Abstract

This ESG Technical Review documents validation of a series of performance tests executed by Micron Solutions Engineers on Hadoop nodes and AI/ML Servers using industry-standard Hadoop and TensorFlow benchmarking tools. The results validate that intelligent placement of Micron SSDs and memory can be used in a variety of applications to help accelerate every phase of the AI/ML workload lifecycle.

The Challenges

Never before has such a bounty of illuminating information been so readily available to guide organizations towards making well-informed decisions that result in quantifiable benefits to the business. According to recent ESG research, using data analytics for real-time business intelligence and customer insight was the IT initiative cited by the second highest percentage of respondents as being the most important to their organization in 2018, behind only strengthening cybersecurity.\(^1\) Traditional business metrics can be combined with a wealth of information generated through recorded metrics, online transactions, mobile devices, social media, and the Internet of Things (IoT), as well as through a diversity of other sources. As organizations continue to stockpile large amounts of data, the challenge of moving, analyzing, and generating insight from the data in a timely manner becomes increasingly important for a quickly changing business landscape. Machine learning and artificial intelligence allow for deeper and more thorough analysis of data and can extract correlations and insight from data that would be near impossible to accomplish with human generated queries. Benefits such as improved operational efficiency, more predictive insights into future scenarios or outcomes, the ability to deliver higher quality products/services, increased customer satisfaction, and lower risk top the long list of objectives that organizations reported they are expecting to accomplish from their investments in the area of ML and AI (see Figure 1).\(^2\)

![Figure 1. Top Ten Most Important Objectives Organizations Expect From Their Investments in ML & AI](image)

In terms of business benefits, what are the most important objectives you expect to accomplish from your organization’s investments in the area of ML & AI? Which is the single most important objective? (Percent of respondents, N=320)

<table>
<thead>
<tr>
<th>Objective</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved operational efficiency</td>
<td>38%</td>
</tr>
<tr>
<td>More predictive insights into future scenarios or outcomes</td>
<td>34%</td>
</tr>
<tr>
<td>Deliver higher quality products/services</td>
<td>33%</td>
</tr>
<tr>
<td>Better customer satisfaction</td>
<td>32%</td>
</tr>
<tr>
<td>Reduced risk around business decisions and strategy</td>
<td>28%</td>
</tr>
<tr>
<td>More insights into historical results</td>
<td>28%</td>
</tr>
<tr>
<td>Better sales/marketing performance</td>
<td>26%</td>
</tr>
<tr>
<td>Improved ability to satisfy compliance requirements</td>
<td>26%</td>
</tr>
<tr>
<td>More efficient/effective R&amp;D</td>
<td>25%</td>
</tr>
<tr>
<td>Improved/fortified cybersecurity</td>
<td>25%</td>
</tr>
</tbody>
</table>

Source: Enterprise Strategy Group


This ESG Lab Review was commissioned by Micron and is distributed under license from ESG.
© 2018 by The Enterprise Strategy Group, Inc. All Rights Reserved.
Machine learning and artificial intelligence undoubtedly provide justifiable benefits but solutions can be costly, and, even more importantly, the time spent ingesting data into systems, transforming the data, and performing deep machine learning analysis can add up and become the bottleneck in the process of generating near-real time actionable insight.

**The Solution: Micron SSDs and Memory for AI/ML Systems**

Artificial intelligence (AI) enables insights from knowledge and uncovers value rapidly and at scale. Micron memory and storage have been foundational in AI’s transformation to highly adaptable, self-training, ubiquitous, machine-learning systems for mainstream use. Micron fast, vast storage and high-performance, high-capacity memory and multi-chip packages power AI training and inference engines, whether they are in the cloud or embedded in mobile and edge devices. Micron innovation accelerates AI to enrich business and private lives beyond what we can yet imagine.

Micron memory products can be leveraged by devices and systems in just about every phase of the AI/ML lifecycle, including:

- **Creation of Data**: Micron DRAM, NAND Flash, and memory cards can be deployed in the devices where data is initially created, including mobile and IoT devices, cameras, tablets, laptops, and servers to provide dense, fast, and reliable memory, ensuring less data is lost at the point of creation.

- **Internet Infrastructure**: Micron DRAM and SSDs are leveraged in many cloud-based servers and appliances, helping to reliably and quickly deliver data across the Internet to public or private clusters where data can be transformed.

- **Data Storage and Transformation**: Data storage nodes, such as for a Hadoop cluster, can deploy Micron DRAM and SSDs to accelerate data movement (ingest into the data store, and egress from the data store into AI/ML servers) as well as reduce the time required to transform data from its raw format into a usable state for AI/ML algorithms.

- **AI Training and Inference**: Micron memory and SSDs leveraged in AI/ML servers, as well as in partner GPU products, can help greatly accelerate staging and training time. Micron products help ensure that AI training pipelines provide data quickly to memory for processing while avoiding potential bottlenecks for CPU-based workloads. They also provide the data speeds required to supply the massive concurrency requirements of high-performance GPU processing.

![Figure 2. Micron Memory Products Can Help Accelerate Every Device and System in the AI/ML Lifecycle](source: Enterprise Strategy Group)
ESG Lab Tested

ESG audited performance testing conducted by Micron that demonstrates and quantifies some of the many ways that intelligent placement of Micron memory products can help accelerate common systems used to support the lifecycle of AI and ML workloads. The tests showed how Micron SSDs and memory products can help to:

- Load data into a Hadoop cluster from an outside source faster.
- Transform data in a Hadoop cluster into a usable state for AI/ML algorithms in less time.
- Speed the ingest of data from a Hadoop cluster into an AI/ML training algorithm.
- Leverage GPUs, such as those offered by Micron’s partners built with Micron memory products, to significantly increase and scale the performance of AI/ML training algorithms.
- Increase performance of training algorithms by using additional Micron memory for CPU- and GPU-based workloads.
- Leverage Micron SSDs to provide the extreme level of I/O bandwidth required to feed the high-performance CPU- and GPU-based workloads.

Accelerating Data Ingest and Processing Performance of Hadoop Nodes

The first set of tests, performed on a Hadoop cluster, demonstrated how Micron SSDs can help to accelerate ingesting data from another source into a Hadoop datastore, minimize the time required to transform data into a state that is usable for AI/ML algorithms, and speed the ingest of transformed data from the Hadoop cluster into an AI/ML algorithm. The Hadoop data nodes for the testbed included:

- **Nodes**: 4 x Supermicro Servers, each with Dual Intel Xeon Gold 6142 CPUs @ 2.6GHz (16 cores, 32 threads).
- **Memory**: 384 GB of Micron DDR4-266 memory per node.
- **Network**: Mellanox Connectx-4 100Gbe.
- **HDD Storage**: 8 x 8TB 7.2K HDDs per node (32 Total).
- **SDD Storage**: 8 x 5210 ION 7.6TB per node (32 Total).

In addition to the four data nodes, the Hortonworks HDP 3.0 big data configuration consisted of VMs including a primary and secondary HDFS name node, four YARN name nodes, resource manager, AppTimeline service, Mapreduce2, and a Zookeeper server. The big data infrastructure VMs were housed on a single Dell R630 server with Dual Intel E5-2690v3 12-core processors @ 2.6Ghz running the KVM virtualization hypervisor. The server was populated with 512 GB of Micron DDR4-2666 memory and two Micron M500DC 800GB SSDs in RAID1.

Performance tests were driven by the built-in Hadoop benchmarking tools and a 2TB data set. The data set size was designed to be larger than the 1.5 TB of total cluster memory to avoid caching of the entire data set and stress different components of the system. All tests were run on 2TB data sets created and run on both HDDs and then on SDDs a total of three times each. Results were then compared for any anomalies and blended into an average across all three runs for each of the disk technologies.

The first benchmarks were performed using TestDFSIO-Read and TestDFSIO-Write to simulate data being ingested into the Hadoop cluster from an outside source for storage and processing (TestDFSIO-Write) and streamed out of the Hadoop cluster into a system that is running an AI/ML algorithm. The tests measured the amount of time required to move a 2TB data set into (writes) and out of (reads) the HDFS filesystem with X3 replication on both the HDDs and Micron SSDs. The results (shown in Figure 3) validated that the Micron SSDs ingested data into the cluster 1.9X faster than the HDDs and provided data to the AI/ML algorithm out of the cluster at a rate that was 3.7X faster.
ESG also validated the results of similar tests using the TeraGen and TeraSort benchmarks to first generate a 2TB random data set into the Hadoop cluster and then sort this data using a map/reduce operation. While these tests take significantly longer to complete than the pure read and write tests shown above due to the work being performed, completion times of these operations are very dependent on the read and write speed that the storage media can deliver to complete them, and thus the results demonstrate near identical improvements to the ingest charts above (1.8X faster TeraGen completion, and 3.7X faster TeraSort completion). The results of TeraGen and TeraSort benchmark tests are shown in Figure 4.

Finally, ESG validated the results of benchmarks that perform significant Hadoop operations on a randomly generated 2TB set of data, resulting in sorted data that has been preprocessed (or transformed) into a format that is ready to be ingested for use by an AI/ML algorithm. The test first uses Randomtextwriter to create a data set by writing out a 2TB list of randomly generated words. Next, the functions Wordcount, Sort, and WordAggHist run map/reduce programs to count the words in the input files, sort the data, and compute the histogram of the words in the input files. The results of the individual components of this test are shown in Figure 5.
Why This Matters

The actionable insight gleaned from AI/ML algorithms can have a significantly positive influence on business decisions, operations, and risk reduction for an organization. But almost all data sources cannot be ingested into AI/ML algorithms until the data has been sorted and processed into a format that is usable by these algorithms.

ESG validated that Micron SSDs can be used in Hadoop nodes in place of HDDs to improve the speed at which data can be ingested, processed, transformed, and fed into AI/ML algorithms. The combined savings means that insights relying on TBs of information can be made available to an organization dozens of hours to days earlier when using Micron SSDs instead of HDDs in Hadoop data nodes. This can result in a significant business advantage, operational efficiency, and reduction in risk for an organization.

Increased Performance for AI/ML Training on CPU-based and GPU-accelerated Workloads

AI/ML self-learning algorithms rely on neural networks that must be trained for specific use cases and require a significant amount of massively parallel processing power to perform deep learning. These workloads can be run on high-performance memory-intensive CPUs but benefit greatly by running on GPUs that can handle millions of operations in parallel. High-performance memory products from Micron can provide significant performance increases in AI/ML servers whether they are in the form of flash memory, SSDs, DRAM, or memory built in to leading partners’ GPU products.

To validate some of the ways that GPUs and Micron memory products can help accelerate AI/ML workloads, ESG reviewed the results of tests performed using the ImageNet data set and ResNet50 image classification model provided by the TensorFlow open source machine learning framework. The tests were performed on an NVIDIA DGX Station configured with a 20-Core Intel Xeon 2698 CPU @ 2.2GHz, 256 GB of Micron DRAM, and 4x NVIDIA Tesla V100 GPUs (a total of 64 GB of...
The GPUs provide 500 TFLOPS of processing power from their 2,560 Tensor Cores and 20,480 CUDA cores. To provide adequate I/O bandwidth to keep data flowing through the AI/ML workload pipeline, 4x Micron 5200PRO 2TB SSDS were configured in RAID10 to store and provide access to the data set.

The ImageNet data set, roughly 150GB in size, was downloaded and preprocessed into the TensorFlow record format prior to running the benchmark. The performance was measured in terms of images per second. ESG reviewed the results of preliminary testing to understand the correlation between batch size (number of images to process before back-propagating the learned information), number of batches, and images per second. Tests were run using the batch size that effectively utilized all of the available GPU memory (batch size 128) or system memory (batch size 2048), and thus provided the optimal performance for the technology under test. This test, summarized in Figure 6, shows that increasing the amount of DRAM for CPU workloads, or DRAM and Onboard GPU memory for GPU workloads, results in increases in the performance of the AI/ML algorithm.

The testing also revealed the benefit of running AI/ML workloads on GPUs versus running the same workload on CPU resources. Using the same test methodology running at the optimal tuned batch size for each technology, ESG validated that a single NVIDIA Tesla V100 GPU processed 43.5X more images per second than running the workload on the 20-core Intel Xeon 2698 CPU @ 2.2GHz. We also validated that GPUs could be added to the environment to produce very near-linear scaling. Four GPUs were able to process 3.7X more images per second than a single GPU, and 160X more images per second than the Intel CPU. It should be noted that only four Micron 5200PRO 2TB SSDS were able to provide ample I/O performance to supply data for the AI/ML algorithm to work on, even when processing over 1,400 images per second. The results of these tests are summarized in Figure 7.
Why This Matters

The massively parallel processing power of GPUs provides an attractive economical solution for the CPU-intensive calculations required to run today’s AI/ML algorithms. ESG validated that $36K of GPUs (four GPUs) could process over 1,400 images per second—a workload that would require over $560K of CPU processors to achieve (a cost savings of 94%).

ESG also validated that performance of AI/ML algorithms can be improved by selecting larger sizes of Micron memory built into GPUs (33% improvement seen), or by adding Micron DRAM to servers (208% improvement seen). However, it is important to remember that high-performing AI/ML servers require adequate I/O performance to provide the data for the algorithm to work on. ESG noted that just four Micron 5200PRO 2TB SSDs in R10 provided data for the processing of over 1,400 images per second.

The Bigger Truth

The value of near real-time insight is undeniable. Organizations that are able to process and extract information about their business in shorter time are better positioned to extract value from potential business opportunities and can react quicker to optimize operations and reduce risk for the company and for their customers. New data is being generated at an enormous rate—but to understand and leverage this data, organizations must move it to the right systems and transform it into a usable format before running their AI/ML workloads to extract insight. Opportunities exist to speed up systems driving every phase of the AI/ML lifecycle, and reducing the time spent in each phase ultimately provides information to key decision makers faster—resulting in a clear advantage for almost all aspects of any operation.

ESG validated the results of performance testing that demonstrates that intelligent placement of Micron memory technologies into Hadoop servers resulted in faster loading and processing of data into a usable format for AI/ML algorithms and allowed for quicker ingest into AI/ML algorithms. Selecting GPUs that leverage greater quantities of Micron memory and adding more Micron DRAM and SSDs to AI/ML workstations resulted in the ability to process data faster. By making the data available to the AI/ML algorithms sooner, feeding the data to the system faster, and taking less time to perform computations, AI/ML algorithms work off of closer to real-time sets of data—providing more accurate and timely insight. For larger sets of data, this total time savings is not measured in minutes, but rather in tens of hours or even days—enough time to make a significant improvement to the security and performance of the business.

Micron does not make the systems that you use to store, transform, process, and extract insight from your data. But the speed, density, efficiency, and reliability of the memory products used in these systems play an irrefutably critical role in nearly every aspect of the AI/ML lifecycle. Micron memory products are well positioned to play a key role in improving every stage of the AI/ML lifecycle, from generation of the data on mobile and IoT devices, through data storage and movement, and big data transformation, ultimately to computational performance of neural networks.

The data that your organization collects is valuable. The memory products deployed in the systems that are required to process and extract insight from this data play a critical role in minimizing the time it takes to extract actionable insight from your data. Micron understands the challenges faced by organizations when deploying high-performance solutions and is ahead of the curve when it comes to providing memory solutions that can improve every phase of the AI/ML workload lifecycle. ESG suggests that, regardless of the technologies you decide to deploy, you contact Micron to see how they can help accelerate the AI/ML lifecycle and minimize your time to actionable insight.
The goal of ESG Validation reports is to educate IT professionals about information technology solutions for companies of all types and sizes. ESG Validation reports are not meant to replace the evaluation process that should be conducted before making purchasing decisions, but rather to provide insight into these emerging technologies. Our objectives are to explore some of the more valuable features and functions of IT solutions, show how they can be used to solve real customer problems, and identify any areas needing improvement. The ESG Validation Team’s expert third-party perspective is based on our own hands-on testing as well as on interviews with customers who use these products in production environments.