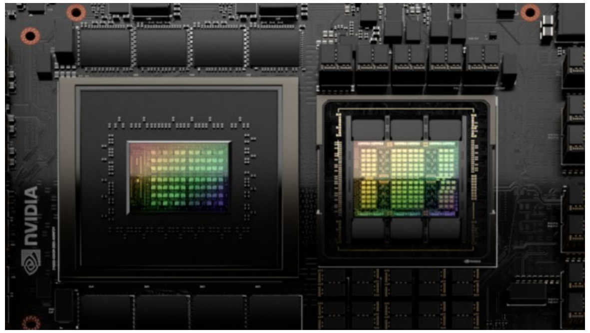
“Reaching the Edge of the Edge: Image Analysis in Space”, DEEM 2024

➔can we utilize these hardware well?

➔can we do more with less?

hardware underutilization

NVIDIA H200



141GB GPU memory

50MB L2 cache

4.8TB/s Memory Bandwidth

- ***@ITU***, many ML jobs utilize ***less than 50% of GPU resources*** e.g., transfer learning, small models
- ***in real-world*, ~52% GPU utilization*** on average for 100,000 jobs

exclusive GPU access is a big part of the problem!

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

workload collocation

= multiple workloads sharing hardware resources

benefits when a single workload cannot utilize available resources well / fully

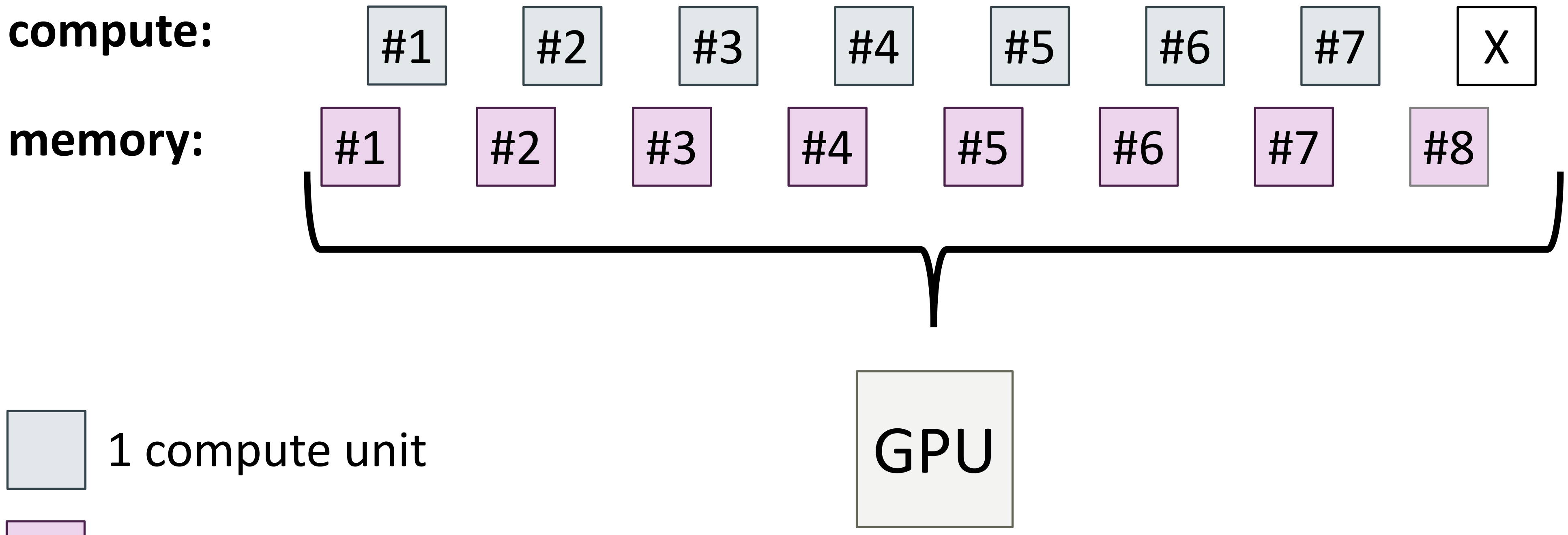
usual challenge → interference across workloads

GPU-specific challenge → no fine-grained & flexible resource sharing mechanism

workload collocation on (NVIDIA) GPUs

- ***naïve collocation & virtualization***
  - kernels of different applications are serialized
  - × provides limited parallelism
- ***multi-process service (MPS)***
  - GPU resources are split (manually or automatically) across applications
  - ✓ kernels of different applications can run simultaneously
  - × allowed for one user only (for safety reasons)
- ***multi-instance GPU (MIG)***
  - hardware support for resource split, introduced with NVIDIA A100
  - ✓ prevents interference & can do all the above in a MIG partition
  - × rigid partitioning of GPU resources

multi-instance GPU



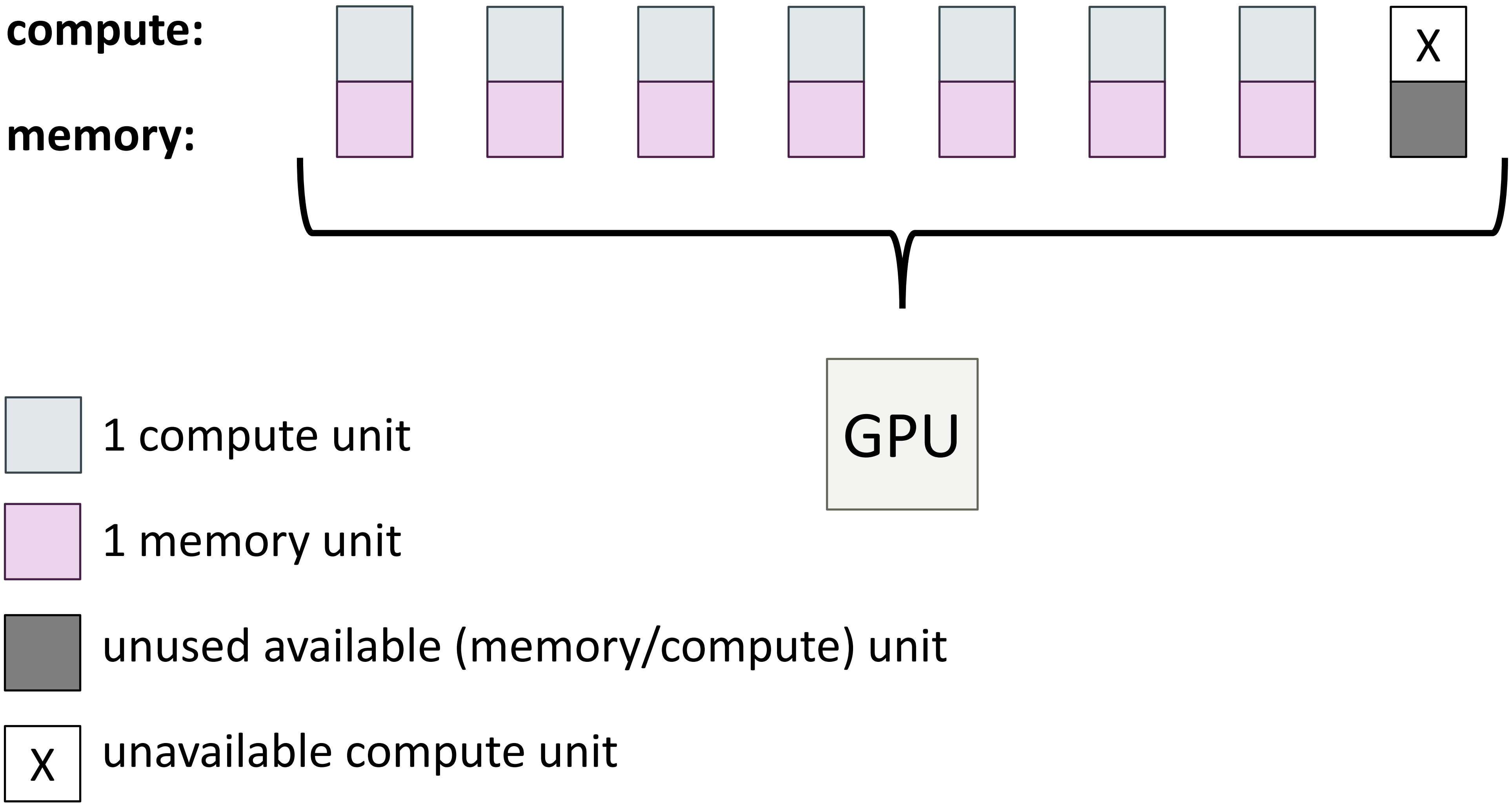
1 compute unit

1 memory unit

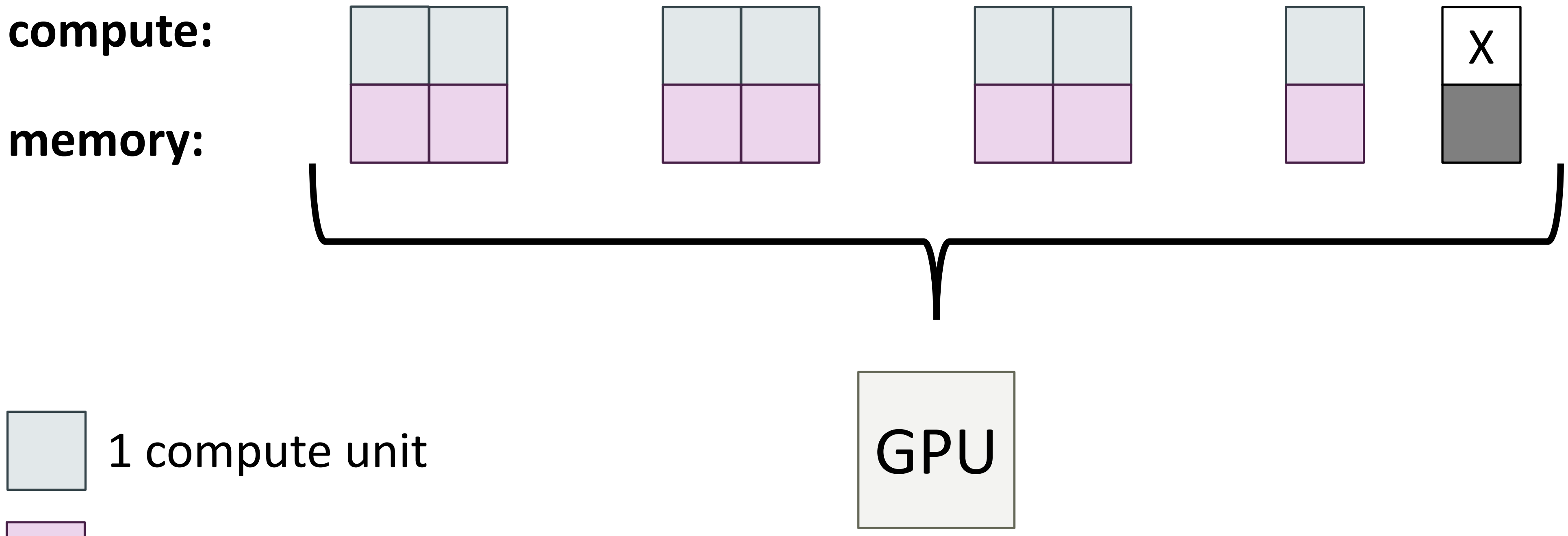
unused available (memory/compute) unit

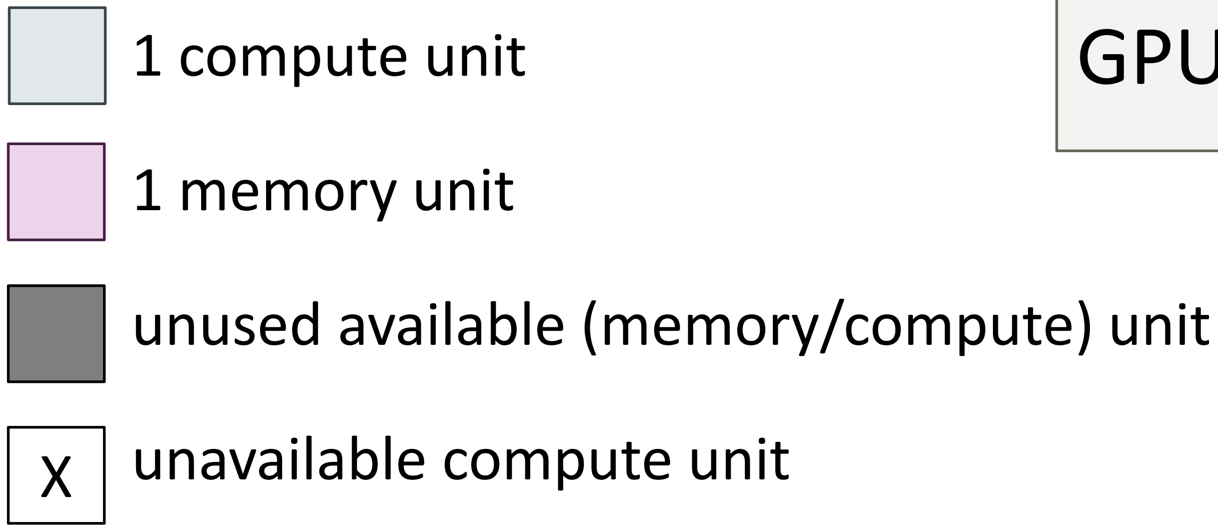
X unavailable compute unit

multi-instance GPU

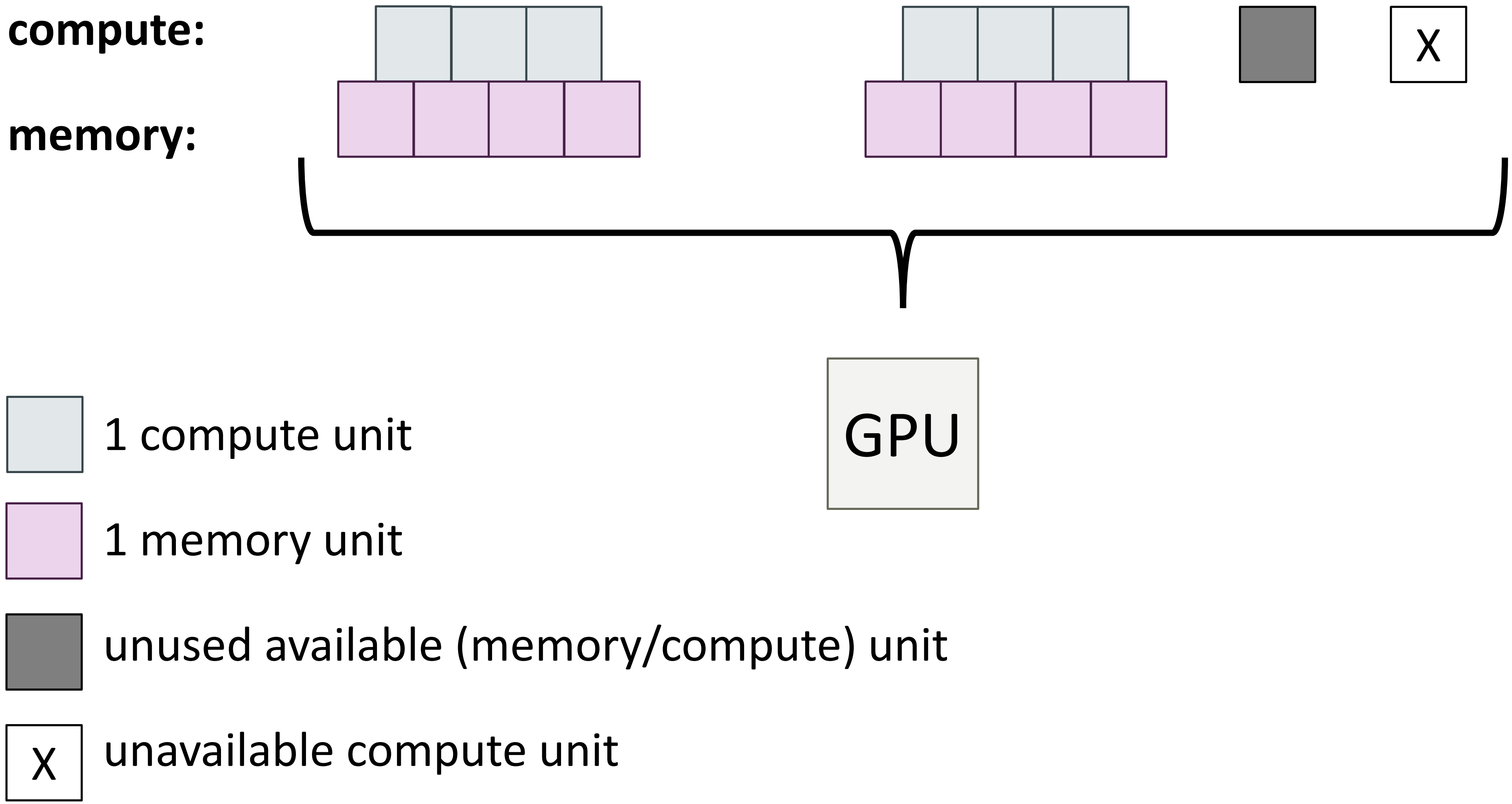


multi-instance GPU

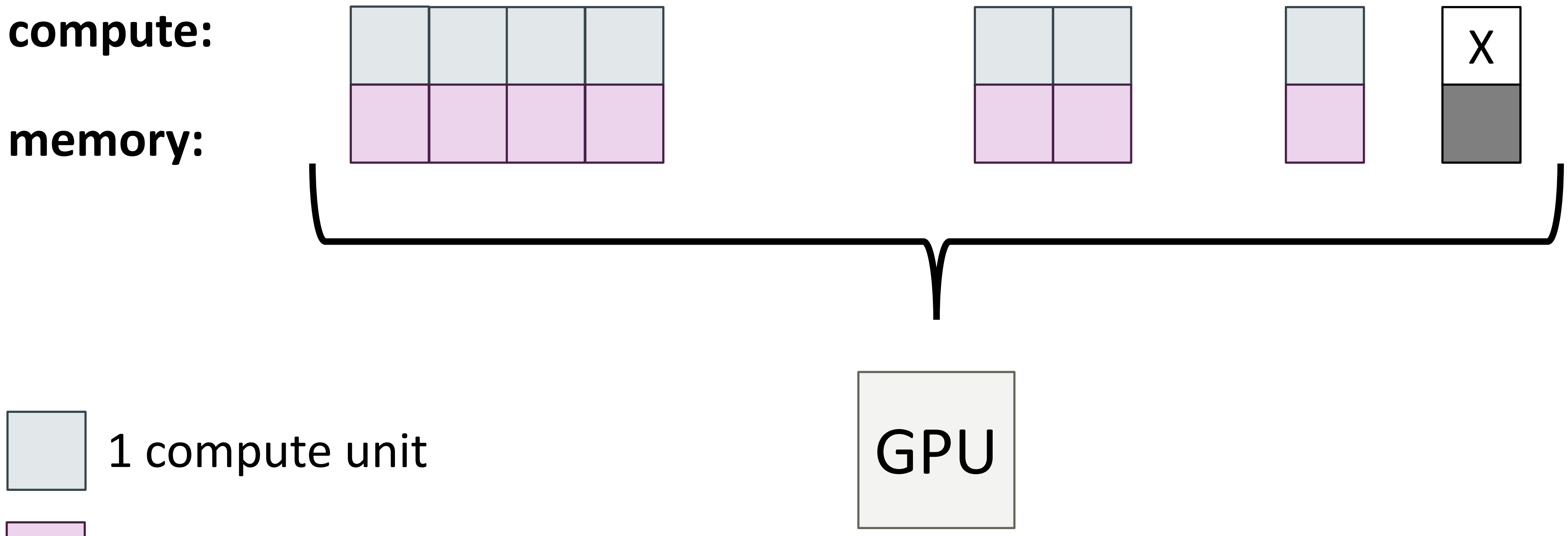




multi-instance GPU



multi-instance GPU



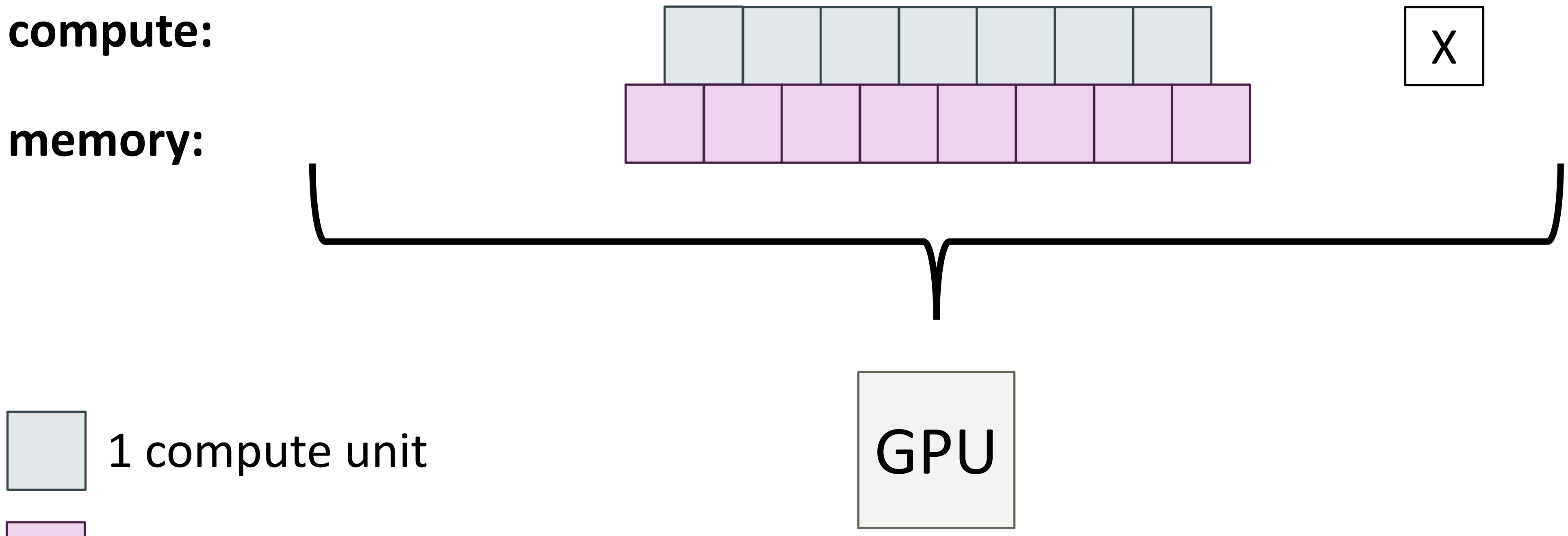
X 1 compute unit

1 memory unit

unused available (memory/compute) unit

X unavailable compute unit

multi-instance GPU



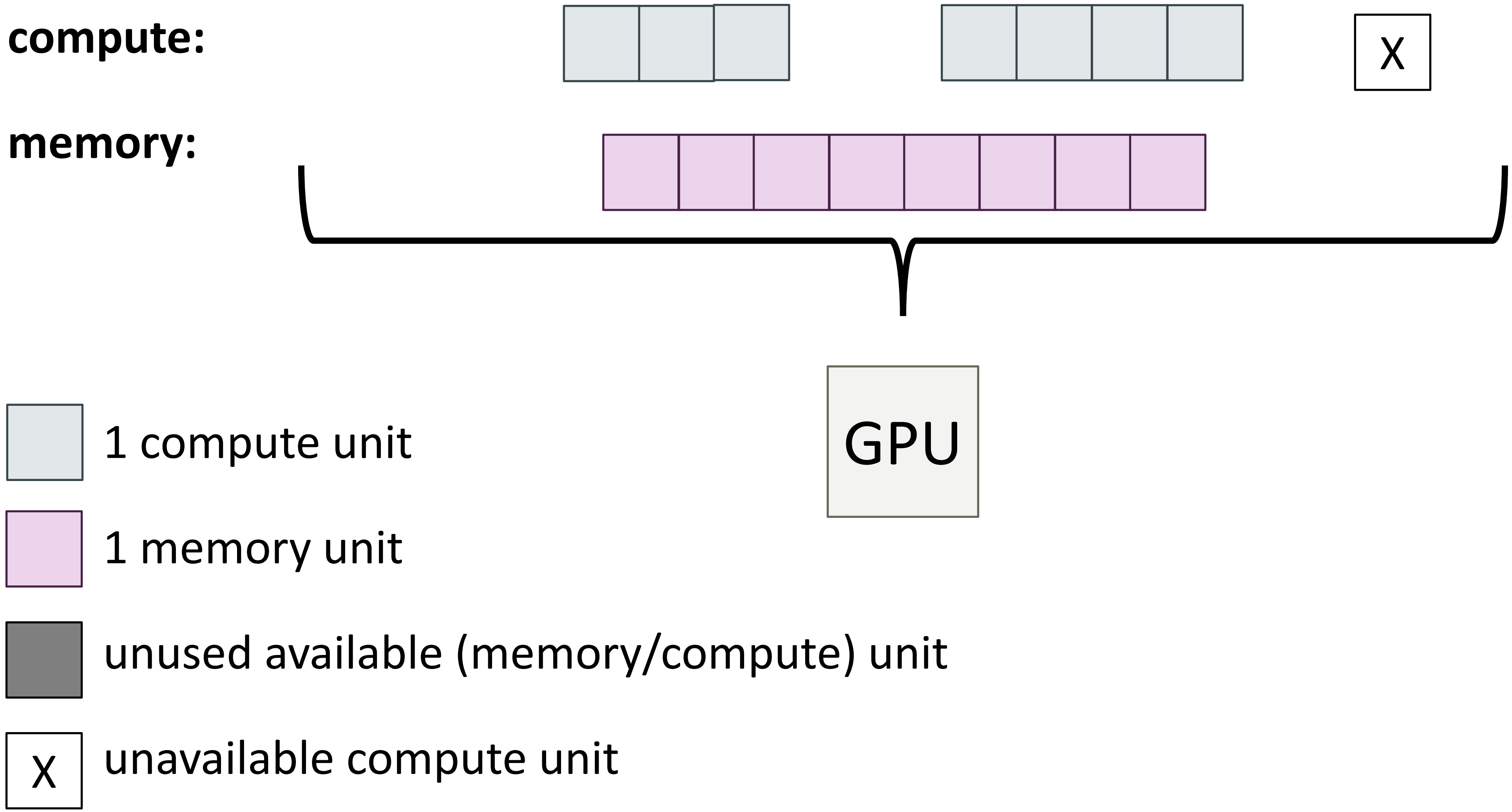
1 compute unit

1 memory unit

unused available (memory/compute) unit

X unavailable compute unit

multi-instance GPU



multi-instance GPU

compute: #1 #2 #3 #4 #5 #6 #7 X

memory: #1 #2 #3 #4 #5 #6 #7 #8

on A100 with 40GB RAM

GPU

- 1 compute unit = 1g = 14 SMs
- 1 memory unit = 5GB
- unused available (memory/compute) unit
- X unavailable compute unit = 10 SMs

SM = streaming multiprocessor

- **available instance profiles differ across GPUs that support MIG**
- **doesn't allow distributed training on A100s**

performance impact of collocation?

NVIDIA DGX Station A100

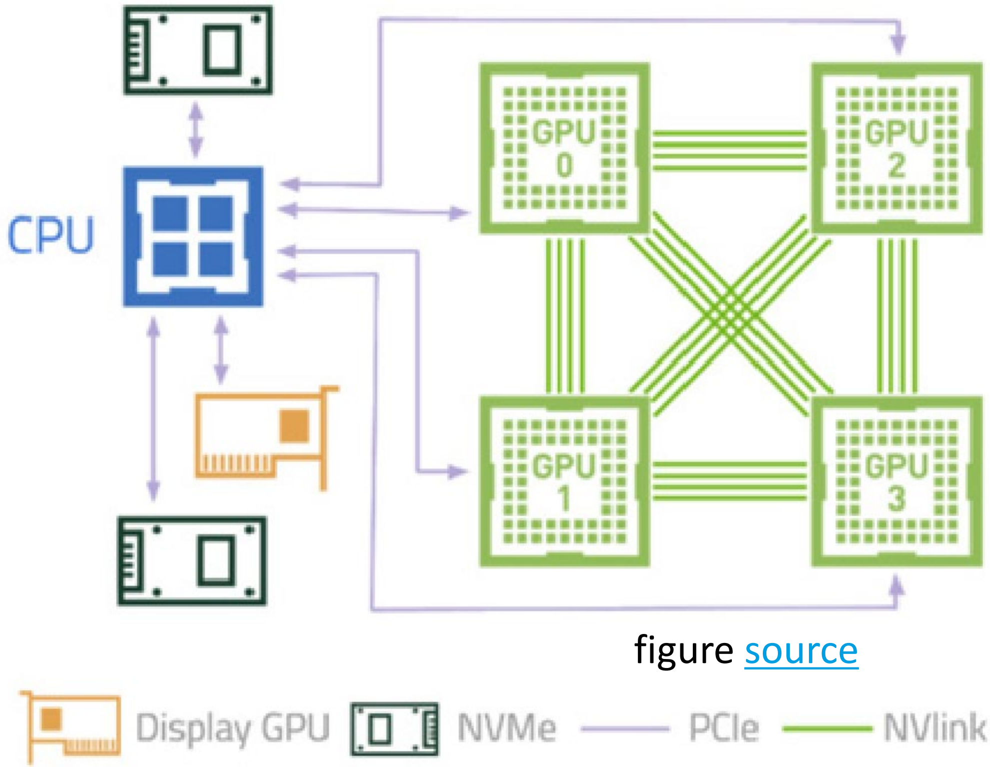


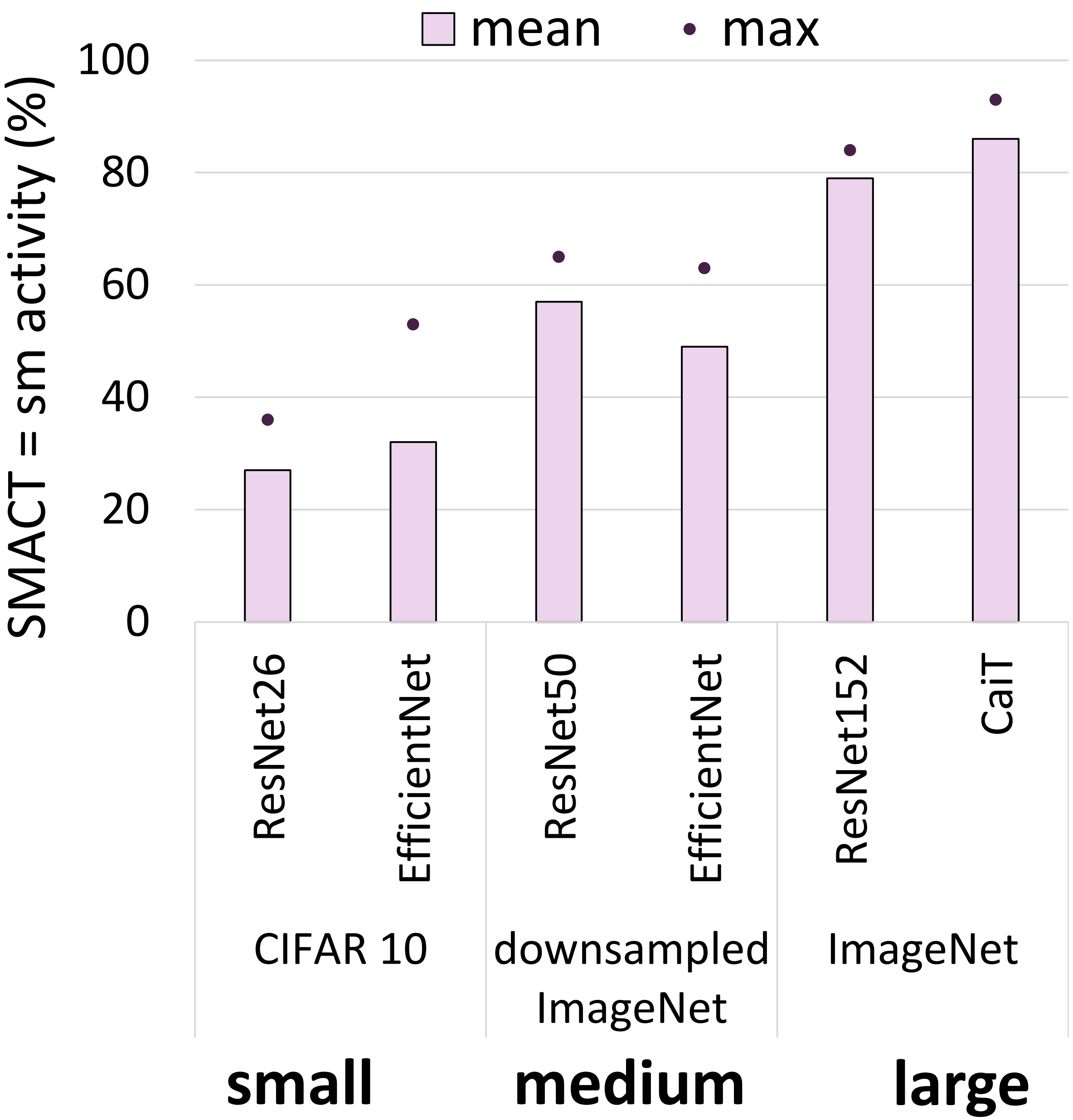
figure source

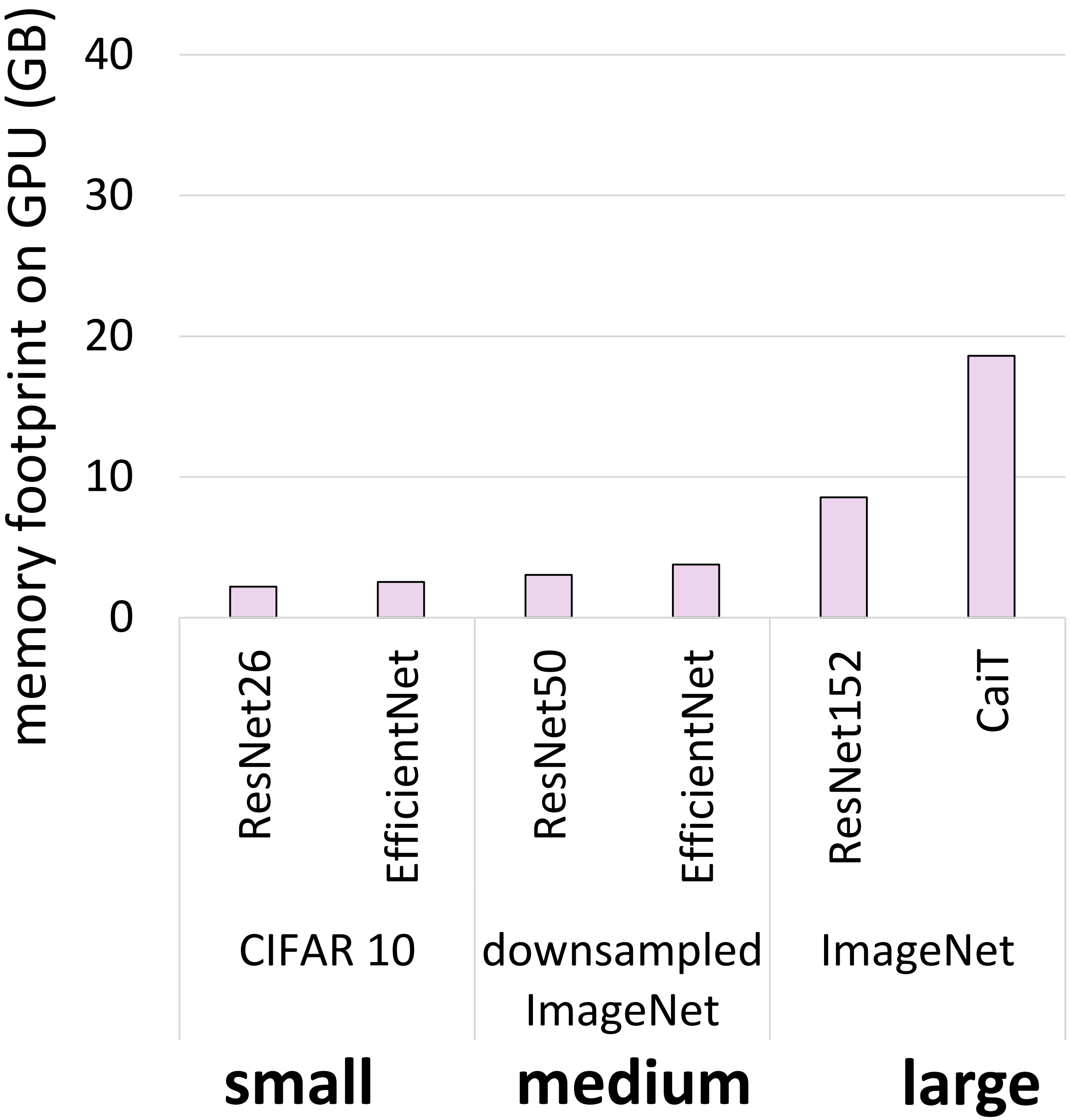
CPU = AMD 7742 – 512 GB RAM
64 physical cores
GPU = NVIDIA A100 – 40 GB RAM

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

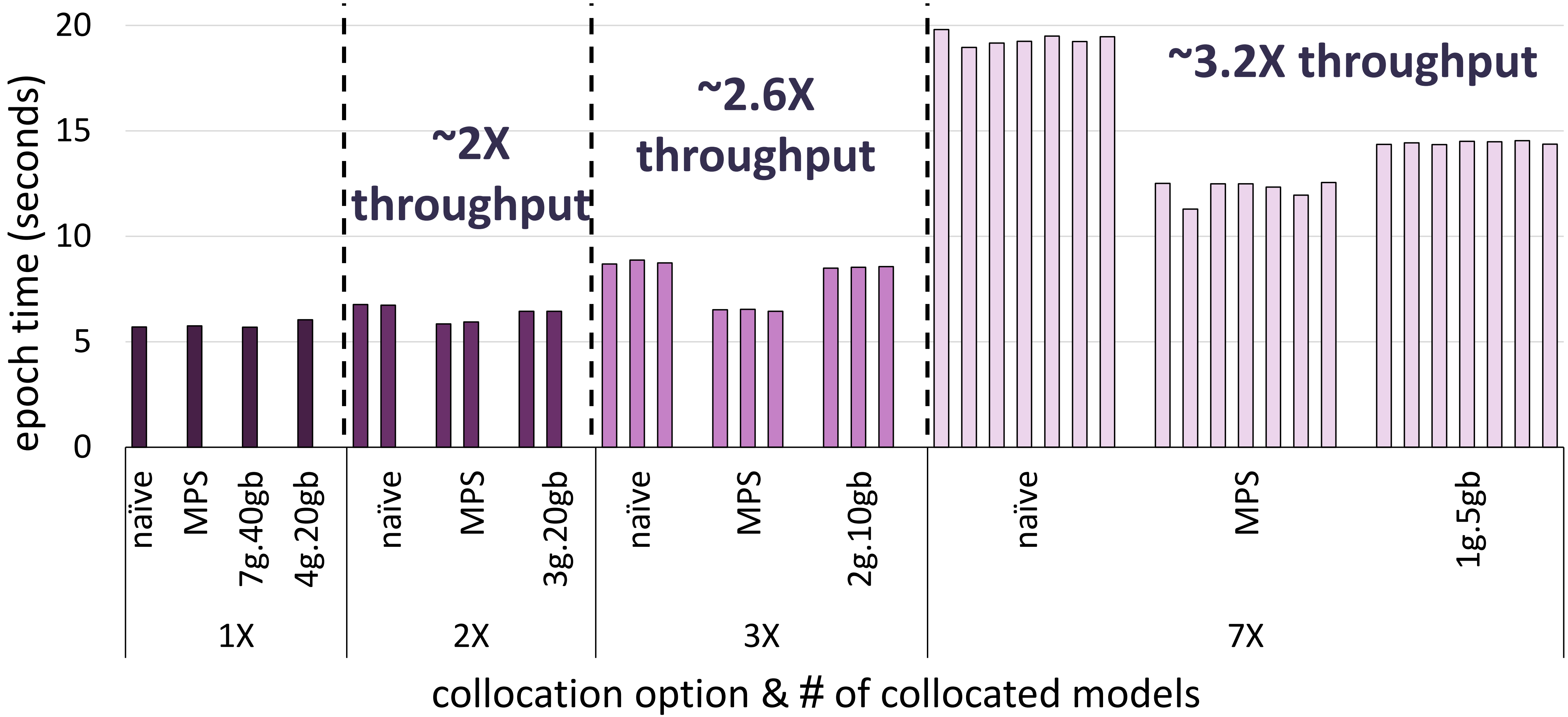
- image models: CNN & transformers + recommender model
- on single GPU with PyTorch v2.0
- results reported from 2nd epoch of training
- nvidia-smi & dcgm as monitoring tools

hardware utilization without collocation





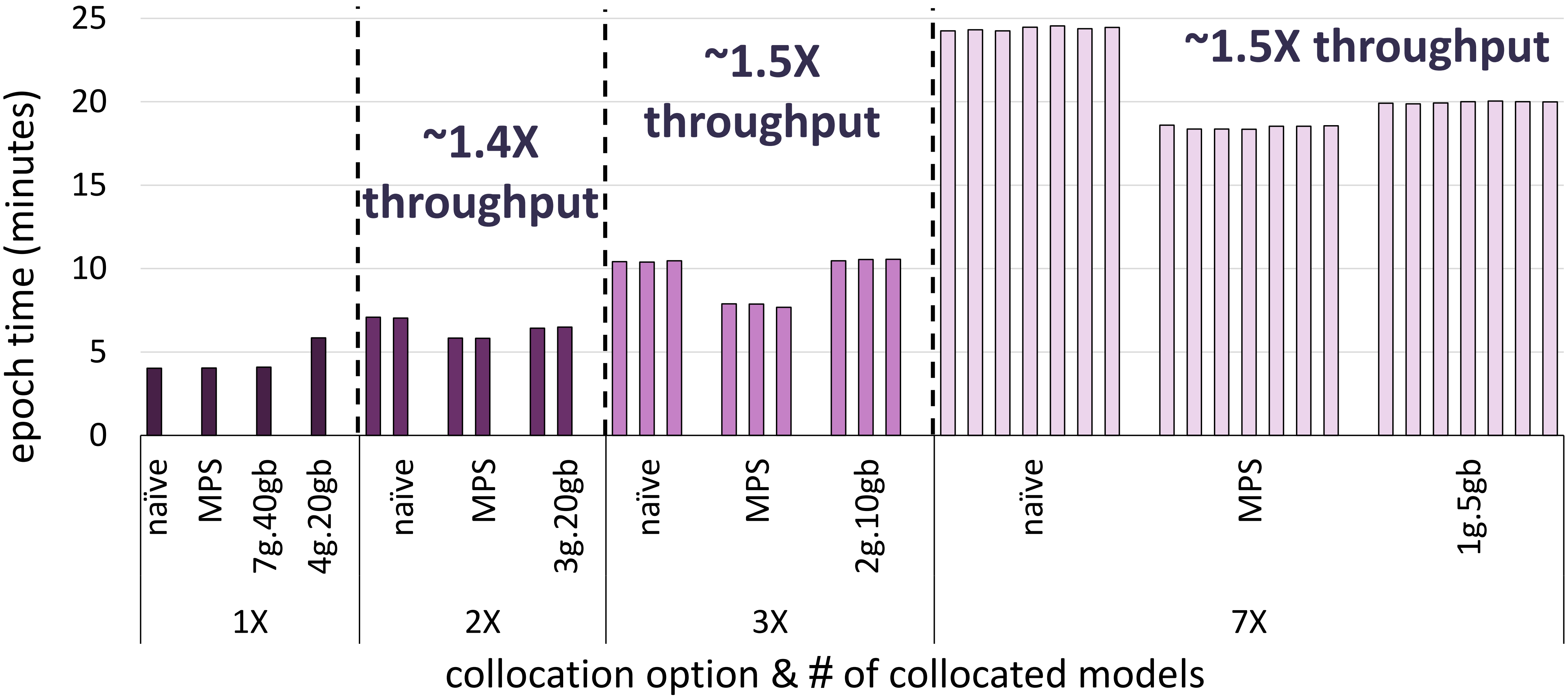
time per epoch – *small case – ResNet26*



collocation benefits despite increased epoch time

MPS > MIG > naïve

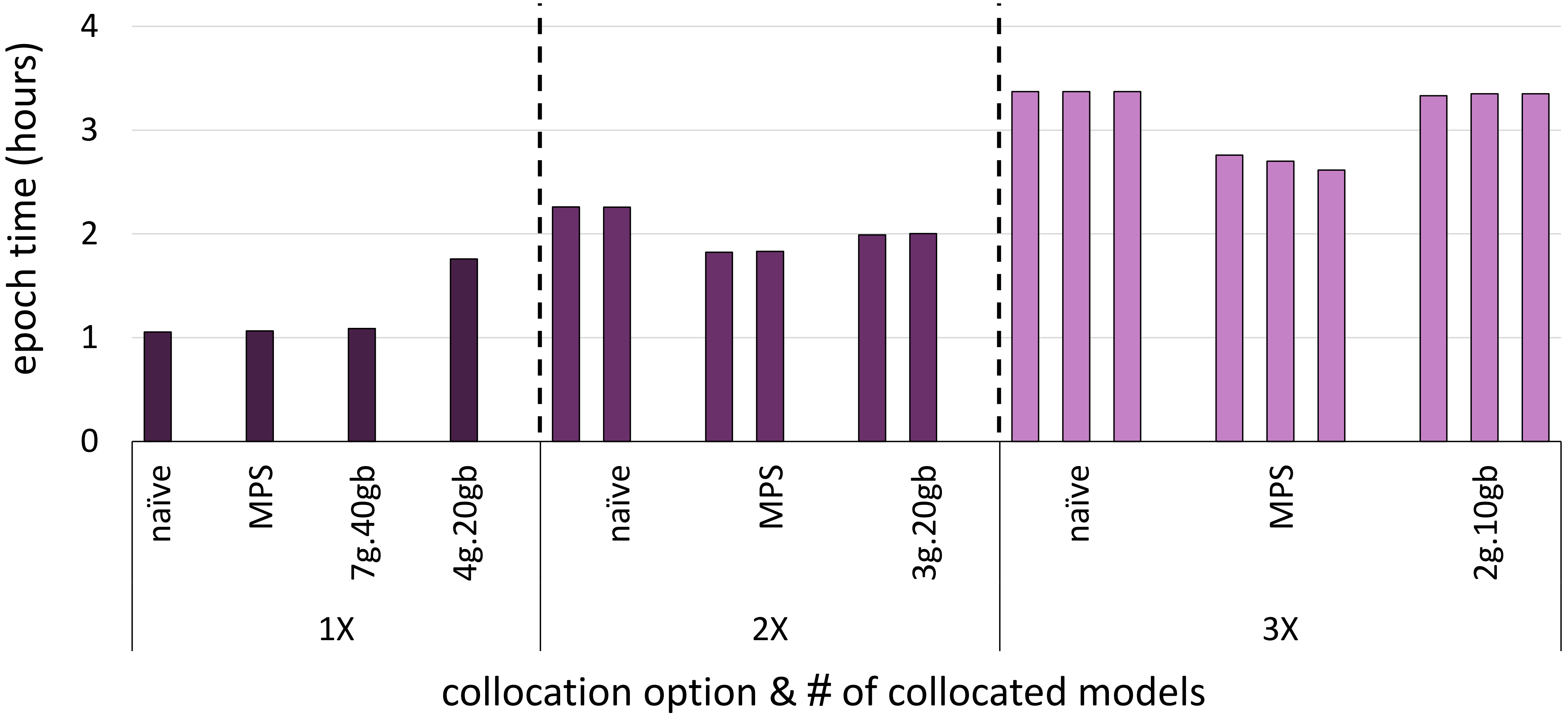
time per epoch – *medium case – ResNet50*



still some throughput benefits

but diminishing returns for increased collocation

time per epoch – *large case – ResNet152*



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

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

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
naïve	6.09 h (+14%)	1.11 h (+5%)	81%	37.75 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
naïve	6.09 h (+14%)	1.11 h (+5%)	81%	37.75 GB
MPS	5.57 h (+5%)	1.10 h (+4%)	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
naïve	6.09 h (+14%)	1.11 h (+5%)	81%	37.75 GB
MPS	5.57 h (+5%)	1.10 h (+4%)	81%	37.62 GB
MIG: 3compute – DLRM 4compute – ResNet shared memory	5.60 h (+5%)	1.40 h (+33%)	39%	37.86 GB

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

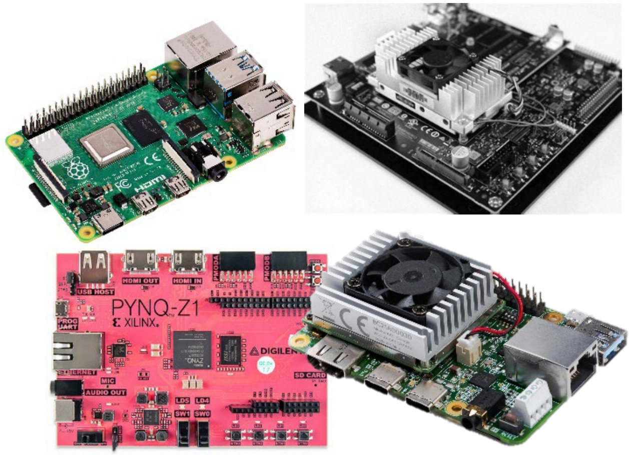
collocation for deep learning

- 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
- MPS performs better thanks to its flexibility
  - wasn’t the case pre-PyTorch v2.0 (with CUDA 11.7)
- MIG is the only option if more strict separation is needed
  - if the workload resource needs known ahead of time, can be configured to achieve performance close to MPS

need to build schedulers that incorporate GPU collocation!

hardware scales for deep learning

large **tiny**

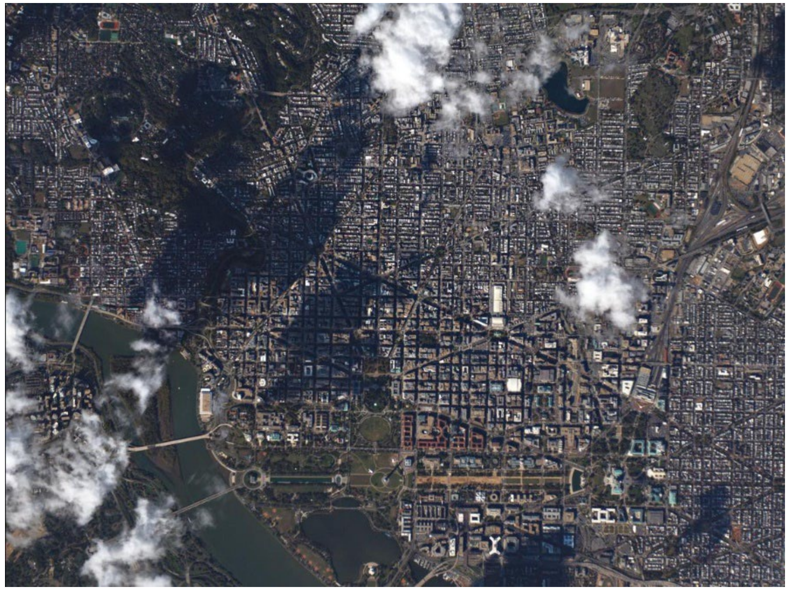


"An Analysis of Collocation on GPUs for Deep Learning Training", EuroMLSys 2024

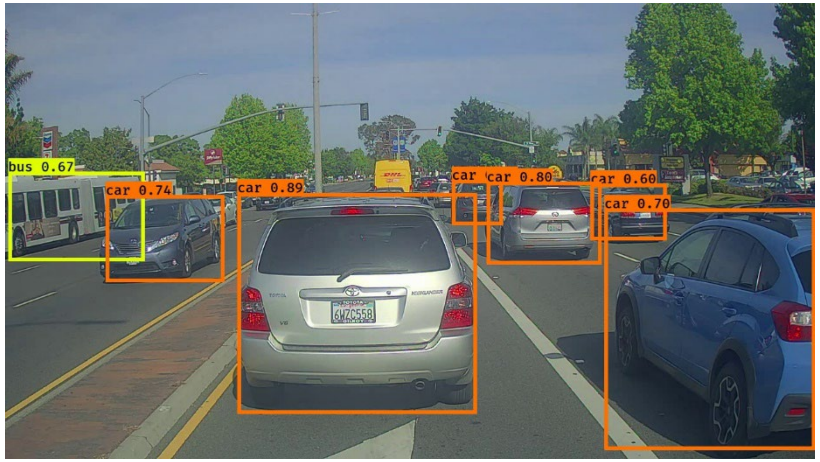
"Reaching the Edge of the Edge: Image Analysis in Space", DEEM 2024

→can we utilize these hardware well?
→can we do more with less?

machine learning @ the edge



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

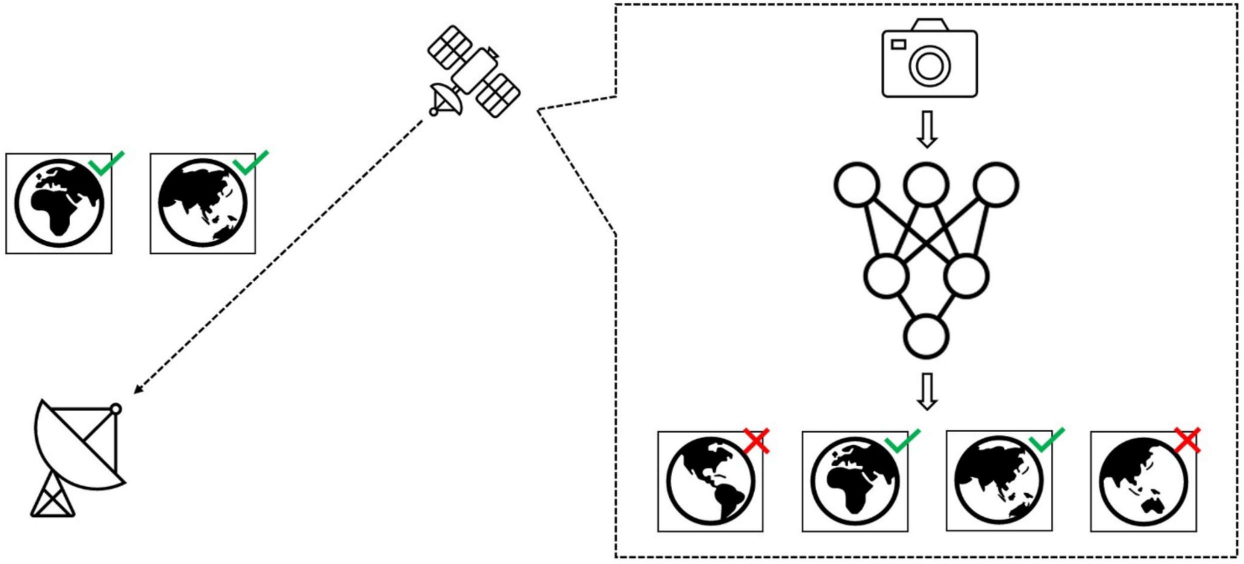


need for efficient & complex data processing closer to data sources!

DISCO: Danish student CubeSat program

https://discosat.dk/

- collaboration across Danish universities
- **use-case:** build a CubeSat satellite for observation of landmasses (especially snow, ice …) in the Arctic
- **goal:** ML-based image classification to send only the relevant images to ground (minimize data movement)



our task: build the image processing unit on the satellite

➔ which edge device can satisfy the *requirements* for this task?

image processing unit requirements

real-time imaging	Arctic region	Greenland
< 4.42 s latency	< 71.74 s latency	< 270 s latency

---

max 49.1 images can be transferred per day!

min 320 images captured per day.

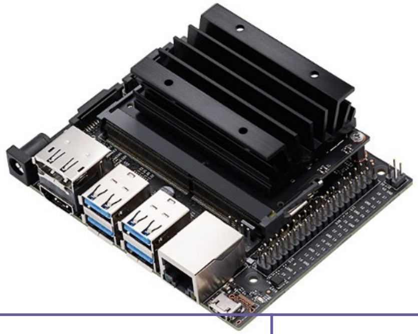
---

max 5 watts

average 2 watts

---

flexible software upload!

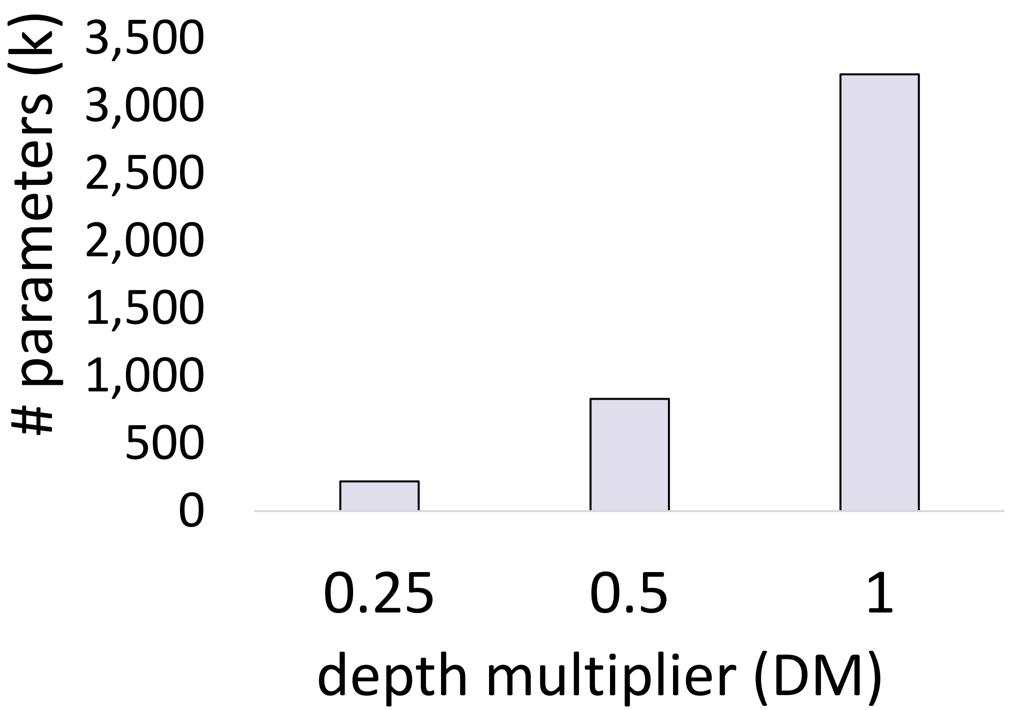


ARM Cortex-M7	Jetson Nano	Toradex Verdin	CoralAI Micro	CoralAI Mini	CoralAI USB Stick	Neural Compute Stick
ARM Cortex-M7 @300MHz	ARM A57 @1.43GHz	ARM Cortex-A53 @1.8GHz, ARM Cortex-M7 @800MHz	ARM Cortex-M7 @800MHz, ARM Cortex-M4 @400MHz	ARM Cortex-A35 @1.5GHz	Raspberry Pi 3 BCM2837 ARM @1.2GHz	
384KB SRAM, 32KB FRAM	4GB	4GB	64MB	2GB	1GB	
none	128-core Maxwell GPU	NPU (2.25 TOPS)	CoralAI Edge TPU (4 TOPS)			Intel Movidius Myriad X VPU

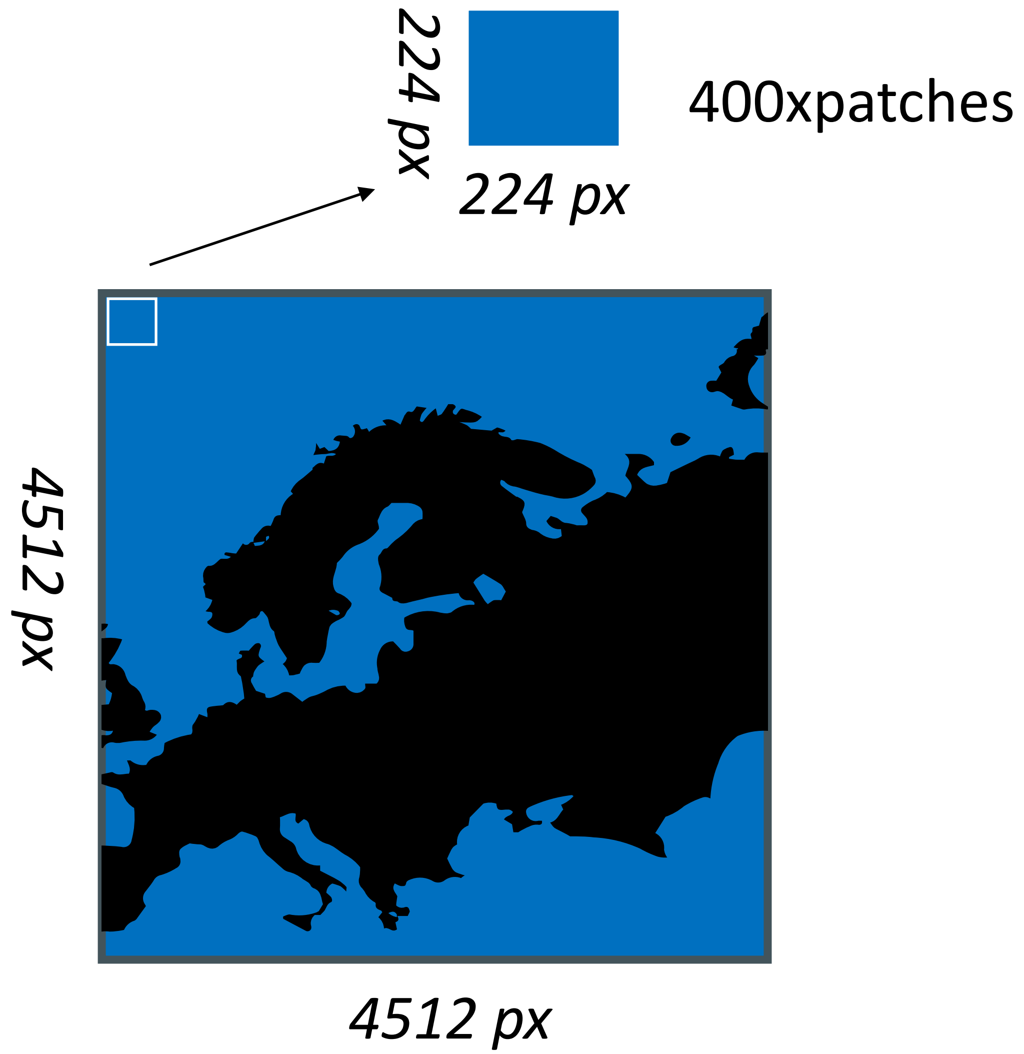
general-purpose ⟷ **higher specialization for ML**

model

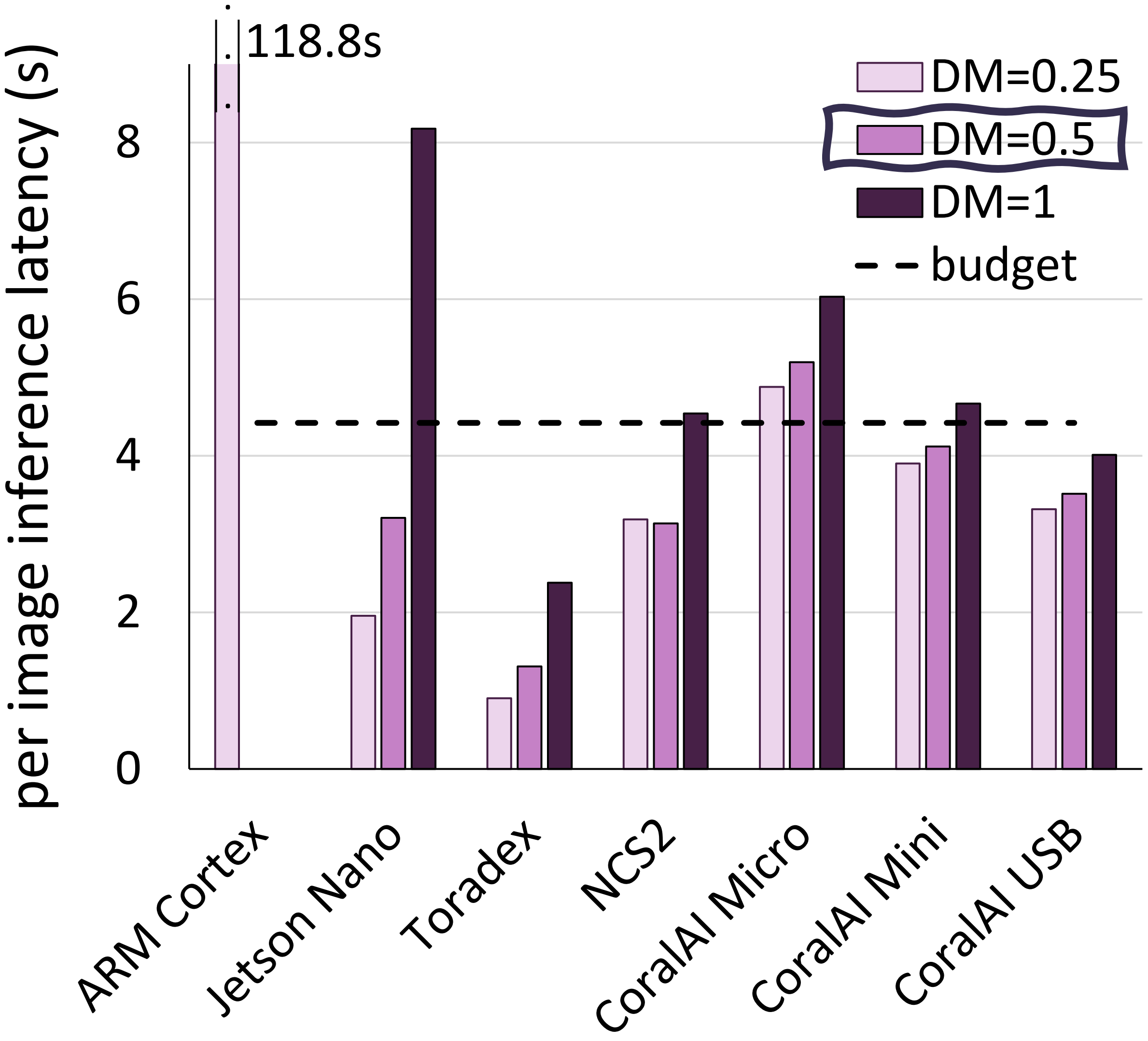
pretrained MobileNetV1

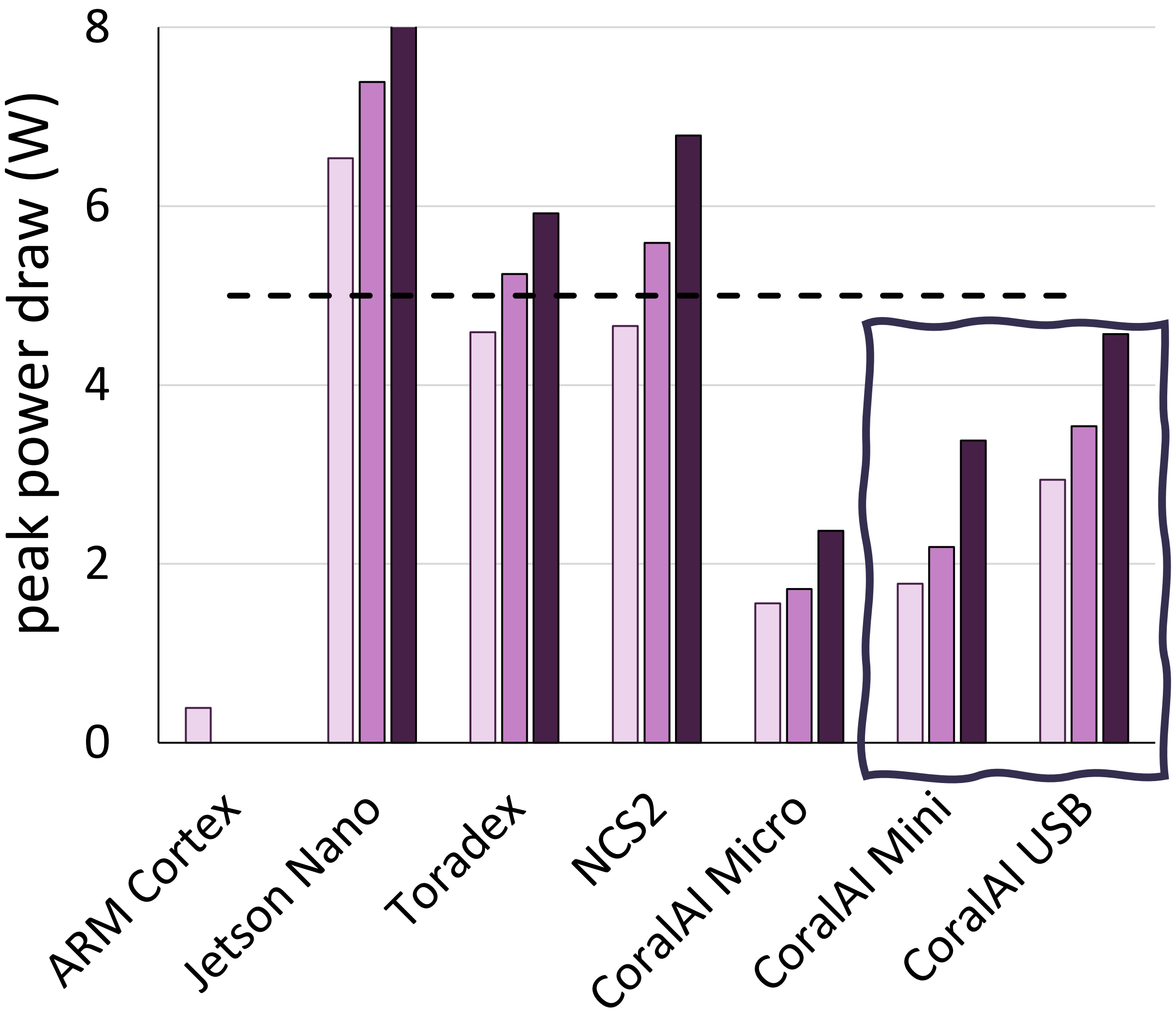


data



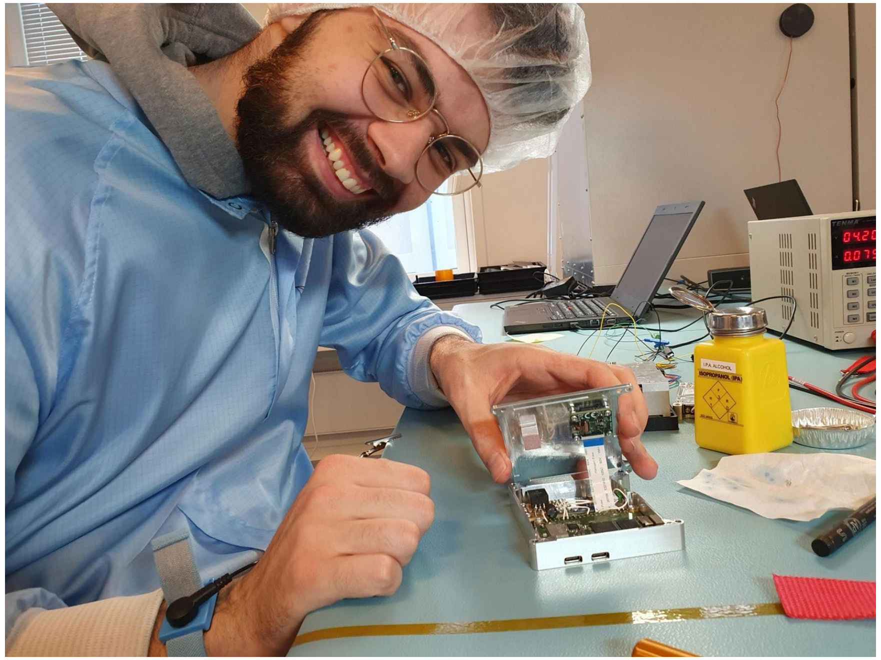
latency & power draw

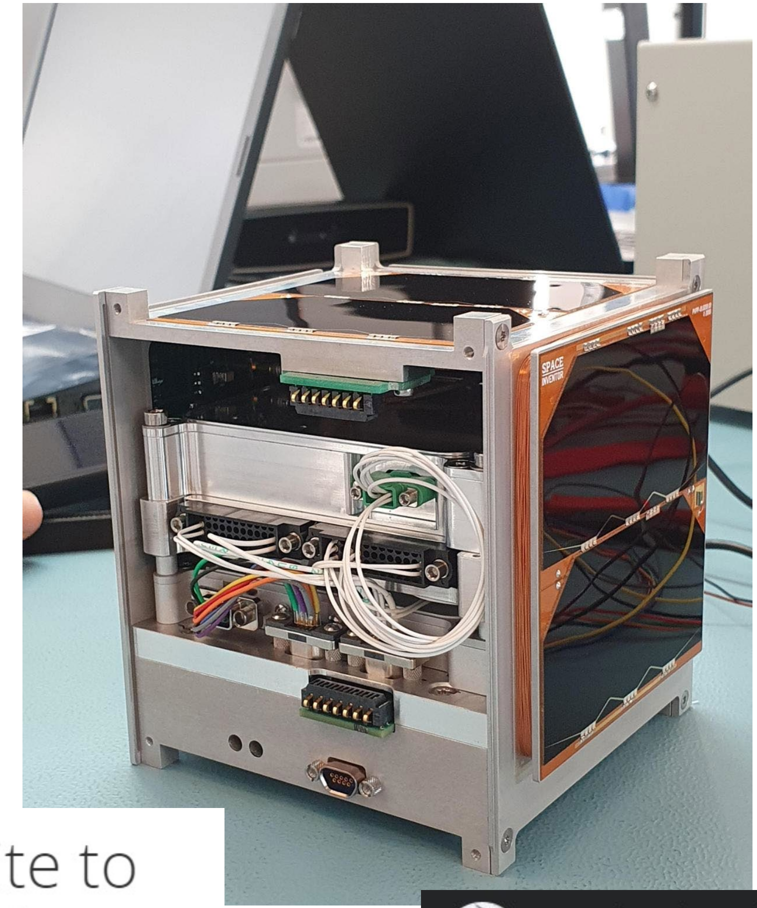




Jetson & NCS2 have low-latency for smaller models, Toradex is best.
Faster devices fail the peak power budget. Corals fit the power budget.

DISCO satellite





Students launch a satellite to test artificial intelligence in space

On April 14, students from ITU will contribute to writing space history. The satellite, DISCO-1, is launched into space and it carries a microcomputer to test artificial intelligence outside the atmosphere. The satellite is developed by the space program, DISCO, which is a collaboration between students from four Danish universities.

IT-Universitetet i København
April 15 ·

Så lykkedes det! 🚀🛰✨

Satellitten DISCO-1, udviklet af danske studerende fra bl.a. ITU, blev her til morgen sendt ud i rummet med SpaceX' raket fra Californien.

Satellitten indeholder en mikrocomputer, der skal teste kunstig intelligens i rummet. 🤗

Læs mere om projektet her 👉 https://www.itu.dk/.../Studerende-opsender-satellit-der...

📷 Julian Priest (CC BY-NC 3.0)

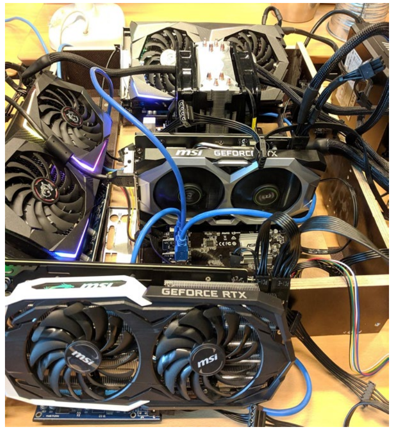
ML @ the edge

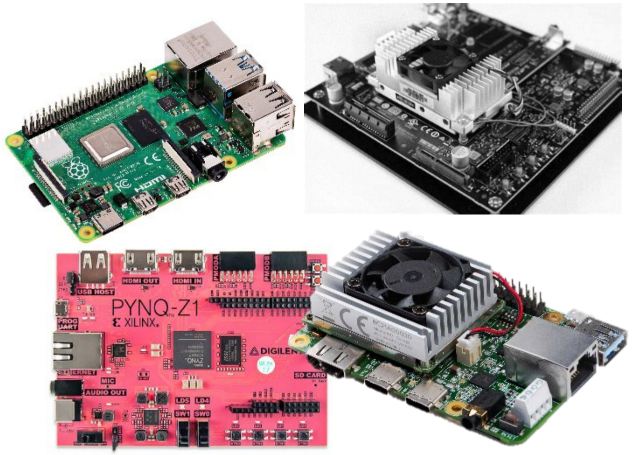
- demand for more data analysis closer to the data source
  - reduces data movement & privacy concerns
  - helps with real-time decisions
- variety of edge devices to choose from offering increasingly powerful hardware but still resource-constrained
  - requires not just latency-efficient,
    but also energy-efficient data processing
- hardware specialization helps with latency & power budget
  - though, we need more flexibility

need for methods that can deal with resource management & program updates at the edge!

hardware scales for deep learning

large ⟷ **tiny**





➔can we utilize these hardware well? ➔ not always

- need more effective workload collocation on accelerators
- energy-efficiency must be part of the utilization analysis

➔can we do more with less? ➔ yes, but it isn’t free lunch

- need to understand better the capabilities of different devices
- every scale requires its own dynamic resource managers

teamRAD - resource-aware data systems

rad.itu.dk

Ties
Robroek

Ehsan
Yousefzadeh-Asl-Miandoab

Robert
Bayer

DASYA

Data-Intensive Systems and Applications

www.dasya.dk

@dasyaITU













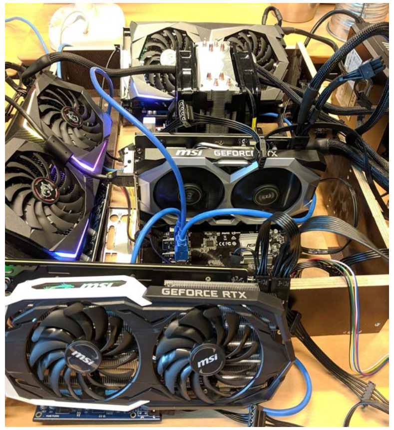


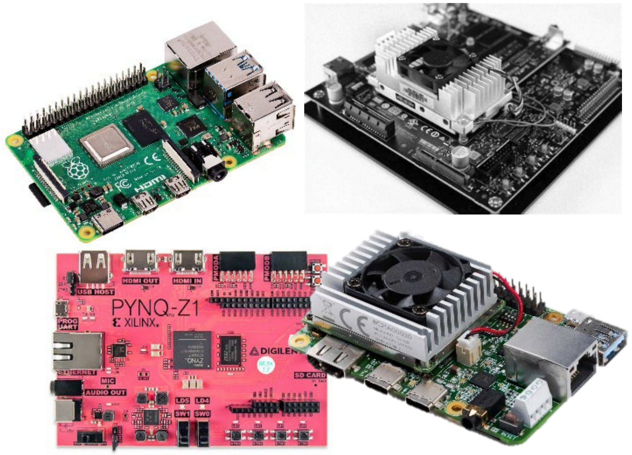



hardware scales for deep learning

thank you!

large ⟷ **tiny**





➔can we utilize these hardware well? ➔ not always

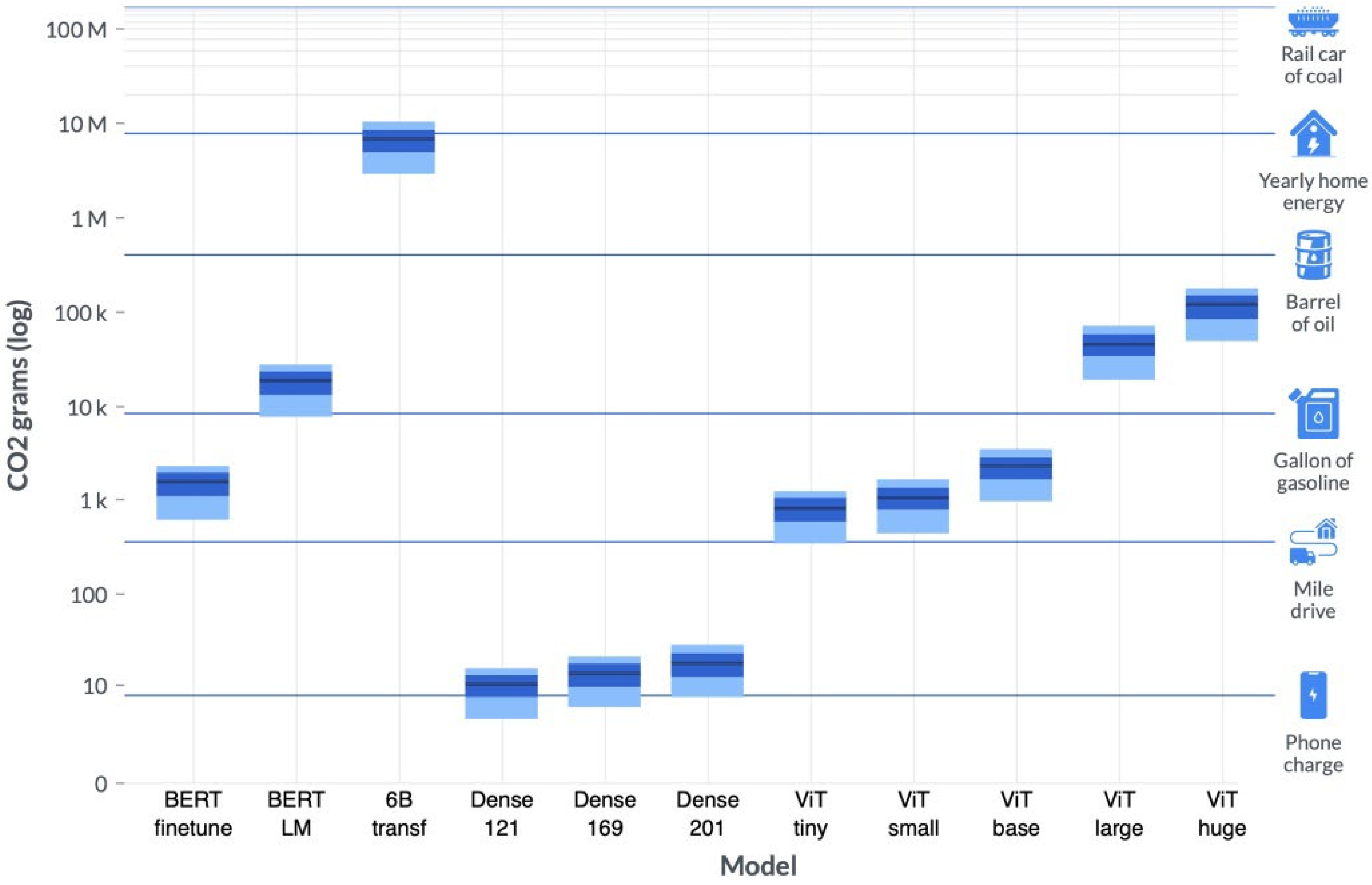
- need more effective workload collocation on accelerators
- energy-efficiency must be part of the utilization analysis

➔can we do more with less? ➔ yes, but it isn't free lunch

- need to understand better the capabilities of different devices
- every scale requires its own dynamic resource managers

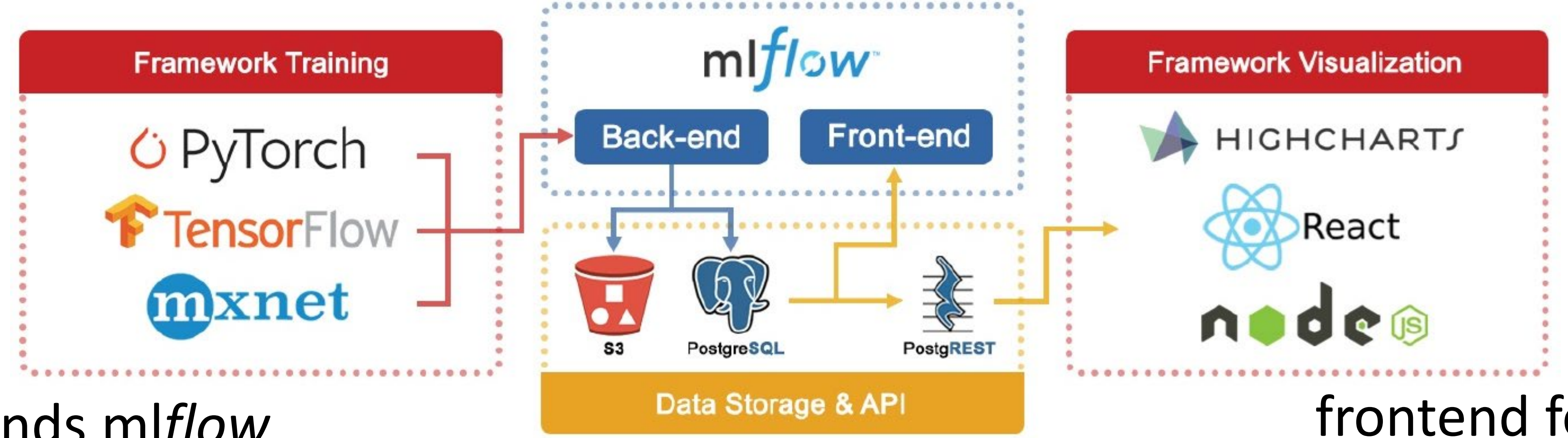
backup

unsustainable growth of deep learning



Dodge et al. "Measuring the Carbon Intensity of AI in Cloud Instances." FAccT 2022

radT



- extends ml*flow*
- incorporates collocation
- allows easy, extensible, and scalable tracking of hardware metrics on CPUs & GPUs
  - listeners for monitoring (dcgm, nvidia-smi, top) & profiling (nsys, ncu, pytorch profiler) tools

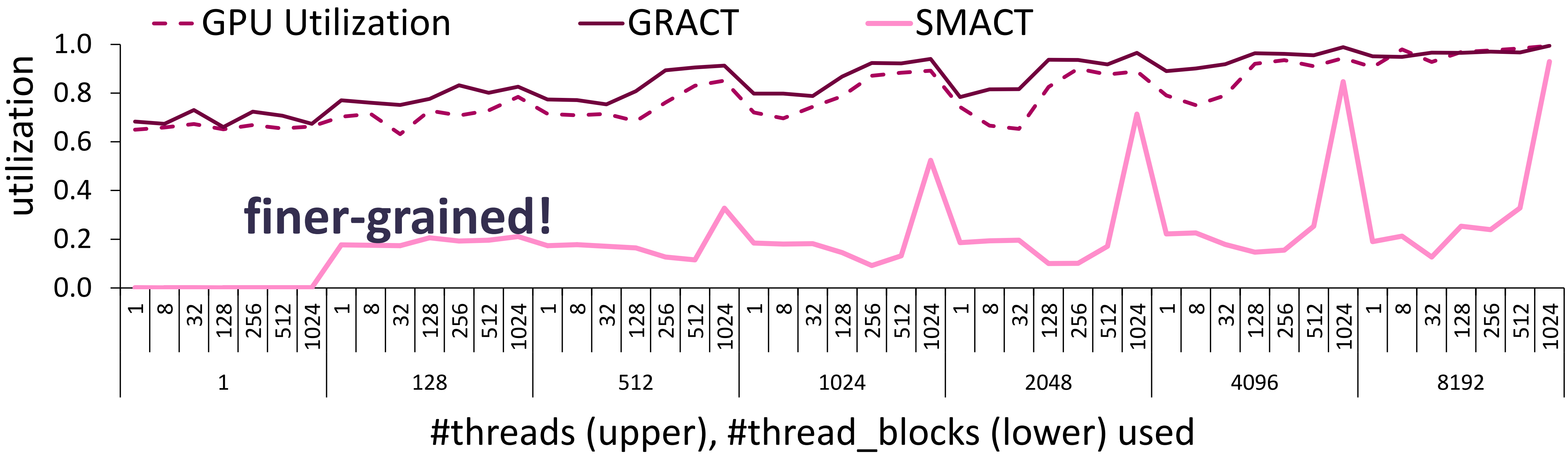
frontend for data exploration

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

Robroek et al. "Data Management and Visualization for Benchmarking Deep Learning Training Systems", DEEM 2023
https://github.com/Resource-Aware-Data-systems-RAD/radt & https://www.youtube.com/watch?v=oaGfzYjKJ1Q

GPU utilization

- **GPU utilization**: % of time one or more kernels were executing on the GPU
- **GRACT**: % of time any portion of the graphics or compute engines were active
- **SMACT:** the fraction of active time on an SM, averaged over all SMs = streaming multiprocessor



coarse-grained GPU utilization metrics could be misleading!

Yousefzadeh-Asl-Miandoab et al. "[Profiling and Monitoring Deep Learning Training Tasks]", EuroMLSys 2023

setup

	ARM Cortex-M7	Jetson Nano	CoralAI *	Neural Compute Stick 2
framework	TensorFlow Lite for Microcontrollers	TensorRT	TensorFlow Lite	OpenVino
quantization	8bit (to fit the device memory)	16bit	8bit (only supports 8bit ints)	16bit (only supports 16bit floats)
batching	not enough memory to do batching	batch size per inference = 16	doesn't support batching	number of concurrent inference requests = 4

accuracy

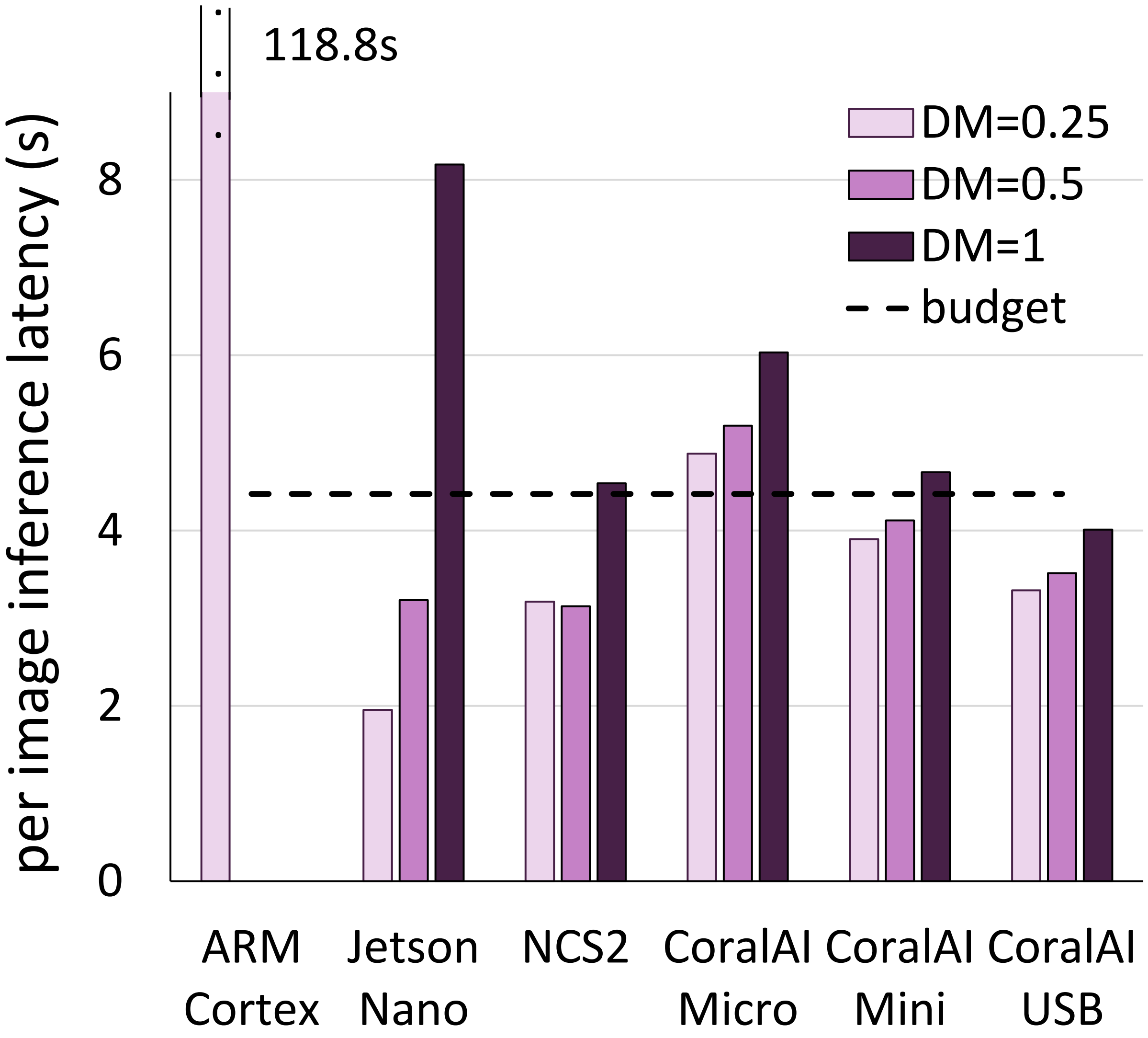
on Flowers dataset, with post-training quantization

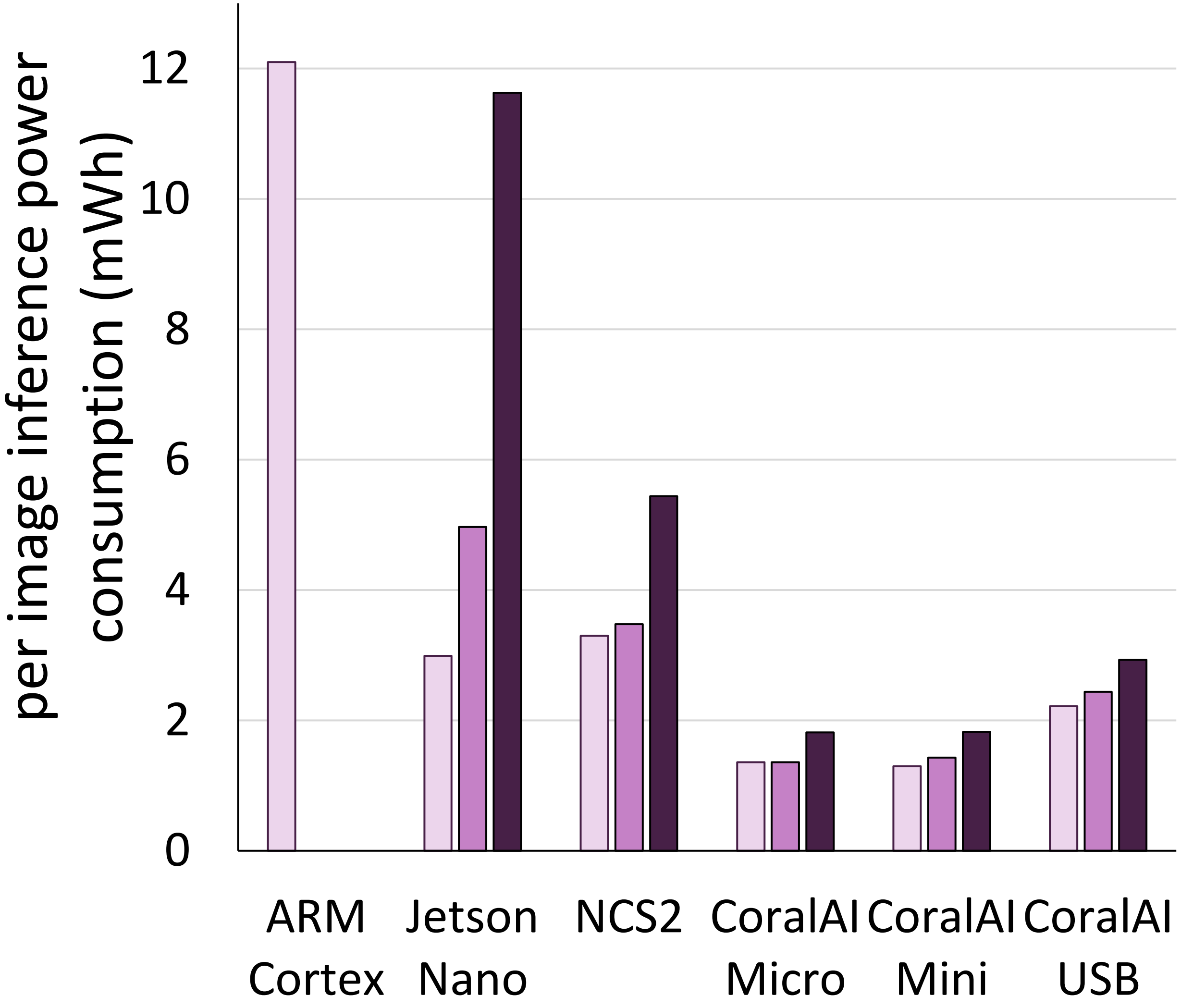
		MobileNet DM = depth multiplier		
		0.25	**0.5**	**1**
accuracy	**32bit float**	86.92%	90.33%	90.74%
	16bit float	86.78%	90.33%	90.74%
	8bit integer	84.33%	89.78%	91.55%
#params		219,829	832,101	3,233,989

accuracy trade-off becomes noticeable

too big & complex for most resource-constrained devices

latency & power draw





ARM-based microcontroller draws little power per unit time but per inference power need is higher than the rest!