MLPerf Storage - Enabling easy Storage for AI benchmarking

Wes Vaske
Senior Member of Technical Staff, Systems Performance Engineer
Micron Technology, Inc.
https://www.linkedin.com/in/wes-vaske-b550988/

the Future of Memory and Storage

Agenda

- MLPerf Storage Overview
- Analysis of MLPerf Storage Workloads

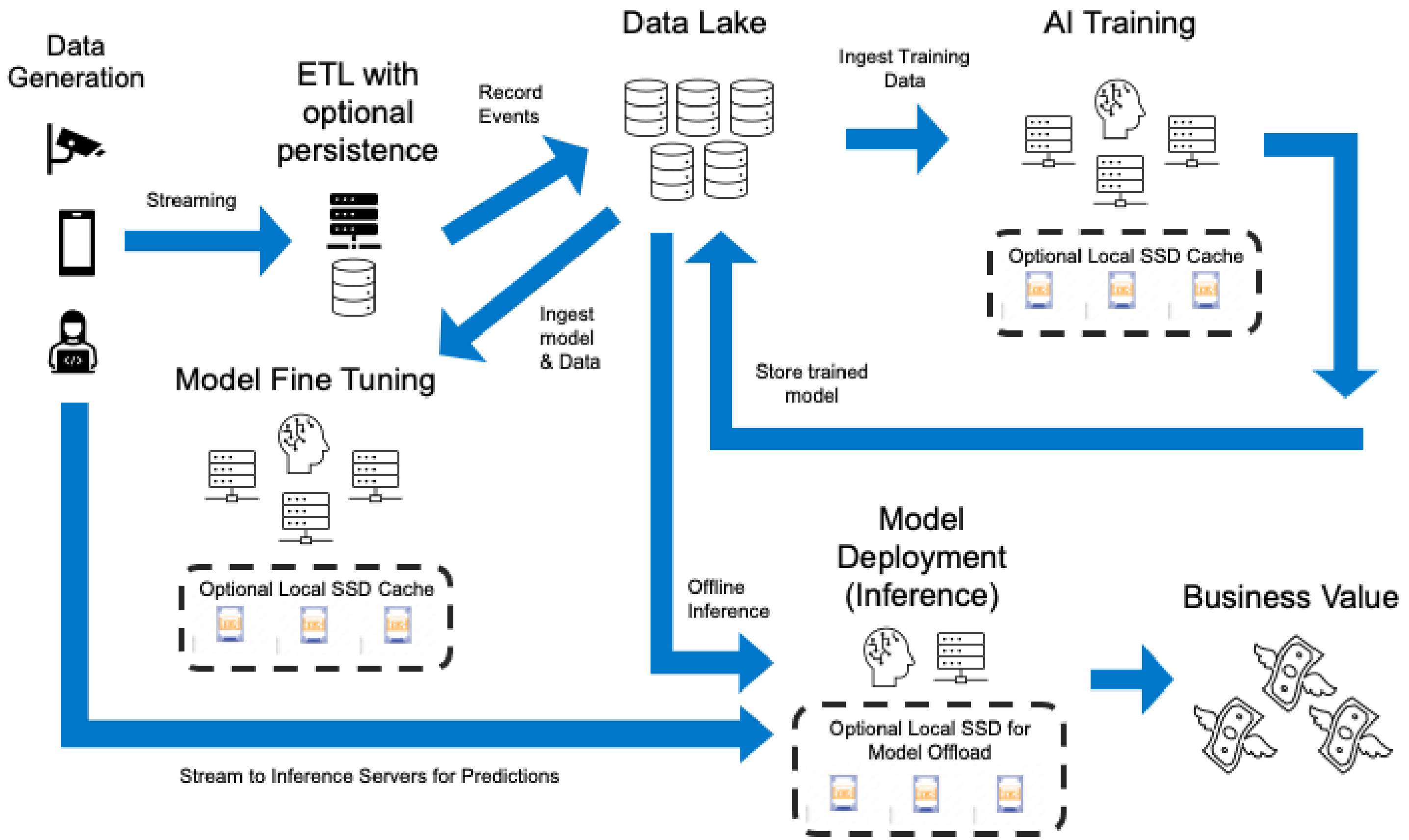
Micron Confidential

MLPerf Overview

Micron Confidential

Storage for AI Overview

- Storage for AI is broad and complex
- Benchmarking AI systems requires costly accelerators
- Available datasets are small compared to datasets used by industry
  - Example: Recommendation Systems
  - Criteo dataset used for AI Training benchmarks is 1TB pre-processed and 300GB post-processing
  - Meta states that a recommendation training job reads 10PB to 100PB of semi-processed data



- [Understanding Data Storage and Ingestion for Large-Scale Deep Recommendation Model Training]

MLCommons and MLPerf Storage

- MLCommons has benchmarks for many parts of the AI pipeline
  - Training
  - Inference
  - Tiny
  - Mobile
  - Algorithms
  - Datasets
  - **Storage**

- **MLPerf Storage** currently focuses on AI Training
  - In the future will add benchmarks for data ingest, preprocessing, and inference
- Version 0.5 supported:
  - BERT (NLP)
  - Unet3D (medical imaging / computer vision)
- Upcoming version 1.0 will support:
  - Unet3D
  - Resnet50 (computer vision)
  - CosmoFlow (HPC)

MLPerf Storage Benchmark

- Emulated Accelerators are defined by:
  - Batch Size
    - Optimal number of samples for training to target accuracy with the real dataset and model
  - Computation Time
    - Found experimentally by running with the real accelerators
    - The time for a forward and backward pass
    - Emulated with a sleep() command

- Defines benchmarks as a set of:
  - Data Format
    - Serialized Numpy, tfrecord, png, etc
  - Data Loader
    - Numpy, DALI, TensorFlow, PyTorch Native
  - Emulated Accelerator Model

Workloads Analysis

Micron Confidential

The Basics (Gen5 NVMe)

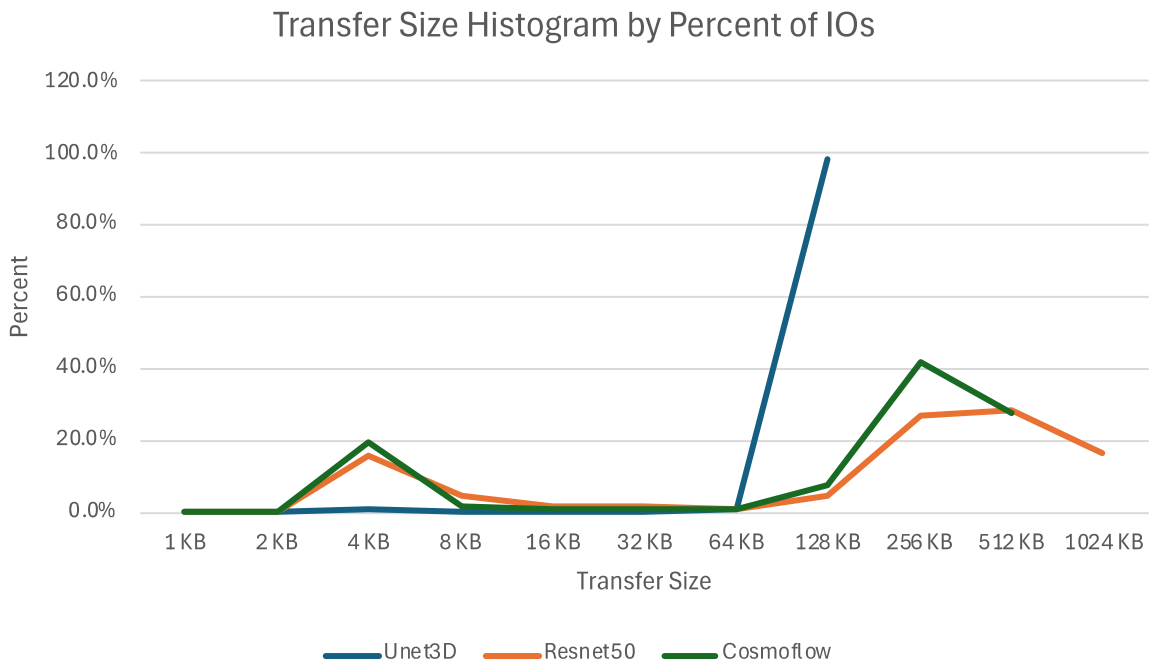
Model	Dataset Size (GB)	Dataset Size (Samples)	MLPerf Storage Config	Data Format & Reader	Mean Throughput (MB /s)
Unet3D	1,300 GB	9,375	Accelerator: H100 Accel QTY: 3 Read Threads: 16	1 sample per file Numpy .npz Pytorch Reader	5,800 MB/s
Resnet50	1,300 GB	95,810,000	Accelerator: H100 Accel QTY: 53 Read Threads: 8	10,000 samples per file TFRecord Tensorflow	4,450 MB/s
CosmoFlow	1,300 GB	485,000	Accelerator: H100 Accel QTY: 13 Read Threads: 8	1 sample per file TFRecord Tensorflow	5,650 MB/s

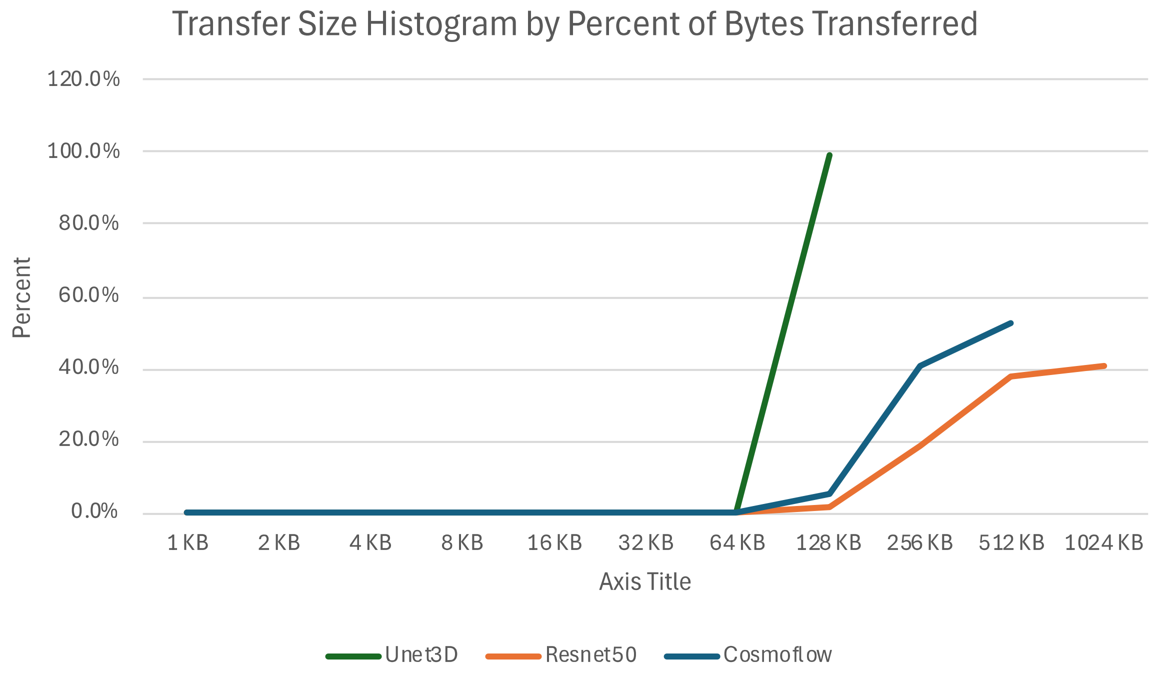
- Performance tuned to be “near” max performance for each benchmark.
- 4x orders of magnitude from fewest samples to most samples
- Throughput requirements per accelerator differ by model

Micron Confidential

Transfer Size Histogram

- Unet3D uses pytorch and shows larger percent of 128KB transfers with none larger
- Resnet50 and CosmoFlow use Tensorflow TFRecords and result in larger IOs
- TFRecord also generates 4KB IOs that make up ~20% of transfers



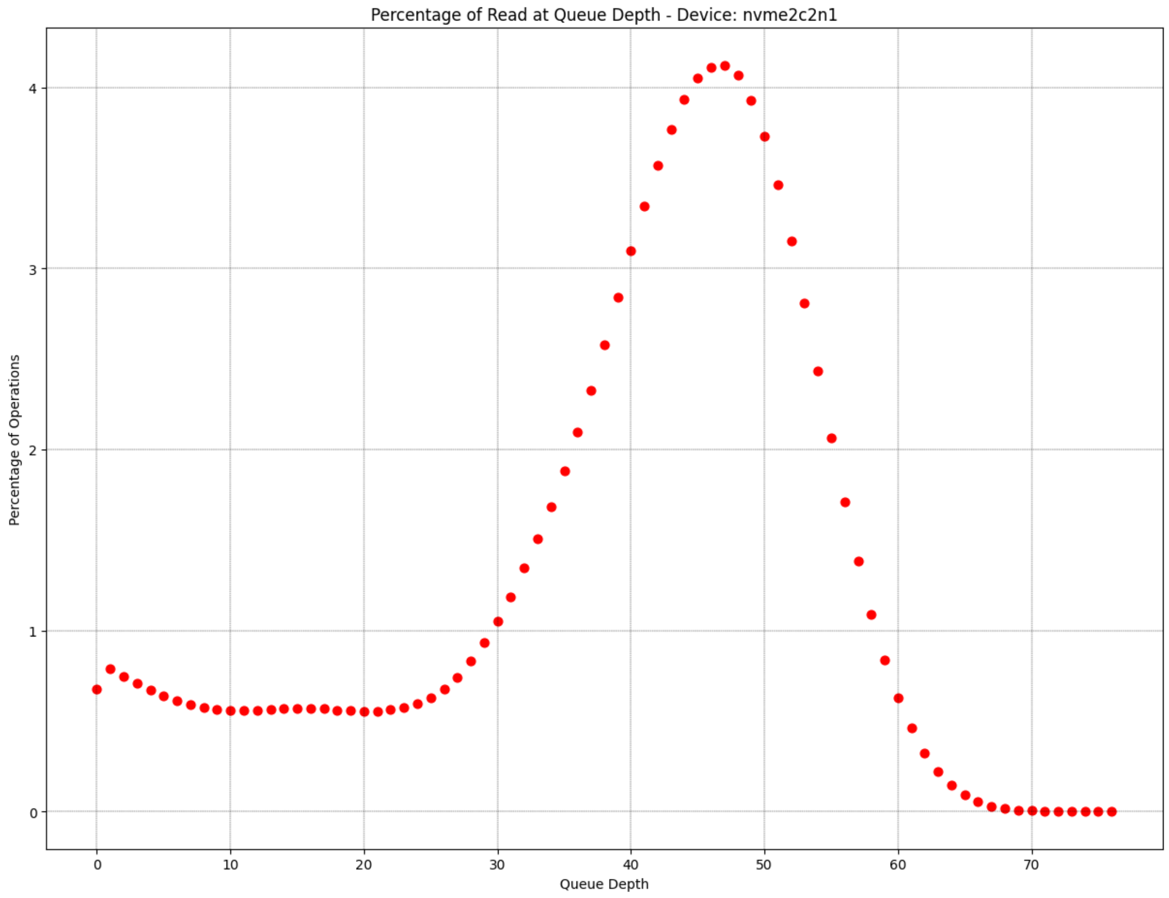


Queue Depth

- Queue Depth is a result of parallelism in the application or storage layer (like file system)
- Each workload generates similar throughput (4.5 – 5.5 GB/s)
- Each workload does large IOs (128k or larger for majority)
- Next slides discuss histograms of Queue Depth
- Each IO is traced with submission and completion times
- Post processing calculates position in queue where each IO was placed
- Histograms will show percent of IOs inserted at a specific Queue Depth

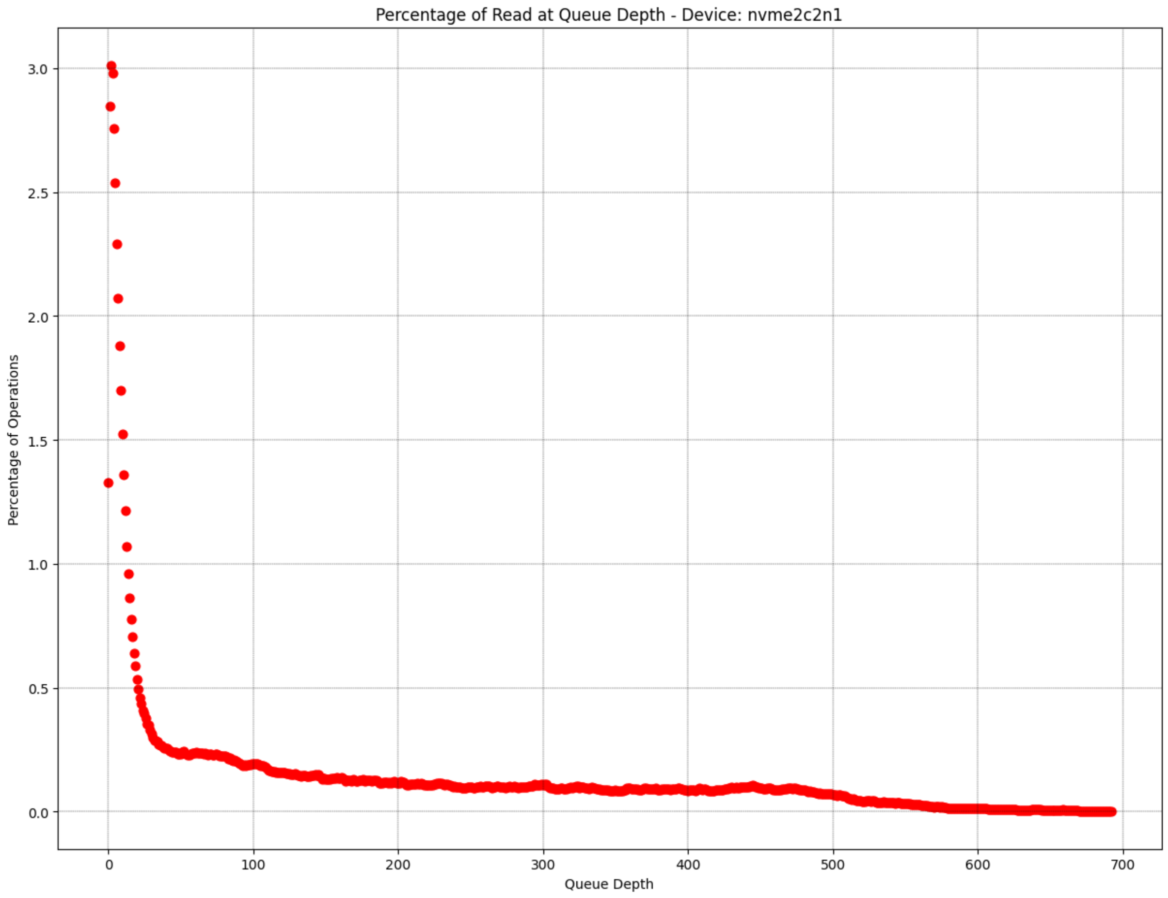
Queue Depth – Unet3D

- Large Peak from 40 to 52 QD
- Moderate number of Low QD IOs (0 to 10 QD is >5% of total IOs)
- Low QD IOs result in latency sensitivity instead of bandwidth sensitivity



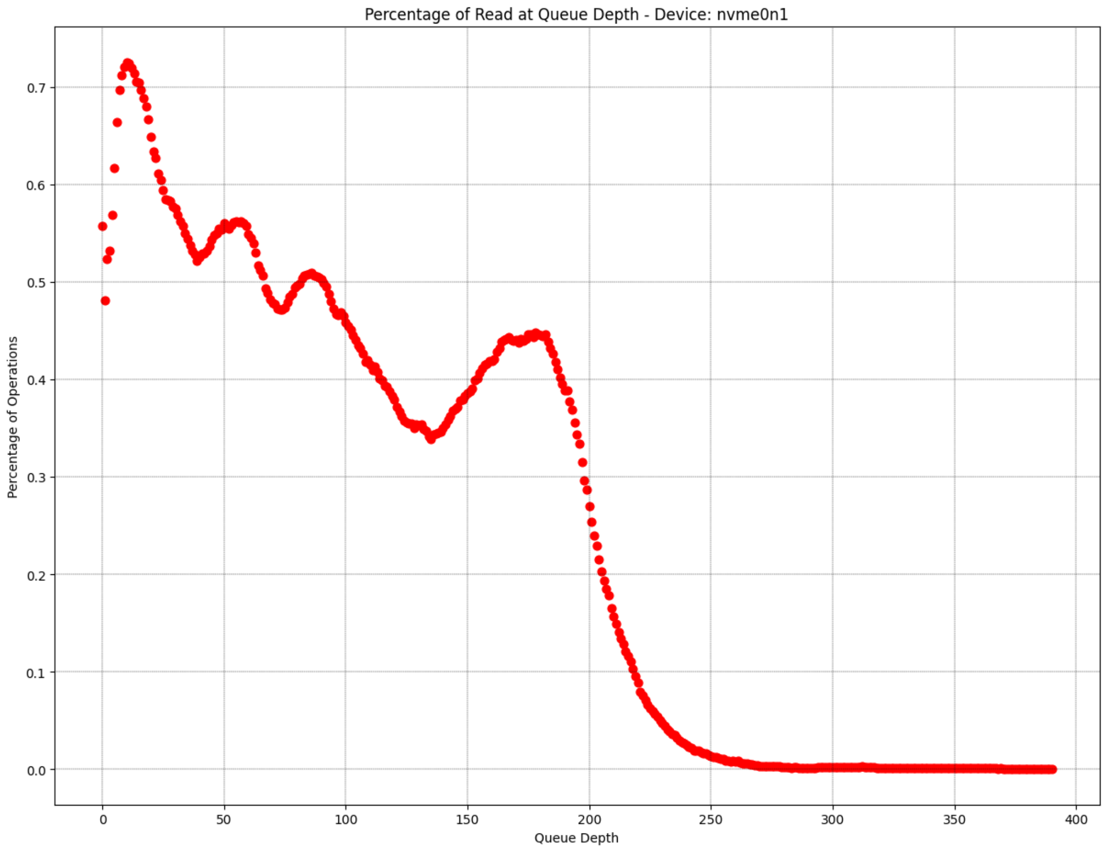
Queue Depth – Resnet50

- Large Peak from 0 to 20 QD
- Loooooong tail to QD 700

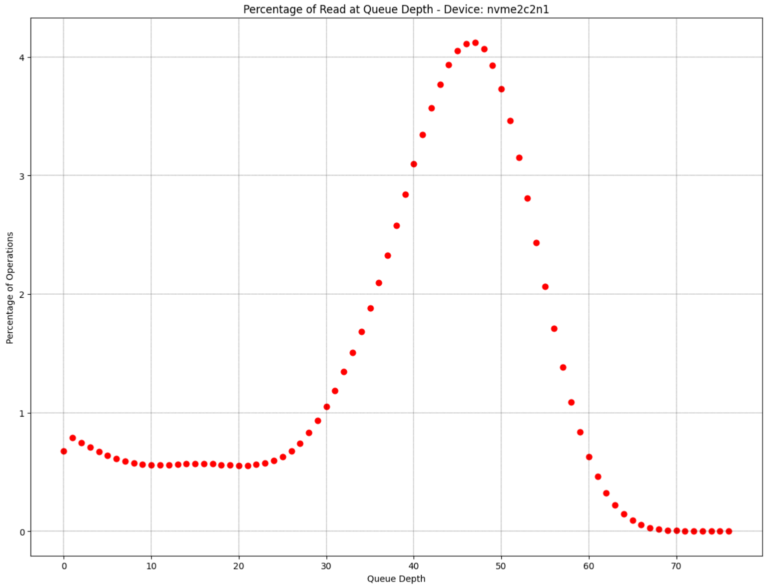


Queue Depth – CosmoFlow

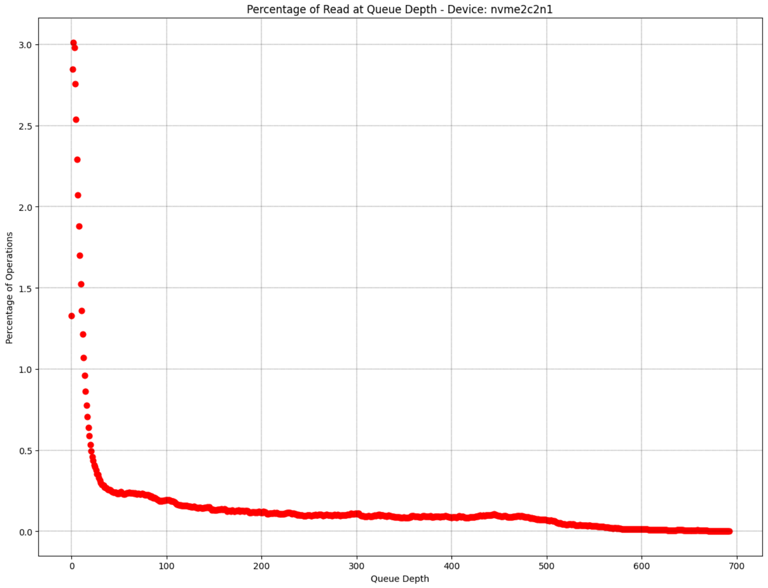
- Multi-modal histogram suggests complex behavior within the application.
- Concentration at low QD show this is highly latency sensitive



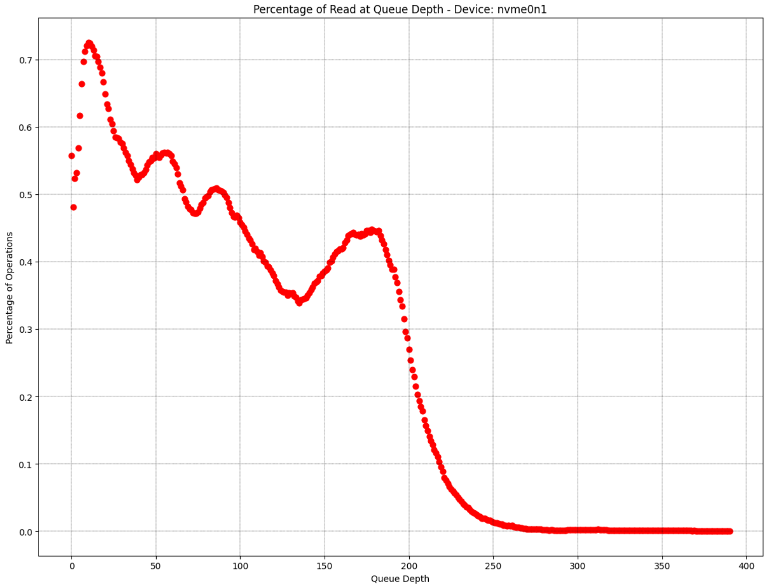
Unet3D



Resnet50



CosmoFlow

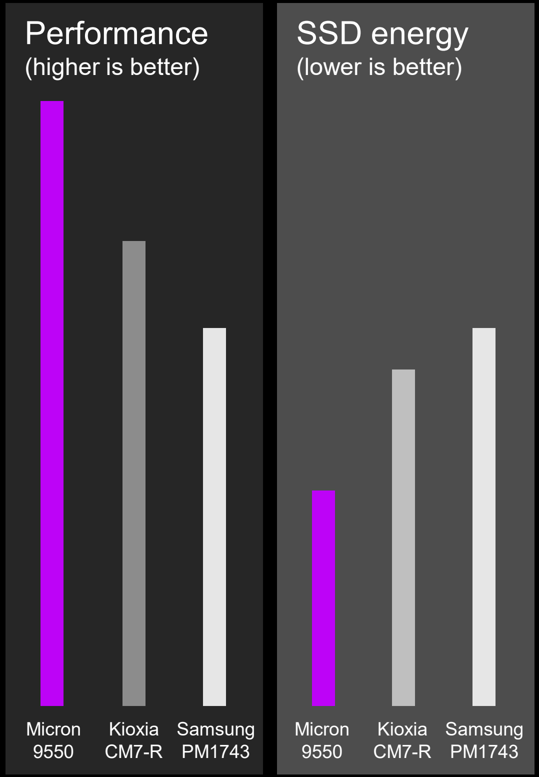


Take Aways:

- Benchmarking storage for AI is expensive and finding datasets is difficult
- MLPerf Storage enables easy(er) testing of AI applications for storage
- Three training workloads each generate significantly different loads to storage
- Highly recommend engaging with MLPerf Storage and using the benchmark in your own environments

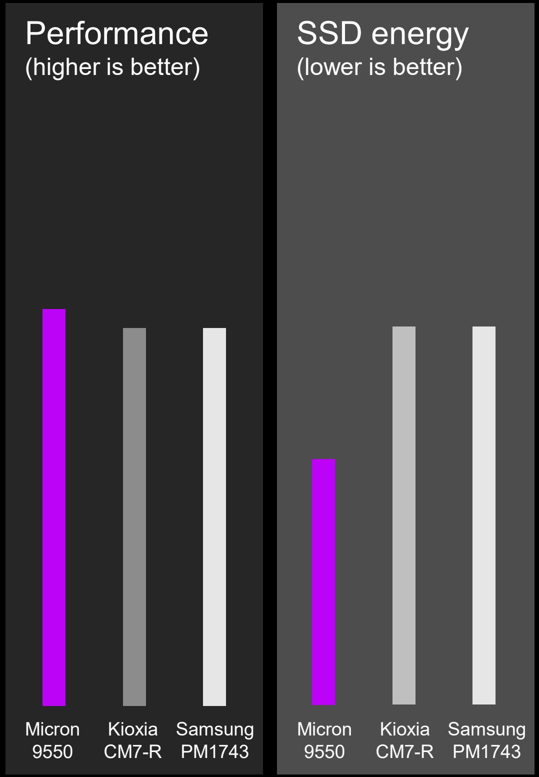
Micron 9550 – built for AI

Graph neural network training
(Big accelerator Memory)



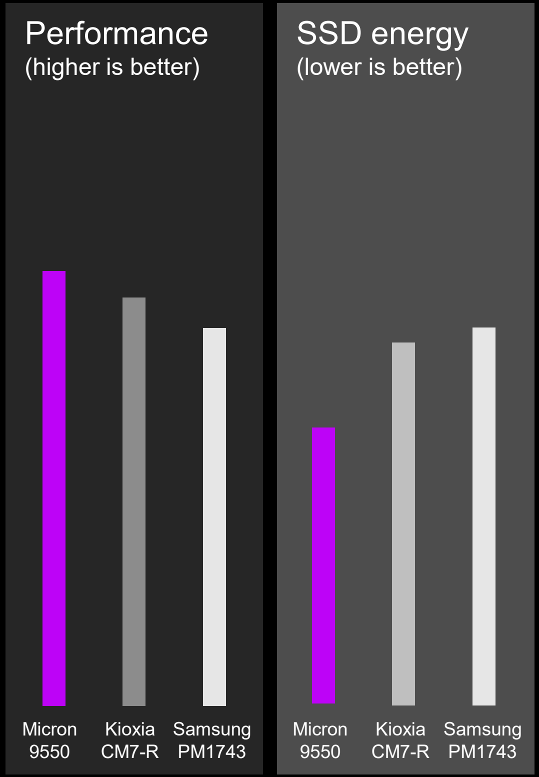
Up to
60% higher performance
43% less energy

Unet3D medical image training
(Deep learning IO)



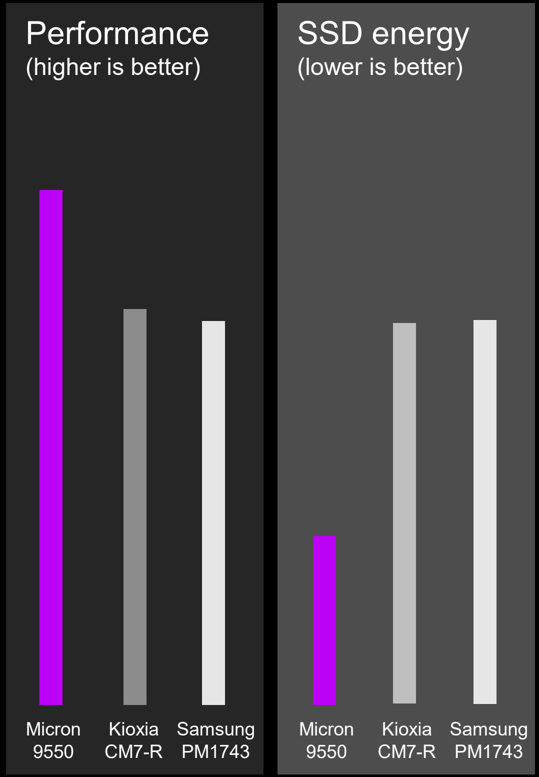
Up to
5% higher performance
35% less energy

Large language model inference
(DeepSpeed ZeRO-Inference LLM)



Up to
15% higher performance
27% less energy

NVIDIA GPUDirect® Storage



Up to
34% higher performance
56% less energy