TECHNOLOGY BRIEF

Intel® Optane™ Technology
Intel® 3D NAND Technology

How Intel Storage Accelerates AI Deployments

Manage your increasingly complicated AI data with Intel® Optane™ Technology and Intel® 3D NAND Technology.

Intel for AI Storage

Intel® Optane™ SSDs optimize compute resources with high mixed performance across a wide variety of workloads, access patterns and block sizes. Intel® 3D NAND SSDs unlock the value of stored data while reducing storage cost, space and operationally efficiencies. The combination of Intel's storage technologies offers an agile and flexible solution that ensures your storage capabilities will keep pace with the evolving, increasingly complicated AI data pipeline needs.

There are two kinds of enterprises: those that are accelerating with AI and those that will be. Even if you or your partners are not currently considering AI, it's time to get a handle on your increasingly complicated data.

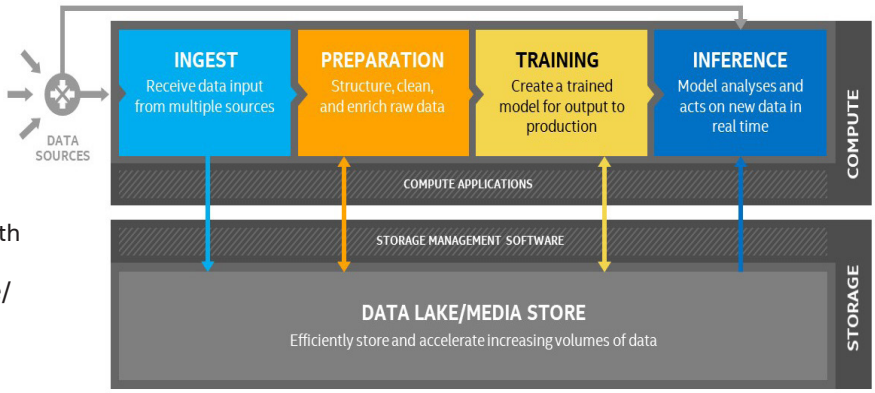
In its most recent research report, Whatech sees the global AI-powered storage market growing "from $10.4 billion in 2019 to $34.5 billion by 2024; growing at a CAGR of 27.1%." Drivers come from far and wide, including "massive growth in data volumes, need for global enterprise infrastructure to refresh their storage architecture, increasing adoption of cloud-based services, and growing demand for AI in HPC data centers."[1]

AI is coming, whether we are ready or not. To take advantage of it, the first step is to be aware of the complexity of AI data. We've all heard of the three Vs and how challenging they can be for storage:

- **Volume**: As AI training data grows, algorithms get smarter. Cloudian points out that "managing these data sets requires storage systems that can scale *without limits*." [2]
- **Velocity**: IDG says that nearly 20% of data will be critical to our daily lives by 2025, and 10% will be "hypercritical."[3] Instant access is a must.
- **Variety**: Variety refers to the format of the data. As businesses look to improve customer experiences, run more efficiently, and stay competitive, they are analyzing data across a wider range of formats, including ingesting data from transactions, social media engagement, and customer service, and involving a wide variety of files, clicks, texts, videos, machine data, and Bluetooth signals.

Navigating the AI Pipeline

To complicate matters, each of these Vs changes dramatically depending on the stage of the AI pipeline. For example, data can be ingested in petabytes, move into training as gigabytes of structured and semi-structured data, and then end up as kilobytes of trained models. In addition, workloads are highly variable, starting with ingest of 100% writes, progressing to preparation where they can reach 50/50 read/write mix, then shifting to training and inference at 100% read.



The solution to the three Vs is to create a common data pipeline underlying the various AI functions with a tier optimized for space-efficient capacity/scaling, and another tier optimized for performance storage/scaling. This includes a data lake or media store to act as a centralized repository for all structured and unstructured data at any scale.

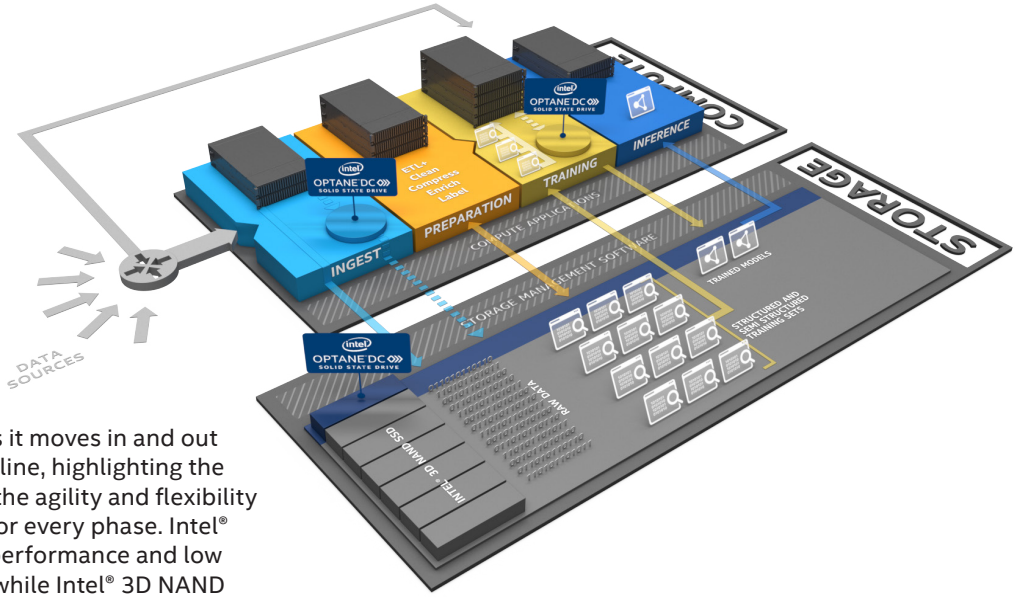
Throughout all of this variability, high throughput (TPT) and extremely low latency are always required. To navigate the pipeline, you need technologies that address these challenges and scale with your evolving AI needs.

Gartner, in a recent report, puts a fine point on the importance of understanding the phases: "The success of ML and AI initiatives relies on orchestrating effective data pipelines that provision the high quality of data in the right formats in a timely manner during the different stages of the AI pipeline."[4]

Compared to a NAND+HDD or all TLC NAND solution, the combination of Intel's storage technologies offers the agility and flexibility needed to meet the right performance for every phase of the AI data pipeline. Intel® Optane™ technology provides the high performance and low latency for fast storage or fast caching, while Intel® 3D NAND SSDs consolidate storage footprints, scale with growing storage needs, and accelerate access.

How Intel Storage Technology Works in the Pipeline

- During the **ingest** phase, the agility to ingest and manage highly variable data formats from a wide variety of sources is of paramount importance. Size of the ingested data can vary and is often in unstructured object or file forms such as videos, images, documents, or conversation transcripts. This stage is heavy on sequential writes with periodic de-staging of the ingest buffer into capacity storage. If data ingest cannot keep up, bad things happen: Data sources can be put on hold, data can be dropped, or the available DRAM may create a bottleneck. With Intel Optane technology, you have an ingest buffer, offering you the ability to scale ingest with high write performance and low latency.
- In **preparation**, data must be moved through labeling, compressing, deduping, transforming and cleansing. Again, ingested data can arrive in a wide variety of sizes, formats, completeness and accuracy. Data that is missing or incomplete needs to be enriched (or ignored) and inconsistencies—such as decimals versus commas in numbered data sets—need standardization. This is an iterative process: fluctuating amounts of data are read and written both randomly and sequentially. Since preparation can consume up to 80% of AI resources, storage devices that deliver low latency, high QoS, and throughput are important to reduce preparation time. Speed of transformation depends on the speed of storage and, as more data sources are added and data variety increases, the demands on storage performance increase. Intel Optane technology delivers unmatched mix of read/write performance to reduce time to prepare data.
- **Training** includes moving repeated, randomized data sets. To do this as efficiently as possible, compute utilization must be optimized. This step is extremely resource-intensive, involving repetitive execution of mathematical functions on prepared data to identify high probability of desired results. Results are then evaluated for accuracy and if not acceptably high, the mathematical functions are modified and tried again. More complex models involving more neural network "layers" results in increased accuracy, but also increased demands on storage. Intel Optane SSD's high random read TPT and low latency can keep critical training resources utilized. Additionally, Intel Optane SSDs can accelerate temp data during model data.
- Finally, in **inference**, the trained model is deployed to execute decisions. Deployment can be 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 accuracy. Real-time edge deployments drive an even higher performance need. And, as edge deployments adopt reinforcement learning—a method where accuracy is evaluated and acted on at the edge—storage performance demands further increase. Instant feedback to evaluate decision accuracy is critical, especially in edge reinforcement. Intel Optane technology offers the acceleration to move these trained models, even at the edge.



The diversity and use of data changes as it moves in and out of the data lake and through the AI pipeline, highlighting the need for storage solutions that provide the agility and flexibility needed to meet the right performance for every phase. Intel® Optane™ technology provides the high performance and low latency for fast storage or fast caching, while Intel® 3D NAND SSDs bring performance and high capacity options.

"The success of Machine Learning and AI initiatives relies on orchestrating effective data pipelines that provision the high quality of data in the right formats in a timely manner during the different stages of the AI pipeline." *-Gartner* [4]

To add another layer of complexity, enterprises are increasingly deploying data lakes or common data pipelines to derive more insights and increase efficiency. They build multiple instances of AI usages, data analytics, reporting and other applications on these data lakes. This further stresses the performance storage layer to support high workload variability and high concurrency while consistently delivering predictably low latency. And, as these enterprises realize more value from AI, the amount of stored data only increases.

Optimize AI Compute Resources

Throughout all phases of the pipeline, Intel Optane technology is optimizing compute resources with high mixed performance across a wide variety of workloads, access patterns and block sizes. And PCIe SSDs with Intel 3D NAND technology unlock the value of stored data while reducing storage cost, space and operational efficiencies.

Venture Beat sums it up, noting that Intel's "decade-long R&D investment in this new memory/storage hybrid" has paid off, saying Intel Optane "should be a net performance gain for massively memory-hungry applications."[5]

To learn more, visit intel.com/aistorage

Sources:

1 https://www.whatech.com/market-research/industrial/627958-major-trends-in-ai-powered-storage-market-discussed-in-a-new-research-report

2 https://www.cbronline.com/emerging-technology/ai-machine-learning-important-data-storage-requirements/

3 https://www.computerweekly.com/news/450416206/Analytics-internet-of-things-to-drive-data-volumes-to-163ZB-by-2025

4 Gartner: Three Ways That AI Will Impact Your Data Management and Storage Strategy, August 2018

5 https://venturebeat.com/2019/11/21/optane-101-memory-or-storage-yes/

Printed in USA 0120/JR/RA/PDF Please Recycle 342334-001US