Exploring the Impact of System Storage on AI & ML Workloads via MLPerf Benchmark Suite

MLPerf Training v0.5

Wes Vaske

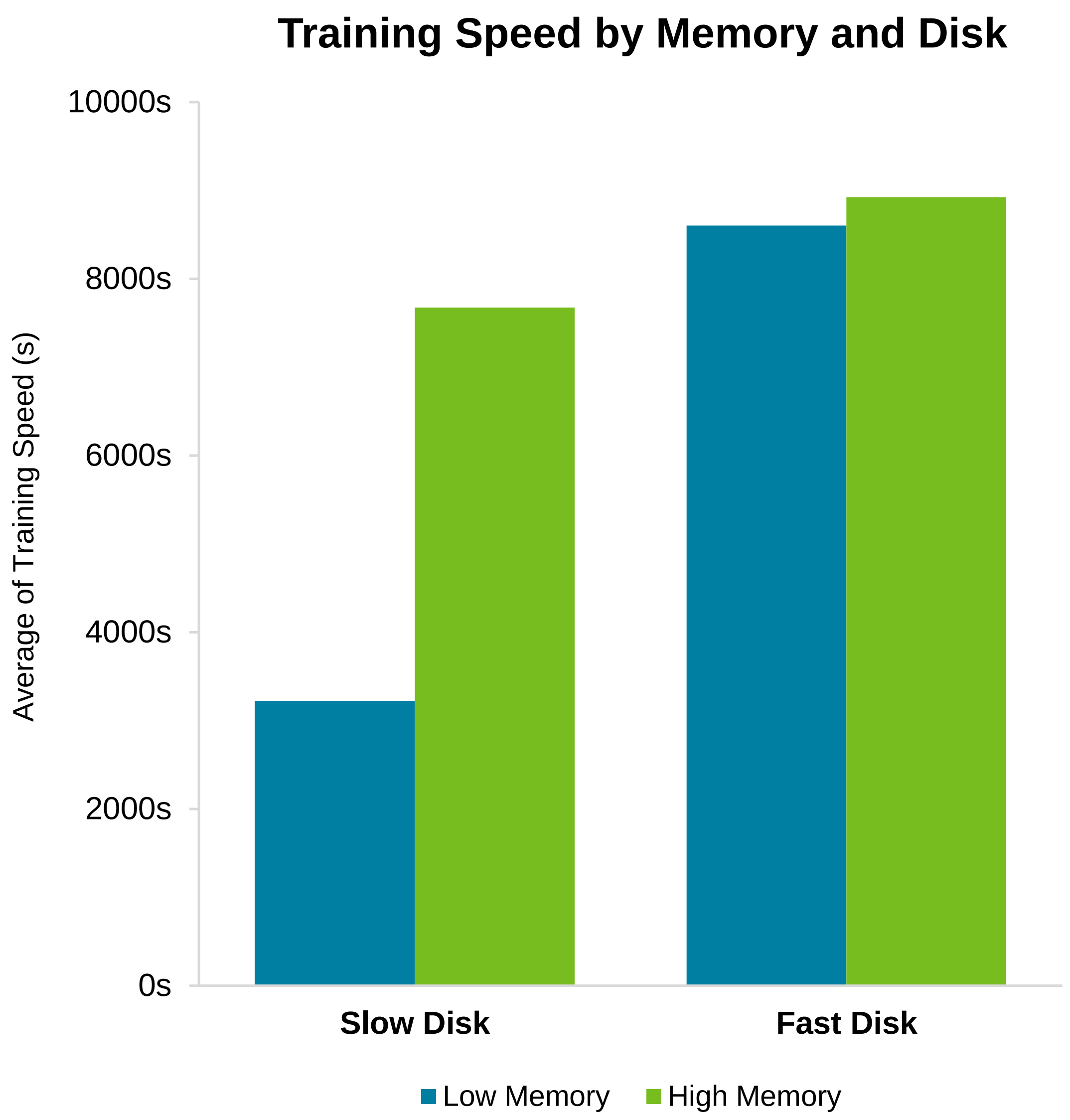
Principal Storage Solutions Engineer

In the edge case, can storage impact training performance?

Training speed of ResNet-50 model with ImageNet.

Container resource limits to show impacts of constrained systems.

- **Memory:**
  High Memory =1TB
  Low Memory = 128GB
- **Disk:**
  Fast Disk = 8x NVMe unlimited
  Slow Disk = 500 MB/s limit



Agenda

- What is MLPerf?
- What system resources do AI/ML apps need?
- Are SATA SSDs fast enough?
- Parallel data ingest and model training.

MLPerf Overview Training

Benchmark suite measuring how fast systems can train models to a target quality metric.

Reference implementation is provided per benchmark:

- Code that implements the model in at least one framework
- A Dockerfile to run the benchmark in a container
- A script to download the dataset
- A script to run and time training

Training Benchmarks:

- Image classification
- Object detection
- Recommendation
- Reinforcement
- RNN translator
- Sentiment analysis
- Single stage detector
- Speech recognition
- Translation

MLPerf Overview | Inference

Benchmark suite measuring how fast a system can perform ML inference.

- Each benchmark is defined by a model, a dataset, a quality target, and a latency constraint
- A LoadGen application is provided to generate queries and measure latencies

Reference implementation is provided per benchmark.

- Code that implements the model in at least one framework
- A Dockerfile to run the benchmark in a container
- A script to download the dataset
- A script to run and time training

Cloud Inference

- Image classification
- Language modeling
- Sentiment analysis
- Single stage detector

Edge Inference

- Face identification
- Object classification
- Object detection
- Object segmentation
- Speech recognition
- Translation

The System Configuration

SuperMicro SYS-4029GP-TVRT

2x Intel Xeon Platinum 8180 CPUs
- Each: 28-core @ 2.50GHz

3TB RAM
- 24x 128GB 2666MHz LRDIMMs

8x Nvidia V100 SXM2 GPUs
- Each: 32GB RAM
- NVLink Cube Mesh gpu-to-gpu fabric

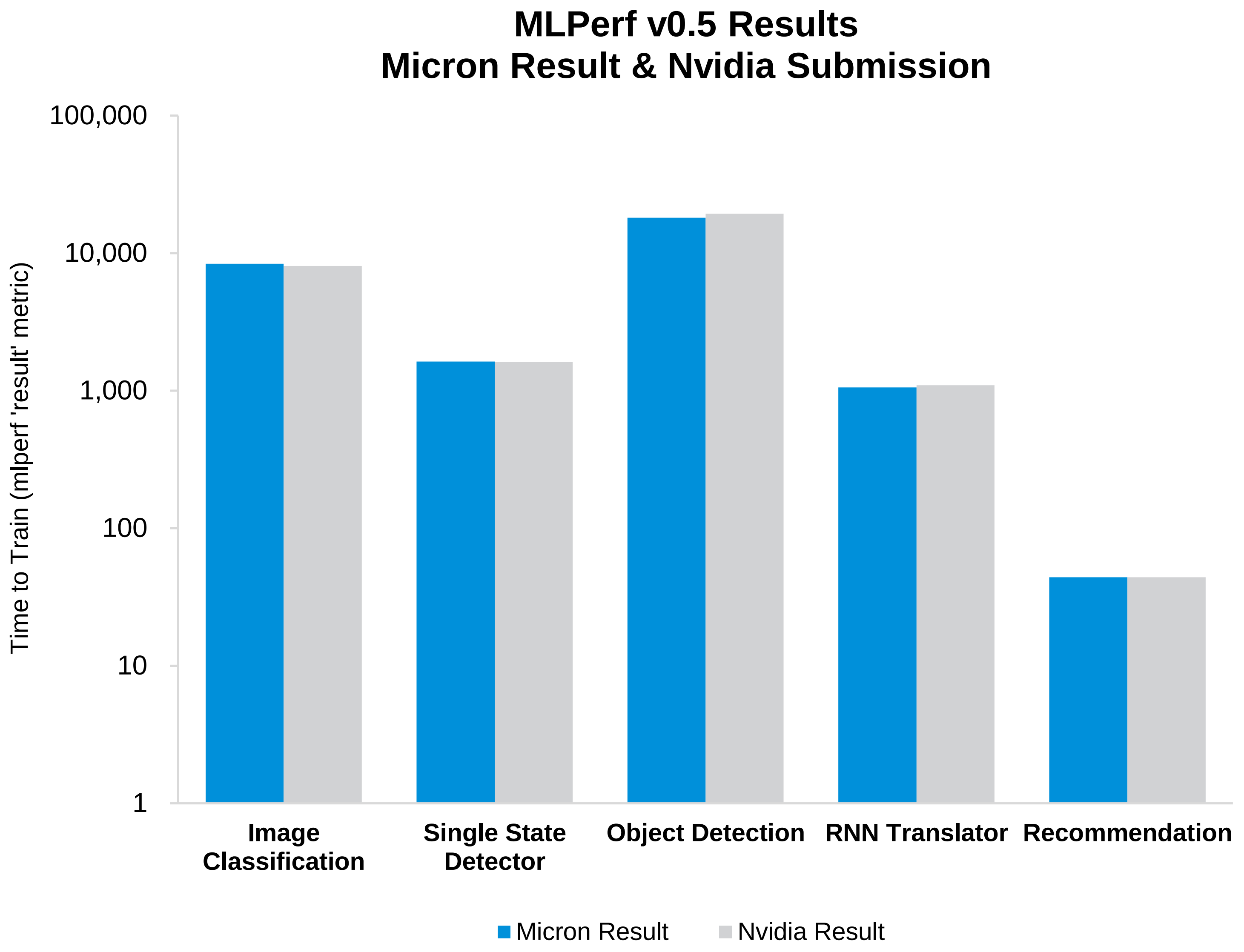
Data Drives:
- 8x SATA SSD
- 8x NVMe SSD

Very similar to an Nvidia DGX-1

How similar to a DGX-1?

Effectively identical performance between the SuperMicro 8x GPU system and the DGX-1 (also an 8-GPU system).

Shows that AI/ML applications are generally compute bound (should not be surprising).

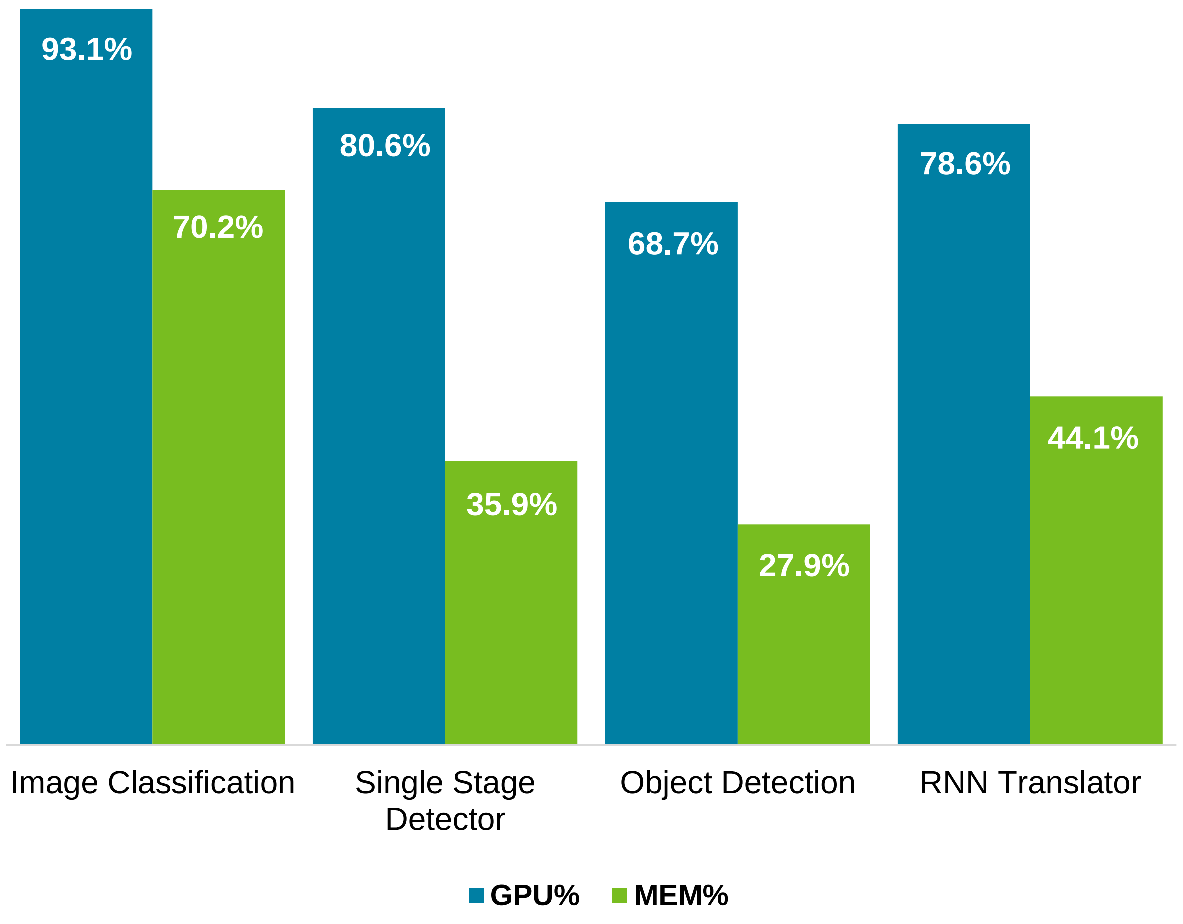


What system resources do AI/ML apps stress?

GPUs (should be obvious)

- High GPU utilization means a well optimized training process.
- The varying memory utilization is an artifact of small datasets used by the benchmarking process.

GPU Core and Memory Utilization by Benchmark



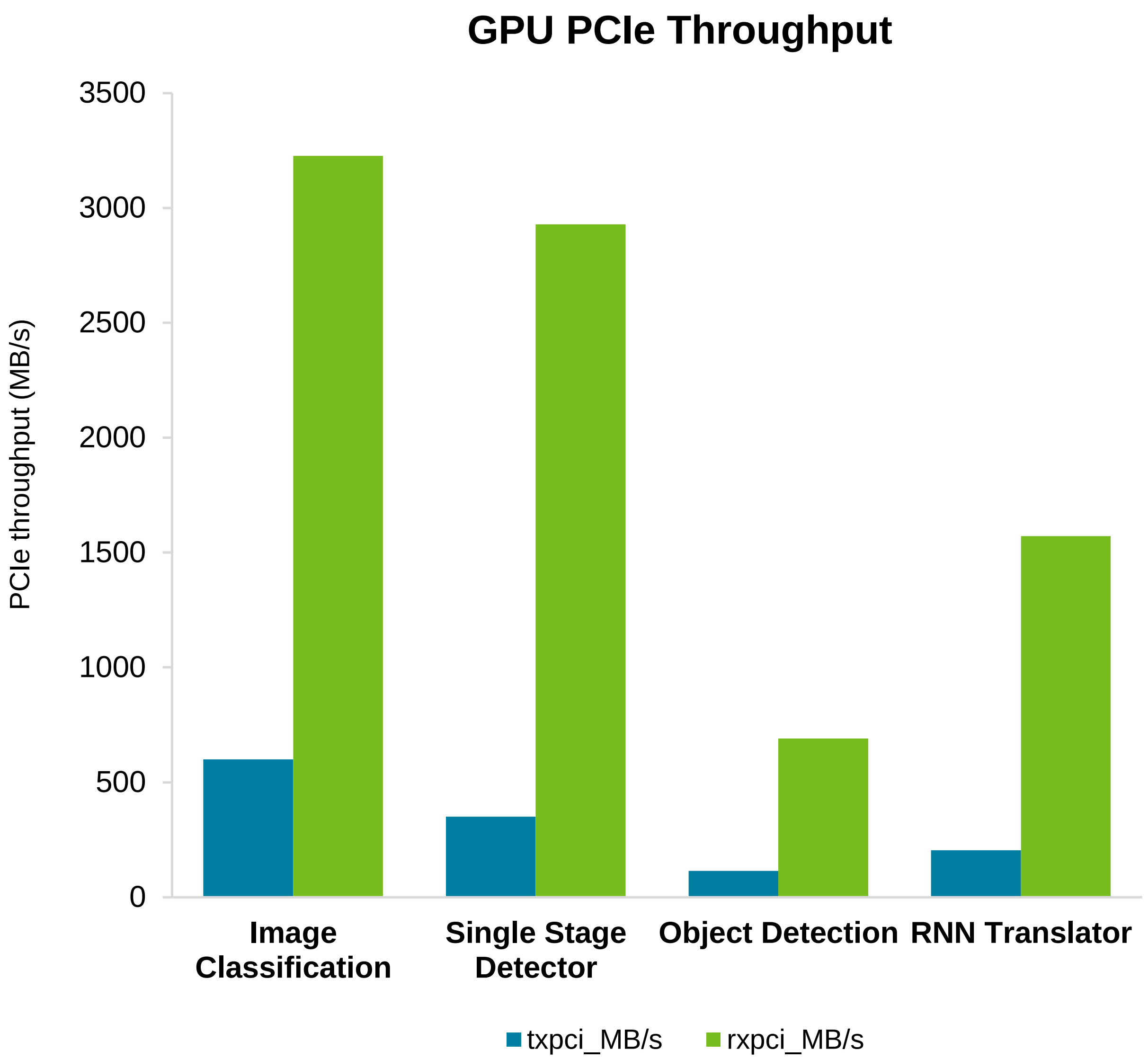
What system resources do AI/ML apps stress?

PCIe Bandwidth

- Data is average per GPU

For Image Classification

- 3,227 * 8 = 25,800 MB/s
- Equivalent to 2x PCIe x16
- There are 4x PCIe x16 lanes connecting GPUs to CPUs
- Significant PCIe utilization but not currently a bottleneck



What system resources do AI/ML apps stress?

CPUs

Max of 3,200% for image classification.

- Non-normalized value is ‘equivalent’ to 32 fully loaded cores or 64 half loaded cores.

Significant requirement but fairly attainable today.

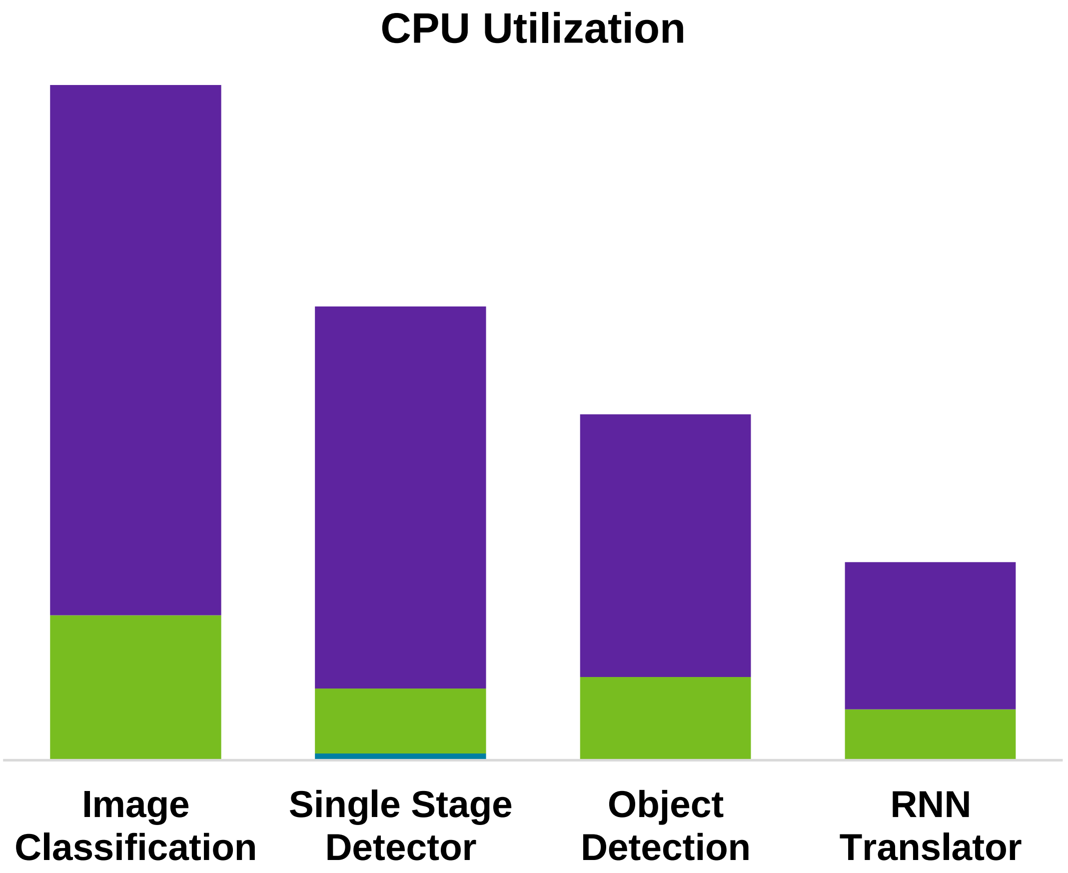


	Image Classification	Single Stage Detector	Object Detection	RNN Translator
User	2563%	1697%	1167%	653%
System	639%	288%	366%	226%
Iowait	5%	31%	4%	1%

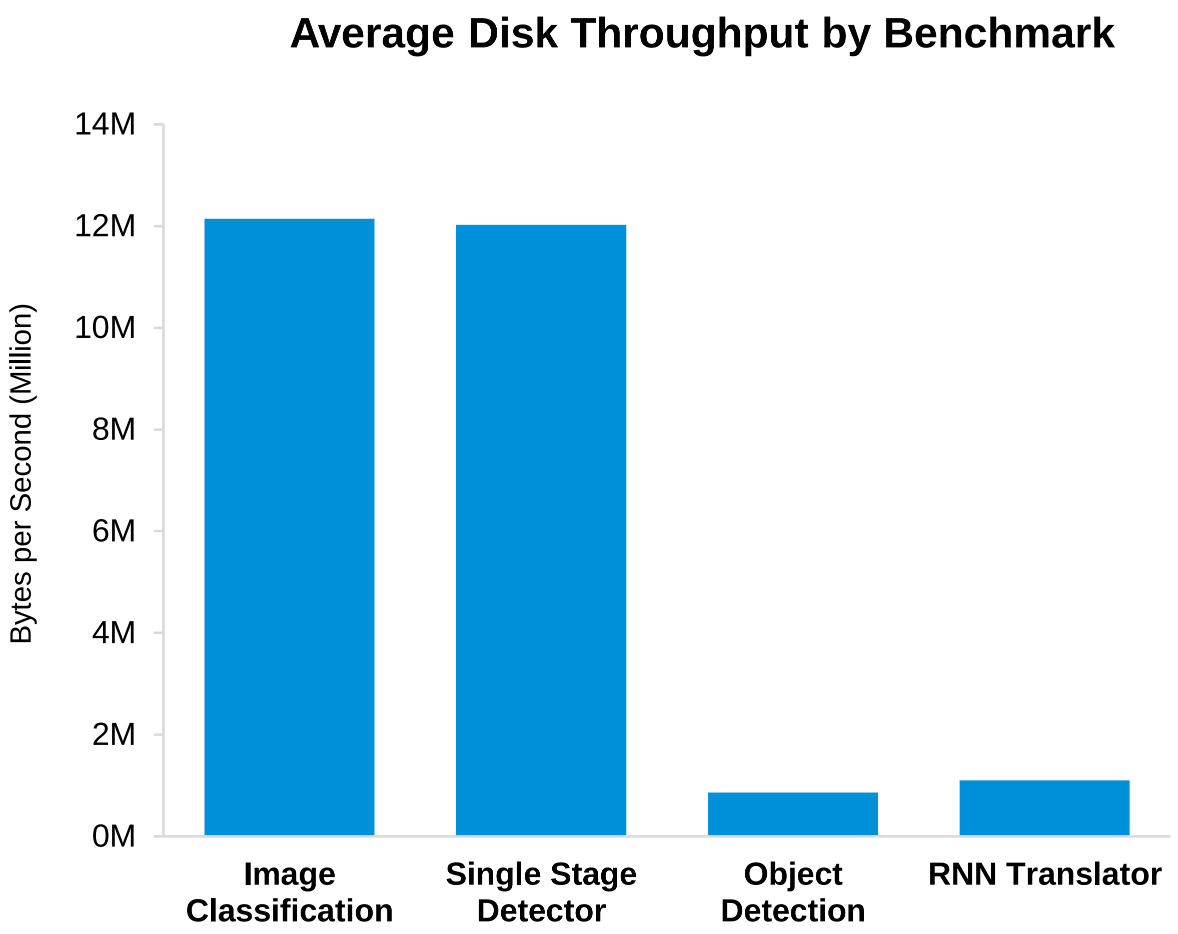
What system resources do AI/ML apps stress?

Disk

- Standard testing shows negligible disk throughput.
- Max of 18 MB/s

What's going on?

- We see high peak disk utilization during first epoch and zero disk utilization on all subsequent epochs.
- Over many epochs this looks like very low disk dependence.



Real World Architecture/Process vs Benchmarks

Training datasets in the real world are significantly larger than those used by these benchmarks.

The datasets for the MLPerf benchmarks will fit in the file system cache.

- Largest benchmark dataset is <150GB

Real world datasets are generally in the TB to PB range.

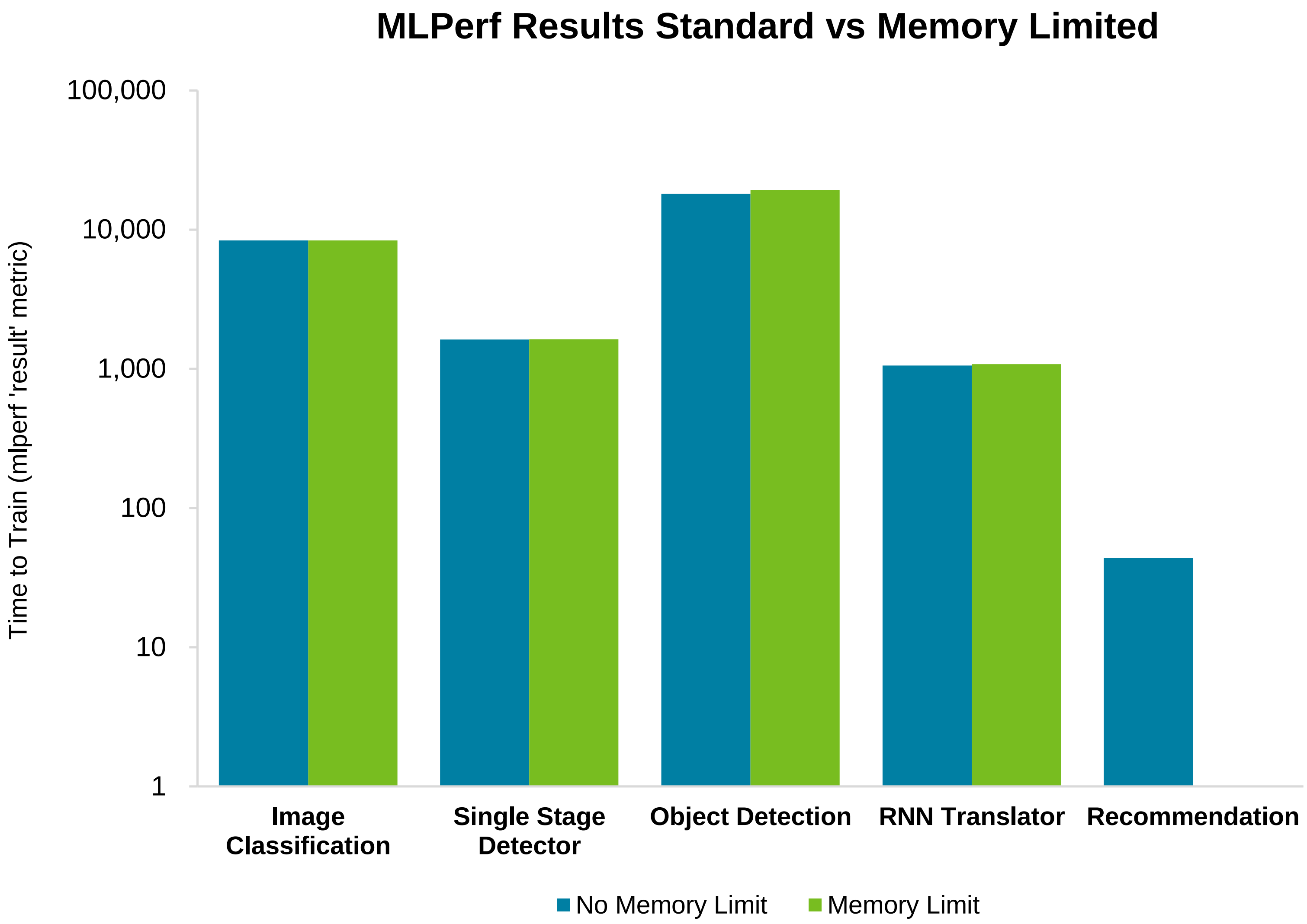
How can we benchmark AI/ML applications in a way that is more representative of customer environments?

Training Datasets Don’t Fit In Memory

Solution:

Limit memory available to the container so that only a small part of the dataset will fit in the filesystem cache.

With proper tuning, the filesystem cache is unable to cache the dataset and the model training performance is unchanged.



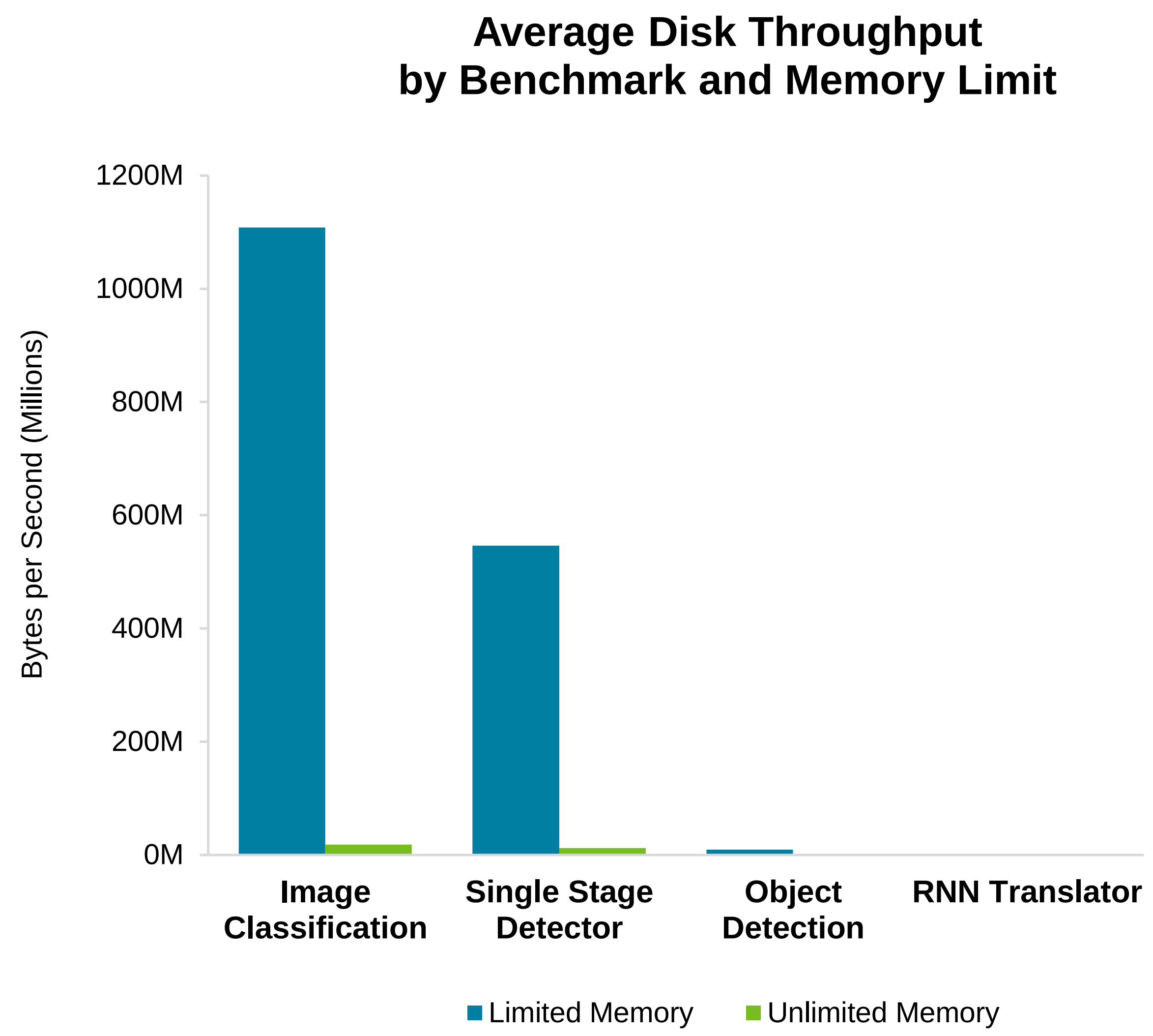
Training Datasets Don’t Fit In Memory

Result:

- Limiting memory does not change the time-to-train (at least with fast enough storage)
- It DOES increase the disk throughput substantially

Image Classification:

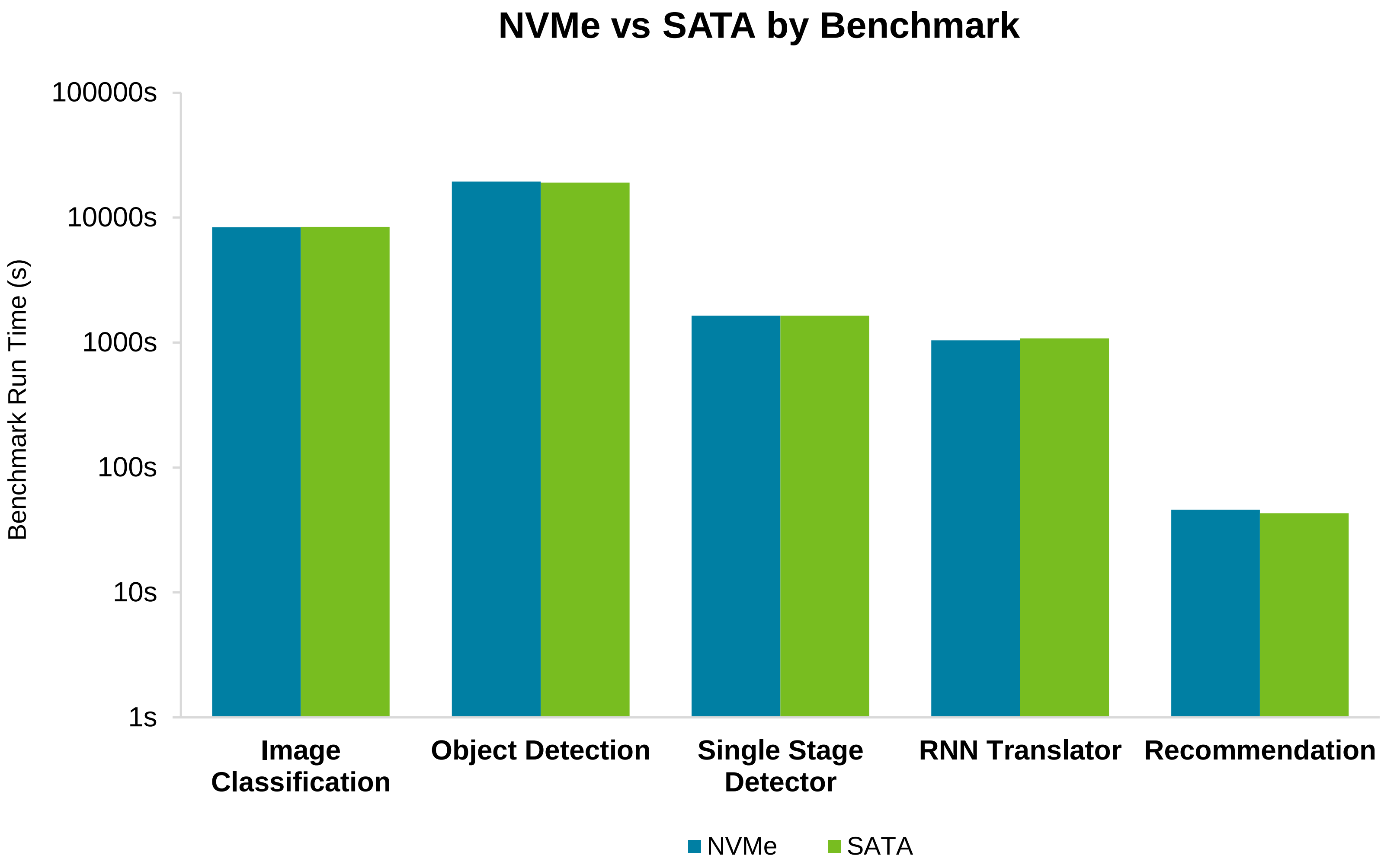
- Disk throughput increased 61x
- Takes 62 epochs to train
- With lots of memory only first epoch reads from disk
- Will less memory 61 more epochs will read from disk



Training Datasets Don’t Fit In Memory

How much performance is “enough”?

When running isolated benchmarks, basic flash based storage is “enough” for full model training performance.



Real World Architecture/ Process vs Benchmarks

Training Jobs Aren't Run In Isolation

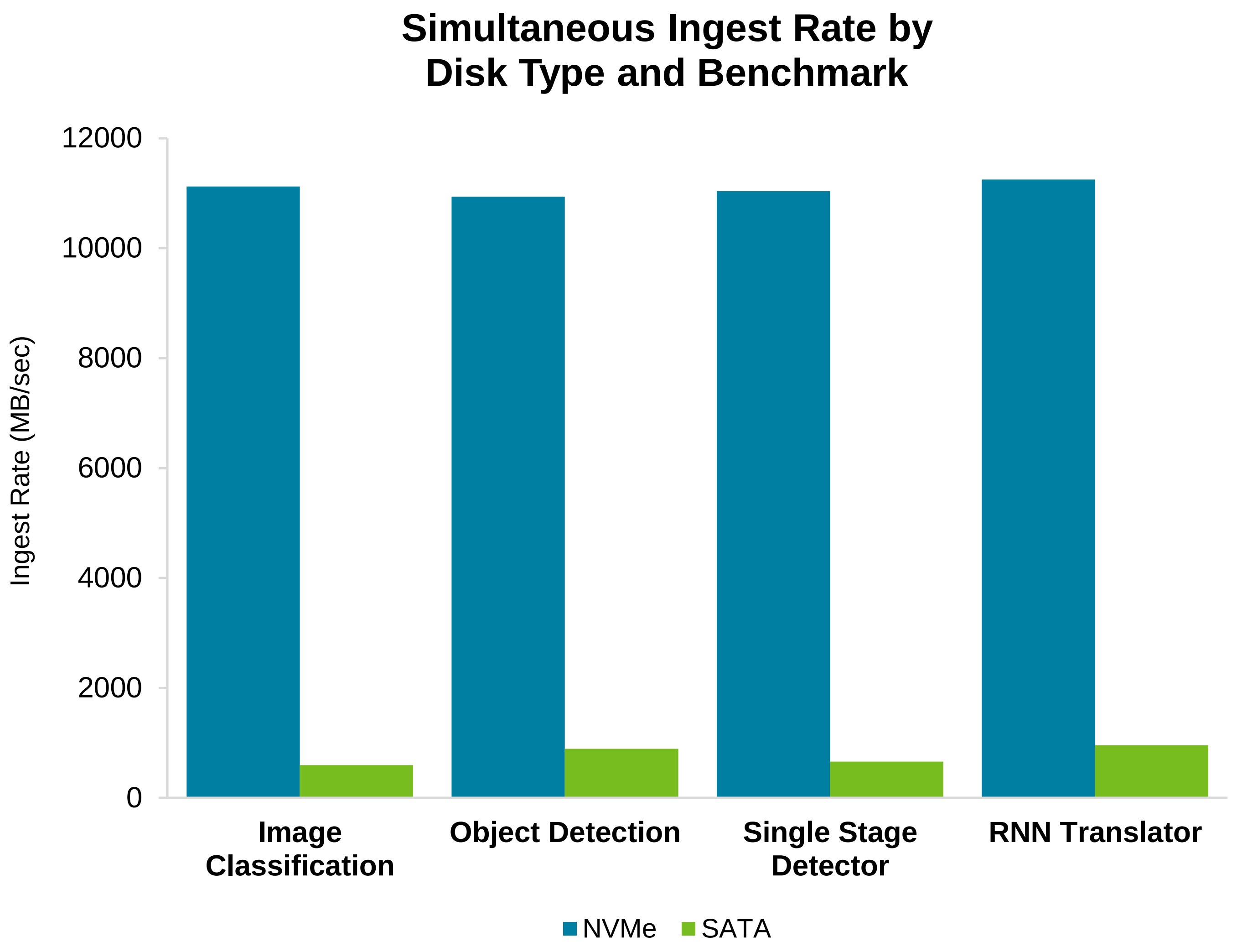
- Training a model is rarely a one-and-done process
- Multiple models need to be trained
- Large datasets need to be copied to the local disk cache for training then removed

NVMe vs SATA Ingest Rates

Ingest Job (fio)

- 32 jobs @ QD1 per job
- 32 files on XFS (32GB per file)
- 128k transfer size

How does ingest affect the model training time?

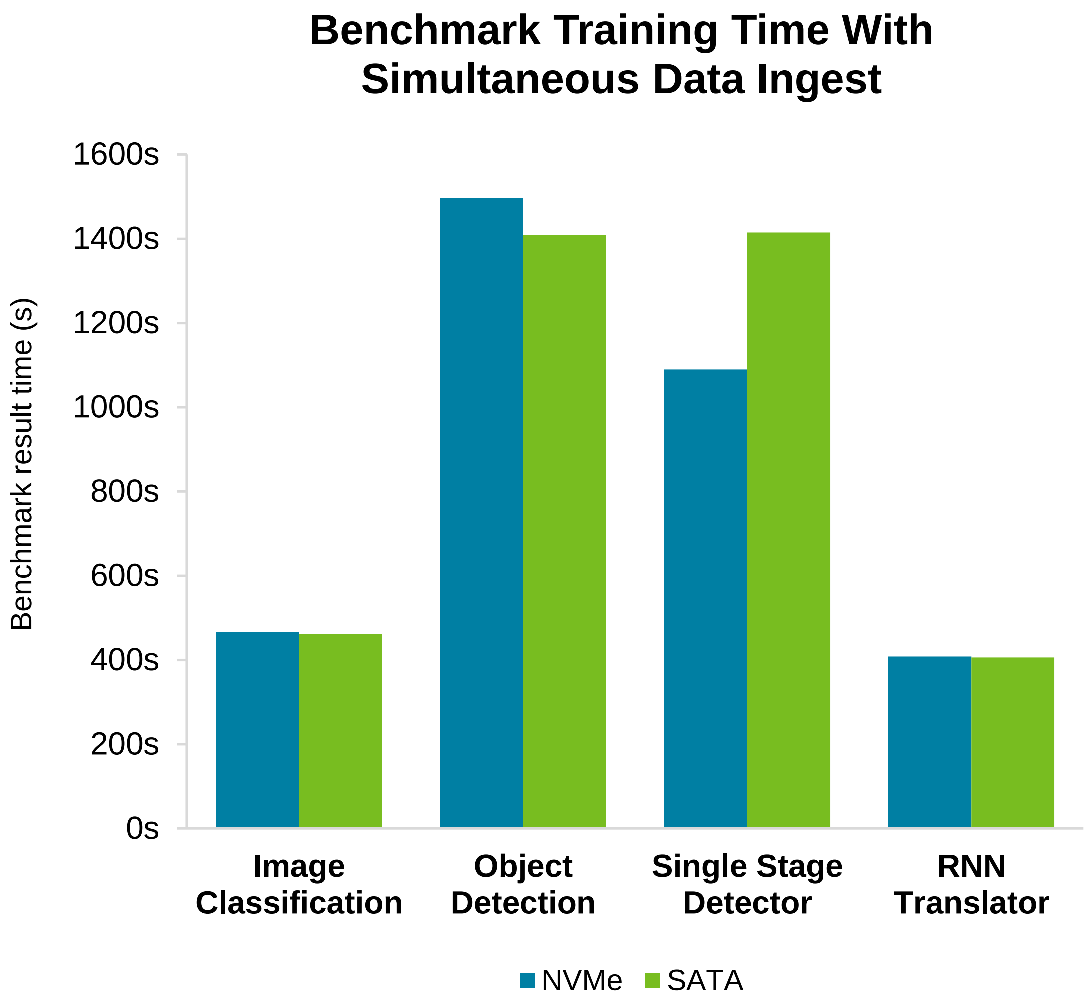


NVMe vs SATA Training Performance With Data Ingest

Single Stage Detector ~30% slower on NVMe than SATA when doing simultaneous training and data ingest.

Interesting facts:

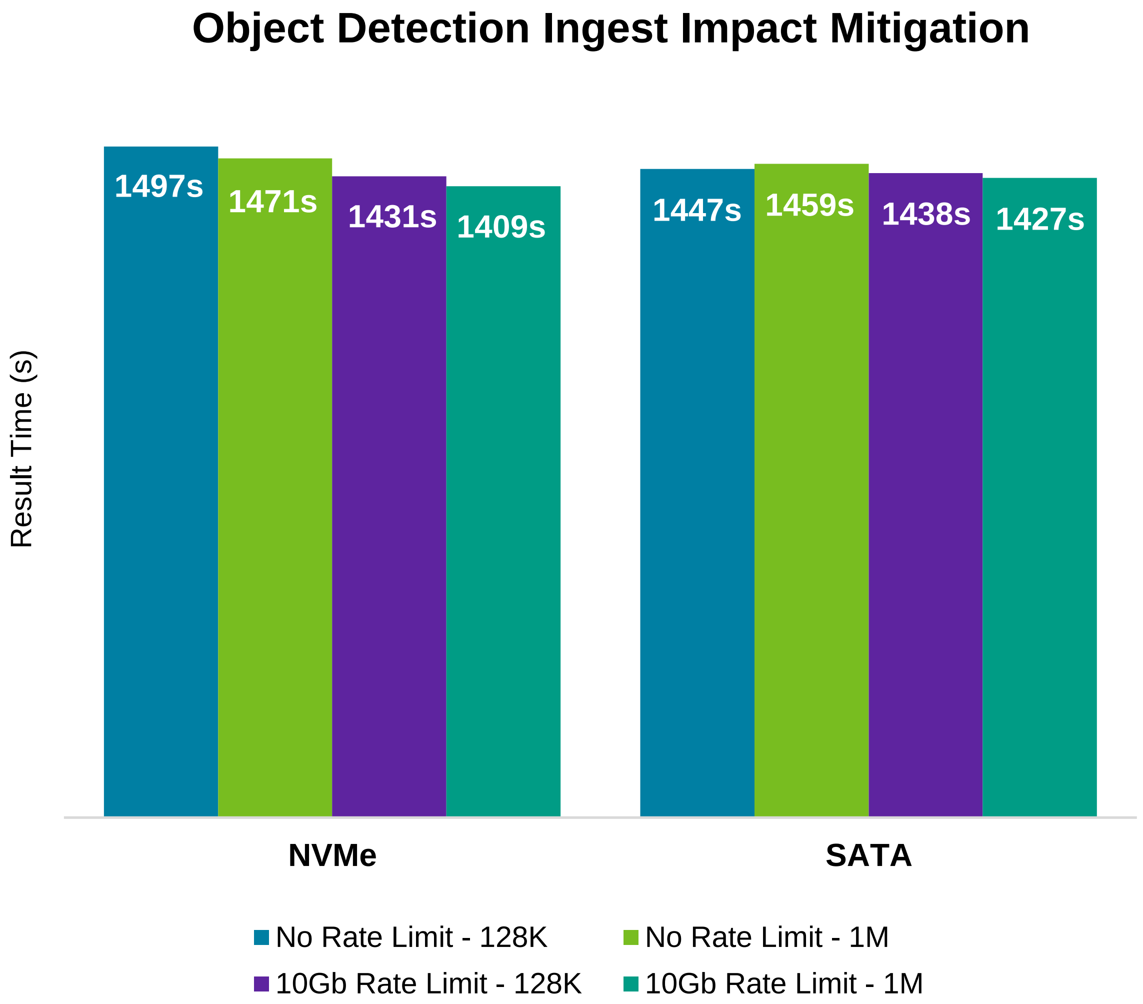
- Single Stage Detector had only the 3rd highest disk utilization during training.
- Dependence on storage performance not 100% correlated with disk activity



Mitigating the Impact of Data Ingest
Object Detection

Tested Mitigations:
- Limit the ingest rate (10Gb)
- Use a larger block size

Both are effective at limiting the impact on training performance of data ingest.

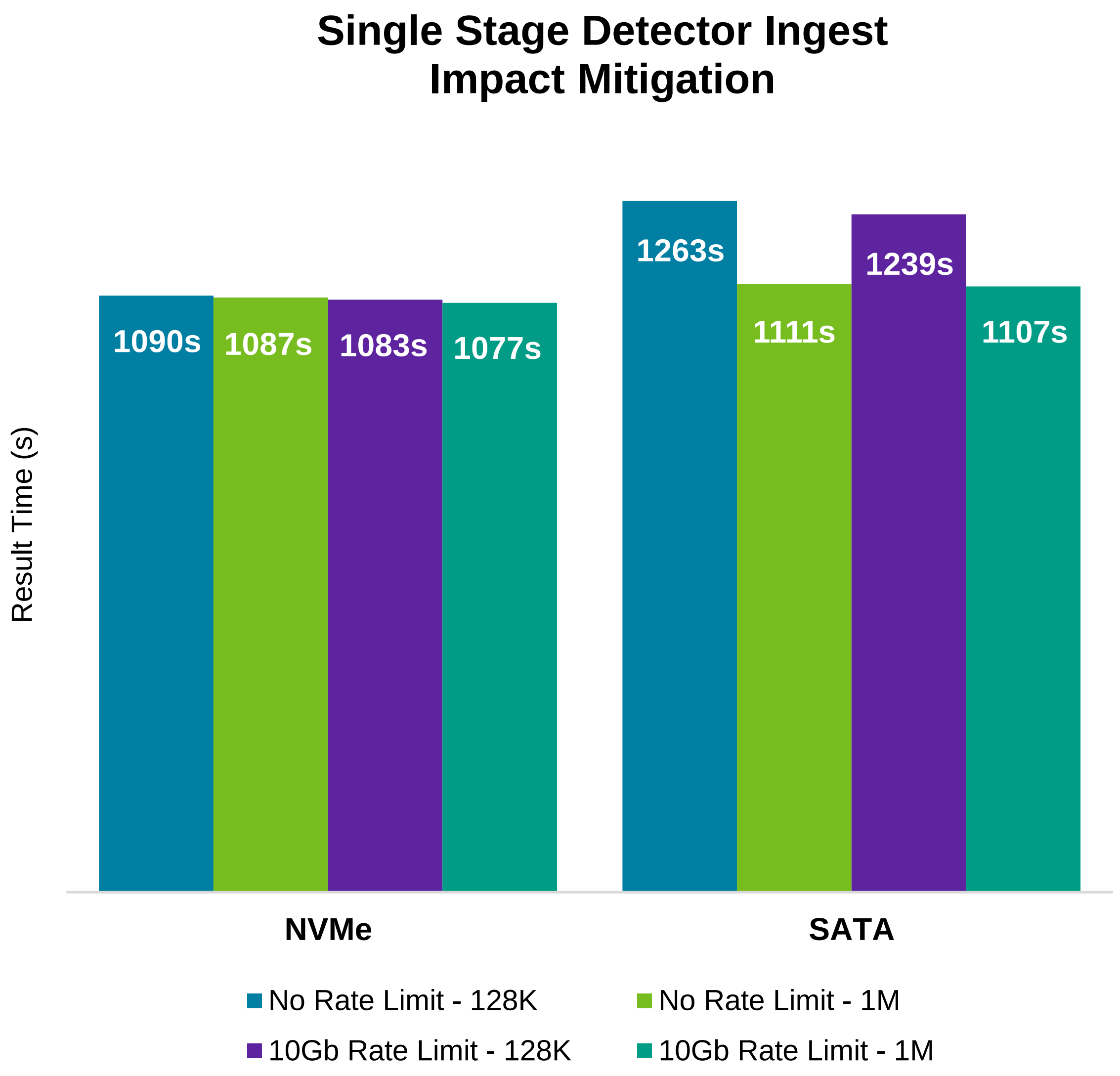


Mitigating the Impact of Data Ingest Single Stage Detector

Tested Mitigations:

- Limit the ingest rate (10Gb)
- Use a larger block size

Using a larger transfer size is more effective than limiting the ingest rate.



High Pace of Software Advancements

Compare MLPerf v0.5 and v0.6

- Up to 36% performance improvements
- Will directly result in higher disk requirements
- 7 months

The data presented here is already out-dated.

Architecting for the future is difficult.

MLPerf	V0.5	V0.6	Diff %
Image Classification	134.6	115.22	14%
Object Detection, Light-Weight	26.9	22.36	17%
Object Detection, Heavy-Weight	322.9	207.48	36%
Translation, Recurrent	18.3	20.55	-12%
Translation, Non-Recurrent	32.7	20.34	38%

For More Information

Twitter:

@wvaske

LinkedIn:

https://www.linkedin.com/in/wes-vaske-b550988

Recent Blog Posts:

- How to architect your System for More Efficient AI Model Training
- AI Matters: Getting to the Heart of Data Intelligence with Memory and Storage
- Artificial Intelligence — Why Now?

https://www.micron.com/solutions/artificial-intelligence

Micron®