Deep Learning Performance and Cost Evaluation

Micron® 5210 ION Quad-Level Cell (QLC) SSDs vs 7200 RPM HDDs in Centralized NAS Storage Repositories

A Technical White Paper

Rene Meyer, Ph.D.

AMAX Corporation

Publish date: October 25, 2018

Abstract . . . 3

Introduction and Problem Statement . . . 4

Tool Setup . . . 4

Storage Drives Compared . . . 5

Table 1: Drive Specifications . . . 5

Test Environment and System Configuration . . . 6

Table 2: Server Specifications . . . 6

Results . . . 7

Table 3: Local drive performance comparison . . . 8

Table 4: Performance comparison of local and remote attached SSD vs.

HDD RAID5 storage array . . . 9

Figure 1: Read Bandwidth Comparison . . . 10

Conclusion . . . 10

Abstract

Near-instant access to training data is critical for most deep learning (DL) workloads to ensure that training durations are not negatively affected by data transfer times. Common practice is to use local NVMe SSDs as a data cache, while the storage back end is a traditional NAS solution using hard disk drives (HDDs). Data is streamed from the back end into the local NVMe for CPU-proximity processing. To update training data, only the back end is updated. Depending on the nature of the update, this process may be lengthy due to limited HDD ingest rate.

Growing data sets, increase in numbers of hidden layers in DL models, larger input data, as well as the need to share training data sets across multiple users and models, have resulted in the demand for a higher performing shared storage solution than HDD NAS can deliver.

In this whitepaper, we present a performance and cost comparison between a 64TB all-flash NAS array utilizing the industry's first quad-level cell (QLC) enterprise SATA SSD, the Micron® 5210 ION, and a 7200 RPM HDD-based array. Of particular interest was how a Mellanox Infiniband EDR remote attached QLC all-flash NAS array performs under DL specific workloads (supplying data directly, without local NVMe) and how that configuration compares to local NVMe flash. Compared to widely adapted three-level cell (TLC) SSDs, QLC SSDs offer a more approachable price point, but with reduced write endurance.

We find an 11x better performance of the QLC flash array over the HDD array for DL specific synthetic FIO tests as well as a significant acceleration of real-world DL applications, in some cases up to 40%. Test results are on par with local NVMe implementations, suggesting that NAS storage solutions built on read-intense QLC SSDs deliver substantial cost-effective performance gains for shared centralized training data repositories over HDD based solution. Given the read-intensive nature of deep learning use cases, reduced write endurance is not a concern, making QLC SSD solutions an attractive alternative to TLC SSD-based solutions.

Introduction and Problem Statement

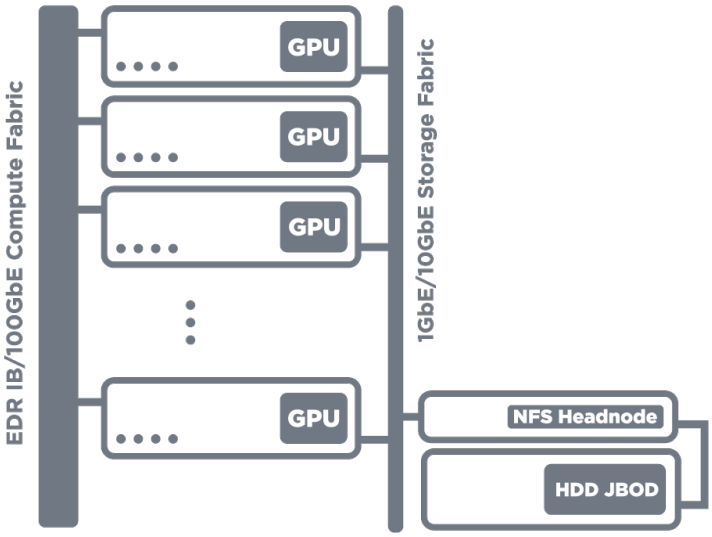
The availability of GPGPU compute accelerators, open DL frameworks and models, extensive data sets and databases combined with high industry demand for data intelligence solutions has triggered exponential growth in the artificial intelligence (AI) and DL segments of the IT market. Common examples are image/video stream recognition and analysis, speech to text, speech synthesis, and super-resolution, among many others. Applications include autonomous vehicles, automatic translation, social media, personal assistants, sports and training analytics, broadcasting, security and surveillance, threat detection, stock market prediction algorithms, consumer analysis, and countless others.

Deep leaning specific workloads can be classified into two categories: model development / training and model / inference deployment. Model development and training requires highly capable compute and storage resources. Compute is commonly GPGPU accelerated and fast access to storage is realized through local NVMe or SATA SSDs. Inference, on the other hand, can run on less specialized hardware and only relies on GPU accelerators in the case of a high number of concurrent users or if real-time analysis is required. Underpinning most training and development-related deep learning workloads is an extremely read-intensive storage IO profile, as vast collections of data have to be fed into the GPGPUs hundreds of times (the number of training epochs) to complete a model training and continuous iterations on model variants to improve model accuracy are trained using the same data sets. Experts note that the read-to-write ratio is as high as 5000:1, which is an abrupt departure from the traditional (pre-AI) data center read-to-write ratio of 4:1.

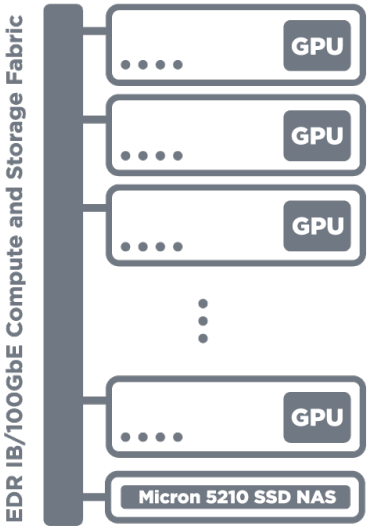
While GPU compute nodes are typically linked together via high-speed fabric – in many cases, 100GbE or EDR IB – the training data resides on an HDD NAS solution that's typically connected via 1GbE or 10GbE. To overcome the performance limitation of HDD NAS storage, training data sets are temporarily migrated to and stored on a local SSD (SATA or NVMe). Increasing deployment sizes, data set sizes, the move to larger models, parallel model training, and multi-user environments all favor a setup that enables faster access to all data in the centralized repository. Flash-based enterprise storage appliances provide fast access to data, but this has traditionally come at a significant price premium.

New QLC enterprise SSDs offer a compelling means to reduce costs, as QLC NAND stores 33% more bits per cell and delivers similar read-performance as traditional TLC-based SSDs. Because the Micron 5210 ION SATA SSD family is targeted as an HDD replacement option, this test compares the results of a 5210 ION deployment to a HDD deployment, while also putting the results in context with TLC-based NVMe all-flash configurations, which carry a significant cost premium.

HDD Based 1GbE/10GbE NFS Storage Solution



SSD Based EDR IB NFS Storage Solution



Test Setup

To study the performance difference between a 7200 RPM SAS HDD versus a Micron 5210 ION SSD solution in the context of DL workloads, we evaluate 3 different configurations: individual drives, local RAID5 volumes consisting of 16 drives and remotely attached NFS RAID5 16x drive volumes connected via EDR Infiniband fabric. For comparison, we also include a local NVMe drive.

We used an FIO test specific to DL workloads and a CNTK training are used as benchmarks. Note that while representative of the platforms and workloads tested in this study, the results presented may not be representative of all DL workloads.

Storage Drives Compared

Datasheet Specification	SSD: Micron® 5210 ION	HDD: SAS 7200 RPM
Form Factor	2.5-inch (7mm height)	3.5-inch (25.4mm height)
Interface	6Gb/s SATA	12Gb/s SAS
Capacity	3.84TB	8TB
Sequential Read (128K transfer)	540 MB/s	215 MB/s

Storage Drives Compared cont...

Sequential Write (128K transfer)	**350 MB/s**	**215 MB/s**
Random Read (4K transfer)	**83,000 IOPS**	**203 IOPS***
Random Write (4K transfer)	**6,500 IOPS**	**273 IOPS***
DWPD (5 year drive lifespan – rated endurance varies by workload)	**Up to 0.8 DWPD**	

Table 1: Drive specifications

Note: While HDD random IOPS performance specifications are not published in HDD datasheets; these values can be measured using the industry standard SNIA Performance Test Specification v1.1 IOPS test, which was used to derive these values.

Test Environment and System Configuration

The test environment consists of a GPU compute server and a storage server, both equipped with a Mellanox EDR IB network adapter and a JBOD attached to the storage server and an Infiniband EDR fabric. Flash drives are located in the front of the 24-bay storage server with passive backplane and direct connected though four 4-port HD mini SAS connector to an internal 16-port HBA. HDD drives are hosted by a 60 drive JBOD and attached to the storage node through an external 4-port 12Gb/s SAS HBA. Details of system configurations and software stacks are given below.

In this study, a 100G Mellanox EDR Infiniband fabric was chosen to maximize network bandwidth and minimize latency.

Component	GPU Compute Server	NAS Storage Server
CPU	**2x Intel® 6146 3.2GHz Scalable**	**2x Intel® E5-2680 v3**
DRAM	**24x 16GB DDR4 2666 MHz**	**8x 32GB DDR4 2666 MHz**
GPU	**8x Nvidia® V100 16GB**	**N/A**
SSD	**1x Micron® 9200 PRO 3.84TB**	**16x Micron® 5210 ION 3.84TB**

HDD	**N/A**	**16x 7200 RPM SAS 8TB**
HBA	**N/A**	**1x 9300-16i (SSD), 1x 9300-8e (HDD)**
Networking	**Mellanox® EDB 100G IB**	**Mellanox® EDB 100G IB**
Switch	**1x Mellanox® 36-port EDR Infiniband**	**1x Mellanox® 36-port EDR Infiniband**
Operating system	**Ubuntu 16.04**	**Ubuntu 16.04**
Test	**CNTK**	**N/A**
FIO version	**3.2**	**N/A**
NFS version	**NFS 4.2 plus RDMA**	

Table 2: Server specifications

The FIO script is shown below. A block size of 128K is chosen as a representative size for e.g. the ImageNet training database.

FIO script:

```
rw=randread
ioengine=libaio
iodepth=32
bs=128k
numjobs=1
size=1000G
direct=1
group _ reporting
```

Results

We first evaluate individual drive perform under deep learning specific workloads. The Micron 5210 ION SATA SSD drive outperforms the 7200 RPM HDD by more 14x in both read bandwidth – 479MB/s vs. 33MB/s – and read IOPS of 3747 IOPS vs. 258 IOPS.

For comparison, we add a locally attached Micron 9200 NVMe drive to the test to get a reference of a best practice implementation. The NVMe drive's read bandwidth, 3291MB/s, and read IOPS performance, 25100 IOPS, are about 7x higher compared to the SATA SSD and 100x higher compared to the SAS HDD. We will later compare the performance of the local NVMe to HDD and SSD-based remote attached shared storage solutions.

Table 3 below show the DL training results, which refer to the duration of the first training epoch. Here the data is loaded from the storage media prior to the training. These results show the impact of storage sub-system performance on the training duration and show how slow drive performance can extend training times. Training times of subsequent epochs may vary depending on if data is partially or fully cached in memory or on a local drive.

	Micron 9200 Pro SSD (NVMe)	Micron 5210 ION (SATA)	7200 RPM (SAS)
	3.84TB	3.84TB	8TB, JBOD
Single Drive Eval (local)			
DL training local drive (sec, per epoch)	850	930	5,536
BW, FIO, read, 128k random, MB/s	3,291	479	33
IOPS, FIO, read, 128k random	25,100	3,747	258
Ave. latency (ms)	1.3	8.5	123

Table 3: Local drive performance comparison

Next, we evaluate the performance of the local storage subsystems featuring 16x Micron 5210 SSDs and 16x 7200 RPM HDD drives in a RAID 5 configuration. On the HDD, 4TB of the 8TB total drive capacity is used to form the HDD RAID volume to match the capacity of the SSD volume. Both storage solutions have a raw capacity of around 64TB and a usable capacity of 53TB.

In Table 4 below, the 16x SSD volume displays a random read bandwidth of 3,075 MB/s and 24,600 IOPS at an average latency of 1.3ms. The performance is about 8x better than the spinning disk-based solution and on par with the local NVMe drive. Results obtained from the FIO benchmark test are summarized in the Table 4 below.

When accessing the storage solutions through an EDR Infiniband fabric via NFS protocol, the performance degrades slightly to 2,934MB/s for the 5210 ION-based storage volume and to 359MB/s for the HDD-based volume. The observed transfer rates are well within the bandwidth of the IB fabric and are expected to result from the overhead of the NFS software stack. Read IOPS performance is 22,400 IOPS for the SSD volume and 2,872 IOPS for the HDD volume. The average latency of the SSD volume increases from 1.3ms to 1.4ms.

	Micron 5210 ION SSD (SATA)	7200 RPM HDD (SAS)
	3.84TB	8TB, JBOD
16x Drive Eval (local, RAID5)		
BW, FIO, read, 128k random, MB/s	3,075	370
IOPS, FIO, read, 128k random	24,600	2,959
Ave. latency (ms)	1.3	10.8
16x Drive Eval (remote, NFS, 100G, RAID5)		
DL training (sec, per epoch)	858	1,248
BW, FIO, read, 128k random, MB/s	2,934	359
IOPS, FIO, read, 128k random	22,400	2,872
Ave. latency (ms)	1.4	11.1
Estimates Costs (System Level)	$302/TB raw*	$53/TB raw*

Table 4: Performance comparison of local and remote attached SSD vs. HDD RAID5 storage array

*2U Storage Node, 24x 8TB drives, $0.22/GB
**1U Head Node, 60x 8TB JBOD

In the studied configuration, the SSD based remote storage array exceeds the performance of a local SATA SSD and is on par with a local NVMe solution. Test results indicate commonly used 1GbE, 10GbE and even 25GbE storage network may not be sufficient to fully support the performance of the 5210 ION-based storage solution. Given the test results, an integration of the storage solution into an EDR Infiniband or 100GbE compute fabric or attachment via separate EDR Infiniband/100GbE storage fabric is recommended.

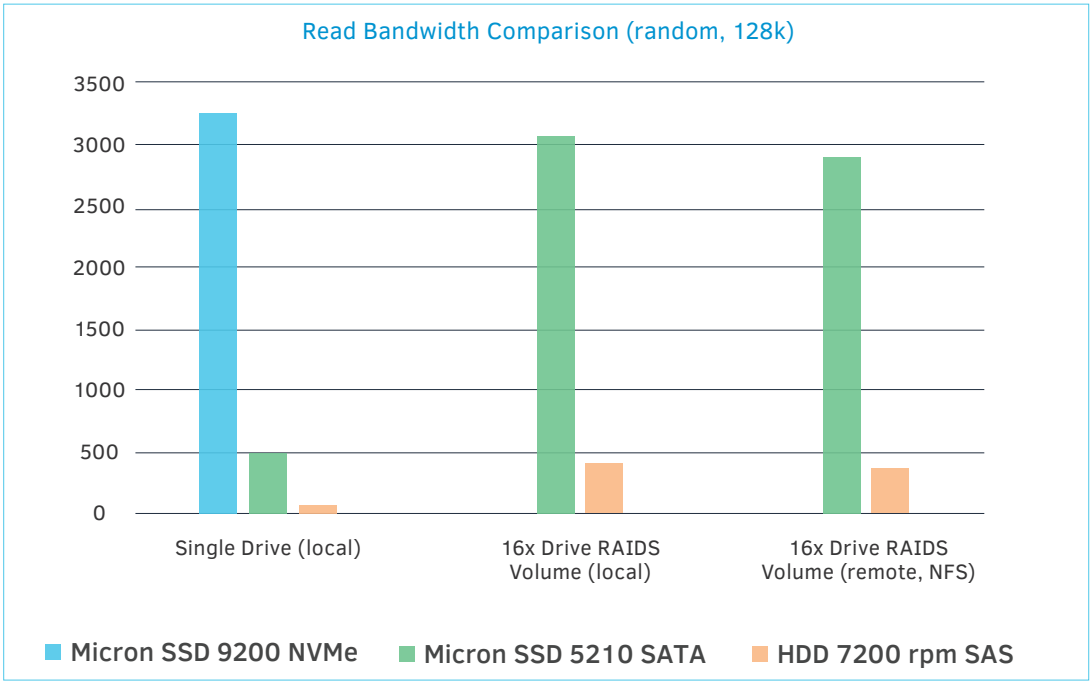


Figure 1: Bandwidth performance summary of local and remote attached SSD vs. HDD RAID5 storage array

From a cost perspective and performance perspective, the Micron 5210 based NFS shared storage solution is well positioned between a lower performing legacy HDD-based storage solution and expensive enterprise-grade all-flash storage appliances.

Conclusion

Our cost and performance comparison of new QLC SSDs vs. HDDs in a NAS array shows that QLC SSDs like Micron's 5210 ION deliver the high performance expected of SSDs – at a price point between that of an HDD NAS and that of a traditional all-flash array built on more expensive TLC (triple-level cell) technology. While TLC SSDs deliver more endurance and greater random write performance, the read/write ratio of most DL workloads may not justify the price premium, which is why this test was conducted on QLC SSDs.

After comparing a 64TB all-flash NAS array built on new QLC SSDs (Micron 5210 ION) to that of a 64TB 7200 RPM HDD-based solution, we find:

- The Micron 5210 ION QLC-based storage array is well suited for DL workloads as these workloads are read intensive (very low amount of write IO traffic) and benefit from high read performance. While many SSDs offer similar read performance, the Micron 5210 use of QLC NAND lowers the cost per GB, making it a more attractive option for this use.

- The QLC SSD NAS solution performs significantly better during normal operation and in single-bit parity RAID mode during the rebuild of a degraded volume over a traditional HDD-based NAS solution. Depending on the current setup and the particular workload, the QLC SSD-based NAS array can reduce training times, increase GPU utilization, increase development efficiency and streamline the development processes through fast centralized shared storage.
- We recommend integrating SSD NAS storage solutions into a 100G high-speed compute fabric instead of attaching it via a separate 1GbE/10GbE/25GbE storage network. The bandwidth realized by high-speed compute fabric attachment enables GPU compute servers to fully benefit from the performance of the storage solution.
- The performance of the evaluated 64TB QLC SSD storage solution is on par with local NVMe storage. Depending on the size of the deployment, we suggest reducing the capacity of the local NVMe drive or even replacing local with remote attached storage.

For more information and full specifications on the Micron 5210 ION SSD, visit www.amax.com/solutions/stormaxnfs/ or contact us at 1(800)800-6238.

Keep up to date with the latest from AMAX visit us: