Analyzing Workloads Using Storage as Memory Replacement

August 6th, 2024

John Mazzie
MTS, Systems Performance Engineer

Micron®

FMS the Future of Memory and Storage

Why do you need storage?

Why do you need storage?

- Models are getting too large to fit in GPU and System Memory
- Storage must be used in some way for large datasets during workloads
- Model Examples
  - Illinois Graph Benchmark Dataset
    - Graph Neural Network Model
    - Heterogenous 600M nodes
    - 2.3TB on Disk
  - Meta Llama3 70B
    - Large Language Inference Model
    - 70B Parameters
    - 142GB on Disk

NVIDIA Big Accelerator Memory

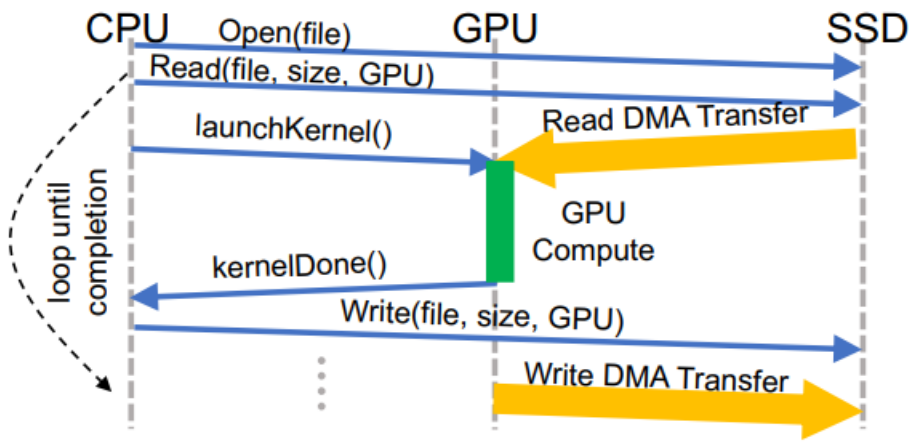
- Prototype project from NVIDIA Research
- Current method
  - Memory map storage to system memory
  - High CPU overhead
  - Low Performance
- New method
  - Proprietary driver
  - GPU coordinates all I/O transfers
    - CPU is completely bypassed
    - Leverages high parallelism of GPU to saturate PCIe bus between GPU and NVMe devices with fine grained I/O
    - High Performance

Microsoft DeepSpeed

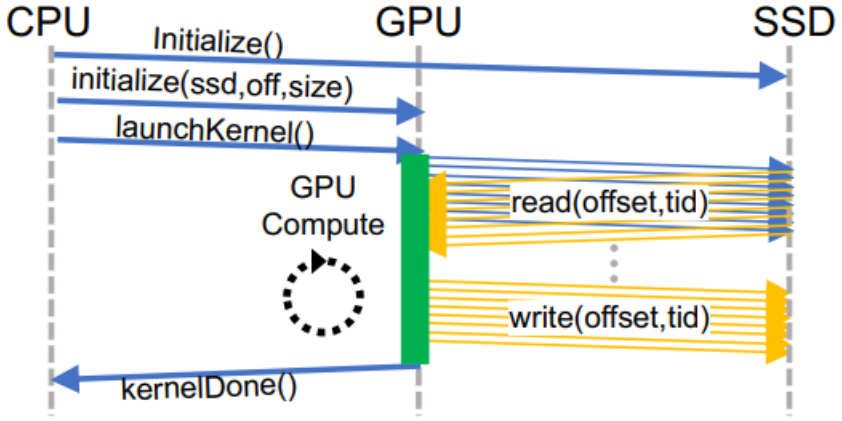
- Part of AI at Scale Initiative at Microsoft
- Collection of powerful memory and parallelism optimizations for efficient large scale model training and inference on modern GPU clusters
- Leverages heterogenous memory (GPU, CPU and NVMe) to scale
- Serves transformer-based PyTorch models
- DeepSpeed provides seamless inference mode for Hugging Face, Megatron and DeepSpeed trained models
  - No change required on the model side to work.

NVIDIA BaM Analysis

Big Accelerator Memory Model



(a) CPU-Centric Model.



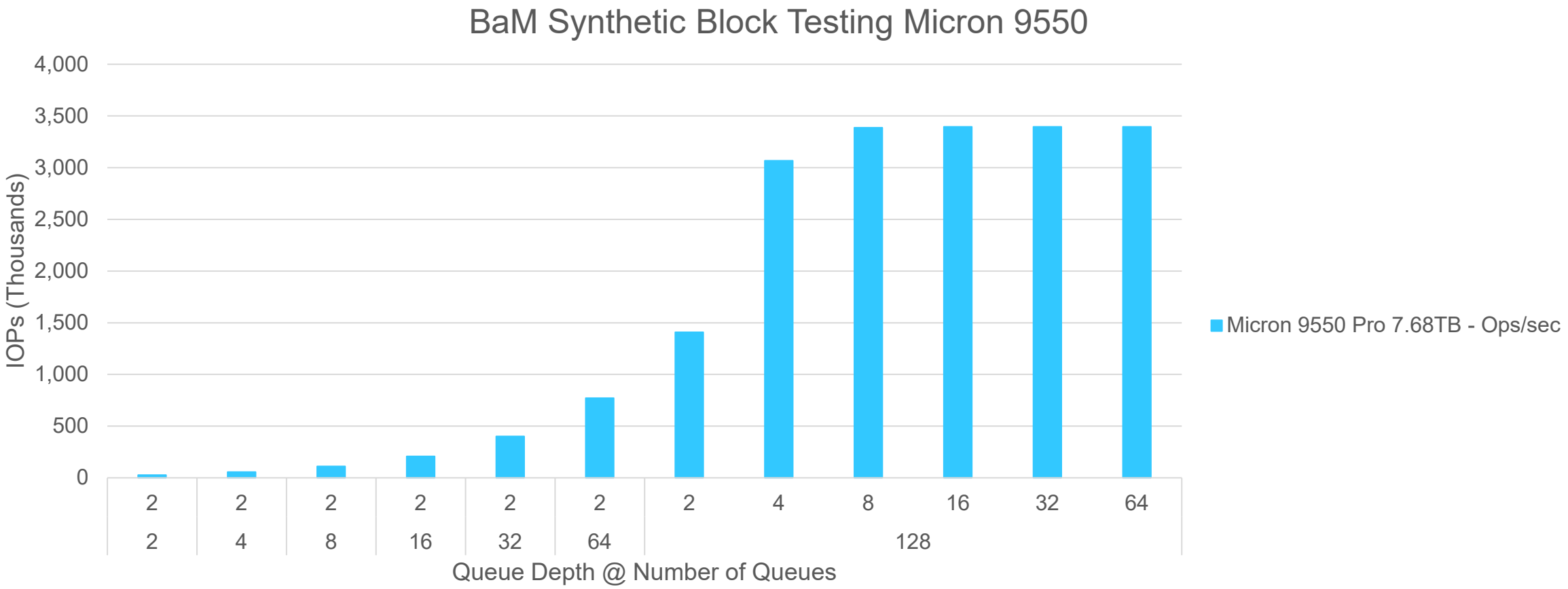
(b) BaM Model.

System Configuration

- System Configuration
  - SuperMicro SYS-512E-TNRT
  - Dual Socket Intel Xeon Platinum 8568Y+ (48 Cores/96 Threads)
  - 16x Micron 96GB DDR5 DIMMs
  - 4x Micron 9550 PCIe Gen5 NVMe SSDs
  - Ubuntu 20.04.6 (Kernel 5.4.0-180-generic)
  - NVIDIA H100 NVL GPU
  - Driver version 535.161.08
  - Cuda version 12.4
- Model (GNN Training)
  - Illinois Graph Benchmark Heterogenous 600 million nodes

BaM Synthetic Testing

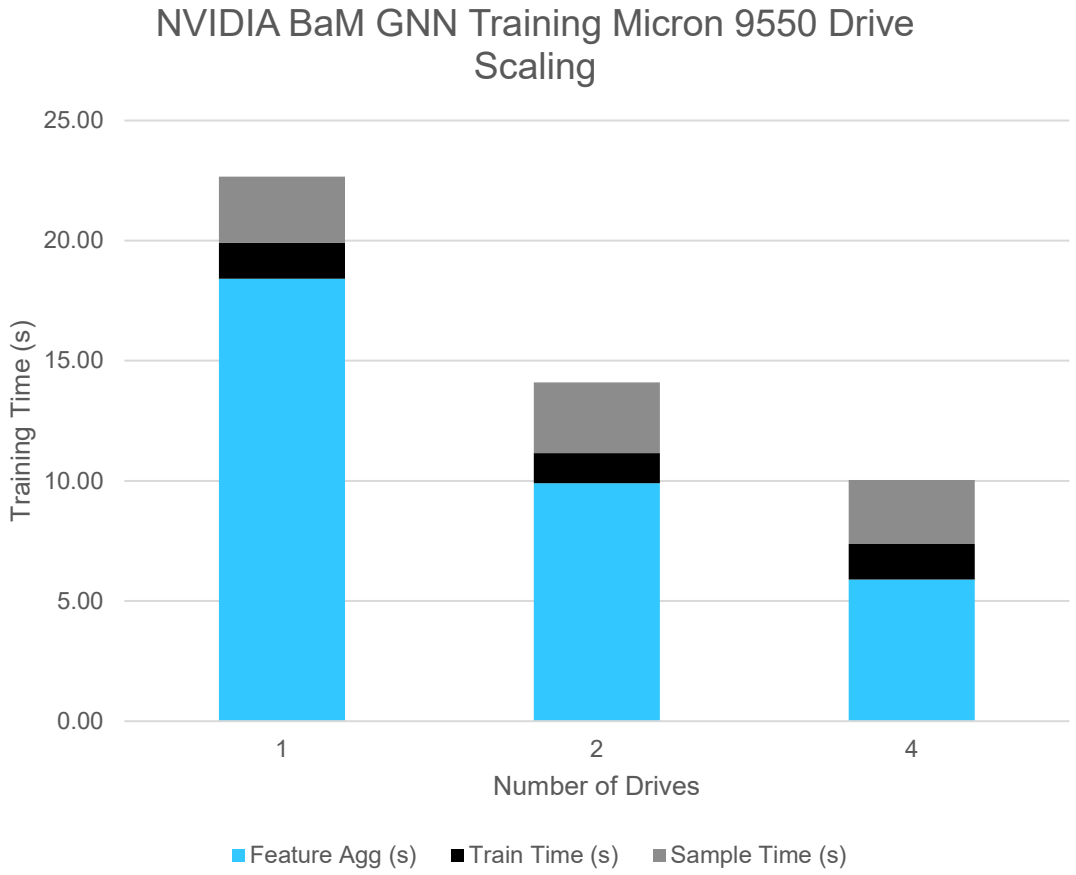
Queue Depth Scaling (Single Drive)



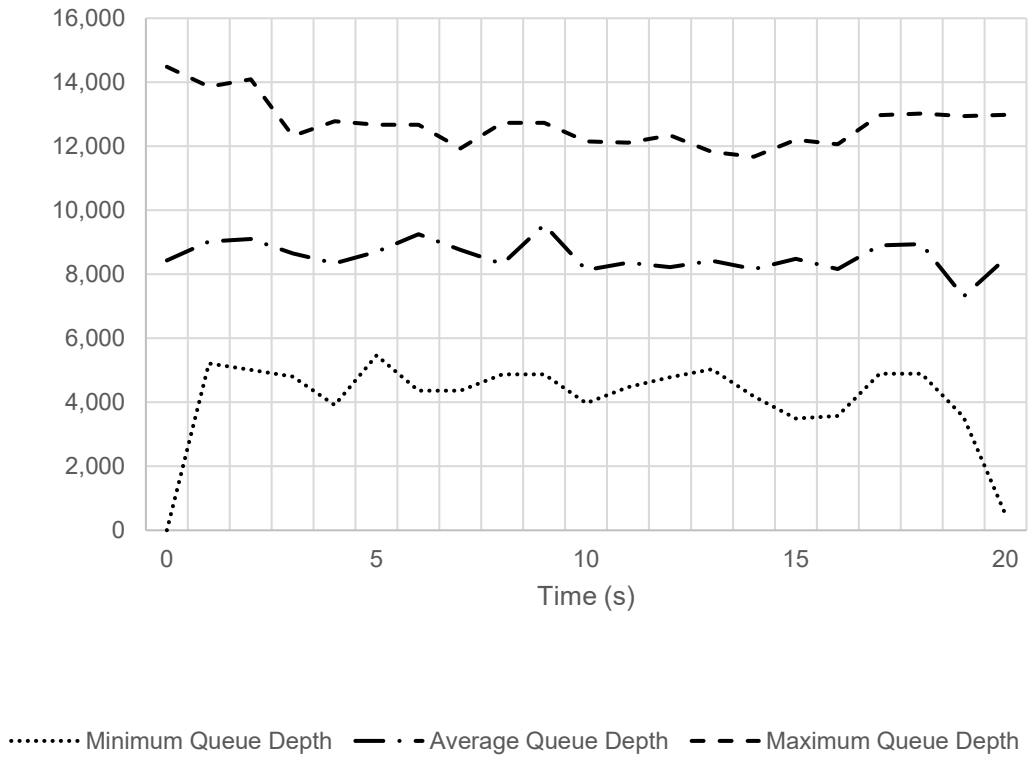
BaM GNN Training Results

4096 Batch Size - 8GiB Cache Size

	4x 9550s	2x 9550s	1x 9550	MMAP (Gen4 NVme)
Sampling	2.67	2.94	2.76	4.65
Feature Aggregation	5.89	9.9	18.41	1,130.12
Training	1.48	1.26	1.49	2.13
End-to-End	10.04	14.1	22.61	1,142.90

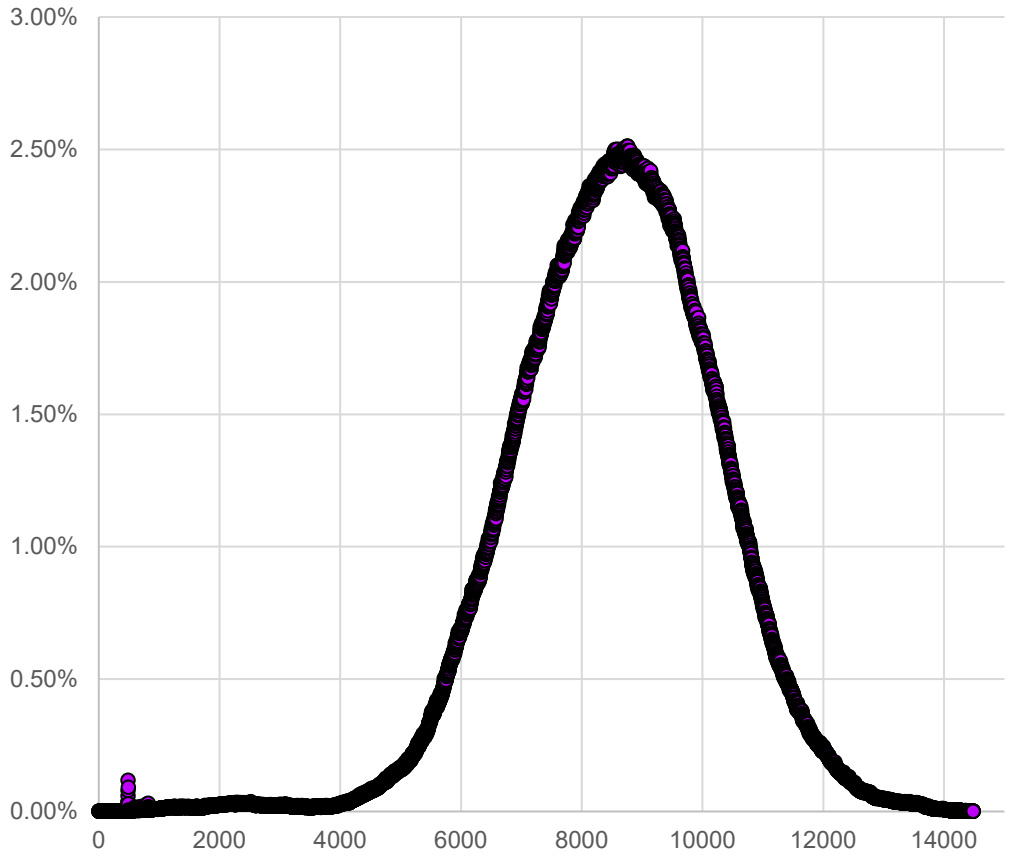


BaM GNN Training Queue Depth over Time



- Typical Enterprise Application
  - Order of 100 QD is considered high
- Big Accelerator Memory – GNN Training
  - Minimum: Between 4K and 6K
  - Average: Between 8K and 10K
  - Peak: Between 12K and 14K

BaM GNN Training Queue Depth Histogram



- Most of the time queue depth is between 6K and 12K
- Much higher queue depth than expected when designing NVMe devices
- Will this workload benefit from more queue pairs?
- What other optimizations can we make to handle these higher number of transactions?

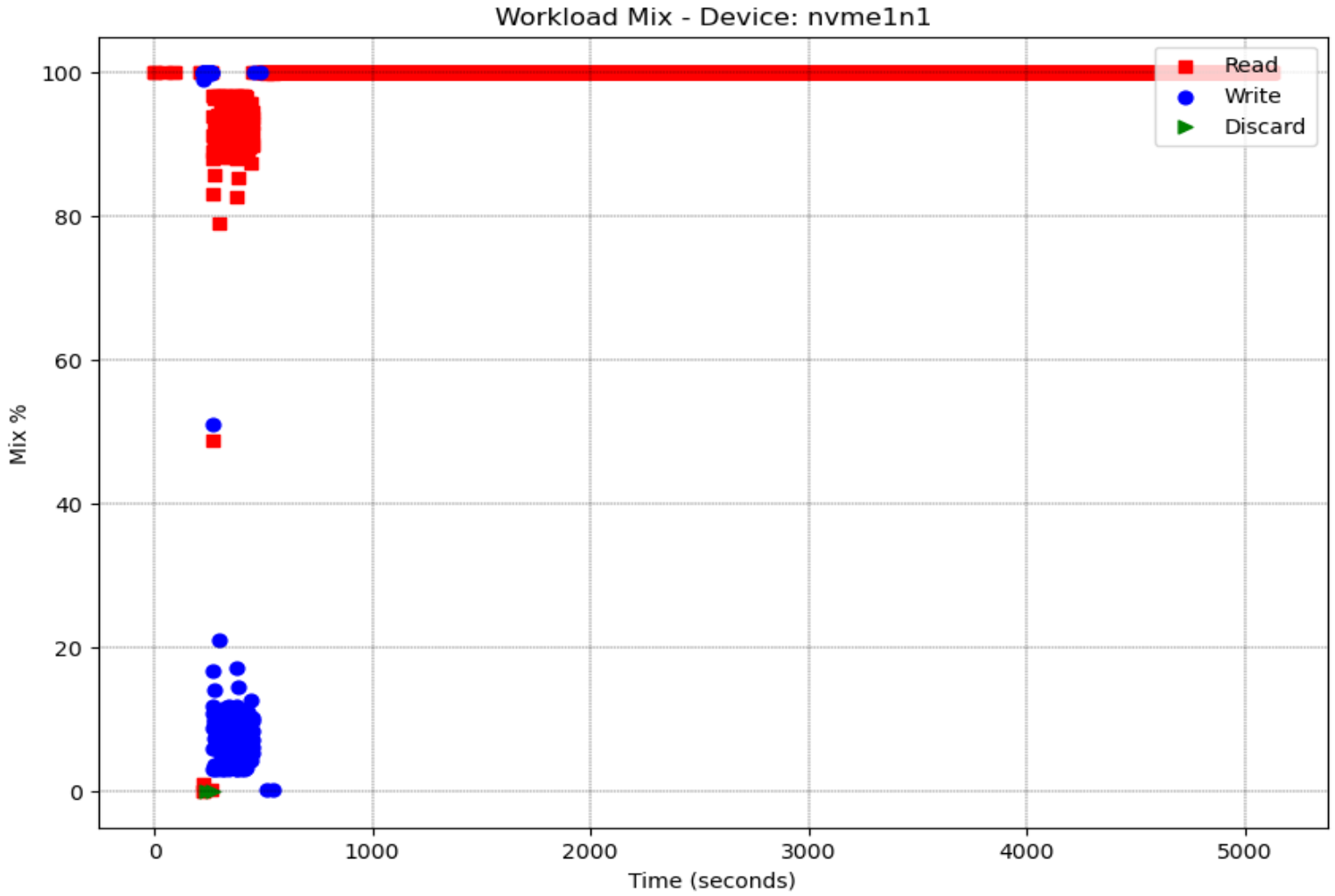
Microsoft DeepSpeed

System Configuration

- System Configuration
  - SuperMicro SYS-512E-TNRT
  - Dual Socket Intel Xeon Platinum 8568Y+ (48 Cores/96 Threads)
  - 16x Micron 96GB DDR5 DIMMs
  - 1x Micron 9550 PCIe Gen5 NVMe SSDs
  - Ubuntu 20.04.6 (Kernel 5.15.0-105-generic)
  - 2x NVIDIA L40S GPU
  - Driver version 550.54.15
- Model
  - Meta Llama 3 – 70b (70 Billion Parameters)

Large Language Model Inferencing with DeepSpeed ZeRO

- Model stored on NVMe
- Model weights and kv-chace offloaded to NVMe SSD
- Llama3 70b Model Zero Inference
  - 98% Reads
  - 2% Writes
- DeepSpeedZero first reads the model from disk (Reads) to GPU
- Updates the parameters on disk (Writes)
- Continues to read from disk during inference



Large Language Model Inferencing with DeepSpeed ZeRO

Drive Performance and Queue Depth

- Reads:
  - Max: 5880 IOPs / 1.6GiB/s
  - Avg: 1180 IOPs / 773 MiB/s
- Writes:
  - Max: 1440 IOPs / 996 MiB/s
  - Avg: 460 IOPs / 318 MiB/s