Technical Validation

Intel Data Center SSD Storage for the AI Data Pipeline

Making Storage Infrastructure 'AI-ready' with Intel SSDs for the Data Center

By Brian Garrett, Vice President, Validation Services; and Tony Palmer, Senior Validation Analyst

April 2020

This ESG Technical Validation was commissioned by Intel and is distributed under license from ESG.

Contents

ESG Technical Validations

The goal of ESG Technical Validations is to educate IT professionals about information technology solutions for companies of all types and sizes. ESG Technical Validations 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.

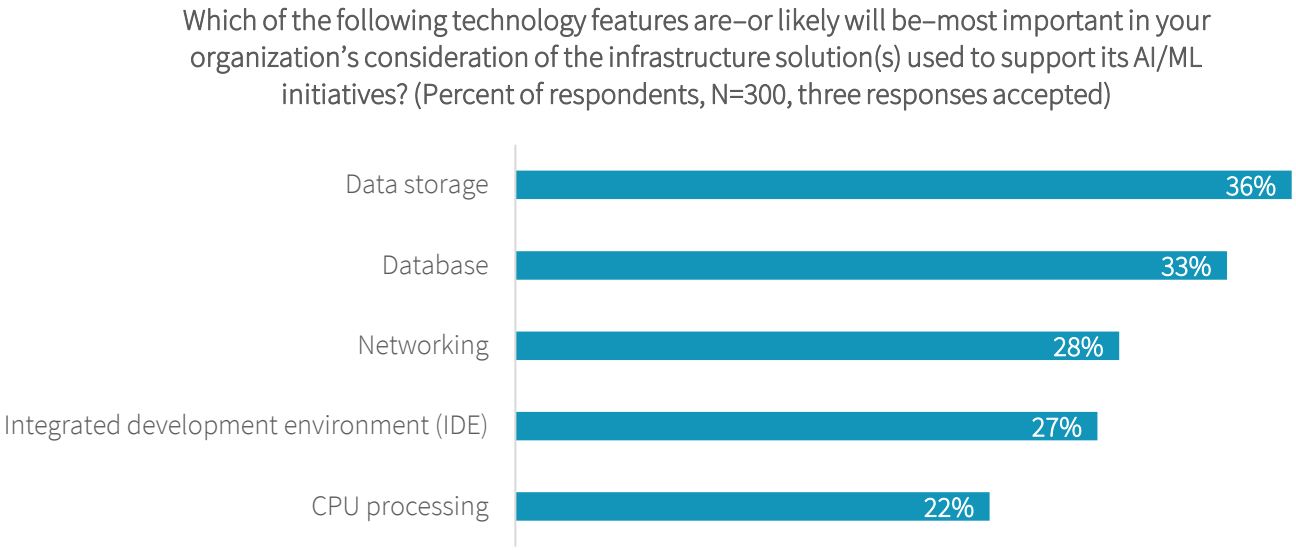
Introduction

This ESG Technical Validation explores the performance and efficiency benefits of Intel Data Center SSDs as a storage media foundation for all of the stages of the AI data pipeline.

Background

ESG research on AI and machine learning (ML) reveals that AI and ML are already widely adopted, and are marching toward ubiquity, with 50% of respondents confirming that they have already adopted AI/ML technology, and 50% expecting to use AI/ML within 12 months of the survey.[1] ESG also asked organizations which technology features are—or likely will be—most important in their consideration of the infrastructure solutions used to support AI/ML initiatives. As seen in Figure 1, storage is the most cited technology consideration for AI infrastructure.

Figure 1. Top Five AI Technology Features Prioritized for AI Infrastructure



Source: Enterprise Strategy Group

With the rising demand for instantaneous, actionable insight, organizations have found that the adoption of data lakes is an effective approach for anchoring modern data platforms that look to serve multiple business units and processes that cross business unit domains. ESG asked organizations about data lake usage and 60% responded that they are either already using a data lake, are planning to implement one, or are evaluating the technology. When asked about their objectives for utilizing a data lake technology solution, the most cited response was to improve scalability, cited by 39% of responses.[2] Clearly, AI/ML is fast becoming a business-critical component of IT, and highly performant and scalable storage is a key to success.

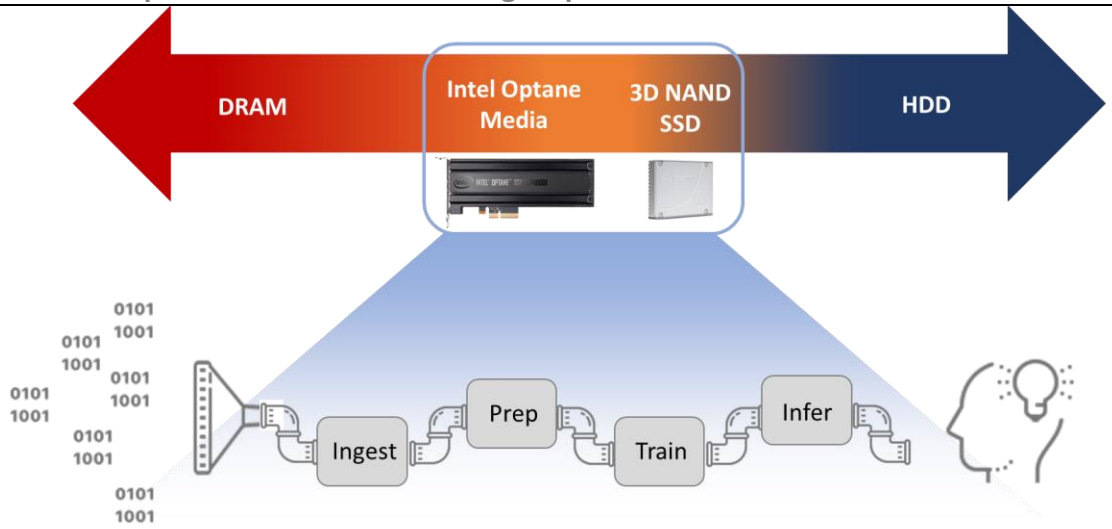
[1] Source: ESG Master Survey Results, *Artificial Intelligence and Machine Learning: Gauging the Value of Infrastructure*, March 2019. All ESG research references and charts in this technical validation have been taken from this research report, unless otherwise noted.
[2] Source: ESG Brief, *Will Data Lakes Drown Enterprise Data Warehouses?*, March 2020.

Intel Data Center SSDs for the AI Data Storage Pipeline

Across the AI data pipeline, I/O requirements are unpredictable, widely variable, and extremely demanding. Data set sizes have a wide range; petabytes of raw ingested data are refined down to gigabytes of structured and semi-structured data for training and complete their journey as kilobyte-sized trained models. Workloads and access patterns vary significantly as well, and AI storage must satisfy the demanding requirements of high throughput and extremely low latency.

AI storage workload requirements start with ingest, where raw data is collected concurrently from various sources. Ingest is mostly sequential writes, periodically destaging to capacity storage. Ingested data is used in the training phase. Preparation is an iterative process that incorporates highly concurrent reads and writes, with random and sequential access patterns. Training needs ultra-low latency, high bandwidth random reads to maximize model accuracy while minimizing training time. Inference is where the trained model is used to make decisions. Storage requirements are complex here as well; low latency reads and writes speed both time to insight and real-time decision making.

Figure 2. Intel Optane for the AI Data Storage Pipeline



Source: Enterprise Strategy Group

Leveraging storage that is efficient, performant, available, and flexible enough to service performance and capacity requirements tiers of the AI data pipeline is a key to success when architecting a future-proof AI infrastructure. Typical HDDs and SSDs are well suited to meet the growing capacity requirement of AI workloads, but pairing Intel NAND SSDs with performance-optimized Intel Optane SSDs yields a tiered approach that's ideally suited to address the AI business objectives across all of the phases of the AI data pipeline (see Table 1).

Table 1. Accelerating AI Business Objectives with Intel Data Center SSDs

	Ingest	Preparation	Train	Infer
AI Business Objectives	Collect more to learn faster	Streamline data preparation	Accelerate learning from existing data	Quicken insight from new data
Intel Optane SSD versus NAND SSD benefits	Better bulk write throughput and endurance	Faster read/modify/write performance	Lower latency for compute and read-intensive workloads	Better performance for read/write workloads
Intel NAND SSD versus HDD benefits	Scale capacity without sacrificing availability or performance per GB			
	Higher storage efficiency with lower power/cooling per GB			

Source: Enterprise Strategy Group

A comparison of the storage media specifications for a pair of prototypical Intel NAND SSD and Intel Optane SSD devices is summarized in Table 2. Intel Optane SSDs are 3.7 times faster for the write workloads during the ingest, preparation, and inference phases of the AI pipeline, and 3.6 times faster for training and inference read workloads.

Table 2. Quantifying the Intel Storage Media Advantage for AI Workloads

	Intel NAND SSD[3]	Intel Optane SSD[4]	AI pipeline impact
Write response (µsec)	37	10	3.7x faster ingest/preparation/inference
Read response (µsec)	36	10	3.6x faster training/inference
Endurance DWPD (drive writes per day)	3	60	20x better for ingest/preparation

Source: Enterprise Strategy Group

Storage media endurance is an especially important consideration for emerging AI workloads that are ingesting massive real-time data sets. The number of full drive writes per day (DWPD) that a storage device can sustain before it "wears out" is a good way to compare the endurance of NAND and Optane SSDs. A 60 DWPD endurance rating of Intel Optane SSD (20x better than NAND SSD) is well suited for the write intensive nature of the ingest and preparation phases of the AI pipeline. The next section of this report explores how these low level device specifications translate into application level performance benefits for each stage of the AI pipeline.

ESG Technical Validation

ESG leveraged performance benchmark results that were documented in previously published and publicly available reports from ESG, Splunk, and the University of California San Diego with a goal of quantifying the performance benefits of Intel Data Center SSDs for each stage of the AI data pipeline. Instead of using AI and machine learning storage performance benchmark tools that weren't generally available when this report was written, the I/O characteristics of each phase of the AI pipeline were cross referenced with existing benchmark results with a goal of quantifying the performance benefits of Intel Data Center SSDs.

Ingest

ESG first looked at AI ingest, where raw data is collected concurrently from various sources with a goal of accelerating the time that's needed to collect data for insight and analysis. The I/O profile for AI ingest is typically 100% sequential writes.

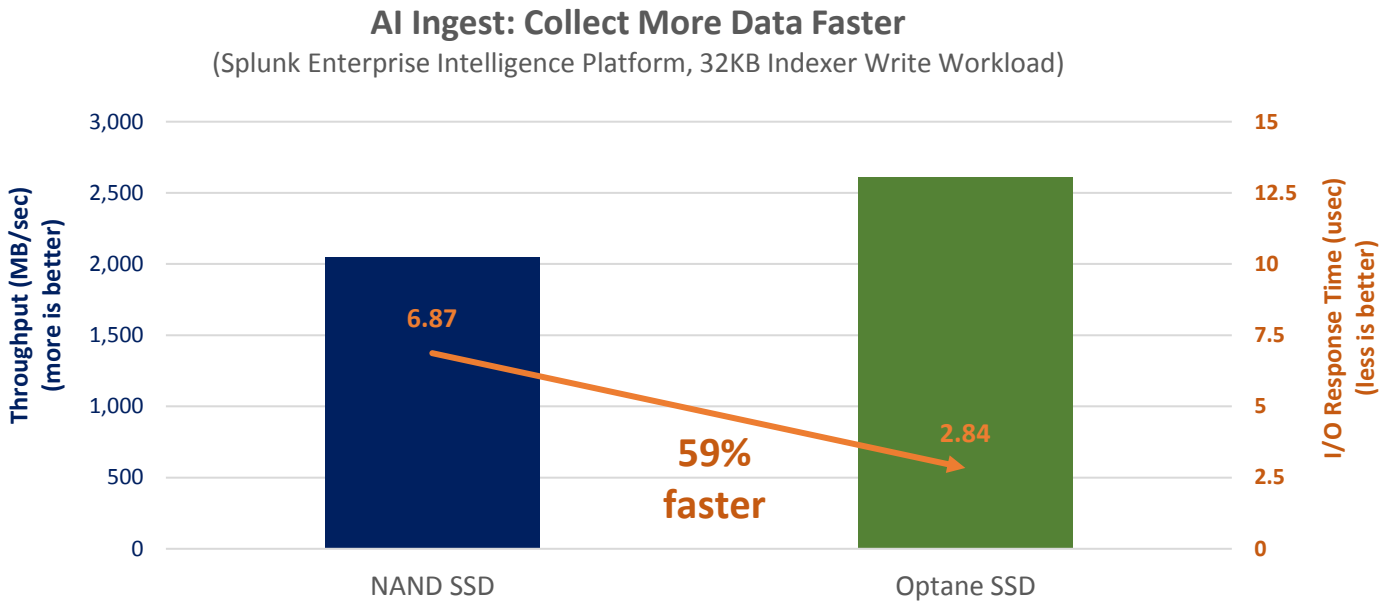
This test compared the performance of Optane SSD to traditional NAND SSD for 32KB index writes during a real-world balanced Splunk Enterprise analytics workload with a mix of ingest, index, and search operations occurring in parallel.[5] As shown in Figure 3, Intel Optane SSDs delivered more IOPS (22%) with significantly faster latency (59%).

[3] Intel D3-S4610

[4] Intel DC P4800X Optane SSD

[5] The configuration and testing methodology is summarized in the Appendix and documented in detail in this report: *High-Performance Data Analytics with Splunk on Intel Hardware*

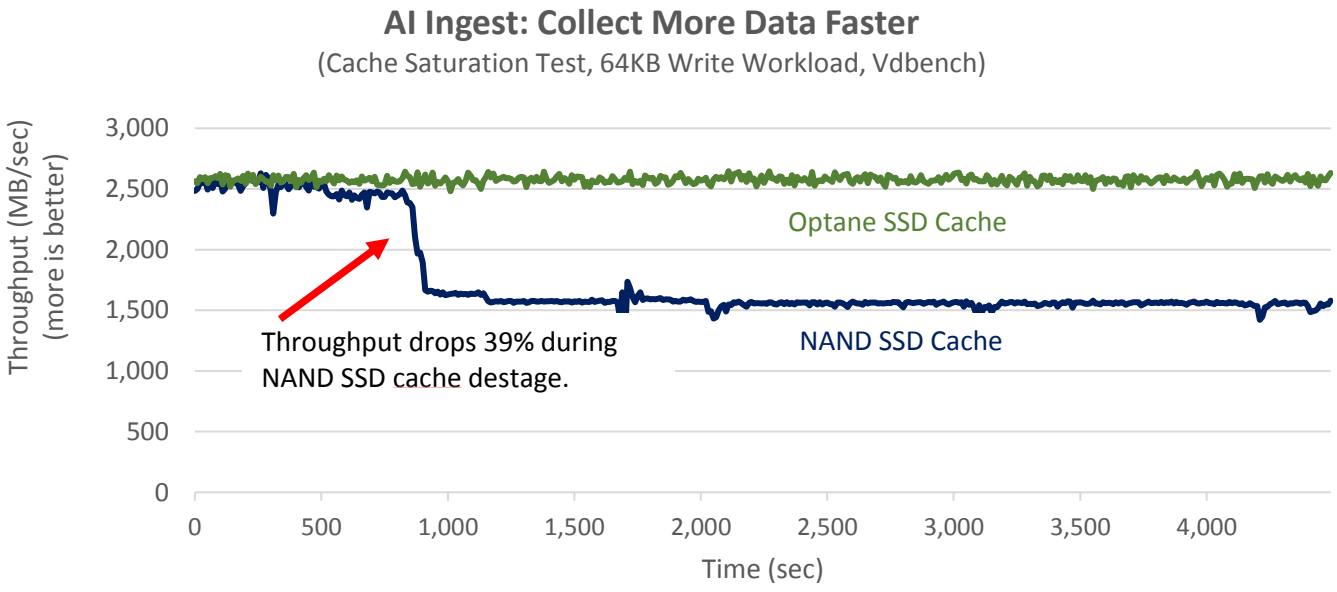
Figure 3. AI Ingest: Collect More Data Faster



Source: Enterprise Strategy Group

A second ingest test compared the performance of Optane SSD and traditional NAND SSD for write caching during a 64KB sequential write workload with a 1.2TB working set for 75 minutes.[6] As shown in Figure 4, Intel Optane SSDs sustained a high level of ingest workload throughput (2,500 MB/sec) during an aggressive cache destage to capacity storage, where NAND SSD performance dropped by 39% and stayed there for the remaining hour of the test.

Figure 4. Ingest Caching: Optane SSD versus NAND SSD



Source: Enterprise Strategy Group

[6] The configuration and testing methodology is summarized in the Appendix and documented in detail in this ESG Technical Validation: *Dell EMC VxRail with Intel Xeon Scalable Processors and Intel Optane SSDs*

Why This Matters

Considering that 54% of organizations surveyed by ESG report that they typically use more than 1 TB of data to train their ML model, and 30% regularly use more than 11 TB, it comes as no surprise that nearly one in four organizations cited data storage as one of the top three parts of the infrastructure stack that would be the weakest links in their ability to deliver an effective AI/ML environment. Speed of ingest is critical, and the faster you can collect data, the faster you can get to insight.

ESG testing and analysis showed that Intel SSDs can ingest data faster than traditional NAND SSDs, and Intel Optane SSDs can do it without the performance impact of destaging, which caused a 39% drop in throughput with NAND SSDs. In short, Intel Optane SSDs enable organizations to collect more data faster, which shortens the time it takes to get answers out of your data.

Preparation

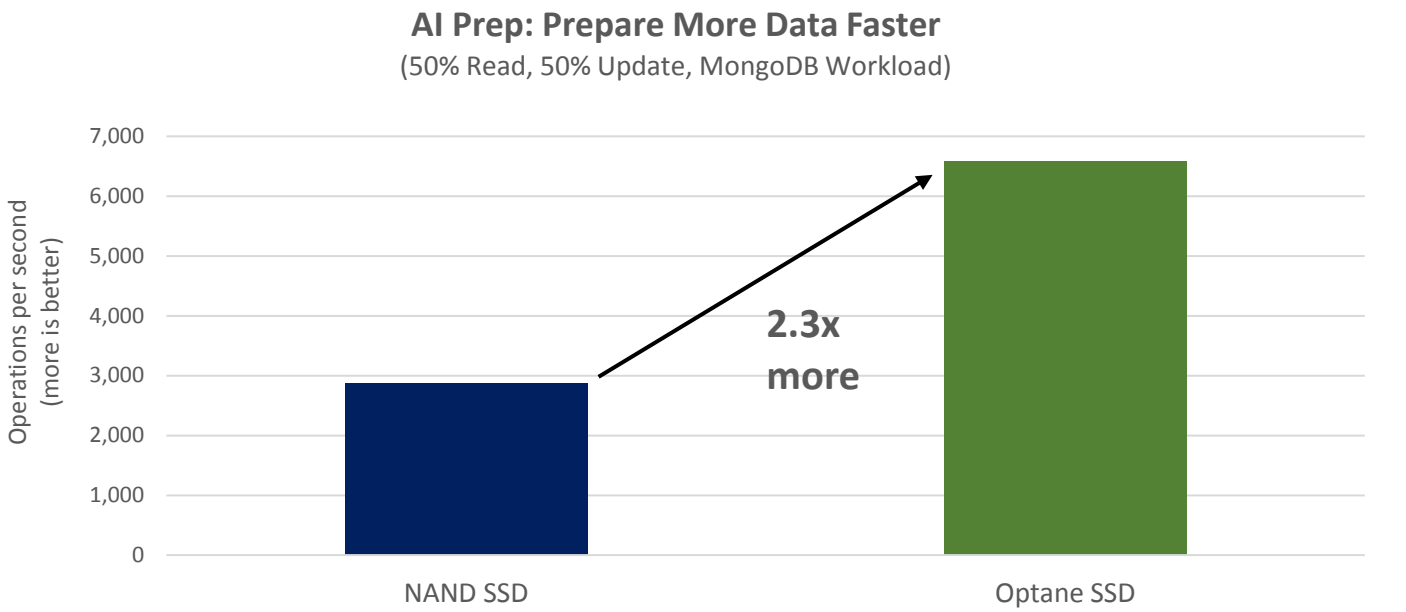
Because of the wide variety of sizes, formats, completeness, and accuracy of raw data, ingested data needs to be prepared for use in training. Data that is missing or incomplete should be enriched or ignored. Data inconsistencies such as decimals versus commas in numbered data sets must be standardized. Data with different attributes such as images for facial recognition must be normalized. Unstructured data requires tagging and annotation. Data may be combined from different sources. Finally, the data will need transformation to the target format such as TensorFlow. This is an iterative process of varying amounts of data that are read and written, both randomly and sequentially.

This iterative process drives a mixed workload with a high degree of concurrency. The read-write ratio will vary depending on the veracity of ingested data and the level of transformation required to achieve the target format. Worst case workloads can approach 50% writes.

This test compared the performance of Optane SSD to traditional NAND SSD for a 50% read, 50% update MongoDB workload.[7] As shown in Figure 5, the Intel Optane SSD sustained 2.3x more operations per second than NAND SSD.

[7] The configuration and testing methodology is summarized in the Appendix and documented in detail in this report: *Basic Performance Measurements of the Intel Optane DC Persistent Memory Module*

Figure 5. AI Prep: Prepare More Data Faster



Source: Enterprise Strategy Group

Why This Matters

Since data preparation can consume up to 80% of AI/ML resources, storage devices that deliver high throughput and low latency with high QoS are key to reduce the time needed to prepare data. Speed of transformation will depend on storage performance. As more varied data sources are added, the demands on storage performance will only increase.

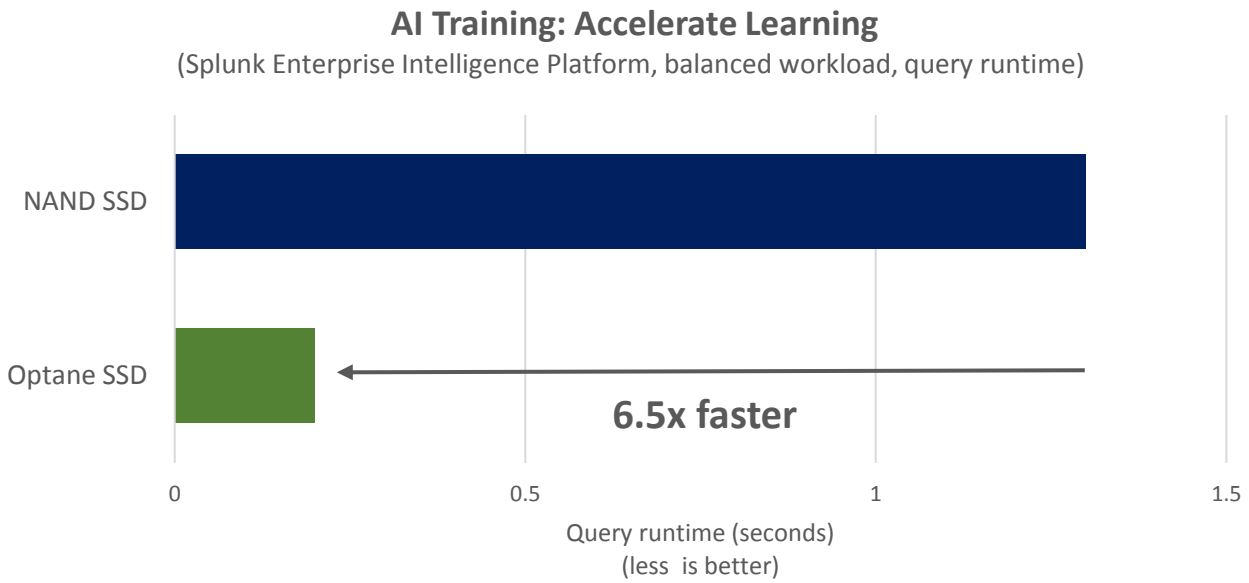
ESG testing and analysis showed that Intel SSDs were able to service a challenging 50% write workload and deliver 2.3x the number of operations per second compared to traditional NAND. ESG observed that Intel SSDs can prepare data faster, which enables organizations to prepare more data in the same amount of time, and more data can drive better, more accurate results.

Training and Inference

Training is extremely resource intensive as it involves a repetitive set of steps that execute a sequence of mathematical functions on prepared data to identify the probability of obtaining the desired result. Results are then evaluated for accuracy and, if not acceptably high, the mathematical functions are modified and then executed again. Training is conducted with mostly random reads and some writes for checkpointing, so training benefits from storage that can sustain ultra-fast, high bandwidth random reads. Faster reads keep valuable training resources utilized and randomness helps improve model accuracy. Mitigating I/O wait time is critical at this phase.

This test compared the performance of Optane SSD to NAND SSD for training workloads by examining the search run times of a Splunk Enterprise analytics workload for search head storage media during a balanced workload test that was running a mix of ingest, index, and search operations in parallel.[8] As seen in Figure 6, Intel Optane SSD completed the query in 0.2 seconds—6.5x faster—with 93% shorter I/O response times (0.05 milliseconds for Optane SSD versus 0.74 for NAND SSD).

Figure 6. AI Training: Accelerate Learning



Source: Enterprise Strategy Group

This test also included performance testing of a typical commodity hard disk drive (HDD) for Splunk Enterprise search head storage media. Compared to a typical HDD, searches completed 3.9 times faster with Intel SSD and 25.5 times faster with Optane SSD.[9]

Next, ESG looked at inference, where the trained model is deployed to execute decisions. Inference can be deployed in the data center or—increasingly—on edge devices. Data movement involves reading the trained model from storage into inference, writing ingested data that is being evaluated into inference, and reading inferred results back into training to improve model accuracy.

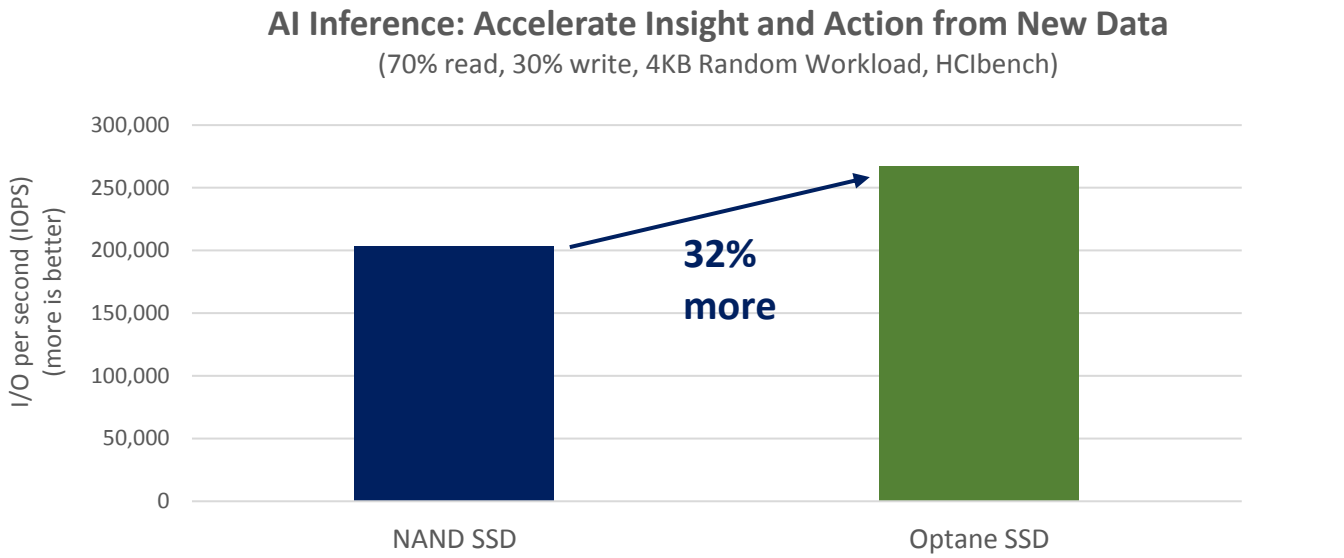
[8] The configuration and testing methodology is summarized in the Appendix and documented in detail in this report: *High-Performance Data Analytics with Splunk on Intel Hardware*

[9] 5.1 second search runtimes for HDD versus 1.2 seconds for Intel SATA SSD and 0.2 seconds for Intel Optane SSD.

This test compared the performance of Optane SSD to NAND SSD for inference workloads training workloads with a 4KB random I/O workload composed of 70% reads and 30% writes.[10]

As shown in Figure 7, the Intel Optane SSD was able to sustain 32% more I/O operations per second with 24% faster I/O response times compared to NAND SSD.

Figure 7. AI Inference: Accelerate Insight and Action



Source: Enterprise Strategy Group

Why This Matters

Training and inference are where the rubber meets the road in AI and ML. Both have a random-access pattern, both are read-intensive, and both require high performance storage with exceptionally low response time. Inference deployment can be in the data center or, increasingly, at the edge. Real-time edge deployments not only need the trained model read into inference quickly, they can also require fast writes of ingested data for real-time decision making. As more edge deployments adopt reinforcement learning—a method where accuracy is evaluated and acted on at the edge—storage performance requirements will increase.

ESG analyzed Intel SSDs running workloads designed to simulate the access patterns and I/O characteristics of both training and inference and found that Intel Optane SSDs were able to provide higher performance, faster run times, and lower response times, which can accelerate an organization's time to insight and action.

[10] The configuration and testing methodology is summarized in the Appendix and documented in detail in this ESG Technical Review: *Hitachi Unified Compute Platform HC*

The Bigger Truth

ESG research reveals that organizations recognize that AI and ML can help them achieve multiple, strategic business objectives, including improved operational efficiency (60%), stronger cybersecurity (47%), better customer satisfaction (45%), and improved quality of products and services (44%). It should come as no surprise that 100% of organizations surveyed are either already leveraging AI/ML or are planning to in 2020.

In addition, enterprises are increasingly deploying data lakes or common data pipelines to reduce complexity, increase efficiency, and most importantly, derive more and better insights by building multiple instances of AI usages, data analytics, and reporting on these data lakes. This approach will challenge the performance storage layer, with the need to support high workload variability and high concurrency, while consistently delivering predictably low latency. As these organizations realize more value from AI, the amount of stored data will continue to increase at a quickening pace, stressing the ability of capacity storage to scale efficiently.

Intel SSDs are designed to offer the performance required for the I/O demands of AI/ML workloads using Intel Optane technology's consistently low latency to enable faster time to trained models, and high throughput, even through buffer destage, to improve ingest performance. Intel NAND-based SSDs can scale the capacity layer much more consistently and efficiently than HDDs, with better availability.

The performance test results and ESG analysis in this report illustrate how a combination of Intel Data Center SSD technologies can be used to meet the demanding storage requirements of all stages of the AI data pipeline. Intel Optane SSDs ingested data 59% faster than traditional NAND SSDs while avoiding a 39% performance impact of write cache destaging. During the AI preparation phase, more than twice the number of operations per second were achieved with Intel Optane SSDs (2.3x more). During the training phase, analytics search runtimes completed 3.9 times faster than a typical HDD with Intel SSDs and 25.5 times faster with Intel Optane SSDs. Workloads that simulate the access patterns and I/O characteristics of both training and inference proved that Intel Data Center SSDs provide higher performance, faster run times, and lower response times.

The results presented in this report were harvested from a variety of publicly available benchmarks from ESG, Splunk, and the University of California San Diego. The I/O characteristics of each phase of the AI pipeline were cross referenced with existing benchmarks with a goal of quantifying the performance benefits of Intel Data Center SSDs. While these discrete results demonstrate how Intel Data Center SSDs improve performance at each stage of the AI data pipeline, ESG looks forward to testing with a single storage solution due to the fact that reference architectures and emerging best practices are using a common storage infrastructure for all stages of AI data pipeline data processing.

If your organization is looking to build an efficient, scalable, future-proof AI infrastructure, leveraging storage that is efficient, performant, available, and flexible enough to service both performance and capacity tiers is a key to success. ESG believes that you should consider the advantages of Intel Data Center SSDs to optimize, store, and move larger, more complicated data sets through the AI pipeline and get to insight quicker.

Appendix

The performance benchmark configurations referenced in this report are summarized in Table 3.

Table 3. Configuration and Workload Summary

	Source	Tested configuration and methodology
Ingest I/O workload	ESG Technical Validation: *Dell EMC VxRail with Intel Xeon Scalable Processors and Intel Optane SSDs*	Media: Toshiba PX5SMB080Y (NAND SSD) versus Intel P4800X (Optane SSD) Test bed: Four-node Dell EMC P570F VxRail cluster Workload: Vdbench 64KB Sequential Write
Prep I/O workload	*UCALSD Research:* *Basic Performance Measurements of the Intel Optane DC Persistent Memory Module*	Media: Intel DC S3610 (NAND SSD) versus Intel P4800X (Optane SSD) Test bed: Intel Xeon Scalable platform, 24 cores at 2.2 GHz, 384GB DRAM Workload: MongoDB, 50% read/50% update
Training I/O workload	*Intel white paper:* *High-Performance Data Analytics with Splunk on Intel Hardware*	Media: Intel P4510 (NAND SSD) versus Intel P4800X (Optane SSD) Test bed: Intel Xeon Platinum 8620, 24 cores at 2.4GHz, 384GB RAM Workload: Splunk Enterprise balanced analytics (Ingest, Index, Search)
Inference I/O workload	ESG Technical Review: Hitachi Unified Compute Platform HC	Media: Intel S4600 (NAND SSD) versus Intel P4800X (Optane SSD) Test bed: Hitachi HC V121F versus HC V124N Workload: HCIbench, 70% read, 30% write, 4KB Random

Source: Enterprise Strategy Group