TECHNOLOGY PAPER

Seagate Edge RX

A Smart Manufacturing Reference Architecture Solution

How Deep Learning Can Boost Efficiency on Factory Floors

The future of manufacturing is here. It has been ushered by data, AI, and machine learning. Its path is being blazed by a unique collaboration among Hewlett Packard Enterprise, NVIDIA, and Seagate.

Today's manufacturing processes are complex and precise—with little or no tolerance for errors. The processes have thus far been automated mostly with robotic arms and other rule-based systems. However, that level of automation alone is not enough, as operational managers face ever increasing challenges:

- How can we reduce investments for clean-room resources?
- How can we cut days or, better yet, weeks from manufacturing time?
- How can we catch more defects before the final assembly step?
- How can we improve current clunky rule-based anomaly detection systems?
- How can we save time and money needed to send data to the cloud for processing?
- Why can't data be stored right on the factory floor—and be put to work there?

Thanks to the adoption of artificial intelligence (AI) technologies—particularly machine learning (ML) and deep learning (DL)—solutions to questions like these are now at hand. Significant productivity and efficiency improvements can become reality.

A 2018 report from the McKinsey Global Institute forecasts that AI will add 16% (or $13 trillion) to the global economic output by 2030. The early adopters of edge-enabled technology stand to benefit the most. Seagate Technology—a global leader in design and manufacturing of disk drives, flash storage, and other data storage and management solutions—has successfully applied AI and ML to build an anomaly-detection solution for quality control of recording head wafer images in the process of disk drive manufacturing. The production of read/write heads and head gimbal assemblies is a highly complex manufacturing process.

This paper presents a tested reference architecture—named Seagate Edge RX—used in Seagate's own smart factories. The hope behind this project is to enable other businesses to also use this architecture in their own AI/ML manufacturing centers. The architecture is based on Seagate's knowledge of the IT infrastructure, both as a provider and a consumer combined with its AI/ML and storage business expertise—as well as Seagate's partnership with Hewlett Packard Enterprise (HPE) and Seagate's continued research on AI/ML with NVIDIA and support from Nexenta.

While the implementation discussed here is specific to the manufacturing process in the data storage industry, it is generally applicable to other types of processes—particularly those with the following characteristics:

- High-volume, high-precision, discrete manufacturing processes producing tools such as semiconductors, electronics, automotive parts, machine parts, etc.
- High-value manufacturing products using high-cost capital equipment
- Verticals generating large volumes of images that cannot be analyzed with traditional methods
- Anomaly detection in security, smart cities, and autonomous vehicles
- Highly complex manufacturing processes with many stages
- Automated manufacturing processes that can collect equipment, process, and inspection data
- Quality control and inspection-driven manufacturing processes
- Lengthy manufacturing processes
- Multisite global manufacturing

Seagate's Industry Expertise

Hard disk drives (HDD) enable active and archival data storage for enterprise and cloud service providers. They can be found in many common consumer products, such as notebook and desktop computers, gaming consoles, set-top box DVRs, personal backup drives, and home network drives. They are widely used in enterprise data centers, cloud service providers, specialized applications such as surveillance, media and entertainment, and content delivery networks. For each usage category, specialized hard drives have been designed to optimize different characteristics such as I/O performance, capacity, and cost. In aggregate, about 400 million HDDs are produced annually by the industry.

Seagate manufactures more disk drives worldwide than any other manufacturer. The company is highly vertically integrated, since it designs and manufactures all of the key components used in its solutions, and its engineers invent many key HDD technologies. For example, the Seagate® Heat Assisted Magnetic Recording (HAMR) technology offers the highest recording density while enabling HDD capacities of 20TB to 40TB in the near future.

Site Overview

Seagate manufactures magnetic (read/write) heads to produce hard disk drives at two wafer fabs located in the US and Europe (see Figure 1). Read/write heads are created on wafers using processes similar to semiconductor manufacturing with comparable dimensions. Processed wafers are sent to two head fabrication sites located in Asia for sectioning, processing, and assembly into head gimbal assemblies.

Disk Drive Manufacturing Overview

The disk drive manufacturing process can be broadly categorized into three areas:

- Read/write heads and head gimbal assemblies
- Head stack assembly

1 Source: "Worldwide Hard Disk Drive Quarterly Update Q4'18 Summary and Updated 4-Quarter Outlook" by IDC's John Rydning

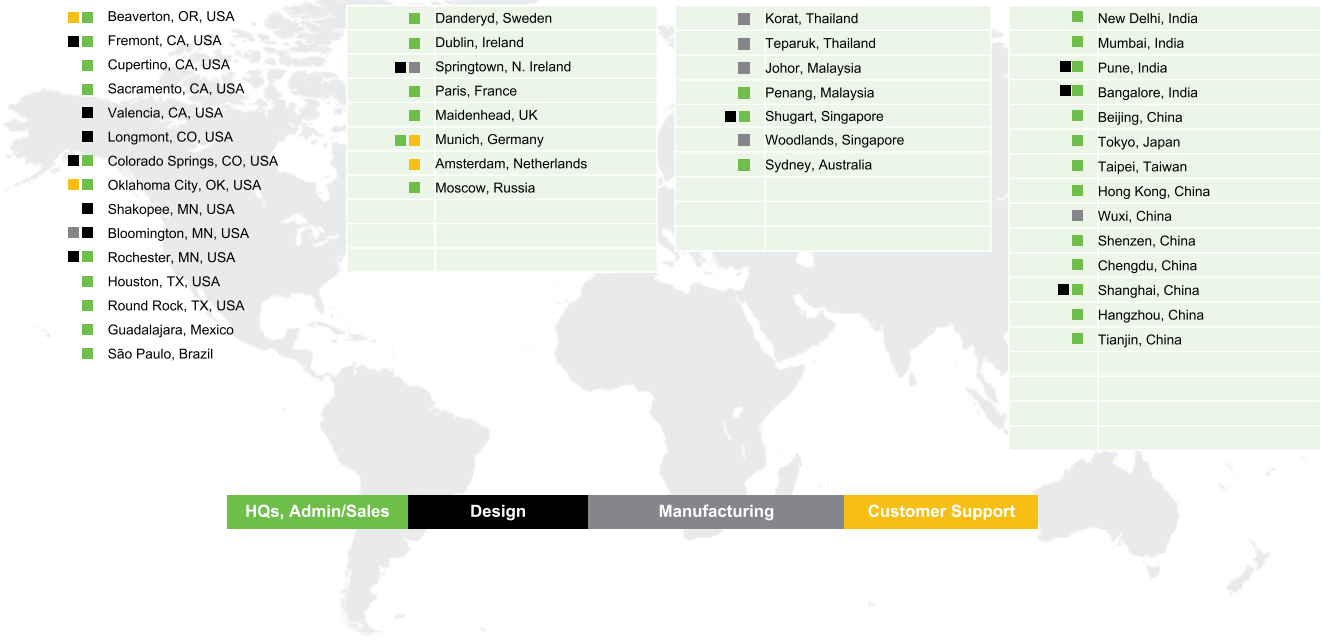

Figure 1. Seagate's manufacturing and design sites

- Final product assembly consisting of head stack assembly, media (disk platter), servo motor, circuit board assembly, etc.

Disk drive heads are fabricated using a process similar to semiconductor wafer fabrication, but with unique materials and processes (see Figure 2). The head fabrication starts with a wafer substrate that is processed in over 1000 fabrication steps during a six-month period. Each wafer yields approximately 100,000 disk heads after the wafer is sectioned into individual bars. The individual bars are then further processed, and the separate read-and-write heads are attached to head gimbal assemblies. Multiple-head gimbal assemblies are inserted into a head stack assembly that allows read/write operations to both sides of multiple disk platters.

Anomaly Detection

Seagate's wafer fabs produce thousands of wafer images each day (like those in Figure 3). Of the over 1000 fabrication steps in the head manufacturing process, several hundred are subject to process

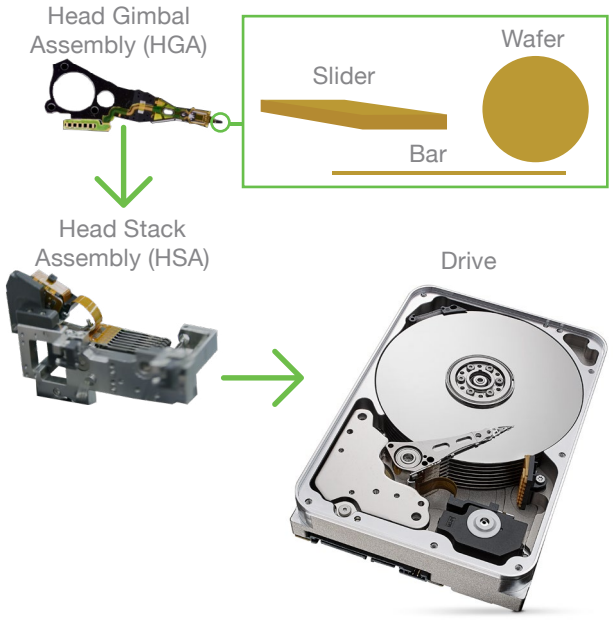

Figure 2. From wafers to heads

control inspection using automated optical and scanning electron microscopy (SEM) imaging.

Inspection stations are positioned at critical stages of wafer production to check for defects. Using both visible light and electron beams, automation tools play a critical role in quality control. They generate more than 10GB of imagery per day for inspectors to review, either manually or with highly specialized and rule-based programs. Both options are extremely labor-intensive and costly to implement for all images.

Maintaining high yield rates is critical to managing product cost and margin objectives in all manufacturing processes, especially in high-value, long processes. Therefore, it is very important to recognize process excursions as soon as possible to remedy the problem and identify random defects.

Erroneous rejection of good quality materials decreases the yield rate of the affected process stage, while allowing defective materials to pass onto the next stage results in higher material costs, higher processing costs, and decreased yield rates in the downstream process stages. For these reasons, Seagate's manufacturing operations focus on early defect detection, recognition, and classification using automated systems.

Seagate previously used conventional rule-based machine vision systems to automate the anomaly detection process. While these systems provide a high accuracy rate, they have several disadvantages as compared to an ML/DL approach. The rule-based system requires that parameters for each defect class are statically coded and the fixed ranges are used to determine the good/reject criteria. Changes in defect appearance or new types of defects require additional rules to be added or changes to be made to existing rules.

A Two-Step Solution

There are two phases to developing the ML/DL inspection system: training and inference (see Figure 4).

The training phase uses classified image data as input into the system. After running the ML model hundreds of times using the training data, the training process produces the weights in the neural network model—based on the way the nervous system works—which classifies different categories of defects. The training process is very computationally and data intensive. It is done as an offline process by data scientists.

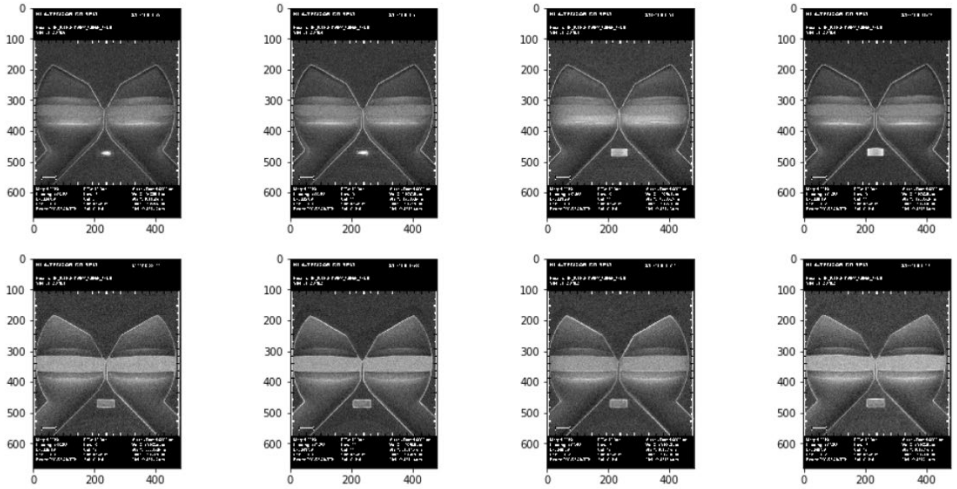

Figure 3. ML training use case with open "good" images from the biggest cluster

The inference phase is done in the factory using real-time imagery to make classification decisions during the inspection process. This process uses the model developed during training to make the classification decisions on the inference server. Running the model in inference mode allows the use of lighter-weight computing resources.

The objective of the training process is to obtain a high level of classification accuracy for the model while the objective of the inference model is real-time classification for each image processed.

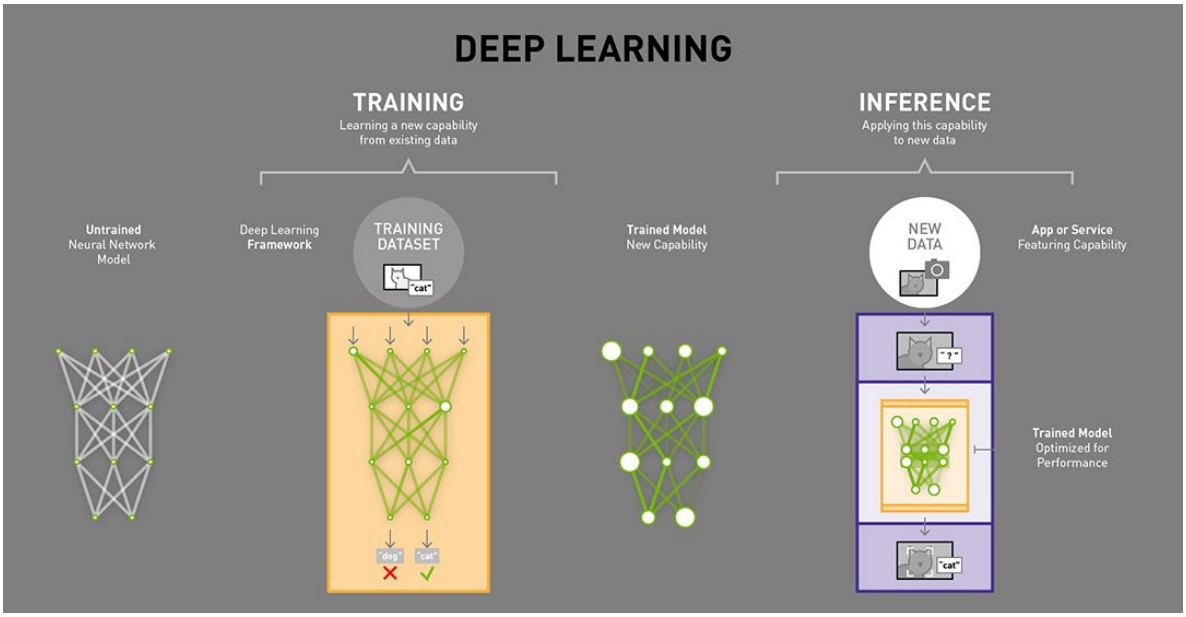

Figure 4. DL overview (Source: NVIDIA)

Storage Requirements and Implementation for AI

Overall, there are three tiers of storage in the factory environment, as shown in Figure 5. The first-level local flash storage is part of the GPU server and is typically not RAID protected to be used for persistent storage. The second-level working storage is used to store training data for the GPU training servers and serves as local storage for the data science team working on model development. The third-level capacity-defined data lake provides a storage repository for all image data produced by factory floor image sensors and retained for use in long-term time-series analysis. This may be optimized for cost and capacity with enterprise-level Seagate disk arrays such as the Seagate® Exos™ 5U84 84-drive disk array.

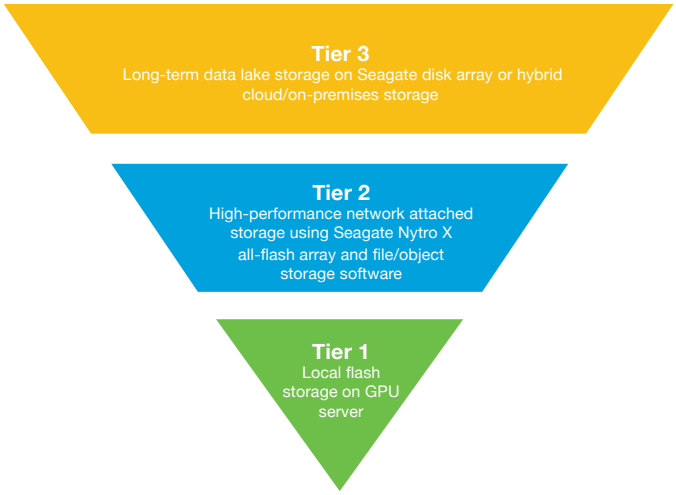

Figure 5. Storage hierarchy for manufacturing and for ML/DL training and operational data

Since the Tier 3 data lake storage was already implemented within Seagate's factory network,

and the Tier 1 storage was integrated on the GPU server, this project focused on the Tier 2 storage requirements.

The primary design goals for Tier 2 storage in designing a globally distributed storage architecture were:

1. Minimize cost, both CapEx, including initial cost of hardware and software, and OpEx, including cost for maintaining and managing the storage systems.
2. Ensure a disaster recovery (DR) capability if one site became unavailable using data replication between sites.
3. Implement software-defined storage with support for multiple storage protocols, including file and block storage with capability for object storage for future applications.

In addition, the following secondary requirements were desirable:

1. Support for hybrid cloud access
2. Docker, Kubernetes, and VMware® vSphere® hypervisor support
3. Global namespace support
4. Easy-to-use data and software management through web and CLI interfaces
5. Support for REST API, preferably through a management framework
6. Local failover for disaster recovery
7. Global technical support

Seagate is not only a major manufacturer of disk and flash media, it also is a major supplier of storage systems and platforms, including disk and flash arrays, storage area network (SAN) products, hyper-converged storage platforms that integrate storage and compute into a single chassis, and data-protected disk and flash array products.

Storage solutions commonly use Software-Defined Storage (SDS). In SDS, industry-standard servers run file, object storage management software, and block storage devices. The underlying storage hardware (flash or disk arrays) is abstracted by software that provides storage services. This abstraction also allows decoupling of storage hardware from computing hardware and enables a best-of-breed solution to be assembled by combining the best storage subsystem with the best SDS software. The disaggregation of the storage subsystem, storage software, and computing hardware allows the system to scale out for higher capacities through adding additional computing nodes and attached storage subsystems and is more cost-effective than tightly coupled appliance-based network-attached storage (NAS) systems.

Seagate engineers used the Seagate Nytro® X 2U24 all-flash array, shown in Figure 6, which combines a high-performance flash storage array with an integrated hardware-based data protection solution. The Nytro X platform, together with storage management software running on industry-standard

servers, forms a cost-effective and scalable storage solution ideally suited for developing AI systems. The Nytro X provides high I/O throughput of 12Gb/s host speed for SAS interfaces with 7Gb/s read throughput and 5.5Gb/s write throughput using the onboard 5005 controller module (see Figure 7). Nytro X's active-active dual controllers and innovative hardware-based RAID and erasure coding provide high reliability and data protection. The system's Seagate-designed ASIC provides high throughput of up to 600K IOPS with the 5005 controller and data protection with Seagate's ADAPT erasure coding or RAID levels 0, 1, 3, 5, 6, 10, and 50. The compact 2U rack-mount array contains 24 bays supporting up to 91TB internally using the 5005 controller.

Figure 6. Seagate Nytro X 2U24 all-flash array with 5005 controller

Specifications	
5005 Controller Performance	600,000 IOPS @ 1ms latency | 500,000 IOPS @ 250 µs latency | 7GB/s read throughput | 5.5GB/s write throughput
Expansion BODs	J1224 (2U24) | Maximum of 3 EBODs
Advanced Features	Thin provisioning | Snapshots | Asynchronous replication
High-Availability Features	Redundant hot-swap controllers | Redundant hot-swap devices, fans, power | Dual power cords | Hot standby spare | Automatic failover | Multi-path support
Device Support	SAS SSD
Data Protection	Seagate ADAPT | RAID levels supported: 0, 1, 3, 5, 6, 10, and 50
System Configuration (24, 2.5-in devices)	91TB max | With 3 EBODs: 364TB (based on 3.8TB SSDs)
Hosts	
External Ports	8 per system
Fibre Channel Models	Host speed: 16Gb/s, 8Gb/s Fibre Channel | Interface type: SFP+
iSCSI Models	Host speed: 10Gb/s, 1Gb/s iSCSI | Interface type: SFP+
SAS Models	Host speed: 12Gb/s, 6Gb/s SAS | Interface type: HD Mini-SAS
System Configuration	
System Memory	16GB per system (4005), 32GB per system (5005)
Volumes per System	1024
Cache	Mirrored cache: Yes | Supercapacitor cache backup: Yes | Cache backup to flash: Yes – nonvolatile
Management	
Interface Types	10/100/1000 Ethernet, Mini USB
Protocols Supported	SNMP, SSL, SSH, SMTP, HTTP(S)
Management Consoles	Web GUI, CLI
Management Software	Seagate Systems storage management console | Remote diagnostics | Nondisruptive updates | Volume expansion

Figure 7. Seagate Nytro X 2U24 all-flash array specifications

Seagate developed and tested two approaches to implementing the ML/DL development and deployment platform: one is a result of partnering with HPE and the other is a collaboration with NVIDIA and Nexenta. Each approach used similar equipment and software and achieved similar results. Both methods, discussed below, are alternatives for readers to consider.

The Architecture Overview

In the first implementation, Seagate partnered with HPE, a leading enterprise systems and solutions provider, using the HPE Edgeline and Apollo hardware platforms and HPE's OneAI software platform (see Figure 8). HPE's Apollo product line is optimized for high-performance computing (HPC) applications including AI and ML. The HPE Edgeline product line is a family of converged edge products designed for industrial IoT, edge computing, and AI inference applications. Seagate also collaborated with HPE as an early alpha tester of HPE's OneAI DL development and deployment automation software platform. This implementation used Seagate's Nytro X all-flash array storage system to provide high-performance persistent storage.

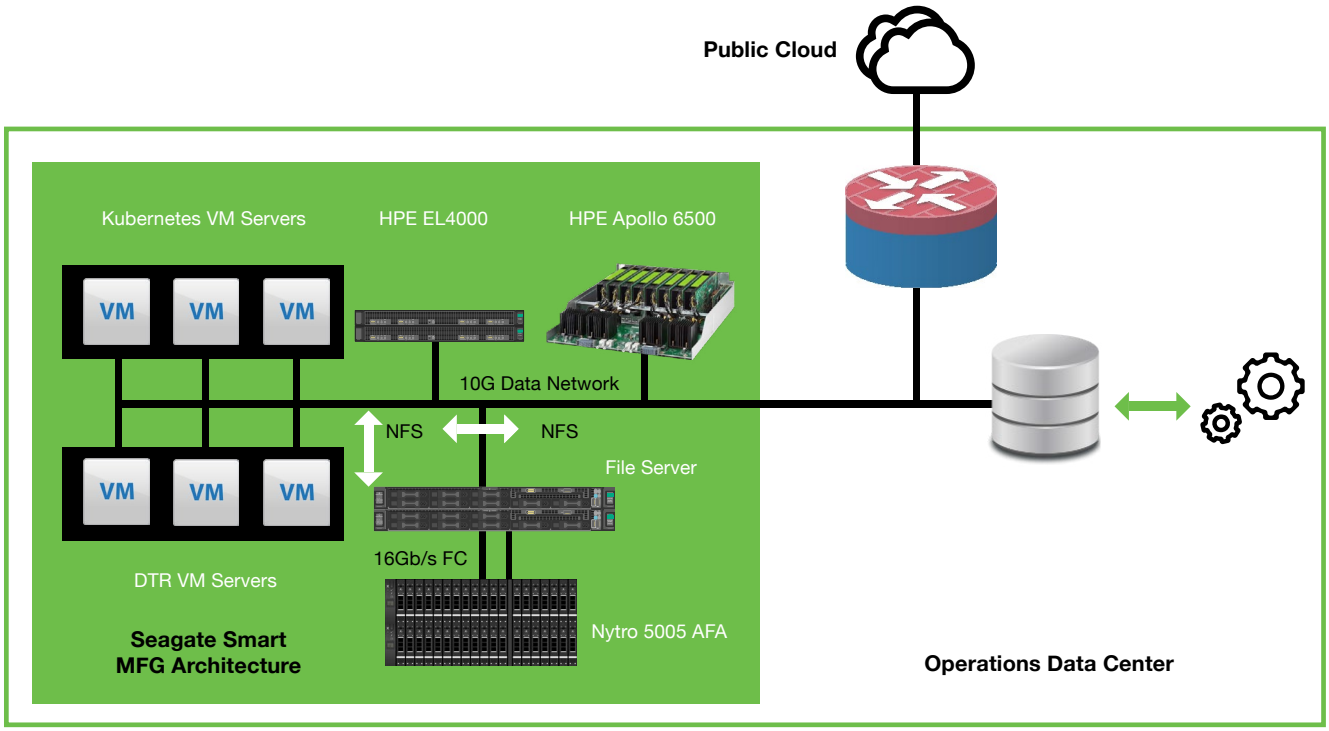

Figure 8. Configuration using HPE Apollo 6500 training server with NVIDIA V100 GPUs, Edgeline EL4000 inference, and NVIDIA P4 including Seagate's Nytro X all-flash array

The HPE Apollo 6500 Gen 10 is an optimized DL training platform containing eight NVIDIA® Tesla® V100 GPUs to accelerate neural network models and deliver up to 125 TFOPS of single precision compute performance. It integrates two server processors supporting up to 3TB of memory and 16 bays for local flash drives. The Seagate Nytro X all-flash array provides high-speed persistent storage using an HPE DL360-configured network file system (NFS) to the Apollo 6500. The HPE Edgeline EL4000 server is designed for large-scale inference workloads for edge computing and factory floor deployments utilizing up to four blade servers, each with one NVIDIA Tesla P4 GPUs, and is capable of running multiple inference models simultaneously.

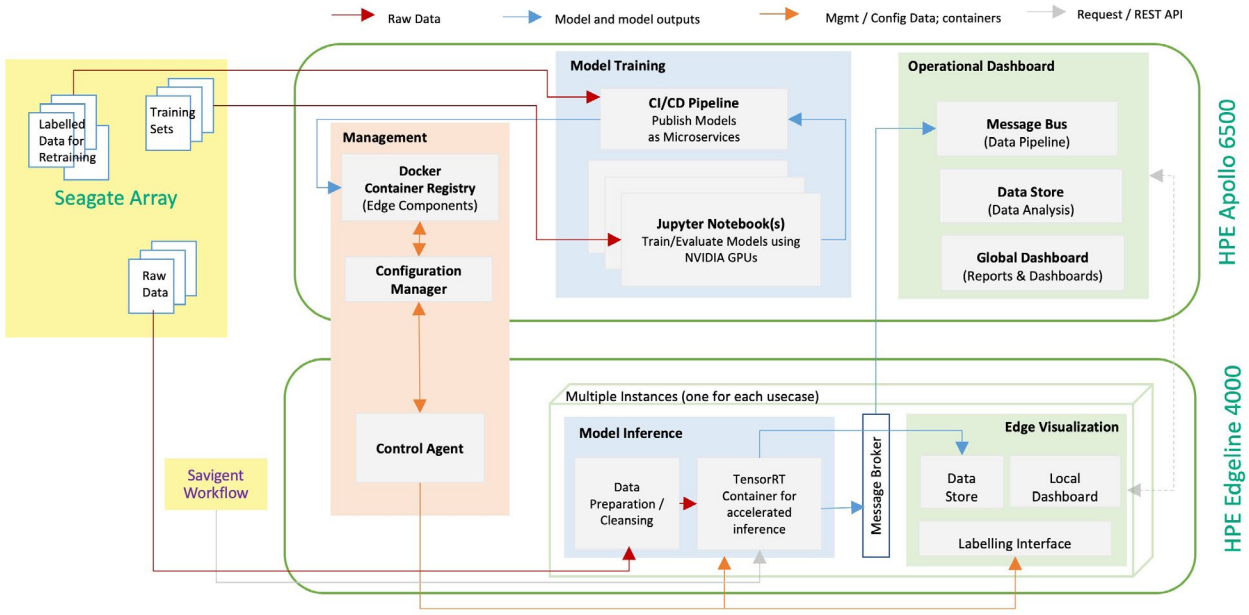

Figure 9. The HPE OneAI software architecture (Image source: HPE)

HPE's OneAI software platform, pictured in Figure 9, is a microservices-based model development and deployment system specifically designed for DL. The microservices environment using Docker containers allows models to be fine-tuned, managed, and easily deployed. The models are developed using Juypter notebooks for creating and sharing DL models within the data science team. The Seagate Nytro X all-flash array is used for storing the training data and inference data.

The DL models developed on the Apollo 6500 are then deployed on the Edgeline EL4000 running in Docker containers using TensorRT run-time and using Apache NiFi for automated dataflow between systems.

An Alternative Architecture

In the second implementation—shown in Figures 10, 11, and 12—Seagate used the NVIDIA DGX-1™ server for training, the NVIDIA Jetson AGX Xavier™ module for inference, and Seagate's Nytro X all-flash array with the Nexenta® NexentaStor™ software-defined storage. The DGX-1 server is a 3U server which contains eight NVIDIA Tesla V100 GPUs, two server processors supporting 512GB memory, and 8TB of local flash storage. The DGX-1 server comes ready to run with a pre-installed software environment running DGX OS, an optimized Docker container run time, access to popular DL frameworks, and third-party accelerated solutions via the NVIDIA GPU Cloud (NGC).

Because the Nytro X is a block-based storage system, it requires a host server to provide the file and object storage protocols. Seagate selected the NexentaStor software-defined storage solution to provide a unified file and block storage solution with the optional NexentaEdge™ providing object storage. NexentaStor, full-featured SDS for enterprise applications, is a proven open-source driven storage software scaling from tens of terabytes to multi-petabytes all-flash, hybrid, and all-disk configurations.

To provide data availability, NexentaStor supports a high-availability configuration with two servers sharing a Nytro X array for active-active failover. NexentaStor also supports continuous or scheduled replication

of snapshots to a remote server to support site-level failover. NexentaFusion™ is Nexenta's included management framework that supports both web-based graphical interface and command line interfaces with REST API's for integration into third-party frameworks. NexentaStor also supports Seagate's other essential requirements, including Docker containers, VMware vSphere and vSAN virtualization, global namespace, as well as hybrid and private cloud implantations.

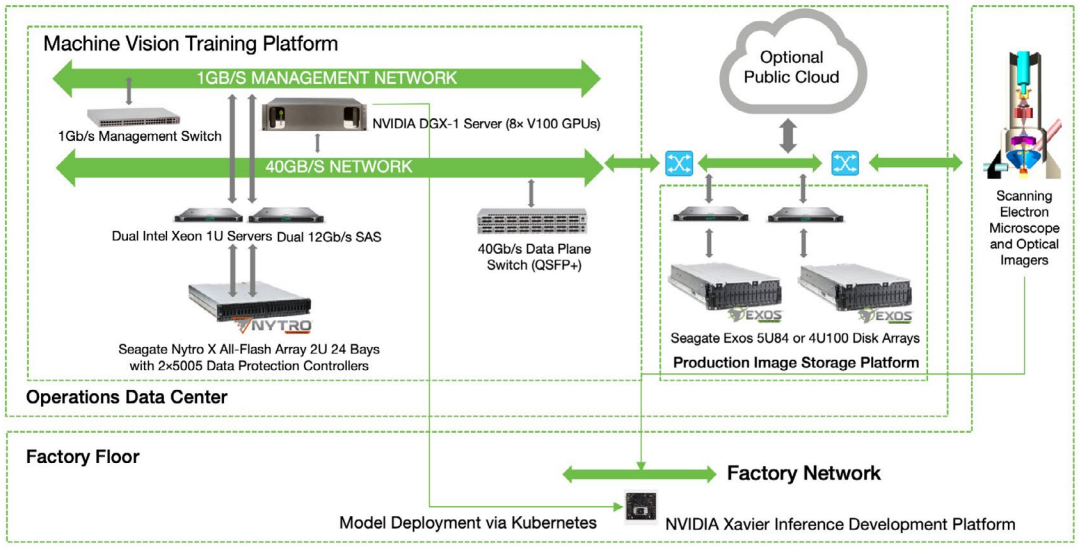

Figure 10. ML platform using NVIDIA DGX-1 server, NVIDIA Jetson AGX Xavier™ module, and Seagate Nytro X all-flash array using the NexentaStor software-defined storage

The training environment is based on Docker containers, which provides a standardized and encapsulated environment for running training workloads. This enables easy setup, process isolation from concurrently running models, and a common environment between the training process and the interference model deployed on end-point hardware.

Each Docker container accommodates NVIDIA CUDA® technology for running deep neural network (DDN) models as well as specific model implementations (application). NVIDIA engineers have developed GPU-enabled Docker containers to allow sharing of the GPU resources among the Docker containers. NVIDIA also provides pre-optimized Docker containers for various DNN models through the NGC platform.

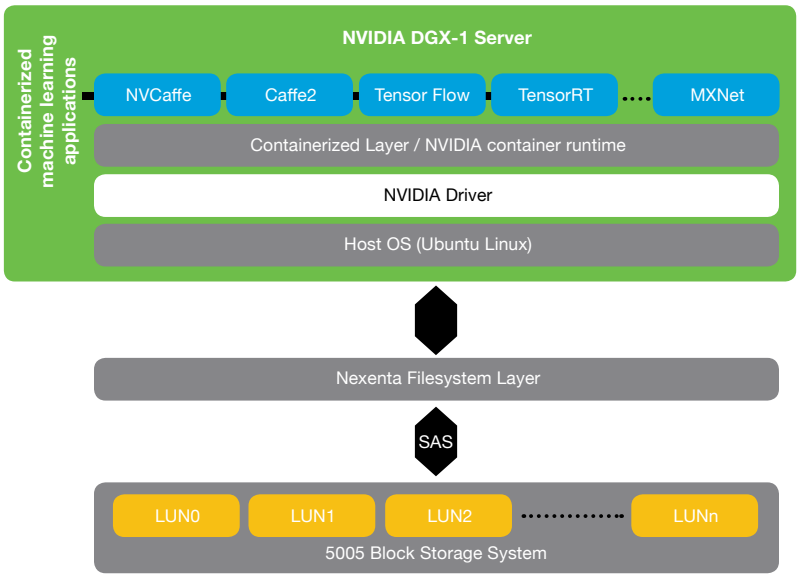

Figure 11. Software architecture of the NVIDIA/Seagate/Nexenta ML training platform

Training Data

The training data set consists of over 100,000 images that have been labeled as True Negative, True Positive, False Negative, and False Positive. The

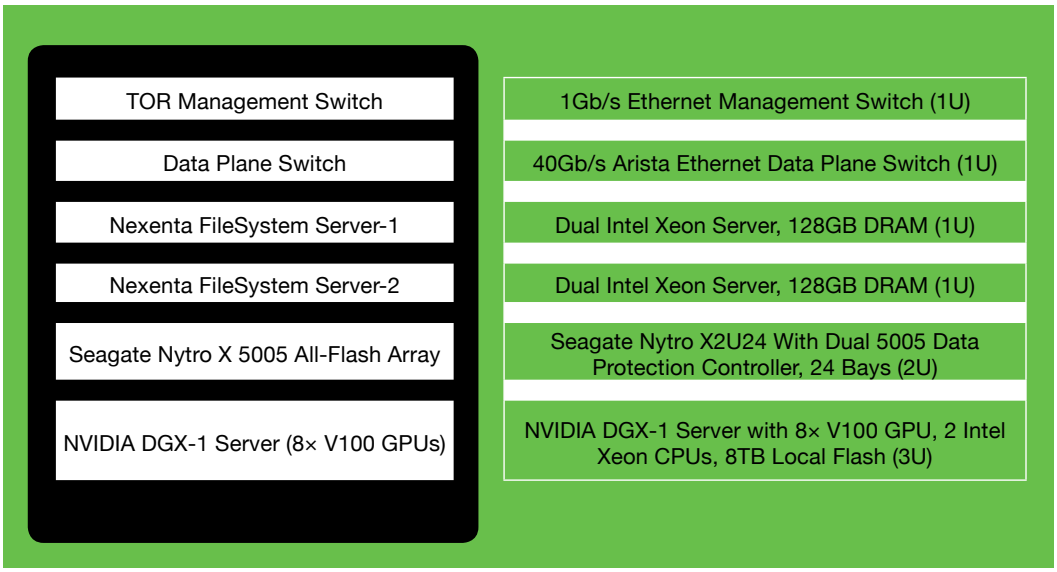

Figure 12. Elevation view of proof-of-concept rack installation for the ML training platform. Inference hardware, factory floor sensors, and image storage archive systems are not shown

original images are 480×680 pixels, 24-bit (three colors, RGB), which were cropped and converted to grayscale as 140×140 pixels, 8-bit images. To balance the categories of the images, Seagate used a 31,000 subset of the training images.

While data labeling remains a challenge for many manufacturing and industrial uses of ML, Seagate has developed a semi-automated method for classifying the training data by preprocessing the data using a clustering similarity algorithm and then presenting sample images to a human expert for validation.

The Seagate ML implementation is typical for image processing applications using a DL approach called supervised learning, which uses convolutional neural networks that must be trained to identify different classes of defects. However, as is typical in most manufacturing processes, the image data produced from optical and SEM images is unlabeled. The training process involves ingesting a large sample image data set into the model that has been labeled as good or with the class of defect to be identified.

The use of DL neural networks for image recognition applications is well covered in the existing literature. While there are many neural network models for image processing, it is left to the implementer to select the best one along with the configuration parameters to optimize the model accuracy. Seagate's team tested 30 different models using the training data to determine the best fit. Each test was run more than 1000 times (each run is called an *epoch*), with the training data randomized for each run to prevent overfitting the model to the data set.

- The size of the training data was limited by several factors:
- Data labeling processing and need for human experts to verify classification
- Periodic changes in process technology and equipment setup limit the training data to less than three months of data
- Time to process data per epoch

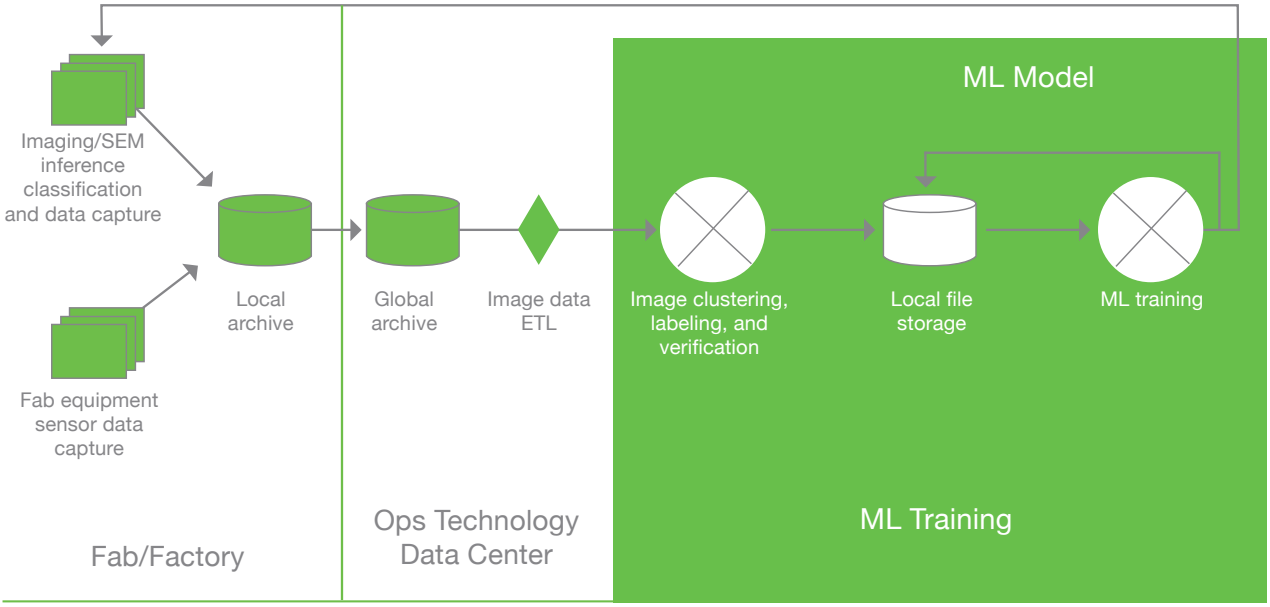

Figure 13. Training data flow

The training data process starts with image data generated by a particular inspection step. This data is stored in a local factory-site data archive and then aggregated to a global data archive. Images used for the training process are cleaned and aggregated in an extract, transform, and load (ETL) process (as shown in Figure 13). The resulting data set is then input into an image recognition clustering algorithm developed by Seagate to preprocess the images and group by similarity. The data set is labeled by defect type and is verified on a sample basis by a quality engineering team at Seagate. The resulting data set is on the order of 30,000 to 50,000 images. A portion of the labeled data set is held from the training data and used to verify the accuracy of the model.

As shown in Figure 14, the process for developing the ML model is iterative and requires repeated testing and tuning to determine the optimal model characteristics and parameters. To shorten the time to

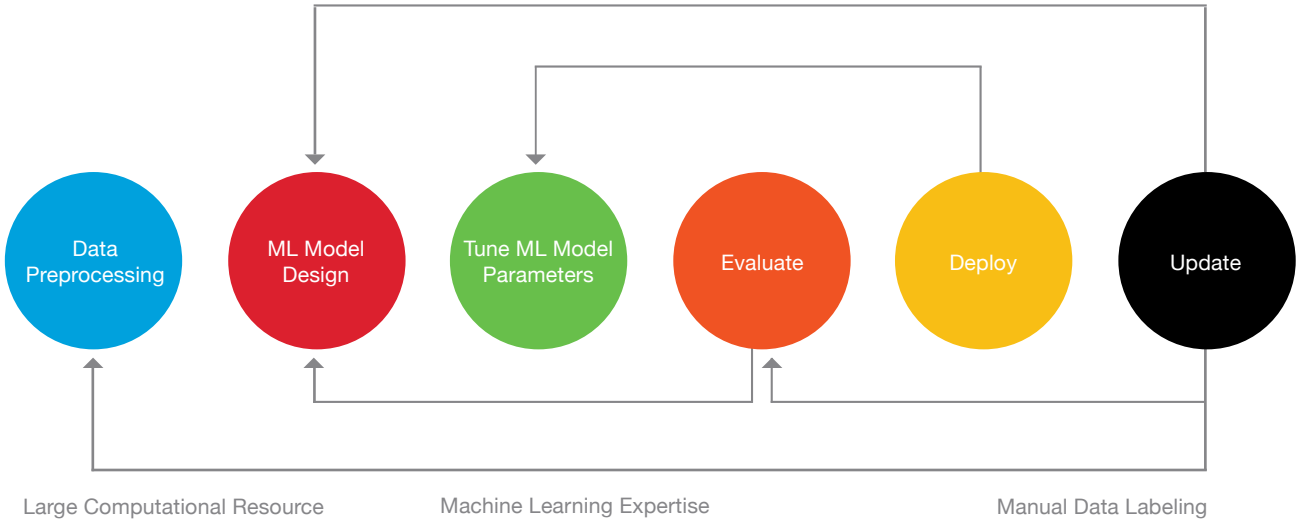

Figure 14. Stages in ML model development

deployment and best use the time of the data science team developing the model, a high-performance training environment is necessary.

Deep Learning Model

Seagate used a DNN model for the classification model. The model uses open source software including the Keras Python Deep Learning Library (https://keras.io/) and TensorFlow (https://www.tensorflow.org/), Seagate used a DNN model for the classification model. The model uses open source software including the Keras Python Deep Learning Library.

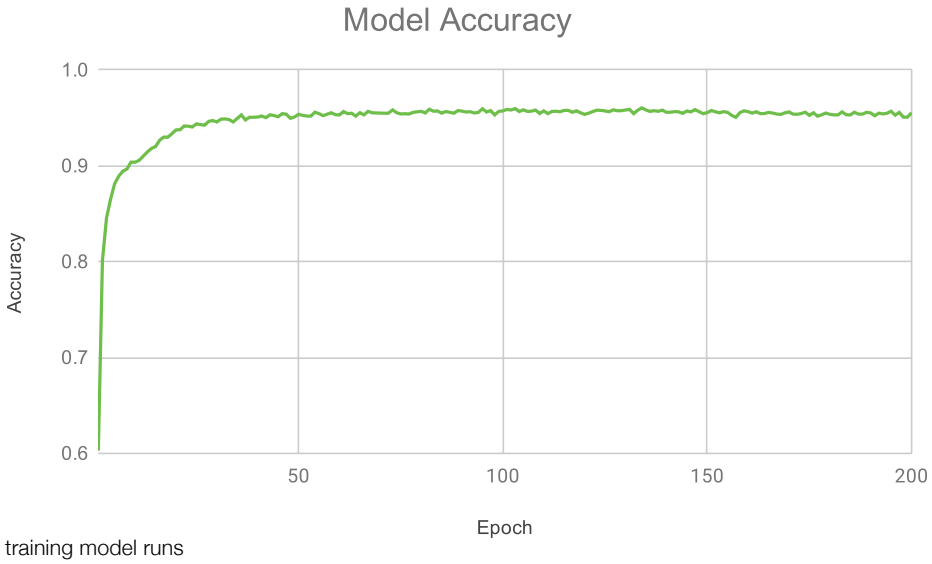

Figure 15. Results of training model runs

Number of Concurrent Models	Image Before Processing/Loading	Time to Run 200 Epochs[2]
1	6.7 minutes	26.3 minutes
2	6.7 minutes	26.3 minutes
4	6.7 minutes	30 minutes
8	7.0 minutes	30.3 minutes

Figure 16. Training processing performance with multiple concurrent models.

Mean	1.82ms
Minimum	1.57ms
Maximum	5.80ms
Time to first inference	1.36 seconds

Figure 17. Interference processing results

Each run of the model, or an epoch, took an average of 7.9 seconds with a minimum of 7.7 and maximum of 8.2 seconds. Seagate ran the model for 200 epochs, which took 26.3 minutes. That model converges after around 50 epochs with an accuracy of over 95%, as shown in Figure 15.

Seagate ran training processing on the DGX-1 server using the Keras model with TensorFlow back end. To simulate the DGX-1 server running concurrent models, multiple training sessions ran in containers where the GPUs were explicitly assigned. Training time depends on many factors, such as the neural network model, the number of classification classes, type and size of training data set, software libraries used, assignment of the training task to GPUs on the DGX-1 server, and number of epochs run.

2 Training time depends on many factors including but not limited to the neural network model, number of classification classes, type and size of training data set, software libraries used, assignment of the training task to GPU's on the DGX-1 server, and number of epochs run.

Conclusion

In collaboration with partners HPE, NVIDIA, and Nexenta, Seagate has developed machine learning solutions focused on large datasets for training ML/DL models.

The Seagate Nytro X all-flash array together with software-defined storage provides a cost-effective solution for persistent storage of training data. The use of a flash-based storage platform provides high-bandwidth storage and low latency while minimizing storage administration overhead. The high-performance storage solution is well matched to the performance of the GPU-based training platforms. This enables the data science team to maximize productivity by running multiple simultaneous models and test various parameters to optimize the model performance during the iterative model development process.

Seagate has deployed this solution in a machine vision defect inspection system used in hard disk read-and-write head manufacturing in one of Seagate's factory sites and is planning to scale the solution to other sites. As this technology becomes incorporated into all of Seagate's manufacturing processes, we expect to see up to a 20% reduction in cleanroom investments, a 10% reduction in manufacturing throughput time and up a 300% ROI from improved efficiency and better quality.

Manufacturing processes are complex. The window of tolerance tends to narrow. Like our factories, images from cameras and similar sensors provide rich sets of information about the process and parts that can be put to use. When we set out to solve our business problem, we did not simply attempt to use something off the shelf. We built this with our partners to solve our unmet need. We are now sharing this to help you with your needs and to improve on the architecture.

We are happy to help with any questions you may have. Send us a note **lyvelabs@seagate.com.**