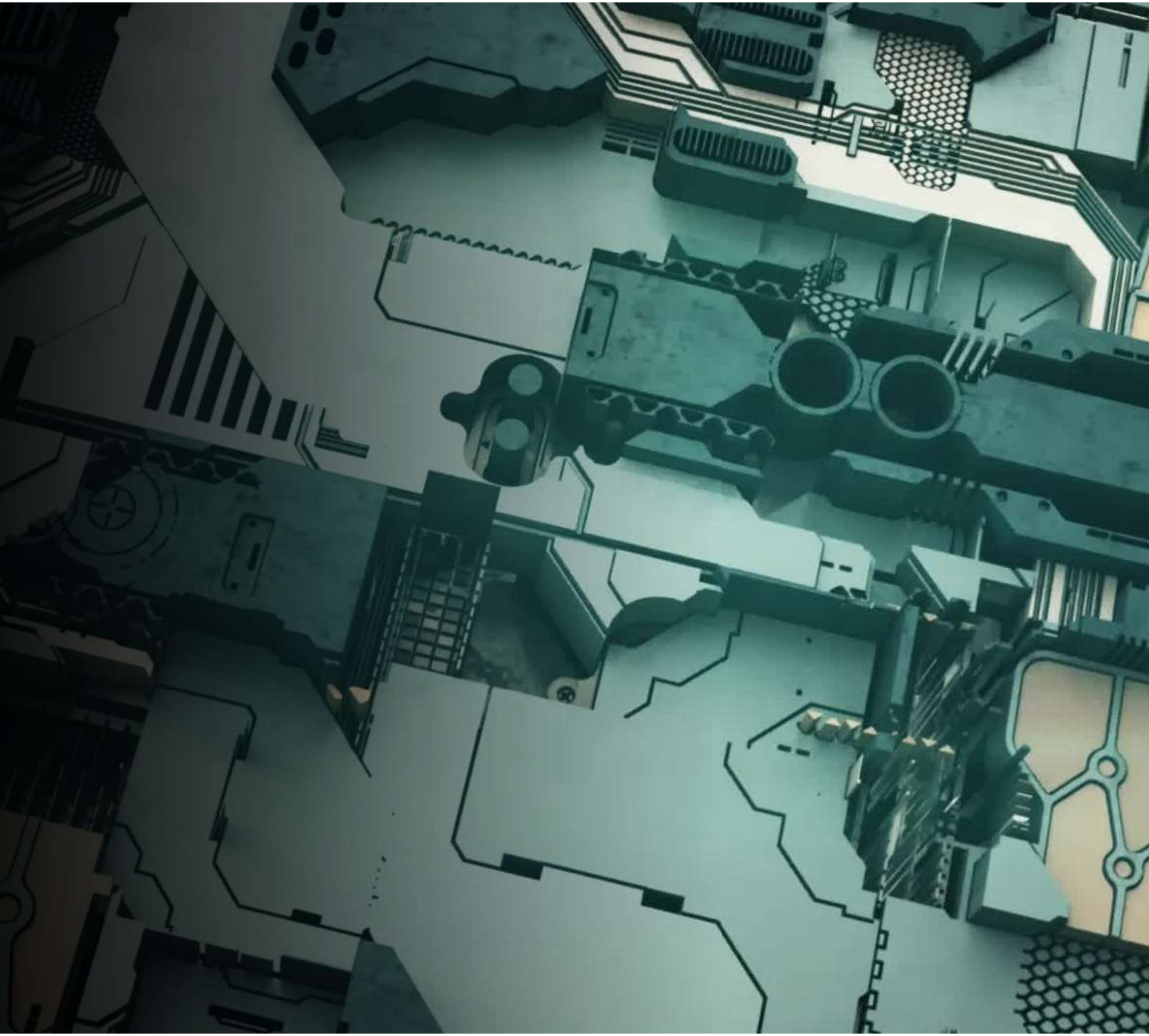
Overprovisioning GPUs in the age of AI

Ehsan Yousefzadeh-Asl-Miandoab

Advanced Data Systems

November 2022

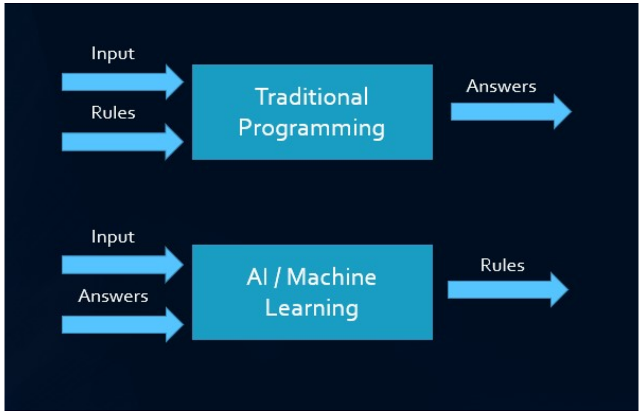


Content

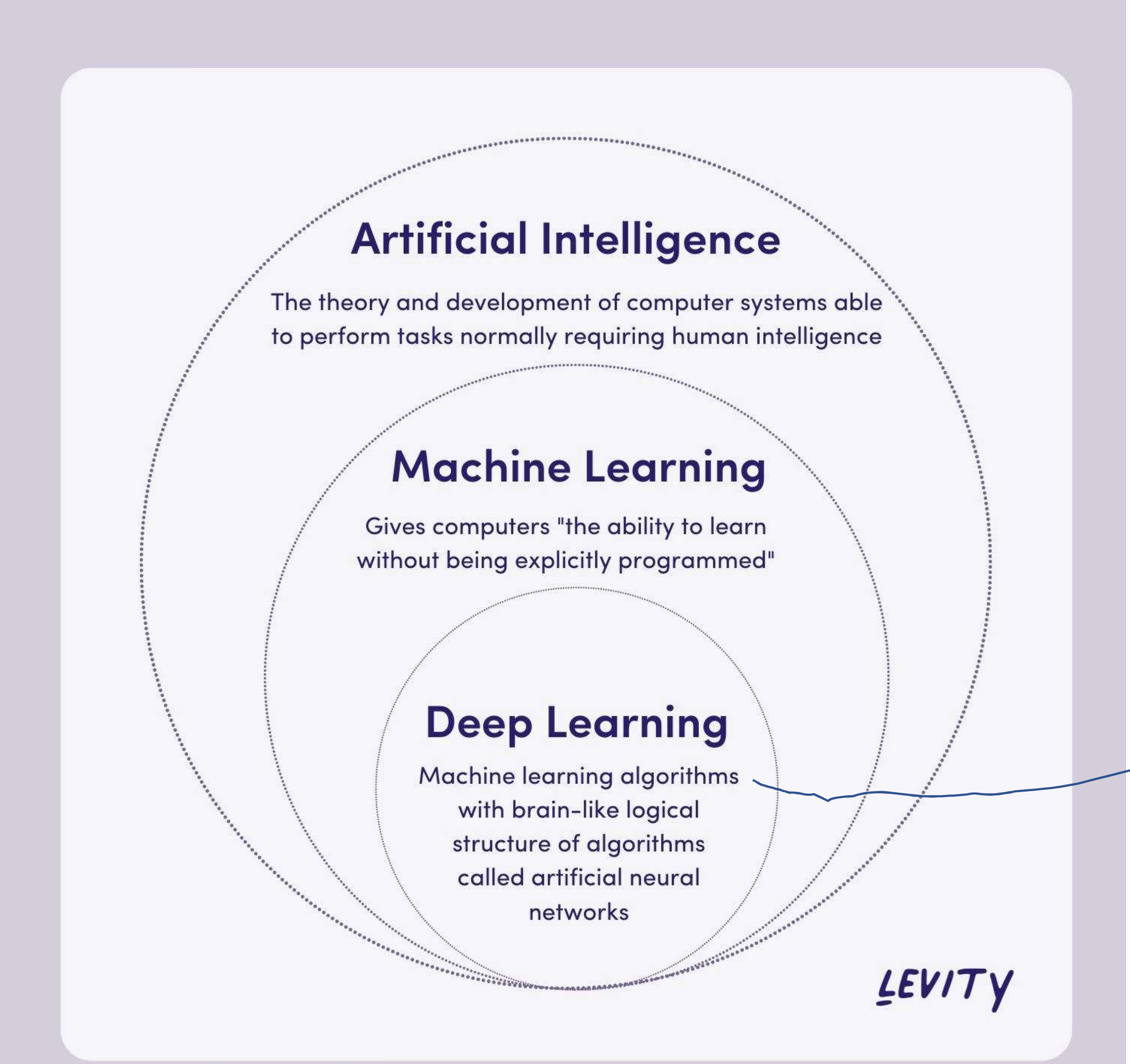
- Machine Learning and Deep Learning
- GPUs as the Primary Deep Learning Accelerators
- GPUs’ Underutilization Challenge
- Provisioning GPUs
- Naive, MPS, MIG Provisioning options
- Experiments, Evaluations, and Results

Machine Learning and Deep Learning

- Machine Learning is a new programming paradigm.
- Instead of specifying deterministic rules in programs, the program finds patterns in data.
- Data is the primary element in Machine Learning.



https://dockship.io/articles/60a9e411bc77f4429e9ddf0d/introduction-with-machine-learning



Artificial Intelligence

The theory and development of computer systems able to perform tasks normally requiring human intelligence

Machine Learning

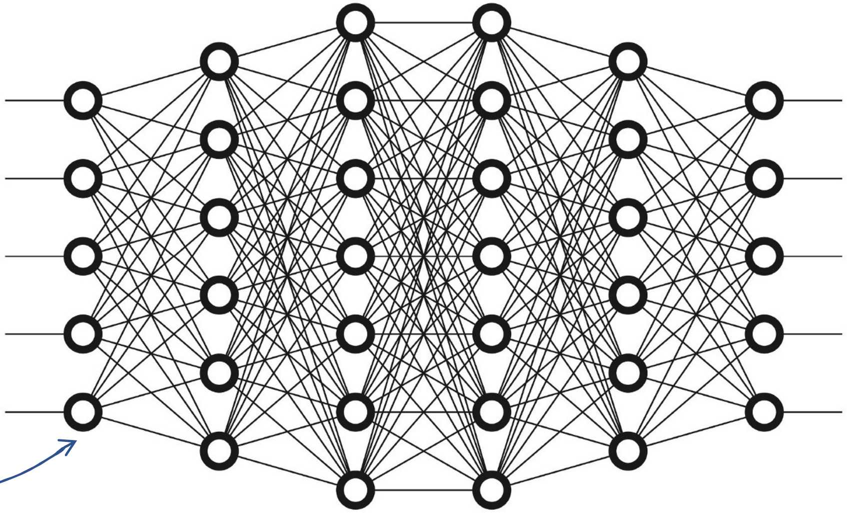
Gives computers "the ability to learn without being explicitly programmed"

Deep Learning

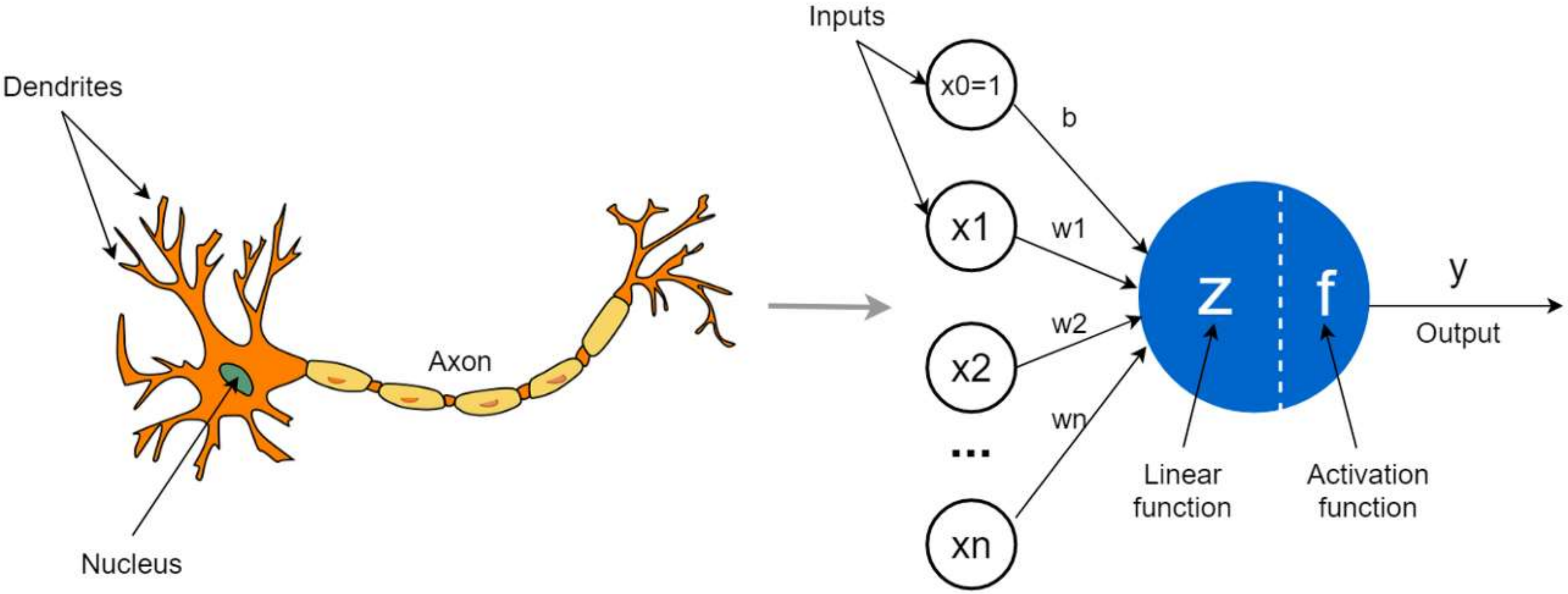
Machine learning algorithms with brain-like logical structure of algorithms called artificial neural networks

https://levity.ai/blog/difference-machine-learning-deep-learning

Training is accomplished by changing Weights, biases



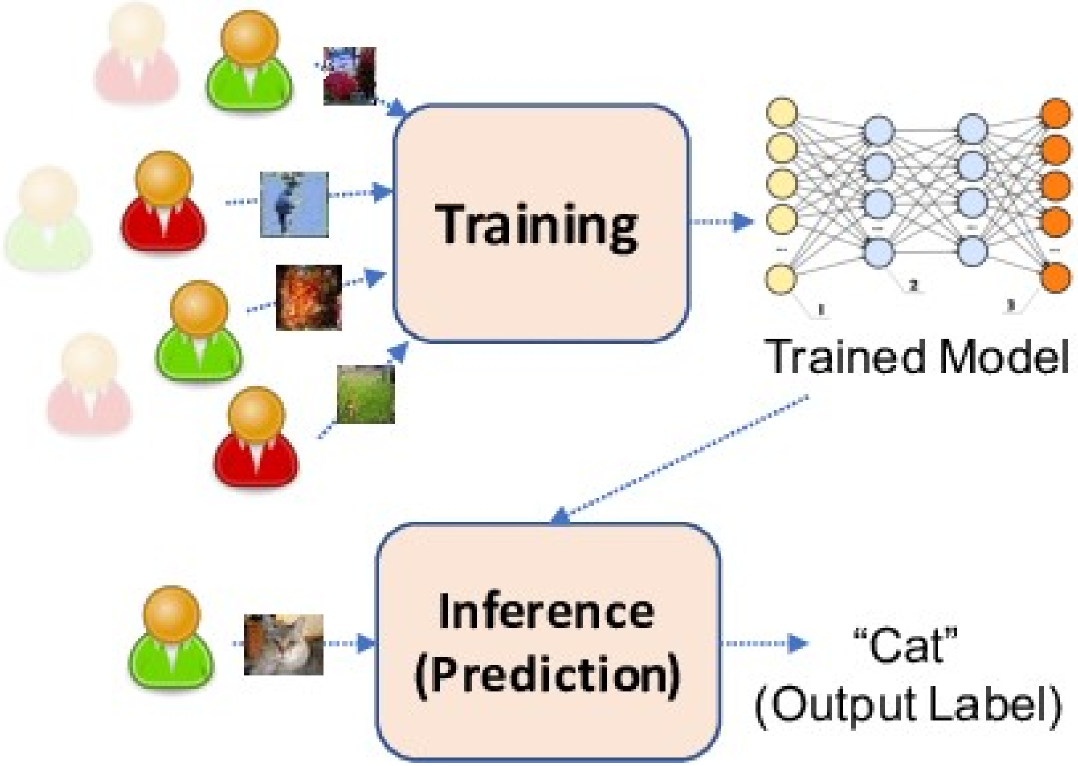
Fully Connected (FC) Neural Network



$$y = f\left(z = \left(\sum_{i=1}^{n} x_i \times w_i\right) + b\right)$$

https://towardsdatascience.com/the-concept-of-artificial-neurons-perceptrons-in-neural-networks-fab22249cbfc

Training and Inference



https://www.researchgate.net/figure/Overview-of-training-and-inference-in-deep-learning_fig1_330842645

Quiz! (True or False?)

1. Machine learning is exactly telling computers what to do.

2. Complex Deep Learning is helpful when we have a large amount of data.

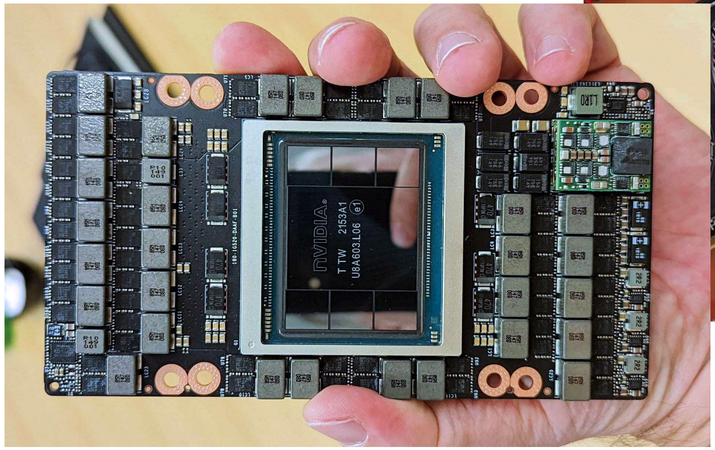
3. AI is a subsection of Deep Learning.

4. Training is computationally lighter than inference.

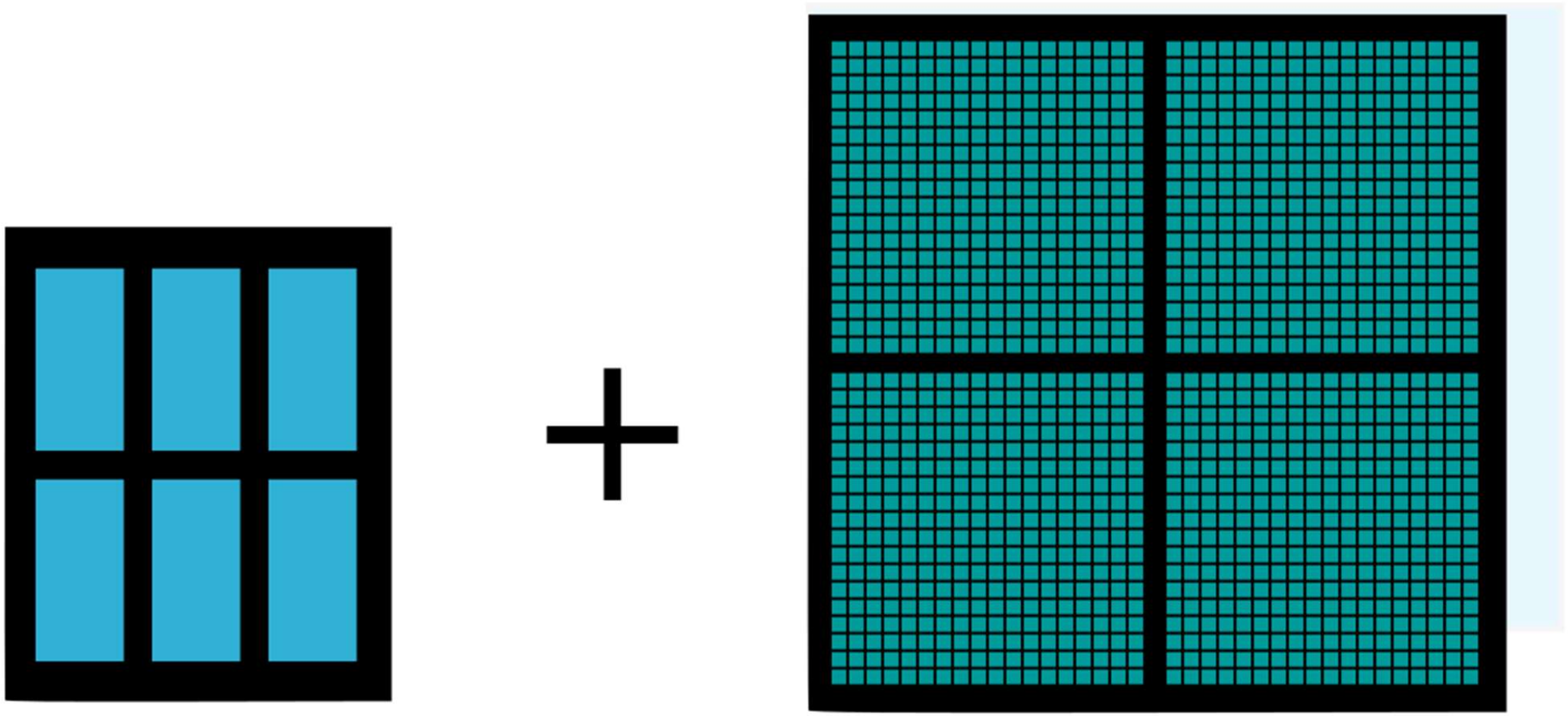
GPUs, Primary Accelerators

https://www.pny.com/nvidia-a100

https://www.pcworld.com/article/416006/the-best-graphics-cards-for-pc-gaming.html



https://www.tomshardware.com/news/nvidia-hopper-h100-sxm5-pictured



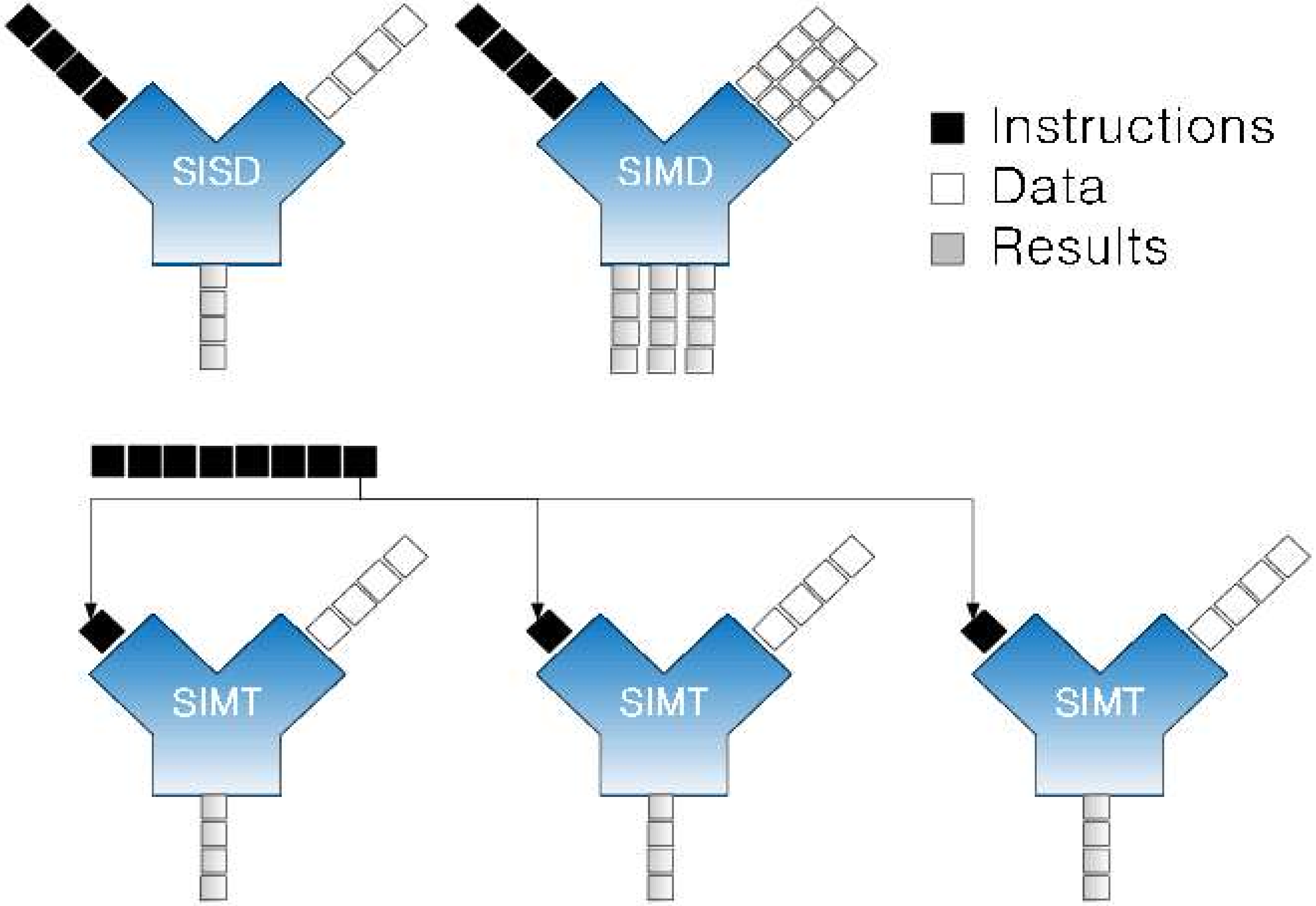
CPU
Multiple Cores

GPU
Thousands of Cores

https://www.cherryservers.com/blog/gpu-vs-cpu-what-are-the-key-differences

SISD: Single Instruction Single Data

SIMD: Single Instruction Multiple Data



SIMT: Single Instruction Multiple Thread

Kyung, Gyutaek et al. "An implementation of a SIMT architecture-based stream processor." TENCON 2014 - 2014 IEEE Region 10 Conference (2014): 1-5.

Quiz! (True or False?)

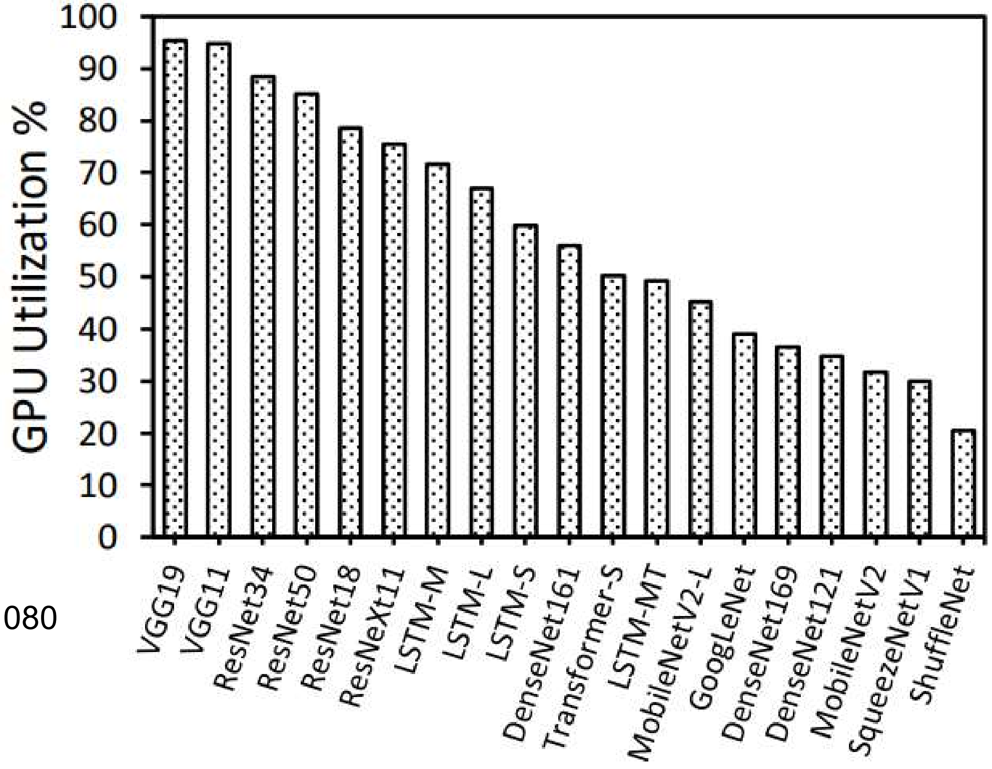
1. CPUs offer less parallelism compared to GPUs.

2. GPUs always execute faster than CPUs.

3. GPUs are the best choice for Deep Learning training.

4. GPUs are primary processors because of cost, programmability, performance tradeoff they offer.

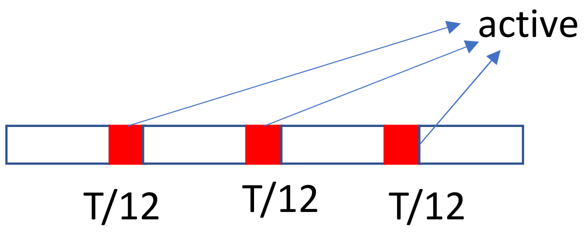
Underutilization Challenge of GPUs in DL Training



Deep Learning Models

Nvidia GeForce 1080
8 GB GDDR5X
1733 MHz

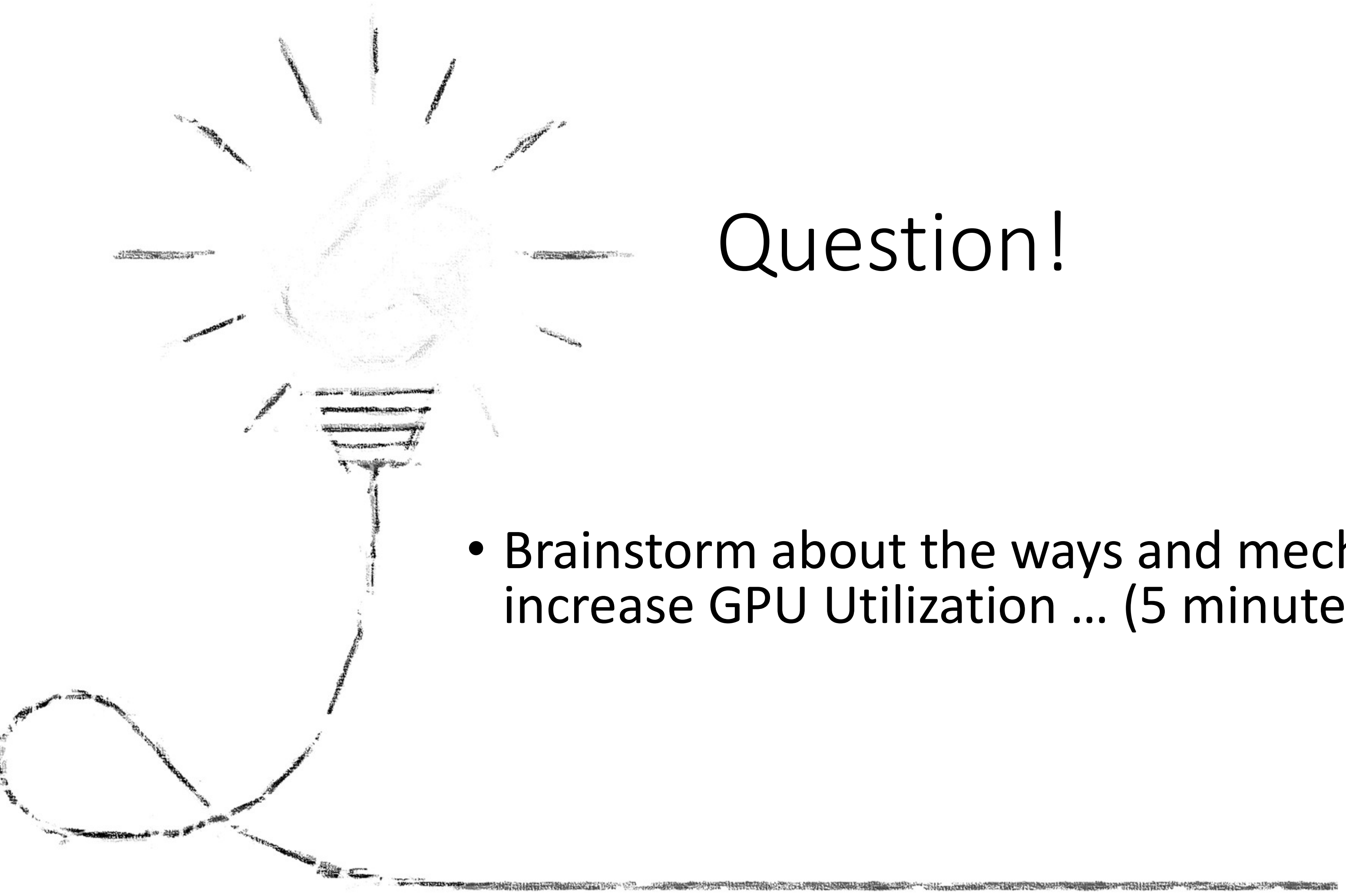
GPU Utilization



$$GPU\ Util = 3 * \frac{T}{12T} = \frac{T}{4T} = 25\%$$

Yeung, Gingfung, et al. "Towards GPU utilization prediction for cloud deep learning." *12th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 20)*. 2020.

Question!

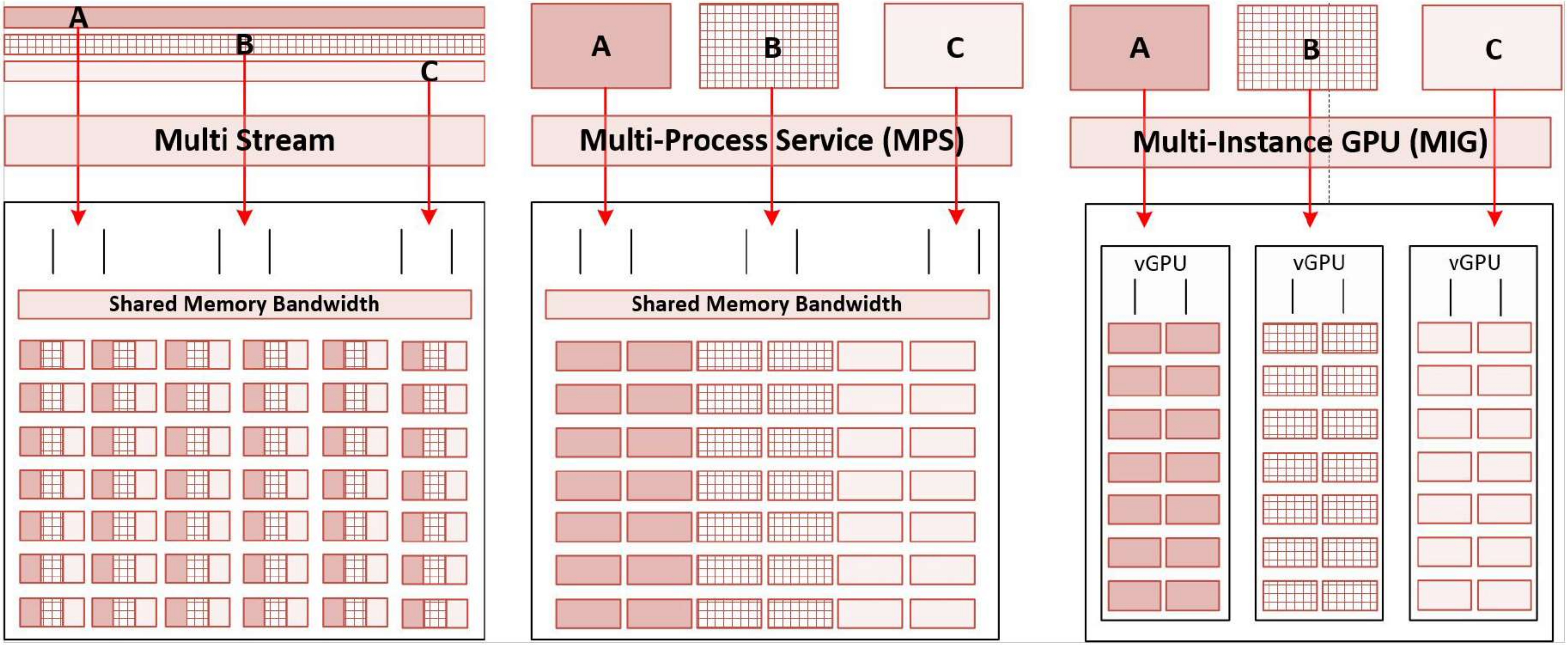


- Brainstorm about the ways and mechanisms to increase GPU Utilization … (5 minutes)

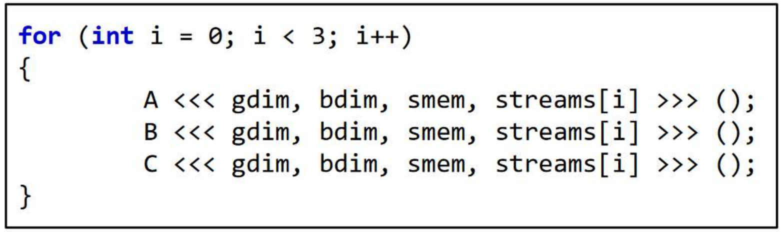
GPU Provisioning

GPU Over-provisioning as the mainstream solution

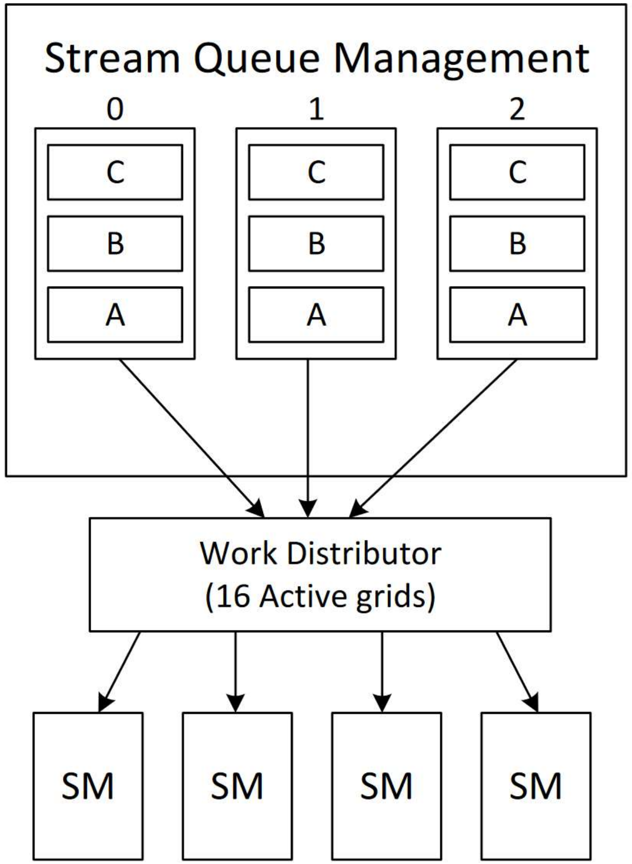
Naïve, MPS, MIG



Naïve or GPU streams

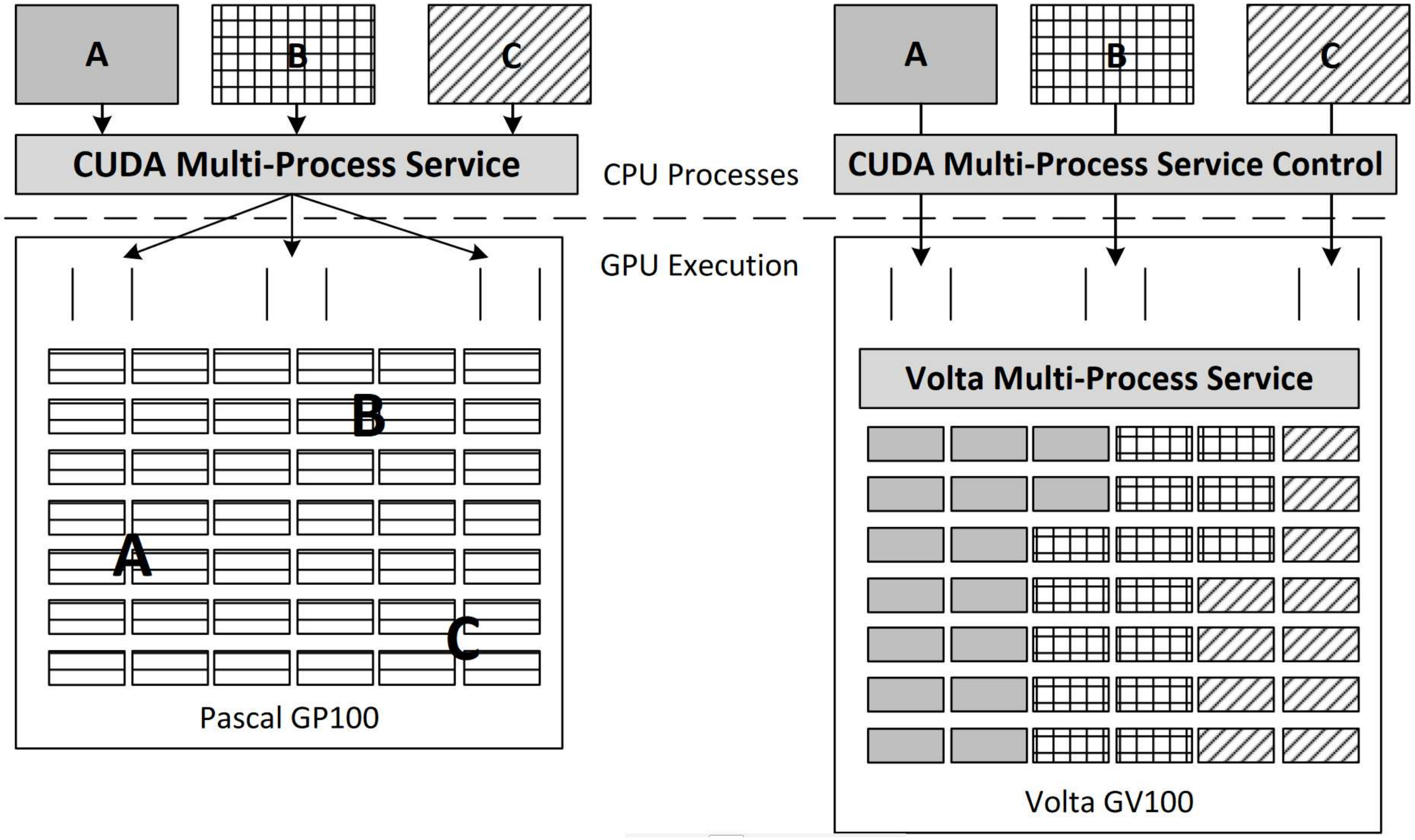


```
for (int i = 0; i < 3; i++)
{
        A <<< gdim, bdim, smem, streams[i] >>> ();
        B <<< gdim, bdim, smem, streams[i] >>> ();
        C <<< gdim, bdim, smem, streams[i] >>> ();
}
```

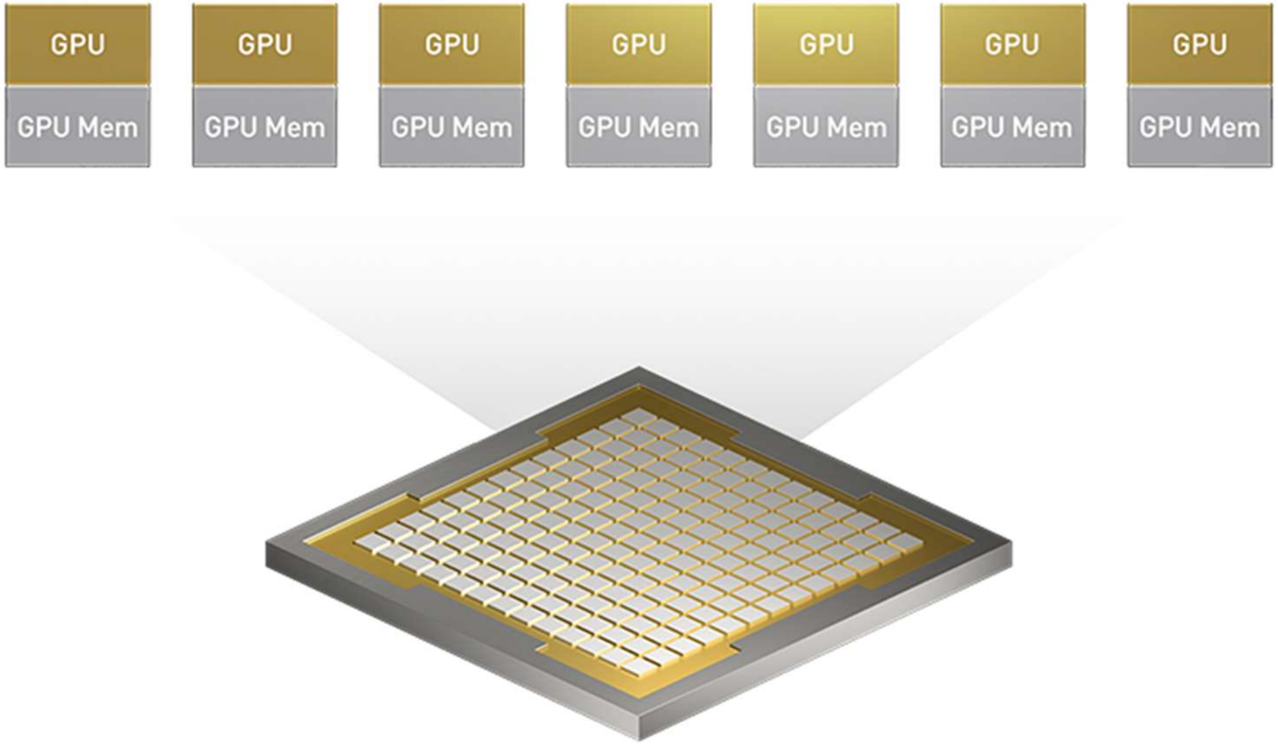


Multi-Process Service (MPS)

More Careful about QoS

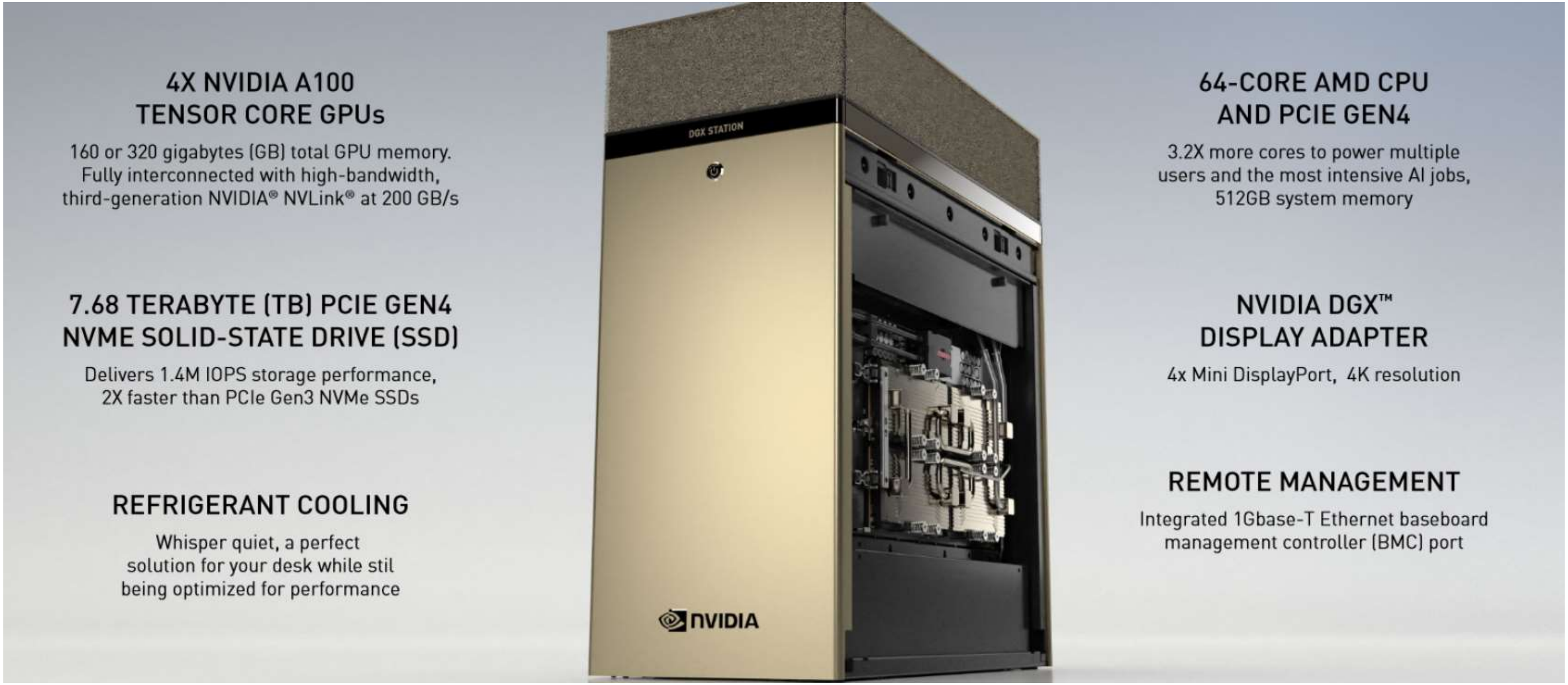

A
B
C
CUDA Multi-Process Service
CPU Processes
GPU Execution
B
A
C
Pascal GP100
A
B
C
CUDA Multi-Process Service Control
Volta Multi-Process Service
Volta GV100

Multi-Instance GPU (MIG)



https://www.nvidia.com/en-us/technologies/multi-instance-gpu/

Experiments



4X NVIDIA A100 TENSOR CORE GPUs

160 or 320 gigabytes (GB) total GPU memory. Fully interconnected with high-bandwidth, third-generation NVIDIA® NVLink® at 200 GB/s

7.68 TERABYTE (TB) PCIE GEN4 NVME SOLID-STATE DRIVE (SSD)

Delivers 1.4M IOPS storage performance, 2X faster than PCIe Gen3 NVMe SSDs

REFRIGERANT COOLING

Whisper quiet, a perfect solution for your desk while stil being optimized for performance

64-CORE AMD CPU AND PCIE GEN4

3.2X more cores to power multiple users and the most intensive AI jobs, 512GB system memory

NVIDIA DGX™ DISPLAY ADAPTER

4x Mini DisplayPort, 4K resolution

REMOTE MANAGEMENT

Integrated 1Gbase-T Ethernet baseboard management controller (BMC) port

https://www.nvidia.com/en-us/technologies/multi-instance-gpu/

Experiments

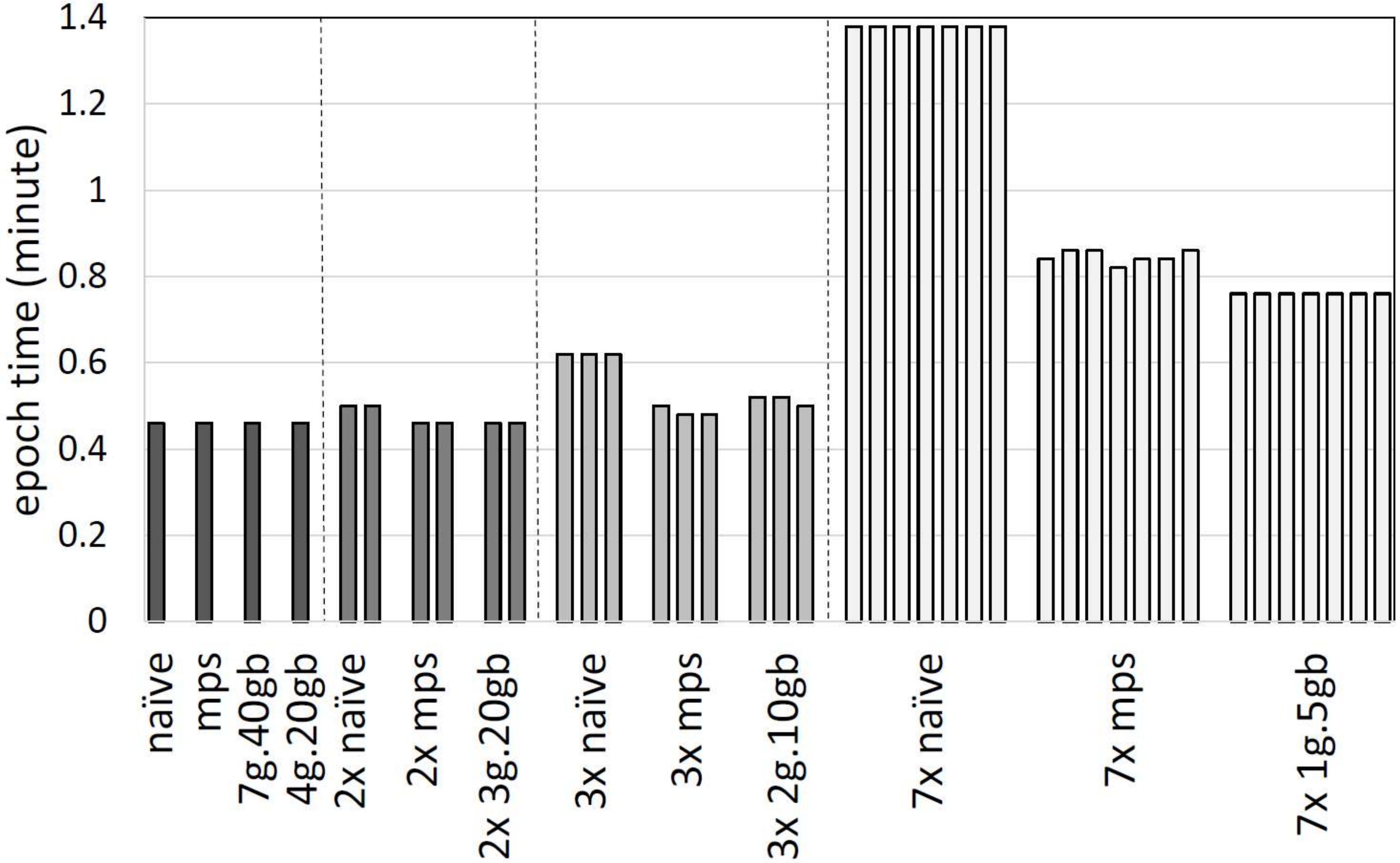
Model	Small	Medium	Large
ResNet	ResNet26 + Cifar10	ResNet50 + ImageNet64x64	Resnet152 + ImageNet
EfficientNet	EfficientNet_S + Cifar10	EfficientNet_S + ImageNet64x64	EfficientNet_S + ImageNet
Cait	x	x	Cait_XXS_24 + ImageNet

Hyperparameter	Value
Batch Size	128, 32 (only resnet)

Small Model

Execution Time or Performance

Resnet26 + Cifar10

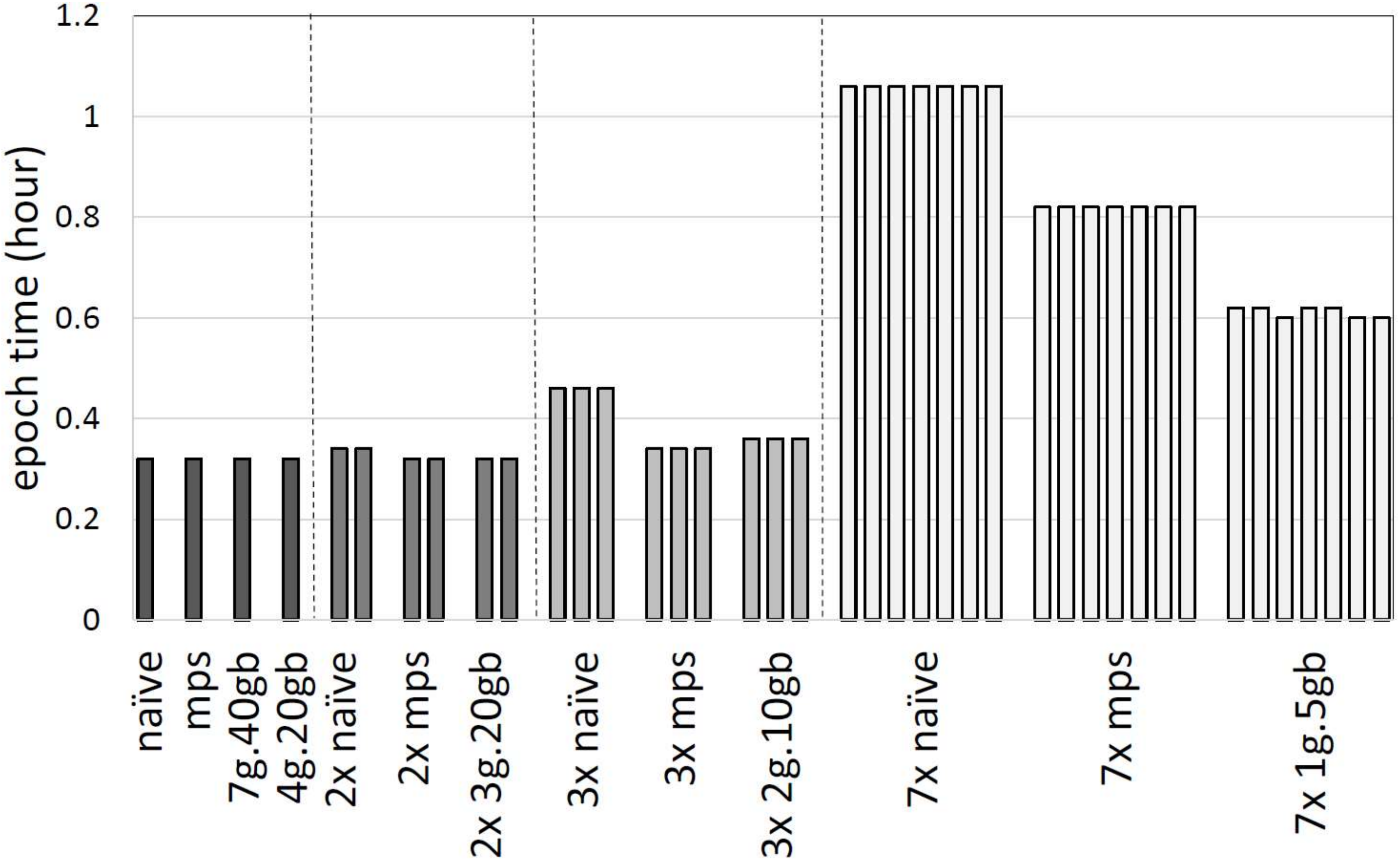


Resource Contention
Stream << MPS < MIG

Medium Model

Execution Time or Performance

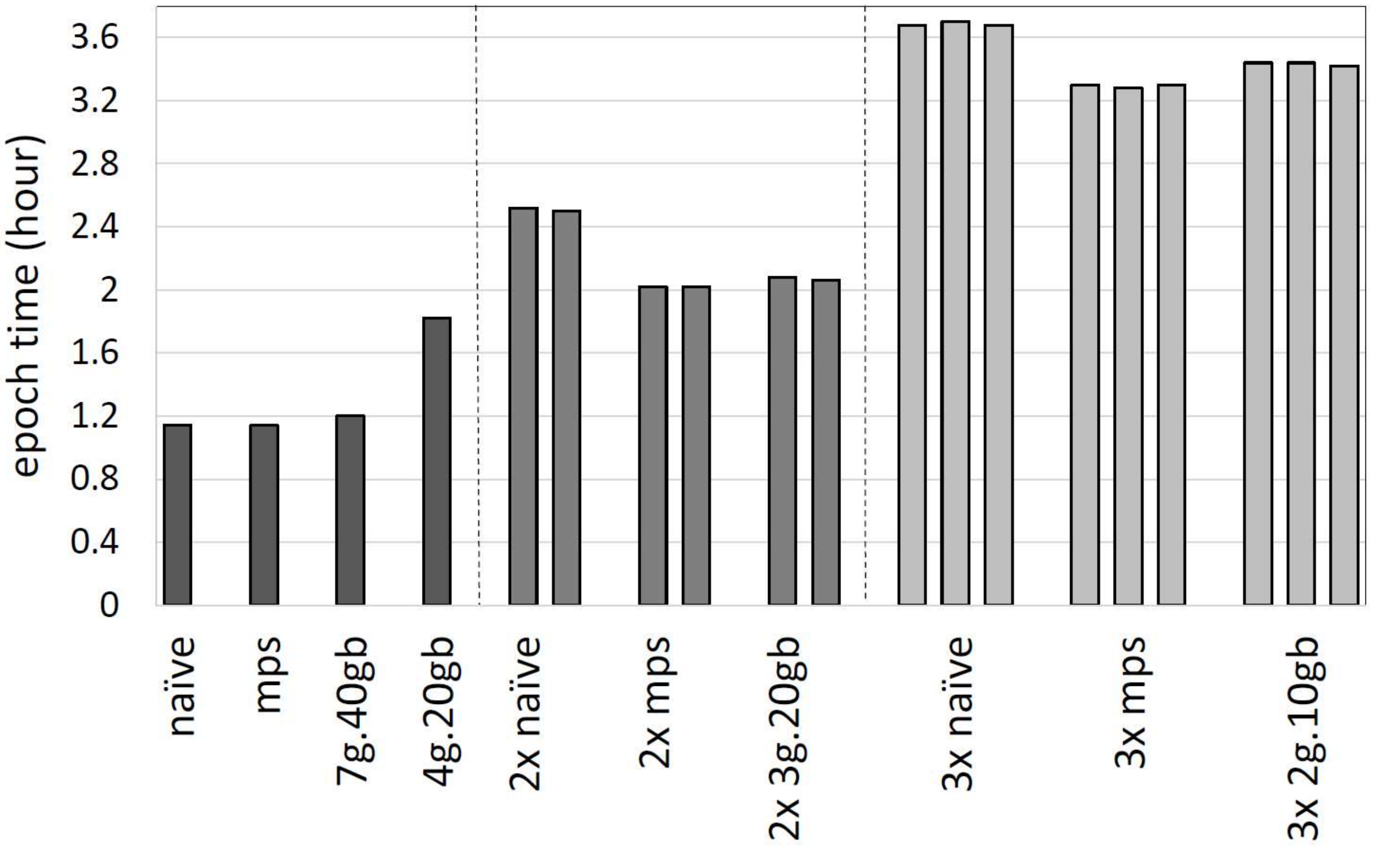
Resnet50 + ImageNet64x64



Execution Time or Performance

Large Model

Resnet152 + ImageNet

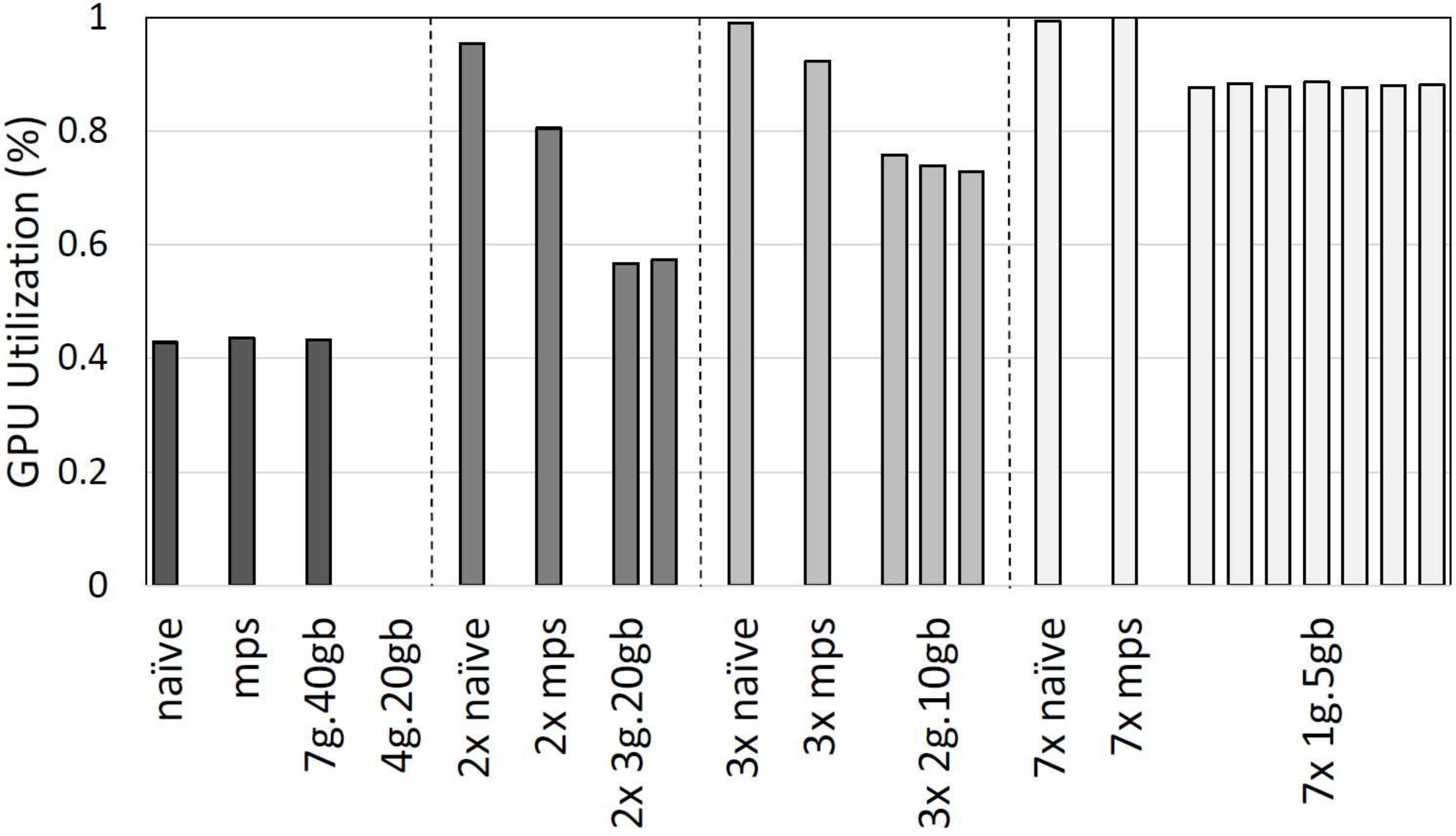


MIG Resources are not enough.

Small Model

GPU Utilization

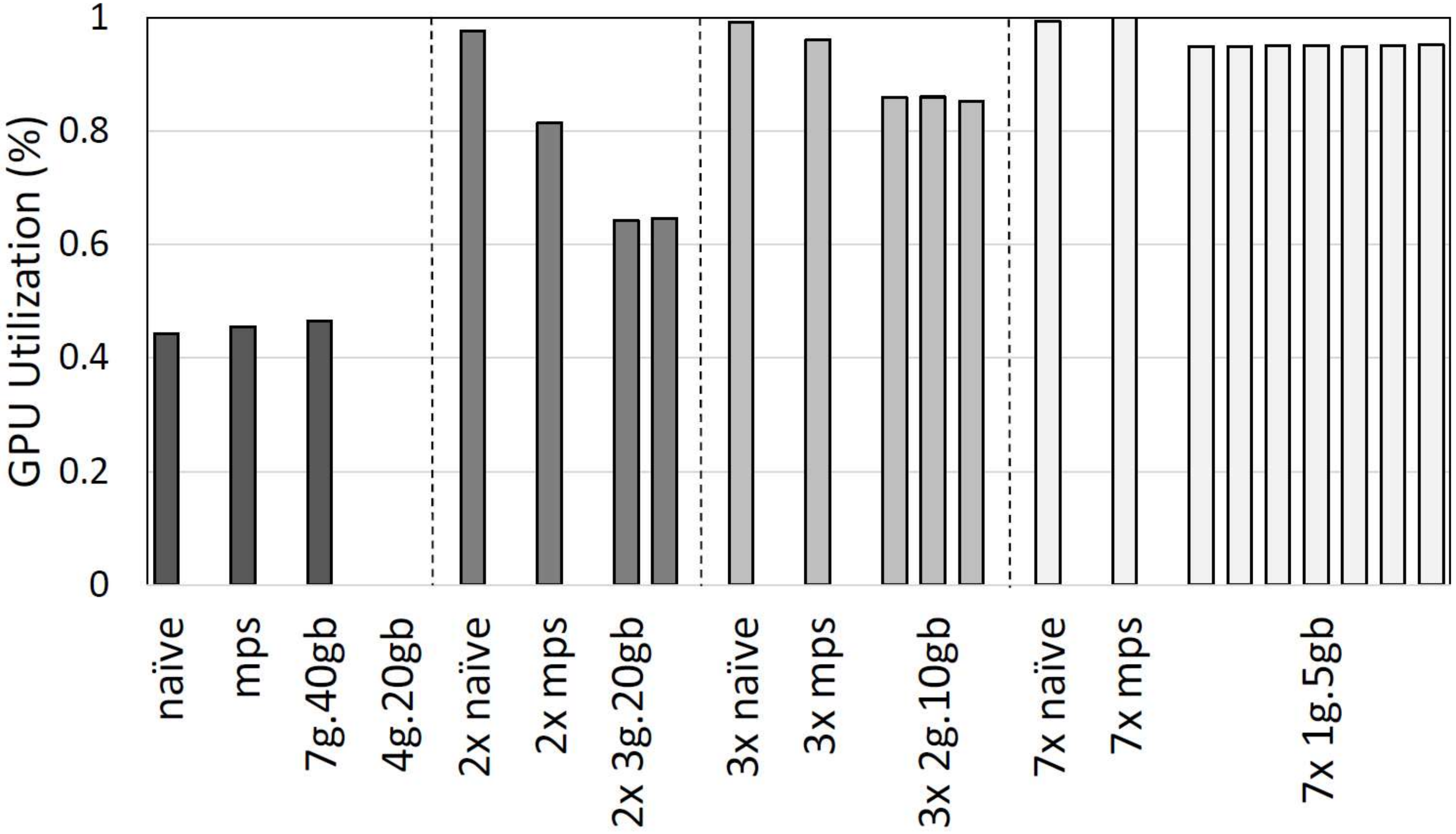
Resnet26 + Cifar10



GPU Utilization

Medium Model

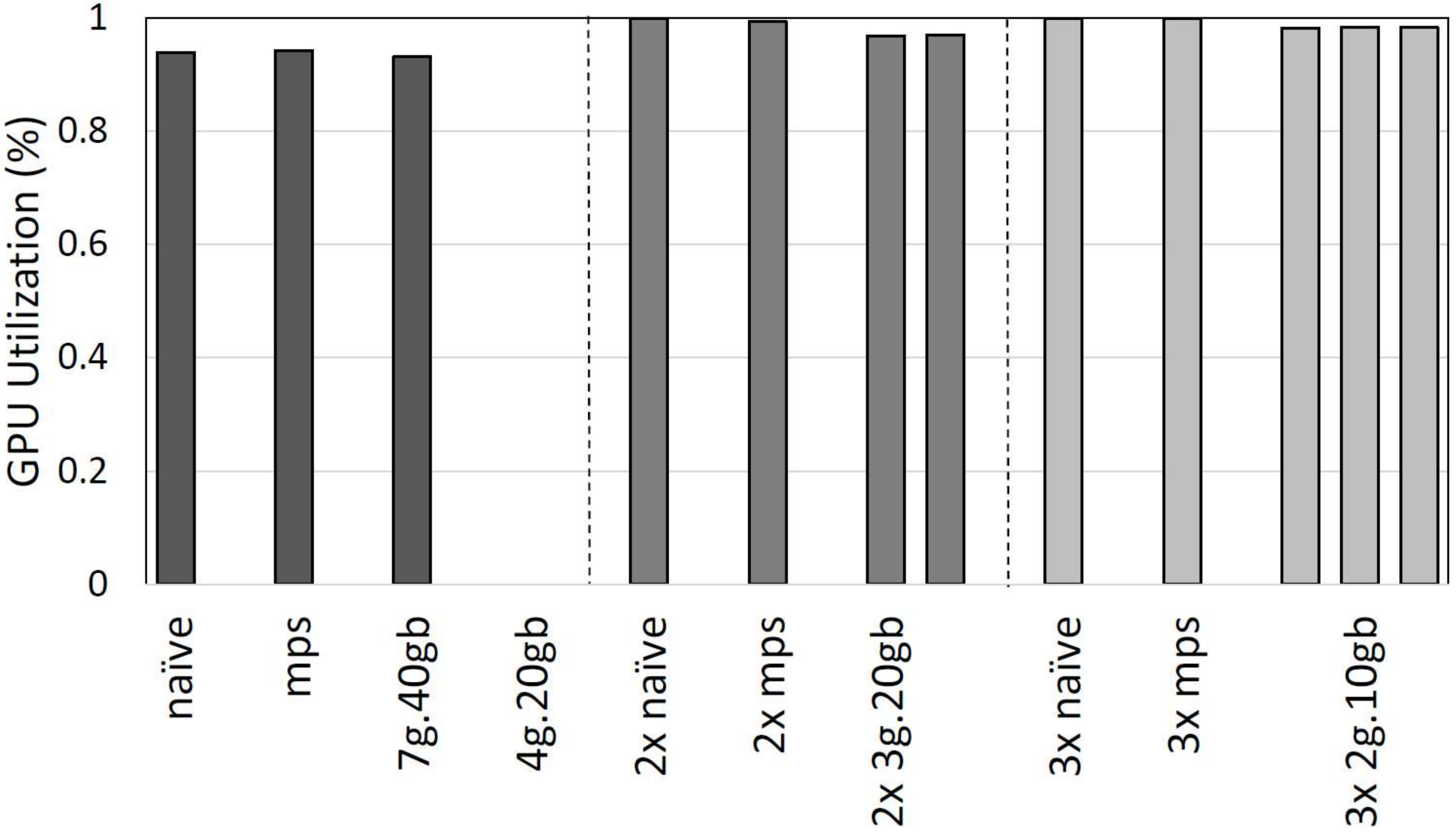
Resnet50 + ImageNet64x64



GPU Utilization

Large Model

Resnet152 + ImageNet



Conclusion

1. Deep Learning offers acceptable solutions to a variety of application
2. Deep Learning training is compute/memory hungry.
3. GPUs are the main accelerators for these applications.
4. GPUs suffer from underutilization in the age of AI.
5. Overprovisioning GPUs as a solution
6. Available workload collocation options: Naïve, MPS, MIG
7. In terms of interference: Naïve < MPS < MIG
8. Intelligent collocation offers energy and performance efficiency