Training Deep Learning Models Using ThinkSystem SR680a V3, SR780a V3, SR685a V3 Compute Nodes with DDN AI400X2 Storage Nodes

Lenovo Analytics Reference Architecture

Last update: **08 September 2024** Version 1.0

Describes the key storage and compute components needed for a high-performance Generative AI solution

Discusses design considerations for DDN storage platform

Provides considerations for training models for optimal performance

Provides bill of materials to allow customer to build their own system

Rob Reviere

Steve Eiland

LENOVO PRESS

Abstract

Training of deep learning models, including Generative AI (GenAI) and its subset Large Language models (LLMs), require data movement through the network to be efficient and rapid with little to no data backups. Lenovo approaches this challenge through powerful high performance data center appliances (SR680a V3, SR780a V3, and SR685a V3) that support NVIDIA's 8-way GPUs and high-performance storage using DDN's AI optimized storage appliances. The architecture described below considers data movement during training and GPU utilization as the primary design considerations. This architecture uses the latest NVIDIA H100 and H200 GPUs along with an InfiniBand network topology to deliver the speeds necessary to train large comprehensive models. In the below sections, the components of the architecture are described, an example of a scalable unit is provided, and the bill of materials for this design are included.

Table of Contents

1 Introduction

Deep learning models are mathematical models consisting of interconnected units called neurons and can be envisioned like the human brain that has neurons and pathways between neurons. An artificial neural network architecture consists of hidden layers of neurons between an input layer and the final output layer. The output of one neuron becomes the input to other neurons in the next layer of the network. The term deep is used to convey the concept of many layers within the architecture. In this architecture, the input layer or the previous neuron layer provides inputs to the subsequent layer via a linear combination of outputs based on matrix algebra. These types of models enable computer programs to autonomously discover patterns and make decisions based on massive amounts of data. Deep learning models are used for the use cases of computer vision, robotics, autonomous driving, and generative artificial intelligence (GenAI).

These models have become powerful tools for discovery and innovation. Lenovo's initiatives on this front are designed to mitigate a company's risks encountered when developing infrastructure supporting deep learning. The massive size of these models and the requirements they impose on an infrastructure push the limits of each of its components. Bottlenecks that offered minimal impacts on performance for more basic algorithms can no longer be tolerated. As such, computing, storage, networking, and data management, all need to be considered as a part of an optimized highly performant design.

Performance is not only a function of the infrastructure but is also highly dependent upon the AI model itself. Generative AI models and LLMs are especially computationally intensive and must be optimized and tuned accordingly. Significant improvements in workload performance and usage cost for compute resources can be gained by using optimized software, libraries, and frameworks, which leverage highly performant accelerators, parallelized operators and fully utilized cores. These challenges and approaches to address them are further detailed in the compute layer GenAI reference architecture.

This document focuses on the role that both the compute layer, network topology, and storage layer play when training deep learning models having parameters in the high 100s of billions or trillions. For these architectures, Lenovo's new 8-way data center appliances: ThinkSystem SR680a V3, SR780 V3, and SR685a V3 serve as the compute layer nodes. In these designs, the AI400X2 Turbo, AI400X2, AI400X2-QLC DDN machines are used for the storage layer. Although the focus of this document is on training, the aforementioned configurations can also be used for inferencing; however, there are more cost-effective configurations if inferencing is the primary goal. The sections that follow address the challenges above, provide optimized reference architectures, and discuss the appropriate solution hardware.

1.1 Audience

The intended audience are CIOs, CTOs, IT architects, system administrators, and those with an AI background who need to be equipped with the knowledge and insights to navigate the complex landscape of AI-powered technologies.

1.2 Purpose

The purpose of this document is to provide a modular foundational design for the compute, storage and network topology required for training deep learning models. The class of deep learning models that constitute

Generative AI especially push the boundaries of all technologies involved and the challenges are further compounded by the massive amounts of data needed to train these models. As such, every bottleneck in the flow of data needs to be analyzed carefully and resolved. This document will address these considerations and it will provide the building blocks for large-scale deployments. The intent of this document is to provide design considerations and best practices for implementing a deep learning solution. The actual use case(s) will dictate the final design; consequently, design workshops are highly recommended at the start of one's AI journey.

1.3 Common Use Cases

Training of foundational models covers use cases for computer vision, robotics, and Generative AI (GenAI) along with its subset large language models (LLMs). In addition, these former separate approaches are being combined to offer new functionality. Computer vision can cover examples from defect identification, smart city applications, inventory management, and image classification to name a few. Robotic examples cover the gambit from manufacturing automation of processes to robotic surgery to automated transportation. The types of GenAI use cases are as varied as the industry verticals to which they pertain. Across industry examples include:

- Generative AI: Large language models, stable diffusion models
- Denoising raw data
- Creating simulated datasets
- Translation of text into various languages
- Classifying and organizing feedback
- Analyzing and summarizing content
- Generation of original art, music, 3D models, audio, video, and code
- Cybersecurity: anomaly detection, malware detection, intrusion detection and encryption
- Chatbots for customer service interfaces
- Natural language processing for call centers
- Fraud and threat detection in the financial industry
- Medical imaging and diagnostic applications

This by no means is comprehensive; however, it provides a landscape of possibilities.

2 Architectural Overview

The architecture developed represents components optimized to meet the computationally intensive deep learning workloads, especially Generative AI (GenAI) and large language models (LLMs). The architecture's compute layer is built with the Lenovo ThinkSystem SR685a V3, SR680a V3, and SR780a V3 8-way GPU Tensor Core appliances using NVIDIA H100, H200 and limited released B200 GPUs. These nodes support NVLink interconnects at 900Gb/s for the H100/H200 and 1800Gb/s for B200. A comparison between NVIDIA's H100 to NVIDIA's H200 can be found here and information on the NVIDIA B200 can be found here. These compute appliances support GPU complexes from NVIDIA, AMD, and Intel; however, in this reference architecture only NVIDIA 8-way GPUs will be presented. Future RA's will discuss these other technologies. Further details on the compute layers hardware are provided in the compute layer section below.

The SR685a uses AMD 4th generation EPYC™ processors, while the SR680a and SR780a are based on Intel's 5th generation Xeon processors. The AMD x86 server processors have up to 96 "Zen 4" cores and 1152 MB of L3 cache per socket. This series of processors delivery high performance per core with the highest thread density and the largest L3 cache. All processors have 12 DDR5 memory channels and 128 PCIe 5.0 I/O with 64 lanes available for PCIe and NVMe devices. Intel's Xeon processors, code named Emerald Rapids, have 64 cores per processor and 105 MB of L3 cache. These processors have 8 DDR5 memory channels and 80 PCIe 5.0 lanes.

All architectural components are connected to a NDR 400 InfiniBand fabric for ultra-low latencies, except for the DDN network, which is based on NDR 200 because NDR 400 is not currently supported by DDN (future support is expected). Although this architecture is not based on high-speed Ethernet, some comments on an all-Ethernet topology are provided where comparisons are warranted. Generally speaking, this architecture can be modified for high-speed Ethernet with minimal changes to switch types and network topology.

The storage layer of this architecture is based on the 3 variants of the fully contained DataDirect Network (DDN) 2U AI400X2 appliances. These are the AI400X2-QLC, the AI400X2 Turbo, and the AI400X2. The AI400X2-QLC is for maximum useable capacity up to 5 PB and provides more storage per node. The AI400X2 Turbo is for maximum read/write performance and provides much more performance density. The AI400X2 is the appliance that balances useable capacity with performance. All appliances support NVIDIA ConnectX-7 NDR200 InfiniBand/Ethernet 200GbE 2-port QSFP112 PCIe5 x8 adapters, which provide 200 Gb/s connectivity for both Ethernet and InfiniBand protocols. These adapters are suitable for providing high-performance connectivity when running HPC and deep learning applications, see the product guide for further details. When deciding upon which DDN appliance to use for a given use case, first the performance calculation to meet the need should be determined. Next, the capacity should be determined. The storage layer is built upon on DDN AI400X systems. Further details on the storage appliances are supplied in the below DDN section.

2.1 Architecture Design

The architecture is a 2-level design consisting of level 1 compute leaves, level 1 storage leaves, and level 2 spines. This all-InfiniBand (IB) network uses NDR400 except for the storage layer that uses NDR200. The management node and the DDN Insight node use either the SR635 V3 (for SR685a V3 compute node design) or the SR630 V3 (for the SR680a V3 and the SR780a V3). For manageability, an all AMD based

system is used for the SR685a V3 design and an all Intel based system is used for the other designs. Said differently, there is no mixing of AMD CPU and Intel CPU technology.

The below figure shows this architecture. In this case, the interconnect topology is based on SR685a V3 servers along with the appropriate NVIDIA IB switches. Also, shown are the major components and their connections along with connections to DDN storage nodes. Included in the overall design is a management/OS node and a node for DDN Insight software. The same configuration can be used for a system based on the SR680a V3 nodes or SR780a V3 nodes.

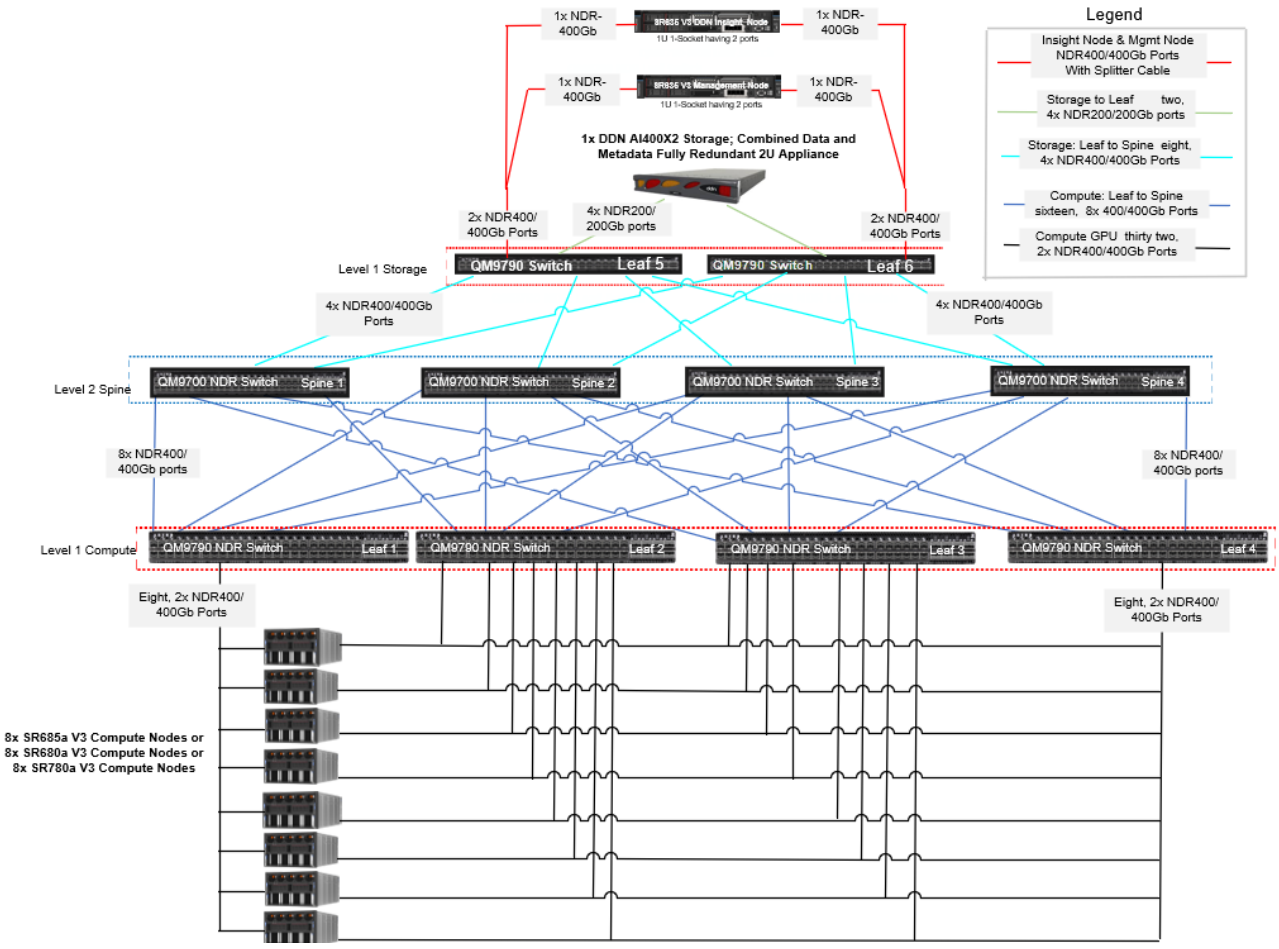

Figure 1: Reference Architecture Showing Compute Nodes SR685a V3 (same connections for SR680a V3, SR780a V3), DDN AI400X2 Storage Nodes and a Unified InfiniBand 400Gb NDR Fabric.

This design supports fail-over redundancy and has a high available network (North/South and East/West bound traffic) that is non-blocking with a Fat Tree Topology. The architecture is a rail optimized design with separate compute and storage networks connected by means of a common spine. This approach to architecture facilitates excellence training performance and growth from small scale to massive scale operations in an economical way.

The compute layer of this design consists of eight 8-way GPU compute nodes connected to four NVIDIA QM9790 unmanaged compute leaf switches, which act as the compute rail. Each server is connected to each compute leaf switch. All compute servers have 2 single-port ConnectX-7 400 Gb/s NDR network connections

for inter-GPU system communication. Each compute leaf switch is connected to four NVIDIA QM9700 managed spine switches. All leaf and spine switches are connected for complete redundancy and high availability.

The storage nodes, based on DDN's AI400X2, are connected to their own storage leaf switches using NVIDIA QM9790 unmanaged switches that act as the storage rail. Each storage server is connected to each storage leaf switch and is configured with ConnectX-7 200 Gb/s NDR InfiniBand ports and PCIe 5.0 x16 slots (400 Gb/s NDR is planned). These nodes support GPU direct functionality to maximize GPU performance.

To deploy the architectural design, a deployment/management server is also included. This management node is a SR635 V3 appliance (for SR685a compute system) or a SR630 V3 appliance (for SR680a V3 or SR 780a V3). These nodes are 1U 1-socket configurations connected with a splitter cable to the 400 Gb/s NDR network and the storage leaf switches. In a similar fashion, the DDN insight node (SR635 V3 or SR 630 V3) is connected via a splitter cable to the same storage leaf switches.

InfiniBand and Ethernet are both popular networking technologies used in high-performance computing (HPC) environments, including generative AI clusters. The choice between the two depends on a range of factors. Cost effectiveness and setup simplicity tilt the scale toward an Ethernet topology. For demanding workloads in terms of low latency communication and high-speed data transfer, InfiniBand is the preferred choice. InfiniBand's benefits become more pronounced when dealing with short-range, high-speed communication within a data center. It is for these reasons that InfiniBand is chosen as the network technology for this design. This network uses the latest generation NVIDIA Quantum InfiniBand (IB), which provide high-speed, high-bandwidth, low CPU overhead, highly efficient, and low-latency data transfers between applications, compute servers, and storage appliances. For this architecture, the InfiniBand Subnet Manager is a centralized entity being run in the switch as opposed to on a server. The subnet manager discovers and configures all the InfiniBand fabric devices to enable traffic flow between those devices. The switches employed provide ultra-high data throughput and density required of highly parallelized algorithms common for deep learning models.

For scaling, the concept of a Lenovo scalable unit (SU) is employed using a rail optimized design. The SU's are based on multiples of 8 compute nodes. Traffic per rail is always one jump from the other nodes in this scalable unit. Traffic between nodes or between rails transverses the spine layer. The storage devices are not part of the SU because often the compute SU will expand while the storage remains the same. Since storage scales linearly, it can be increased separately as needed. The Lenovo scalable unit (SU) consists of the following:

- 8 compute nodes (either the SR685a V3, SR680a V3 or SR780a V3),
- 4 unmanaged NDR compute leaf switches (NVIDIA QM9790s),
- 4 managed spine switches (NVIDIA QM9700),
- 2 unmanaged storage leaf switches (NVIDIA QM9790s),
- 1 DDN AI400X2 storage nodes.

The scaling is illustrated in the below diagram. The 8 leaf switches provide 1 leaf per GPU in the hosts. Once there are enough hosts to fill that leaf, a cookie cutter approach is used for up to 16 leaf switches. There are only 8 links per host; therefore, the original hosts remain on the original switches with the next N nodes being on the new 8 leaf switches. At this point in the expansion, more spines are needed. These reusable blocks

can be scaled any number of times to match the use case demands. For NVIDIA partners, there is a scaling tool that provides additional configurations consistent with the one presented here.

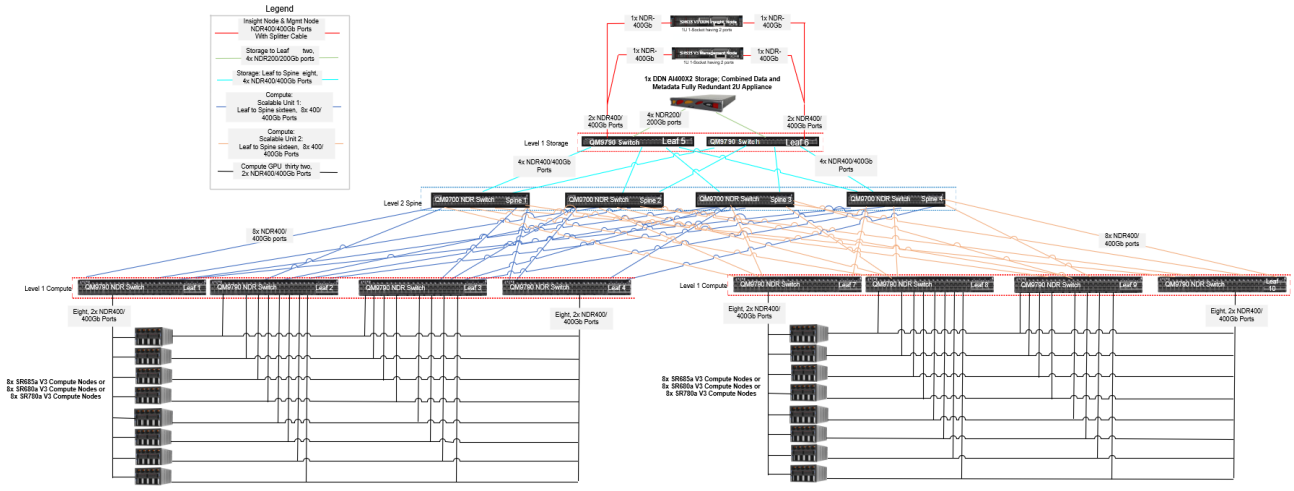

Figure 2: Scale out; Two Lenovo Scalable Units(SU) Based on Each Unit Having 8 GPU Nodes with 4 Compute Leaves, 4 Spine Switches, 2 Storage Leaves and 1 DDN AI400X2 Box.

Guidelines for sizing are provided in the below 2 tables. The first table provides MLPerf results for the various GPUs including those that can be utilized in this design as well as GPUs that can be used for inferencing on other appliances like workstations (e.g., RTX4090). These results are from MLPerf standardized testing and are not hardware specific. All results are for one GenAI model type, resnet50, with the results for GH200, H100 NVL, and H200-SXM-141GB-CTS being extrapolated from Llama2 model benchmark results because resnet50 results were not available. Duration refers to total training time under the benchmark test conditions. The price per GPU was based on available pricing at time of publishing and is used as an order of magnitude estimation.

NVIDIA GPUs	GPU count	GenAI Model	Duration (min)	Duration extrapolated/ GPU	Memory (GB)	TDP (W)	Price (US$)/ GPU	Consumed power to train (W)
RTX 4090	6	resnet50	122.076	732.453	24	450		5493.40
RTX A5000	2	resnet50	271.193	542.386	24	230		2079.15
A30	8	resnet50	58.731	469.850	24	165		1292.09
L4	6	resnet50	170.378	1022.270	24	72	$ 1,747	1226.72
L40	8	resnet50	48.353	386.822	48	300	$ 5,628	1934.11
L40S	8	resnet50	50.592	404.736	48	350	$ 5,702	2360.96
A100-PCIe-80GB	8	resnet50	29.862	238.898	80	300	$ 10,974	1194.49
GH200	1	resnet50	76.179	76.179	144	1000		1269.64
H100 NVL	1	resnet50	131.385	131.385	94	800	$ 21,113	1751.80
H100-PCIe-80GB	8	resnet50	20.831	166.648	80	350	$ 21,100	972.11
H100-SXM-80GB	8	resnet50	13.303	106.422	80	700	$ 171,711	1241.59
H200-SXM-141GB	8	resnet50	12.078	96.621	141	700	$ 173,916	1127.24
H200-SXM-141GB-CTS	1	resnet50	73.344	73.344	141	1000		1222.39

The second table provides a scaling guideline for this architecture based on the DDN AI400X2 storage node and the Lenovo scalable units (SUs) consisting of 8 compute nodes per SU.

Item	Feature	Lenovo Scalable units (SUs), 8 compute node basis						
		1	2	3	4	5	6	7
Compute components	ThinkSystem SR680a V3, SR780a V3, SR685a V3	8	16	32	64	96	128	256
	GPUs: H100, H200	64	128	256	512	768	1024	2048
DDN storage components (scale linearly)	AI400X2 combo appliances	1	2	4	8	12		
	AI400X2 metadata appliances						3	6
	AI400X2 data appliances						16	32
DDN storage specifications	Aggregate read throughput (GB/s)	90	180	360	720	1000	1400	2800
	Aggregate write throughput (GB/s)	65	130	260	520	780	1000	2000
	Per GPU read throughput (GB/s)	1.4	1.4	1.4	1.4	1.4	1.4	1.4
	Per GPU write throughput (GB/s)	1	1	1	1	1	1	1
	Useable capacity (500TB appliance option) (PB)	0.5	1	2	4	6	8	16
	Useable capacity (200TB appliance option) (PB)	0.25	0.5	1	2	3	4	8
	NDR200/200 GbE ports	8	16	32	64	96	152	304
	1 GbE ports	4	8	16	32	48	76	152
	Nominal Power (KW)	2.2	4.3	8.6	17.3	25.9	40.2	80.5
	Nominal cooling (KBTU/hr)	7.4	14.7	29.5	58.9	88.4	137.2	274.5

The racks employed for these builds should utilize the Lenovo Hercules 42U or similar racks with the power PDU parallel to the cage and their power outlets positioned out of the way. Each rack should consist of a complete SU without splitting scalable units between separate Hercules racks. Final design is dependent upon available space within a cage.

The BOMs for this design are provided in the BOM sections below. The actual comprehensive configuration implemented depends on the use case, the organization's existing infrastructure, the existing technical staffing, and the budget.

2.2 Solution Components

2.2.1 Optional Software

Optional software can be included in the design. The choice of software is dependent upon the use cases and the customer's needs. Some of the software components are part of an NVIDIA offering and there need depends on whether the customer already has these in-house capabilities. Other components are dependent upon the in-house IT infrastructure, for example NVIDIA Base Command is advised if multi-tenancy is required. Other components are open-source and have a management role. All these components are designed to optimize performance for deep learning loads. The below table provides the software options.

Component	Features
Kubernetes (K8s)	Open-source: container management and orchestration
Slurm Workload Manager	Open-source: fault tolerant, Linux cluster management and job scheduling system
AI Enterprise	NVIDIA: development tools, production-ready containers, framework for rapid deployment of AI workloads
Base Command Manager Essentials	NVIDIA: Comprehensive cluster management solution that automates provisioning and administration on clusters into thousands of nodes
NGC	NVIDIA: Collection of GPU-optimized containers for AI and HPC
NIM	NVIDIA: Inference Microservices (NIM) providing instant access to AI models that can run anywhere.
Magnum IO	NVIDIA: Enables AI performance increases through and SDK
UFM	NVIDIA: Platform to maintain high-performance InfiniBand networking fabric, which allows data center operators to efficiently provision, monitor, manage, and troubleshoot their fabric
NCCL (required) RCCL (only for AMD GPU networks)	NVIDIA: collective communication library implements multi-GPU and multi-node communication primitives optimized for NVDIA GPUs and networking. This is a requirement for optimal performance of a rail optimized design and provides multi-GPU path optimization. AMD: has similar capability if AMD GPUs are used. *stand-alone library that provides* multi-GPU and multi-node collective communication primitives optimized for AMD GPUs.

Insight	DDN: Software to monitor and maintain the spectrum of IO resources and workloads on DDN storage network
LiCO	Lenovo: Software to simplify the use of clustered compute resources for AI development

2.2.2 GPU Design

The primary advantages of GPUs, in the context of model training, are high memory bandwidth, parallelization, and faster (and more) memory access. These characteristics heavily influence model training time. To exploit these capabilities, this architecture is based on the latest NVIDIA GPUs. A detailed description of this GPUs can be found in the NVIDIA H100 Tensor Core GPU Architecture document. A comparison of GPUs commonly used for deep learning training is provided below with further details on the workhorse NVIDIA H200 GPU found here.

	INSTANCE TYPE	GPU	GPU MEMORY	vCPUs	STORAGE	NETWORK BANDWIDTH
NVIDIA H100	8x NVIDIA H100	H100 SXM	80 GB	224	30 TB local per 8x H100	3200 Gbps per 8x H100
NVIDIA H200	8x NVIDIA H200	H200 SXM	141 GB	224	30 TB local per 8x H200	3200 Gbps per 8x H200
NVIDIA GH200	1x NVIDIA GH200	GH200 Superchip	96 GB	72	30 TB local per GH200	400 Gbps per GH200
NVIDIA B200	8x NVIDIA B200	B200 SXM	180 GB	224	60 TB local per 8x B200	3200 Gbps per 8x B200

Figure 3: GPU Specifications for Widely Used NVIDIA GPUs for AI Training

Any GPU processor has three core functions: compute, memory, and communication. These capabilities are not enough for the high demands of training operations, which is why added capabilities like NVIDIA GPUDirect® are needed. The family of DDN AI400X2 (discussed below in DDN section) appliances integrate with GPUDirect, which provides a direct path between the GPU memory and the external storage. This is shown in the below image that is based on an Ethernet topology; however, a similar configuration exists for InfiniBand networks. For this technology, data movement is routed through an optimal network interface to the nearest available GPU.

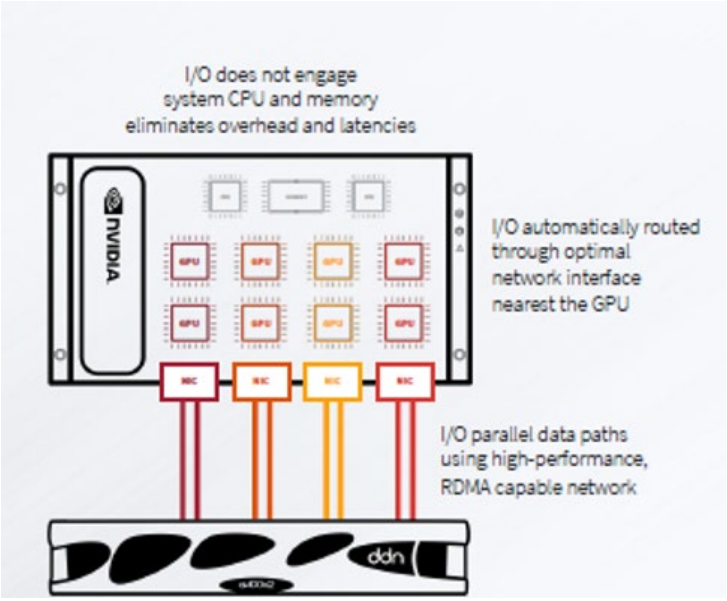

Figure 4: DDN AI400X2 with NVIDIA GPUDirect Storage over Ethernet

This technology enables data movement in and out of GPU memory without the need to go through the CPU; consequently, avoiding extra copies of data through a bounce buffer in the CPU's memory. Without GPUDirect, a bottleneck is created resulting in adverse effects on application performance and increased training time resulting from slower data loading. This challenge becomes especially prevalent with increases in dataset size and model complexity, which are the hallmark of deep learning training.

The NVIDIA 8-way GPU design is another enhancing capability for training deep learning models. The figure below illustrates this design. This design allows all GPUs to work in parallel. Every GPU is available to every other GPU supporting advanced GPU-to-GPU communications and high-speed NVLink/PCIe 5.0 connectivity between the GPUs, processors, and networking.

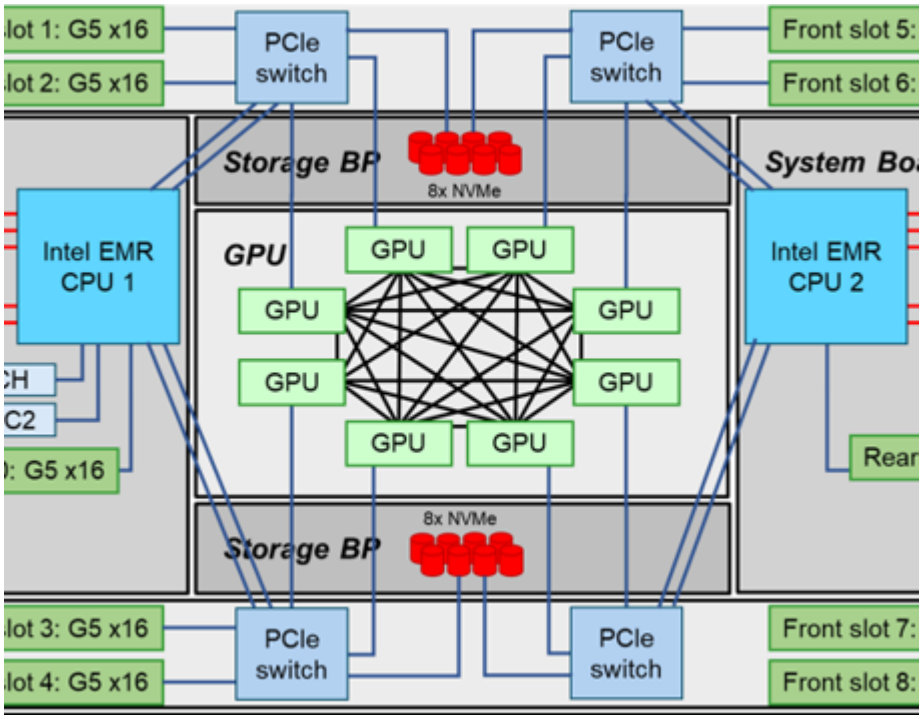

Figure 5: Excerpt from the 680a V3 Block Diagram Showing the 8-way GPU Design.

With all GPUs running simultaneously and in parallel, the training workload can be distributed resulting in large reductions to training time. In addition, large batches of data can be processed in parallel resulting in overall throughput increases. Further consideration regarding GPU utilization for deep learning training can be found here.

2.2.3 Model Training Implications on Architectural Design

For training deep learning models, especially GenAI models, it is all about data movement and matrix calculations. Architectural designs must consider these for training time to be substantially decreased. Both hardware and software considerations are needed. On the hardware side, special purpose accelerators, like the H100/H200/B200, and GPU architectures, like 8-way GPUs, are used. On the software side, optimized libraries, and frameworks, like PyTorch are used to improve training operations.

Training times are also improved by strategies that improve GPU utilization and performance. These strategies include mixed precision training, optimizing data transfer and processing, and breaking the training operations into smaller blocks. Of these the most dominate driver is data loading during the training process, which is limited by the rate at which data can be read and reread from storage. The key to performance is the ability to read data multiple times, ideally from local storage. The closer the data is cached to the GPU, the faster it can be read. Caching data in local RAM provides the best performance for reads. While local storage is fast, it is not practical to manage a dynamic environment with local disk alone. It is not always possible to load the needed training data entirely into fast local storage. In this case, data chucks are needed from external storage that feed the modeling process. As such, both persistent and nonpersistent storage needs to be designed to balance the needs of performance, capacity, and cost while minimizing training time. For a practical guide see, How do GPUs speed up Neural Network training? (youtube.com).

2.2.3.1 GPU starvation

A rate limiting step occurs when GPUs become starved for data because of limits to data transfer from storage. In this case, the GPU processing must wait for data to arrive. For training complex models, this becomes especially apparent. Studies by Google and Microsoft have shown that up to 70% of the training time is associated with waiting for data. At the start of each training epoch, described below, training data is kept on high-capacity object storage and then copies are produced, and these are sent to a Lustre (DDN EXAScaler) storage system. The next step in the training data pipeline is to copy data to local GPU storage (NVMe) before it is finally moved into the GPU. Each of these data hops adds time to the overall training cycle.

Other causes of starvation are unoptimized data pipelines and inefficient parallel processing with uneven workload distribution across GPU cores. Deep learning training models require high memory bandwidth; therefore, anything that causes insufficient bandwidth will slow training cycles. Resource clashing where multiple processes/applications compete for the same GPU resources can cause GPU starvation and thus slow the training process.

2.2.3.2 Training Epoch

To appropriately design a deep learning training architecture, one must, at least have a high-level, understand the operations that occur during the training process, especially factors that influence the training epoch operations. First, there are some concepts related to deep learning modelling that need understanding. All deep learning models consist of an activation function, input layers, output layers, hidden layers, loss functions and other attributes. Deep learning models try to generalize the data using an algorithm that to make predictions about new related data. This algorithm maps inputs to outputs with parameter values (weights) that minimize the error when mapping the two. An optimization algorithm finds parameter values that minimize the error between the inputs and the predicted outputs.

A training epoch is defined as one complete pass through the training data by a training algorithm. Within an epoch, the training model processes, using matrix multiplications, a subset of the training data called a batch. Once a batch is completely processed, the model updates its internal parameters, which is referred to as a model iteration. The number of epochs can be 100s or 1000s and is a key hyperparameter in training neural networks. The number of epochs is set to allowing the learning algorithm to run until the error of the model is minimized. Too few epochs and the model can be underfit. Too many epochs and the model can be overfitted.

The outcome of a training process is stochastic and is dependent upon the effectiveness of the optimizer utilized. The overall training time and the individual epoch step times are dependent upon whether the model converges to the global minima of the loss function. This is influenced by many factors. If the system's components (compute/network/storage) produce fluctuations in performance or there is inconsistent peak performance, errors in the intermediate calculations can occur resulting in obstacles to model convergence and thus lengthened training time.

To achieve optimal results, it is crucial to consider the following:

1. **Optimizer selection**: The choice of optimizer significantly impacts the outcome of the training process. Popular optimizers like Stochastic Gradient Descent (SGD), Adam, RMSProp, and Adagrad have different strengths and weaknesses.

2. **Convergence criteria**: Setting appropriate convergence criteria is essential to ensure that the model converges to a stable solution. Common metrics include mean squared error (MSE) or cross-entropy loss.
3. **System performance fluctuations**: Variations in system components can affect training time and accuracy. Ensuring consistent peak performance from compute, network, and storage resources is vital for efficient training.

To mitigate these challenges:

1. **Monitor and adjust**: Regularly monitor the training process and adjust hyperparameters or optimizers as needed to achieve convergence.
2. **Use robust optimizers**: Select optimizers that are less prone to getting stuck in local minima, such as Adam or RMSProp.
3. **Implement regularization techniques**: Use techniques like dropout, L1/L2 regularization, or early stopping to prevent overfitting and improve model generalization.

By acknowledging these factors and taking steps to mitigate their impact, the deep learning training process can be optimized leading to better performance and faster model convergence.

Training deep learning models require datasets to be re-read many times during the multi-epoch learning process placing heavy loads on compute nodes, networking, and storage.

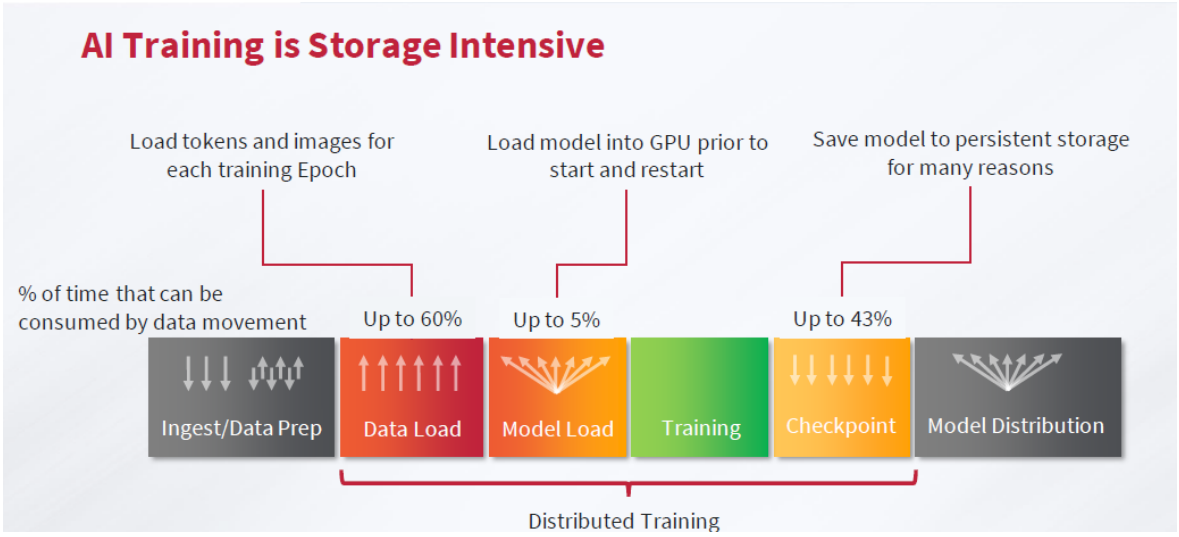

Figure 6: Distribution of Training Time by Activity

During training, waiting on data can be a critical bottleneck. As models scale the data movement component for multi-Epoch training becomes the dominate factor in training time. The DDN system has an approximate 3x increase speed in data loads and 15x faster checkpoints training time is drastically reduced by using the DDN system.

2.2.3.3 Checkpoints

The checkpoints are controlled typically by code when training starts. Checkpointing is a crucial strategy in machine learning and deep learning that provides HW Failure Protection by creating frequent checkpoints that

allow for rapid restarts in the event of hardware failures, minimizing downtime and loss of progress. Beyond this, checkpoints are valuable for improving prediction accuracy, multi-system training, transfer learning, better fine-tuning, and early stopping. Very fast write performance in the storage environment allows for more frequent and less costly checkpointing, making the process more efficient and less disruptive, especially in environments that require continuous training or are susceptible to interruptions. The below figure shows the relationship between training operations and the number of epochs and it relationship to checkpoints.

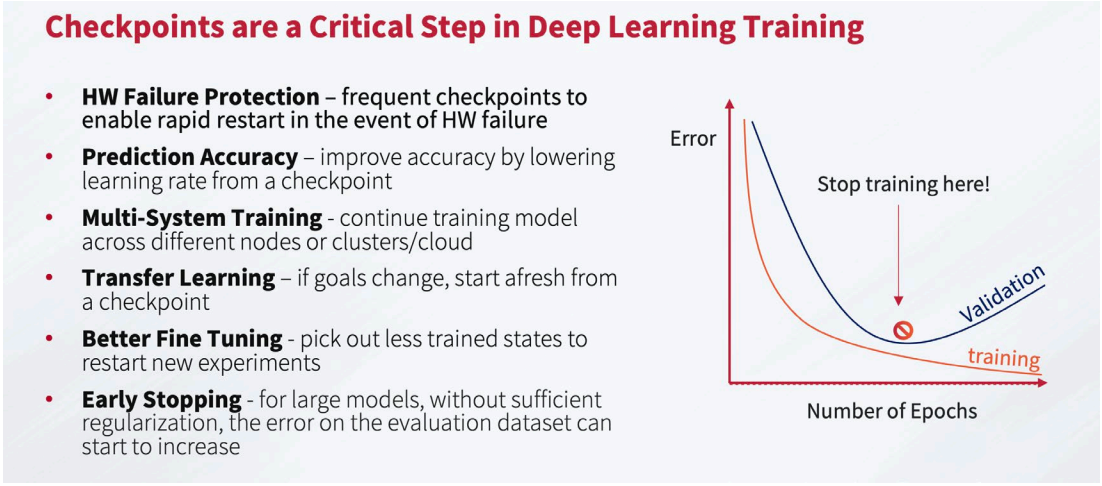

Figure 7: Checkpoints, Training Epoch and Model Error Minimization for Model Convergence

3 Compute Layer

The compute and data-intensive applications that train deep learning workloads demand maximum performance. It is important to find a balance between performance, power consumption and scalability. The SR680 V3 (Monza), SR780 V3 (Marina Bay), and SR685 V3 (Monaco) Think Systems are all AI ready nodes that meet this balance. These appliances are versatile GPU-rich rack servers that support NVIDIA's high-performance 8-way Tensor Core GPUs: H100, H200 and B200 (limited) and the AMD CDNA3 GPU architecture -- MI300X GPU (planned).

The following sections contain information on the Lenovo compute layer devices: ThinkSystem SR685a V3 (Monaco), ThinkSystem SR680a V3 (Monza), and ThinkSystem SR780a V3 (Marina Bay). All appliances support the 8-way GPUs complex. These appliances support CPUs from AMD or Intel and all high-performance GPUs complexes from NVIDIA, AMD, or Intel. Monaco and Monza are 100% air cooled thermal solutions while Marina Bay is an 80% water/20% air cooled thermal solution. The SR680a V3 and SR685a V3 servers feature two Intel® Xeon® and two AMD EPYC™ processors, respectively, and are 100% air-cooled servers suitable for most data centers.

These products support customer AI applications by engaging the following features:

- 8U Chassis (Monaco and Monza) or 5U Chassis (Marina Bay)
- AMD CPU complex (Monaco) or Intel CPU complex (Monza and Marina Bay) with system board and BMC I/O board
- GPU Complexes from AMD (Monaco and Monza), NVIDIA (Monaco, Monza, and Marina Bay) and Intel (Monza)
- PCIe Gen5 switch module for high speed PCIe chassis interconnect to support PCIe slots, GPU complexes and storage.
- Air cooled thermal infrastructure (Monaco, Monza) and water-cooled thermal infrastructure (Marina Bay)
- Power infrastructure and distribution with support for up to 8x CFFv4 and CFFv5 (Future) and N+N redundancy

Monza and Marina Bay use the MGX HPM system board and support 2x Intel EMR processors, 32x DDR5 4800MHz DIMMs and multiple PCIe Gen5 ports for high-speed system PCIe expansion. The PCIe Gen5 ports support direct connect to M.2 boot drive, cabled connections to the rear PCIe slots, and the switch board for additional PCIe expansion to the GPU complex, GPU NIC slots, and NVMe drives. All systems have power infrastructure and distribution with support for up to 8x CFFv4 and CFFv5 (Future) and N+N redundancy. The table below provides a quick compare and contrast between systems.

ThinkSystem	SR680a V3	SR685a V3	SR780a V3
Form Factor	8U, horizontal orientation only	8U, horizontal orientation only	5U in standard 19" rack
CPU	Intel Emerald Rapid 2P, up to 350W TDP VROC, 6438M, 6448H ,8462, 8468, 8480 (56 cores)	2S AMD Geno/Turin, up to 350W CPU TDP 48 cores, 64 cores, 96 cores	Intel Emerald Rapid 2P, up to 350W TDP
Memory	32x DDR5 slots with max frequency supports RDIMM/3DS 128GB memory	24x DDR5 slots both with max frequency supports RDIMM/3DS 64GB / 96GB at 4800 MHz (any GPU combination) 64GB / 96 GB/128 memory at 6400MHz	32x DDR5 DIMMs slots with max frequency at 5600MHz supports RDIMM/3DS Up to 4TB (32x 128GB)
Front Drives	16x 2.5" HS SAS/SATA/NVME	16x 2.5" HS SAS/SATA/NVME	up to 8x U.2/U.3 NVMe Direct Connect - Hot swappable
GPU Support	Intel Gaudi3 1000W liquid assist 8x NVIDIA H100 (700W/GPU); 8x NVIDIA H200 (700W/GPU); 8x NVIDIA B200 (1000W/GPU) (limited release); 8x AMD MI300X (air cooled) 700W -750W; with NVLink interconnects at 900GB/s, 1800GB/s for B200	8x NVIDIA H100 (700W/GPU); 8x NVIDIA H200 (700W/GPU); 8x NVIDIA B200 (1000W/GPU) (limited release); 8x AMD MI300X (air cooled) 700W -750W; with NVLink interconnects at 900GB/s, 1800GB/s for B200	8x NVIDIA H100 (700W/GPU); 8x NVIDIA H200 (700W/GPU); 8x NVIDIA B200 (1000W/GPU) (limited release); 8x AMD MI300X (air cooled) 700W -750W; with NVLink interconnects at 900GB/s, 1800GB/s for B200
IO Support	8x PCIe Gen5 x16 FHHL (Network-direct connect to GPU) 2x PCIe Gen5 x16 FHHL or 1x PCIe Gen5 x16 + 1x OCP 3.0	8x PCIe Gen5 x16 FHHL (Network-direct connect to GPU) 2x PCIe Gen5 x16 FHHL or 1x PCIe Gen5 x16 + 1x OCP 3.0	8x (front) PCIe Gen5 x16 FHHL (Network-direct connect to GPU) 2x (rear) PCIe Gen5 x16 FHHL slot with ability to support DPU via CPU
Rear IO	1x VGA + 2x USB 3.1 (Gen 1) 1x 1GbE for dedicated management 1x UID LED / System Health LED 1x OCP3 Adapter Slot	1x VGA + 2x USB 3.1 (Gen 1) 1x 1GbE for dedicated management 1x UID LED / System Health LED	1x VGA + 2x USB 3.1 (Gen 1) 1x 1GbE for dedicated management 1x UID LED / System Health LED
Boot Drive	2x M.2 non-hot swap 960GB RAID	2x M.2 non-hot swap 960GB RAID	2x NVMe M.2 non-hot swap • 960GB/1.92TB PM9A3 M2 NVMe SSD • RAID via VROC
DIMM	32Gb, 64GB, or 128GB	32Gb, 64GB, or 128GB	Up to 4TB (32x – 128GB)
NVMe	Up to 16x U.2/U.3 1.6TB, 3.2TB, 3.84TB, 6.4TB	M.2 adapters with RAID support enabled by 2 NVMe M2 drives 1.6TB, 3.2TB, 3.84TB, 6.4TB	Up to 16x U.2 or U.3 hot-swap NVMe SSDs
Cooling	Air Cooled – N+1 Hot Swap Fan solution Liquid to Air Cooling for front 4x GPUs	Air Cooled – N+1 Hot Swap Fan solution Liquid to Air Cooling for front 4x GPUs	Direct water cooled for GPU,CPU and NVLink switch, Air cooled: Networking, DIMMS, storage and PSU
Slot	1x OCP3 x16 PCIe Gen5	1x OCP3 x16 PCIe Gen5	No OCP slot
Switch	PCIe Gen5 Switch (8x SXM5 NVLink GPU)	PCIe Gen5 Switch (8x SXM5 NVLink GPU)	PCIe Gen5 Switch (8x SXM5 NVLink GPU)
PSU	5x TT CRPS Premium for GPU's (N+1) 2x TT CRPS Premium for CPU MB	8x Titanium CFFV4 PSU, N+N, Hot swap – 2600W 5x TT CRPS Premium for GPU's (N+1) 2x TT CRPS Premium for CPU MB	8x Titanium CFFV4 PSU, N+N, Hot swap – 2600W OR 8x Titanium CFFV5 PSU, N+N, Hot swap – 3200W 8x non-redundant, N+N
BMP Chip		AST2600 on SCM	
TPM (security)	TPM2.0	TPM2.0	TPM2.0
UEFI & BMC	XCC-based	XCC-based	XCC-based
OS	RHEL, Ubuntu, Alma Linux, Rocky Linux, ESXi	RHEL, Ubuntu, Alma Linux, Rocky Linux, ESXi	RHEL, Ubuntu, Alma Linux, Rocky Linux, ESXi

Monaco and Monza have similar 8U chassis design with an 8U GPU shuttle removable from the front and a 2U CPU Complex shuttle removable from the rear. The SR780a V3 server features two Intel Xeon processors and uses a hybrid water-air cooling system to take advantage of the efficiencies of cooling the CPUs and GPUs using a direct water-cooled infrastructure.

Further information can be found at:

https://www.storagereview.com/review/lenovo-thinksystem-sr685a-v3-and-sr680a-v3-gpu-servers

https://lenovopress.lenovo.com/lp1921-new-8-gpu-ai-servers-from-lenovo

3.1 Compute layer: SR680a V3 (Monza) product highlights.

The ThinkSystem SR680a V3 is an air-cooled enterprise class server, which offers a choice of accelerators featuring NVIDIA Tensor Core H100 or H200 GPUs with planned support for the NVIDIA Blackwell platform and the AMD MI300X. With high-speed interconnects between the GPUs, the system delivers optimal computational power for demanding AI and HPC workloads. The SR680a has an Intel MGX HPM system board that supports 2x Intel EMR processors, 32x DDR5 4800MHz DIMMs and multiple PCIe Gen5 ports for high-speed system PCIe expansion. The PCIe Gen5 ports support direct connect to M.2 boot drive and cabled connections to the rear PCIe slots and the switch board for additional PCIe expansion to the GPU complex, GPU NIC slots and NVMe drives. Provided in the figure below is an

an exterior image of the SR680a V3.

Figure 8: ThinkSystem SR680a V3

The key features of this appliance are as follows:

- 8U chassis
- Two 5th Gen Intel® Xeon® Scalable processors up to 350W TDP with a maximum of 48 cores and 96 threads
- 8 GPUs with high-speed interconnects and choice of:
    - NVIDIA H100 700W SXM5 GPUs with 80GB HBM3 GPU memory per GPU with NVLink interconnects at 900 GB/s
    - NVIDIA H200 700W SXM5 GPUs with 141GB HBM3 GPU memory per GPU with NVLink interconnects at 900 GB/s
    - AMD MI300X 750W OAM GPUs with 192GB HBM3 memory per GPU (planned) with Infinity Fabric interconnects at 896 GB/s
- TruDDR5 memory operating at a maximum 5600 MHz frequency
- Up to 4TB memory using 32x DDR5 DIMM sockets for system memory, 16 DIMMs per processor
- System memory 2TB using 32x 64GB RDIMMs
- 8 memory channels per processor and supports 2 DIMMs per channel (DPC)
- PCIe 5.0 Gen5 x16 FHHL interfaces to all GPUs and network adapters
- Support for high-speed networking up to 400 Gb/s, directly connected to the GPU complex
- NVIDIA BlueField-3 DPU adapters supported for hardware accelerator infrastructures
- Up to 16x 2.5” high-speed hot-swap NVMe SSD drives and up to 2x M.2 RAID via VROC
- OS supported: RHEL, Ubuntu, Alma Linux, Rocky Linux, ESXi

Further details can be found in the technical specification document, the SR680a V3 datasheet and the product guide.

The SR680a V3 and its NVIDIA Tensor Core GPUs can efficiently be scaled or partitioned into seven isolated GPU instances. A second-generation Multi-Instance GPU (MIG) provides a unified platform enabling elastic data centers to dynamically adjust to varying workloads.

Alternatively, the SR680a V3 can use the Instinct™ MI300X by AMD, which is a discrete GPU based on the AMD CDNA™ 3 architecture, designed for high-throughput compute units, AI-specific functions including support for new data types and decoding of photos and videos. These accelerators provide 192GB of HBM3 memory per GPU.

The high-level system block diagram is shown below. This figure shows the 8-way GPU configuration, the CPUs and the PCIe switches and their interrelationship to each other.

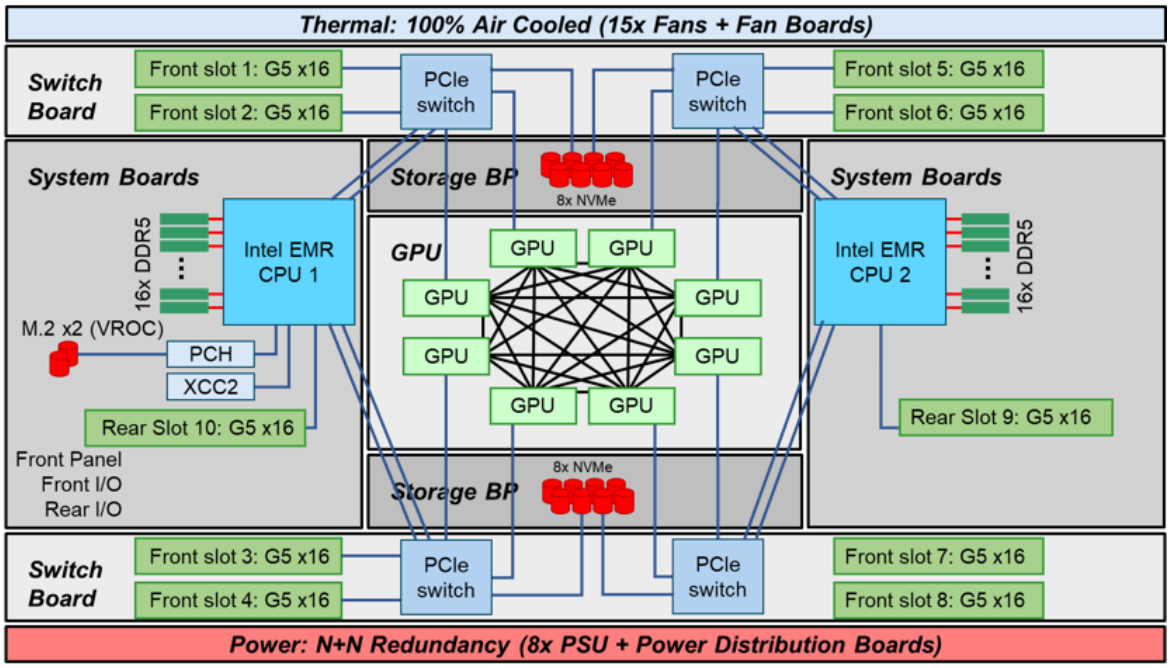

Figure 9: SR680a V3 (Monza) High-level System Block Diagram

The PCI express connectivity is illustrated in the below figure.

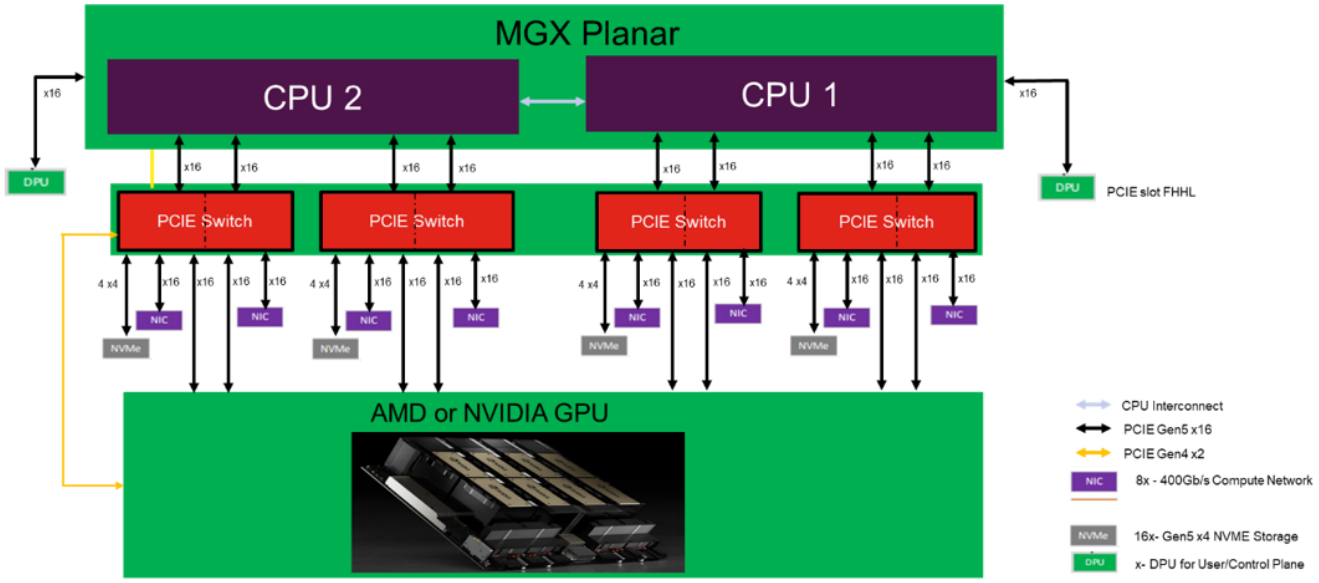

Figure 10: PCI Express Block Diagram

The power subsystem supports 8x Common Form Factor CFFv4 power supplies configuration only. The power supplies are independently powered by line cords and are hot swappable with system rear access. Hot swappable is defined as being able to plug or unplug a power supply into an active system, provided the power supply being plugged or unplugged is not connected to a live line cord.

3.2 Compute layer: SR780a V3 (Marina Bay) product highlights.

The SR780a is a high performance 5U 2-Socket 8-way GPU (top cover accessibility) system that is based on Intel's CPU processor technology. It takes advantages of water-cooled CPUs, GPUs, and NVLink switches using Lenovo's Neptune™ liquid cooling system. This system supports a rear removable CPU compute module with 2 Intel Emerald Rapids processors and an integrated GPU compute module in an 8U rack server. The SR780a V3 cooling system provides 80% water-cooling and 20% air-cooling to this enterprise class server. This appliance is housed in a compact design, which seamlessly fits into a standard 19-inch rack with a water manifold attached at the rear of the rack.

Shown in the figure below is a front exterior view of the SR780a V3 showing the 8 slots (bottom) and 8 drives (top). The subsequent figure provides a cross sectional view of the water-cooling system with cold (blue) and hot (red) lines.

Figure 11: ThinkSystem SR780a V3

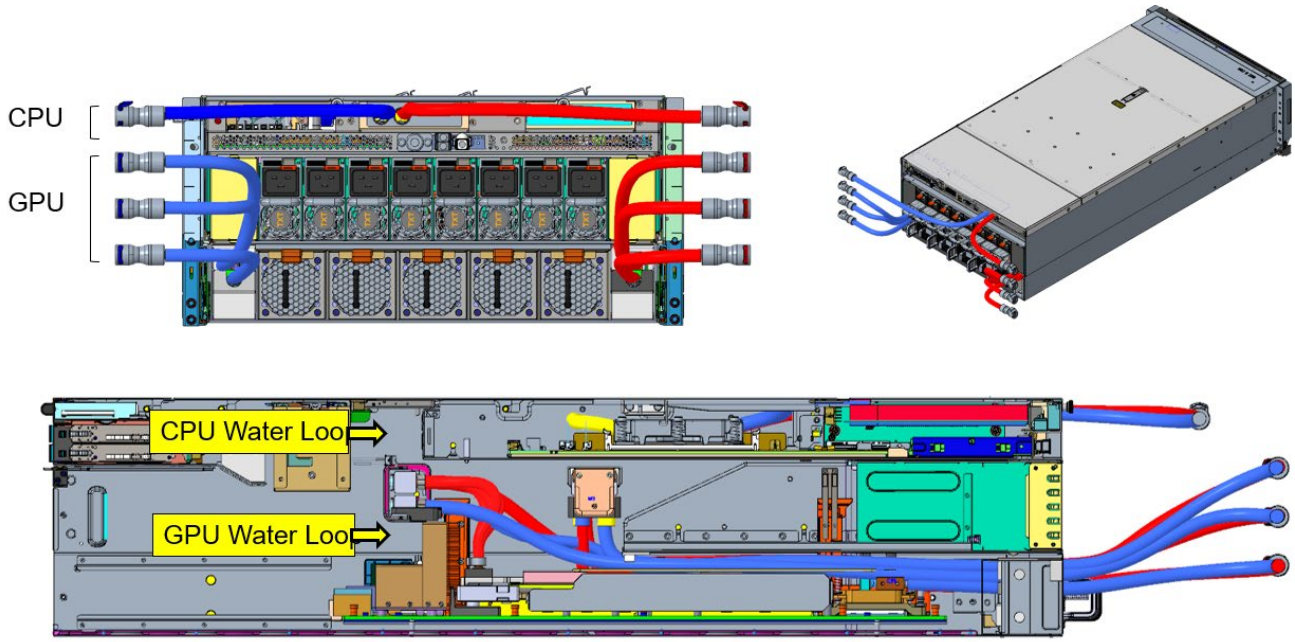

Figure 12: ThinkSystem SR780a V3 Showing CPU and GPU Water Cooling

The key features of this appliance are as follows:

- 5U chassis with up 80% water-cooling (CPUs & GPUs), 20% air-cooling (other components)
- Two 5th Gen Intel® Xeon® Scalable processors up to 350W TDP with a maximum of 48 cores and 96 threads
- 8 GPUs with high-speed interconnect and choice of:
    - NVIDIA H100 700W SXM5 GPUs with 80GB HBM3 GPU memory per GPU with NVLink interconnects at 900 GB/s
    - NVIDIA H200 700W SXM5 GPUs with 141GB HBM3 GPU memory per GPU with NVLink interconnects at 900 GB/s
    - NVIDIA B200 SXM6 GPUs with 175GB HBM3 GPU memory per GPU (planned)
    - AMD MI300X 750W OAM GPUs with 192GB HBM3 memory per GPU (planned) with Infinity Fabric interconnects at 896 GB/s
- Up to 4TB memory using 32x DDR5 DIMM sockets for system memory with maximum frequency at 5600MHz. One memory channel per memory controller supports 2 DIMMs, while the other memory channels have 1 DIMM per channel.
- Up to 10x PCIe 5.0 Gen5 x16 FHHL interfaces to all GPUs and network adapters
    - 8x front connected PCIe switch for GPU connectivity
    - 2x rear connected for CPU for CPU connectivity
- Support for high-speed networking up to 400 Gb/s, directly connected to the GPU complex
- Up to 8x high-speed U.2/U.3 hot-swap NVMe SSDs
- Up to 2x M.2 for boot (RAID via VROC)
- OS supported: RHEL, Ubuntu, Alma Linux, Rocky Linux, ESXi

The SR780a V3 uses the MGX HPM System Board, which supports 2x Intel EMR processors, 32x DDR5 5600MHz DIMMs and multiple PCIe Gen5 ports for high-speed system PCIe expansion. The PCIe Gen5 ports support direct connect to M.2 boot drive and cabled connections to the rear PCIe slots and the switch board for additional PCIe expansion to the GPU complex, GPU NIC slots and NVMe drives.

The high-level system block diagram is shown below. This figure shows the 8-way GPU configuration, the CPUs and the PCIe switches and their interrelationship to each other.

The SR780a V3 and its NVIDIA Tensor Core GPUs can efficiently be scaled or partitioned into seven isolated GPU instances. A second-generation Multi-Instance GPU (MIG) provides a unified platform enabling elastic data centers to dynamically adjust to varying workloads.

Alternatively, the SR780a V3 can use the Instinct™ MI300X by AMD, which is a discrete GPU based on the AMD CDNA™ 3 architecture, designed for high-throughput compute units, AI-specific functions including support for new data types and decoding of photos and videos. These accelerators provide 192GB of HBM3 memory per GPU.

The SR780a V3 was designed with energy thresholds in mind. Its Neptune™ liquid cooling system provides high performance results lower energy thresholds. The Neptune™ system enables the SR780a V3 to run 8 of the latest power-hungry GPUs with sustained performance over longer periods of time. Neptune's™ liquid cooling removes heat more efficiently than traditional air cooling while requiring less space than that with traditional fans. This allows the GPUs and CPUs to run in accelerated mode for longer periods of time.

Further details can be found in the technical specifications, SR780a V3 datasheet and the product guide.

3.3 Compute layer: SR685a V3 (Monaco) product highlights.

The Lenovo ThinkSystem SR685a V3 (Monaco) is a powerful 8U server that features two 4th generation AMD EPYC™ 9004 processors and 8 high-performance GPUs. This system offers a choice of accelerators featuring NVIDIA Tensor Core H100 or H200 GPUs with planned support for the NVIDIA Blackwell platform and the AMD MI300X. This completely air-cooled server is designed as a deep learning and Generative AI training server, with advanced GPU-to-GPU communications and high-speed PCIe 5.0 connectivity between the GPUs, CPUs, and networking. This server has interconnects with the fastest transfer rate utilizing an AMD Infinity Fabric or NVIDIA NVLink. It was designed in-house from ground up to provide best-in-class modularity, thermal performance, and reliability. The below figure shows an exterior image of this server.

Figure 13: ThinkSystem SR685a V3

The key features of the ThinkSystem SR685a are as follows:

- 8U chassis with 100% air-cooling
- Two 4th Gen AMD EPYC™ processors with up to 64 cores and 128 threads having a max TDP rating up to 400W
- 8 GPUs with high-speed interconnect, choice of:
    - AMD Instinct™ MI300X 750W TBP OAM GPUs with 192GB HBM3 memory per GPU (planned) with Infinity Fabric interconnects at 896 GB/s. Each GPU running at 5.2TB/s with a total bi-directional GPU-to-GPU bandwidth of 6.4TB/s
    - NVIDIA H100 700W SXM5 GPUs with 80GB HBM3 GPU memory per GPU with NVLink interconnects at 900 GB/s. Each GPU running at 3.3TB/s with a total bi-directional GPU-to-GPU bandwidth of 7.2TB/s

    - NVIDIA H200 700W SXM5 GPUs with 141GB HBM3 GPU memory per GPU with NVLink interconnects at 900 GB/s. Each GPU running at 4.8TB/s with a total bi-directional GPU-to-GPU bandwidth of 7.2TB/s

- 4800 MHz DDR5 system memory DIMMs
    - Up to 24 TruDDR5 memory RDIMMs, 12 DIMMs per processor
    - 12 memory channels per processor and supporting 1 DIMM per channel (DPC)
    - Using 24x 64GB RDIMMs, the server supports 1.5TB of system memory
    - Using 24x 96GB RDIMMs, the server supports 2.25TB of system memory
- Up to 10 PCIe Gen5 x16 FHHL interfaces to all GPUs and network adapters. Front: 8x PCIe 5.0 x16 FHHL slots with GPU Direct support. Rear: 1x PCIe 5.0 x16 FHHL slot + 1x OCP 3.0 slot with PCIe 5.0 x16 interface.
- Supports eight NDR 400Gb/s InfiniBand adapters with high-speed GPU Direct connections, installed in front-accessible PCIe 5.0 x16 slots.
- NVIDIA BlueField-3 DPU adapters supported for hardware accelerator infrastructures
- Up to 16x PCIe 5.0 2.5” high-speed hot-swap NVMe SSD drives maximizing drive I/O performance for throughput, bandwidth, and latency
- Up to 2x NVMe drives on M.2 drives install in an M.2internal adapter with integrated RAID for OS boot or additional storage.
- OS supported: RHEL, Ubuntu, Alma Linux, Rocky Linux, ESXi

The SR685a has eight GPUs and eight front slots, all connected to PCIe switches using PCIe 5.0 x16 links. Each GPU is connected to all other GPUs using XGMI x16 connections, each link with 128 GB/s bidirectional bandwidth. The server also supports 16 NVMe drives, each connected to a PCIe switch using a PCIe 5.0 x4 link. The server uses Lenovo TruDDR5 memory operating at up to 4800 MHz and supports up to 24 DIMMs with 2 processors. The processors have 12 memory channels and support 1 DIMM per channel. The server supports 1.5TB of memory using 24x 64GB RDIMMs and two processors, or 2.25TB of memory using 24x 96GB RDIMMs and two processors.

The SR685a supports 2x AMD Genoa processors, 24x DDR5 4800MHz DIMMs and multiple PCIe Gen5 ports for high-speed system PCIe expansion. The PCIe Gen5 ports support direct connect to PCIe slots, OCP slot, M.2 boot module and cabled connections to the PCIe switch board for additional PCIe expansion to the GPU complex, GPU NIC slots and NVMe drives.

The high-level system block diagram is shown below. This figure shows the 8-way GPU configuration, the CPUs and the PCIe switches and their interrelationship to each other.

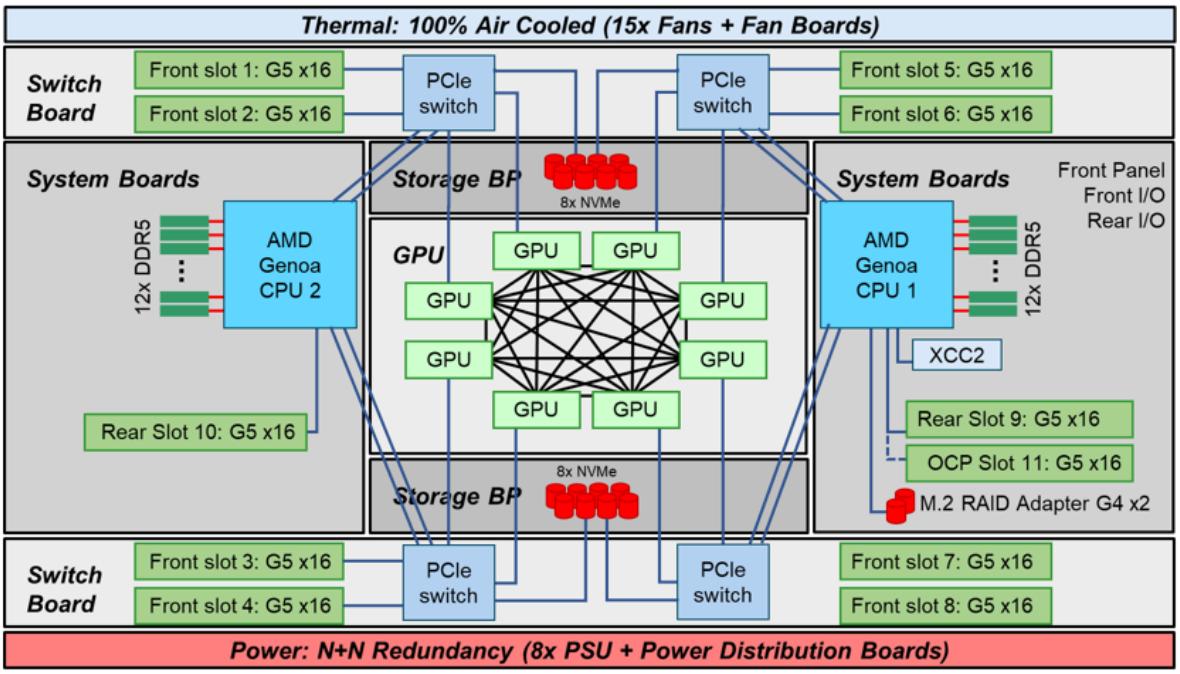

Figure 14: 685a V3 (Monaco) System Block Diagram

The PCI express connectivity is illustrated in the below figure.

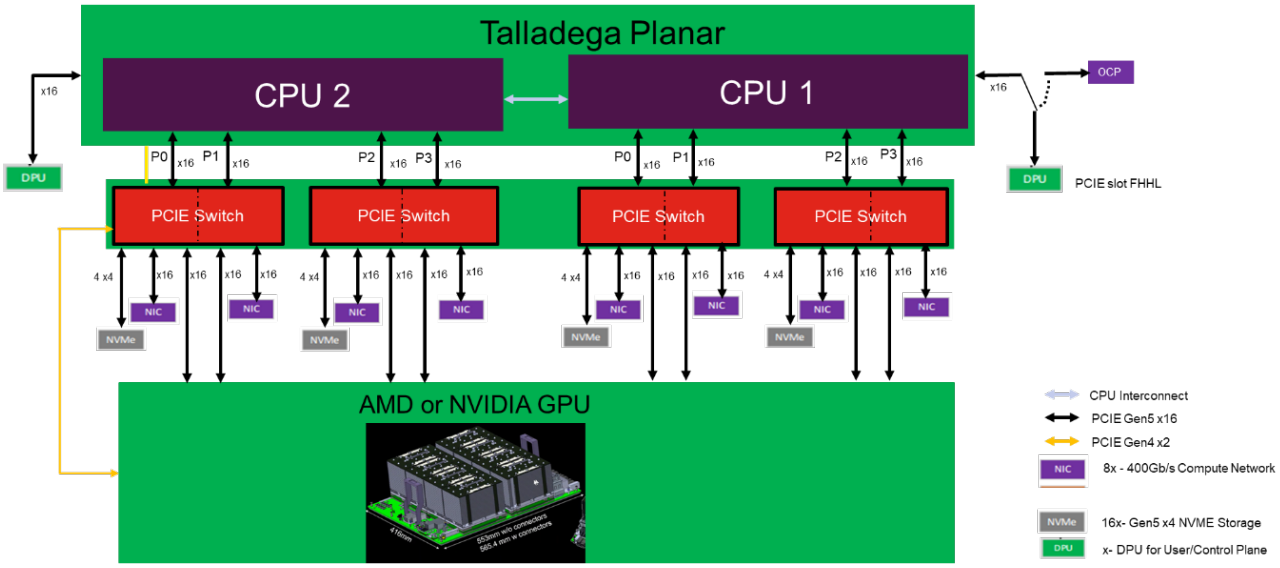

Figure 15: Monaco PCI Express Block Diagram

This server's 8U chassis design has an 8U GPU shuttle removable from the front and a 2U CPU Complex shuttle removable from the rear. The power subsystem supports 8x Common Form Factor CFFv4 power supplies configuration only. The power supplies are independently powered by line cords and are hot swappable with system rear access. Hot swappable is defined as being able to plug or unplug a power supply into an active system, provided the power supply being plugged or unplugged is not connected to a live line cord.

Further details can be found in the technical specifications, SR685a V3 datasheet and the product guide.

4 DDN Storage Layer

DataDirect Networks (DDN) provides powerful external storage systems that deliver efficient use of GPU resources for AI workloads, DDN blogs. These systems support unstructured, structured, and semi-structured data for deep learning and high-performance computing environments across governments, enterprises, and academic sectors. DDN devices have a fully containerized architecture with elastic scalability and supports dynamic multi-tenancy. DataDirect Networks storage is highly efficient storage at scale and is NVIDIA's only recommended storage for their SuperPOD deployments. The DDN storage system is used for all of NVIDIA's internal systems.

The DDN system has a shared parallel architecture that delivers data with high-throughput, low-latency, and enormous concurrency, for maximum application acceleration. This architecture is massively scalable in performance and capability. The DDN nodes deliver fast random IO performance to every GPU client while having high filesystem metadata responsiveness. DDN solutions deliver full multi-dimensional performance and scaling with GPUs providing peak per client performance. They also scale linearly with AI application performance.

DDN provides true end-to-end parallelism to balance IO from each application to all servers eliminating hot spots while delivering fast read/writes and providing predictable linear scaling. Storage hot spots, which slow applications, can be caused by 1:1 client/server relationships and uneven workloads. These can be eliminated via full parallelism. Hot spots in storage can also occur when one server has client only access resulting in all data from peer servers needing to be read/written before the client can access the data again. This too can be eliminated through complete parallelism.

This storage system supports virtualized multi-tenancy for dynamic provisioning of key encrypted storage based on users' capacity and performance needs. These systems allow for client-based compression to provide efficient use of space without sacrificing performance. Using the latest EXA client-side compression, DDN can provide up to a 5:1 compression for AI training data sets and a 2:1 compression for AI checkpoints.

The relationship between Lenovo, DDN, and Databricks can be leveraged to build a complete data strategy deployed on DDN nodes. As part of this relationship, Lenovo and DDN can work with organizations to help them take advantage of Databricks' enhanced data management capabilities, their ability to streamline access to data from many sources, and their ability to remove data barriers that impede successful AI implementations.

4.1 DDN Solutions and Software

The DDN product line includes the integrated A^3I storage solution based on DDN's EXAScalar (EXA6) Lustre file system and the Insight monitoring tool, which monitors the storage hardware and filesystem.

4.1.1 EXAScalar (EXA6)

The EXAScalar product is based on the open-source Lustre filesystem enhanced for manageability, performance, scalability, capacity, and reliability. It is a highly available, massively parallel, clustered file system solution for Linux-based applications. It provides near-wire-speed data transfer capabilities and unbounded data capacity and scalability while supporting enterprise-level security features. It is capable of GPU direct for direct RDMA to and from NVIDIA GPU memory. EXAScalar is built upon a cutting-edge, open-source Lustre file system, and as such can deliver high-performance solutions by scaling to tens of thousands

of clients and petabytes of storage. EXAScaler is engineered to reduce the complexity of deploying High Performance Computing (HPC) storage and to scale with evolving application requirements. Its other uses are as workspaces for long term dataset storage or as a centralized repository for manipulation, sharing, and acquisition of results from NFS, SMB and S3.

The EXA6 system is delivered as a hardware and software solution that can start small, with a single storage appliance, and then scale for performance and storage by adding additional appliances. The EXA6 system integrates with DDN's SFA (Storage Fusion Architecture) storage appliances to simplify the deployment and management of file system storage.

As a parallel file system, EXAScaler stores data across multiple networked servers providing files and directories that are spread across a storage infrastructure. Additionally, high bandwidth is achieved by clients being able to read and write to all servers simultaneously.

The EXA6 system has the below additional attributes:

- Combines multiple storage servers into a single global namespace exported to clients.
- Separates object IO from metadata IO for performance and scalability.
- Each client has a separate and independent connection to each file system server.
- Separate servers for configuration, metadata, and object storage management.
- Each server runs a client process to talk to other servers.
- Each server mounts one or more Storage Targets.
- Packaging and management tools
- Extends Lustre with unique features for data storage and management, file system security, integration, and usability. Examples include:
    - SFA for data storage
    - S3 Data Services for file system integration
    - EXAScaler Management Framework for usability
    - CLI administration tools for support, file system configuration, management, and tuning.
    - Data management and usability capabilities, such as hot pools for performance/infrastructure optimization.
    - File system management framework with GUI, CLI, and API layers for usability.
- Hardware integration with DDN's SFA storage appliances
- Large namespaces can be created by aggregating multiple boxes into a single filesystem.
- Fully optimizes for deep learning training through simplifying data access, which is especially important for fast data access required during training operations where local caching is inadequate.
- Fast read performance for high multi-threaded and single threaded applications like fast staging of data to local disk and training on large lossless compressed images, uncompressed images, or large individual data objects.

Unlike Lustre, EXAScaler manages file system resources to ensure that their failures, or the failures of

servers hosting them, do not affect continuous file system availability. EXAScaler runs a High Availability stack, which allows for MDS and OSS failover pairs using Linux HA services. When file system resources fail, EXAScaler uses the HA services to initiate fencing and migrate the resources from a failed file system server to its failover pair, and to resume normal operation as soon as the migration is completed. This HA capability is reinforced with SFA storage appliances' component redundancy (redundant controllers, power supply units, and cooling modules) and support for hot swappable drives.

EXAScaler includes data management and integrity filesystem features developed by DDN and only available in its appliances and cloud offerings. EXAScaler's has a powerful data orchestration engine that gives users comprehensive data residency controls using policy-based placement.

The EXA6 system has acceleration technologies with Hot Nodes capability that accelerates data access by automatically caching data on the local NVMe of GPU systems, reducing IO latency and traffic by avoiding network round trips, and freeing up infrastructure to serve additional workloads. This capability significantly improves the performance of applications if a dataset is accessed multiple times during a particular workflow, like during training operations.

EXA6 Hot Nodes includes extensive data management tools and performance monitoring facilities. These tools enable user-driven local cache management and make integration simple with task schedulers. For example, training input data can be loaded to the local cache on a ThinkSystem compute appliance as a pre-flight task before the AI training application is engaged.

During deep learning training, the same input dataset or portions of the same input are repeatedly accessed over multiple training iterations (epochs). Commonly, an application reads the input dataset from shared storage directly, thereby continuously consuming shared storage resources. With Hot Nodes, as the input data is read during the first training iteration, the DDN software automatically writes a copy of the data on the local NVME devices. During subsequent reads, data is delivered to the application from the local cache rather than the shared storage. This entire process is managed by the DDN client software. Data access is seamless, and the cache is fully transparent to users and applications. The use of the local cache eliminates network traffic and reduces the load on the shared storage system. This allows other critical deep learning training operations, like checkpointing, to complete faster by engaging the full capabilities of the shared storage system.

The Hot Node tools can expose insightful information about cache utilization and performance, enabling system administrators to further optimize their data loading and maximize application and infrastructure efficiency gains.

Other capabilities of Hot Nodes are listed below:
. Leverage Persistent Client Cache on file system clients with local flash storage.
. Gain performance from fast (NVMe-type) client storage for intensive IO.

. Reduce IO overhead for OSTs to avoid contesting IO to the OSTs from the clients.

Another capability of EXAScalar is Hot Pools that intelligently move data between high-performance flash and large capacity disk to ensures efficient use of storage. These pools optimize file storage to improve performance and keep hot files on fast data pool (NVMe OST pool) and move cool files to a slow data pool

(HDD OST pool), from where the files are still directly accessible. Hot pools automatically or manually move cool files back to the hot pool if the files become frequently used again.

4.1.2 DDN A^{3}I

A special breed of pre-configured EXAScaler solutions called A^{3}I is specifically designed to facilitate solutions that rely on AI, deep learning, and Big Data. DDN A^{3}I solutions enable and accelerate end-to-end data pipelines for deep learning workflows of all scale running on ThinkSystem appliances. The DDN shared parallel architecture enables concurrent and continuous execution of all phases of deep learning workflows across multiple ThinkSystem compute nodes. This eliminates the management overhead and risks of moving data between storage locations. At the application level, data is accessed through a standard highly interoperable file interface, for a familiar and intuitive user experience.

Significant acceleration can be achieved by executing an application across multiple ThinkSystem nodes in a cluster of nodes simultaneously and engaging parallel training efforts of candidate neural networks variants. These advanced optimizations maximize the potential of deep learning frameworks.

DDN A^{3}I solutions integrate a wide range of networking technologies and topologies to ensure streamlined deployment and optimal performance for AI infrastructure. The latest generation NVIDIA Quantum InfiniBand (IB) and Spectrum Ethernet technology provide both high-bandwidth and low-latency data transfers between applications, compute servers and storage appliances.

DDN A^{3}I Multi-rail enables grouping of multiple network interfaces to achieve faster aggregate data transfer capabilities. The feature balances traffic dynamically across all the interfaces, and actively monitors link health for rapid failure detection and automatic recovery. DDN A^{3}I Multi-rail makes designing, deploying, and managing high-performance networks straightforward, and is proven to deliver complete connectivity for at-scale infrastructure for ThinkSystem cluster deployments.

4.1.3 DDN Insight

DDN Insight is a centralized monitoring tool for the DDN storage solutions. It provides end-to-end data management, enabling health and performance monitoring across DDN's product portfolio from a single web-based user interface. Typically, storage networks are made up of local block-based arrays, file systems, business analytics repositories, and distributed object stores, each with their own unique management requirements. DDN Insight is a single application that monitors and manages all types of storage resources. It is available as a software-installable package on customer-supplied management servers, designated as the DDN Insight head node within the storage network. DDN Insight consists of different software components for monitoring storage products such as Storage Fusion Architecture (**SFA**) and EXAScaler (**ES**).

DDN Insight greatly simplifies IT operations and enables automated and proactive storage platform management guided by analytics and intelligent software. Insight monitors several key variables that affect data I/O performance and identifies bottlenecks. It provides deep real-time analysis across the entire cluster, tracking I/O transactions from applications running on compute nodes all the way through individual drives in the AI400X2 appliances. The embedded analytics engine makes it simple for operators to visualize I/O

performance across their entire infrastructure through intuitive user interfaces. These include extensive logging, trending, and comparison tools, for analyzing I/O performance of specific applications and users over time. Insight's open backend database makes it simple to extend Insight's benefits and integrate other AI infrastructure components within the engine, or export data to third party monitoring systems.

Other features of Insight:

- Advanced EXAScaler monitoring with historical data collection
- Real-time scalable visibility into software and appliances
- Multi-solutions monitoring for EXAScaler, GRIDScaler, SFA, and IME Platforms from DDN
- Extensive customization
- Open backend database and integration with Grafana
- Integration with other systems, switches, SW

Insight requires specific OS builds that may interfere with patching schedules of other services running on those systems. As such, Insight should not run on servers with other duties; therefore, it should be on its own server.

Through the combination of DDN's Insight along with Lenovo's LiCO monitoring/orchestration software a complete view of system wide performance is gained.

4.2 DDN AI400X2-QLC, AI400X2-Turbo, and AI400X2 highlights

There are three variants of the DDN A^{3}I appliance: the AI400X2-QLC, AI400X2-Turbo, and AI400X2 each virtualized into 2RU. All appliance types have the features described above including the data management and security features of hot pools, hot nodes, and encryption. This appliance group integrates the DDN A^{3}I shared parallel architecture and includes a wide range of capabilities including extensive monitoring. Shared performance scales linearly as additional appliances are integrated into a network. All types are fully integrated turnkey appliances with no external switches. A high-level comparison of these variants is provided in the below figure.

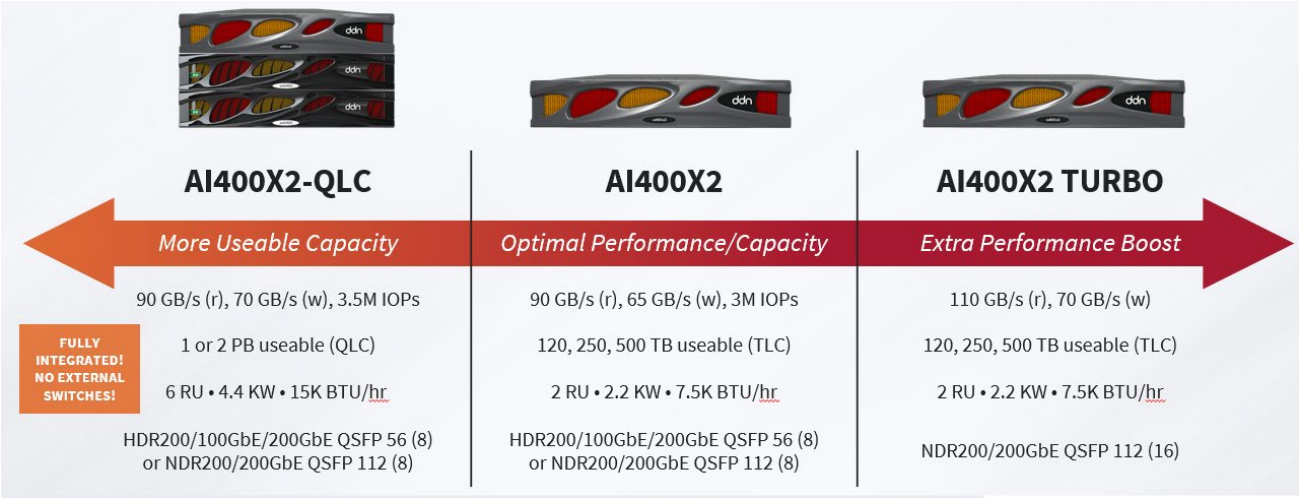

Figure 16: High level Comparison of DDN A^{3}I Appliance Classes

This appliance family has predictable capacity, capability, and performance. The all-NVMe configuration provides optimal performance for a wide variety of workloads and data types. This ensures that network operators achieve the most from at-scale GPU applications, while maintaining a single, shared, centralized data platform.

These appliances are low-power compact storage boxes providing extreme performance using NVMe storage drives and network transport with either InfiniBand or RoCE Ethernet. They are available in 60, 120, 250 and 500 TB capacity configurations. Optional hybrid configurations with integrated HDDs are also available for deployments requiring high density deep capacity storage. The DDN H100/H200 recommended storage for up to a 10-node configuration is provided below.

DDN Recommended Storage Based on Application IO Requirements

	Light IO Workload	Medium IO Workload	High IO Workload
AI400X2 Appliances	1	2	4
Useable NVME Capacity (250/500 TB)	250 TB / 500 TB	500 TB / 1 PB	1 PB / 2 PB
Aggregate Shared Read	90 GB/s	180 GB/s	360 TB/s
Aggregate Shared Write	65 GB/s	130 GB/s	260 GB/s
Per GPU Shared Read	1.1 GB/s	2.2 GB/s	4.4 GB/s
Per GPU Shared Write	800 MB/s	1.6 GB/s	3.2 GB/s
Per DGX Shared Read	9 GB/s	18 GB/s	36 GB/s
Per DGX Shared Write	6.5 GB/s	13 GB/s	26 GB/s

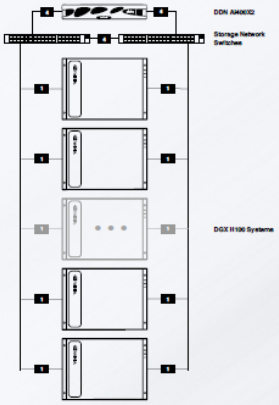

Figure 17: AI400X2 IO Profile Data Pipelines as Function of Process Function. Details found here.

Additional features of the DDN A^{3}I appliances include:

- Full integration streamlining deployment and simplifying management operations.
- Configured at several different capacities ranging from 30 TB to 500 TB for NVME and can be expanded with up to 5 PB of capacity disk for hybrid configuration.
- Multiple appliances can be aggregated into a single namespace that provides up to several hundreds of petabytes of useable capacity.
- Communication with other parts of network topology using multiple NDR200 InfiniBand or 200 GbE network connections for performance, load balancing, and resiliency.
- Native InfiniBand support that maximizes performance and minimizes CPU overhead.
- Ability for multiple clients to access data on the appliance simultaneously, with performance available to all systems and distributed dynamically. Clients can be optimized for ultra-low latency workloads.
- An all-NVME architecture to provide excellent random read performance, often as fast as sequential read patterns.
- A small form factor that is power efficient with dense performance and capacity, achieving optimal data center utilization and better operational economics.

Contrasts between appliance types:

Generally, the AI400X2-QLC and AI400X2 have the same performance. The difference is the underlying capacity. With the AI400X2, you can only use TLC drives that are faster and have better write durability but are only available up to 500TB per appliance. The QLC version has the same performance when equipped with at least 2 enclosures (its minimum config) but provides more capacity for the same performance up to 5.5 PB. Depending on the capacity and performance targets, TLC or QLC can be used to provide the appropriate amount of performance for a given capacity. The A^{3}I systems traditionally use TLC flash; however, QLC is becoming an option.

The Turbo appliance provides better read performance and more RAM per virtual machine on the same hardware at the expense of serial attached SCSI and QLC expansion. It is available only as a TLC appliance (500TB max per appliance) but provides 20-30% higher read performance than the non-Turbo version. Because of the increased RAM per VM, these appliances can be better suited for applications such as dedicated metadata server appliances and can provide read performance with 20% less hardware; however, there is no increase in write performance as that is dependent on the NVME devices themselves. The Turbo appliance also has 16x NDR200 ports rather than 8, but only 8 ports are needed to get full performance. Unlike the non-Turbo appliance, rather than connecting IB/Ethernet cables to both ports on the same CX7 card, only a single port is used to maximize PCIe bandwidth.

Lenovo provides 10 DDN options as shown in the below table all have installation services and 3yr. PROS.

Storage Name	Useable Storage	Machine Type	Feature Code
AI400X2 Turbo	500TB	7DFACTO2WW x1	C5UL
AI400X2 Turbo	250TB	7DFACTO2WW x1	C5UK
AI400X2 Turbo - OSS	500TB	7DFACTO2WW x1	C5UH
AI400X2 Turbo - OSS	250TB	7DFACTO2WW x1	C5UG
AI400X2 Turbo - MDS	3.84TB	7DFACTO2WW x1	C5UJ
AI400X2	500TB	7DFACTO2WW x1	C5UF
AI400X2	250TB	7DFACTO2WW x1	C5UE
AI400X2 - OSS	500TB	7DFACTO2WW x1	C5UC
AI400X2 - OSS	250TB	7DFACTO2WW x1	C5UB
AI400X2 - MDS	3.84TB	7DFACTO2WW x1	C5UD

4.3 DDN Node Sizing and Performance

Storage sizing guides and information can be found on the DDN website and in DDN's RA with the ThinkSystem appliances. These links (FIO, MDTest, MLPerf, 2024 per3s NCP architecture) provide some performance data based on the filesystem testing tool FIO, used for all single- and multi-node bandwidth performance; the MDTest measuring multi-node metadata performance; and MLPerf benchmarks.

Performance testing on the DDN A^{3}I architecture is conducted with industry standard synthetic throughput and IOPS applications, as well as widely used deep learning frameworks and data types. The results demonstrate that with the DDN A^{3}I shared parallel architecture, GPU-accelerated applications can engage the full

capabilities of the compute and storage infrastructure. Performance is found to be evenly distributed across all components and scales linearly as more systems are engaged.

5 Neptune Water Cooled Technology

Training deep learning models push power consumption above 20KW per rack. The power requirements are drastically increasing across technologies for CPU, memory, GPUs, and network adapters, as seen in the below figure. In addition, server power requirements have tripled since 2014.

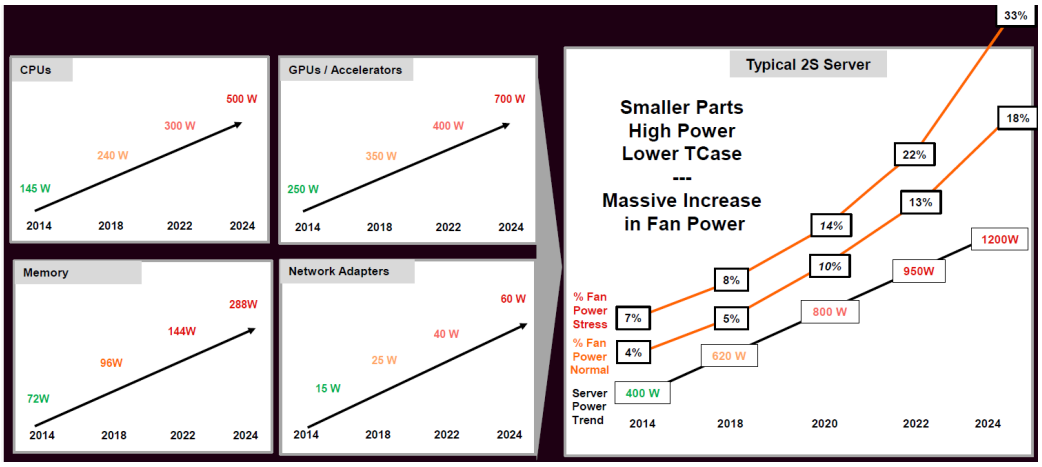

Figure 18: Power Trends Across Technologies

The power densities are now at levels requiring highly efficient heat extraction. At these levels, the heat generated has pushed air cooling to its limits. As such, water cooling technologies are being required. Lenovo's water-cooling technologies have existed for a decade and the enhancements with the latest Neptune™ Water Cooling technology are designed to water cool not only power hunger GPUs, but also water cool CPUs and NVLink switches. This technology removes up to 95% of the heat generated during training while simultaneously improving efficiency to allow energy consumption to remain within thresholds.

These Neptune systems have direct open-loop deionized water cooling and utilize a new cold plate design to maximize thermal extraction. The below figure illustrates this design.

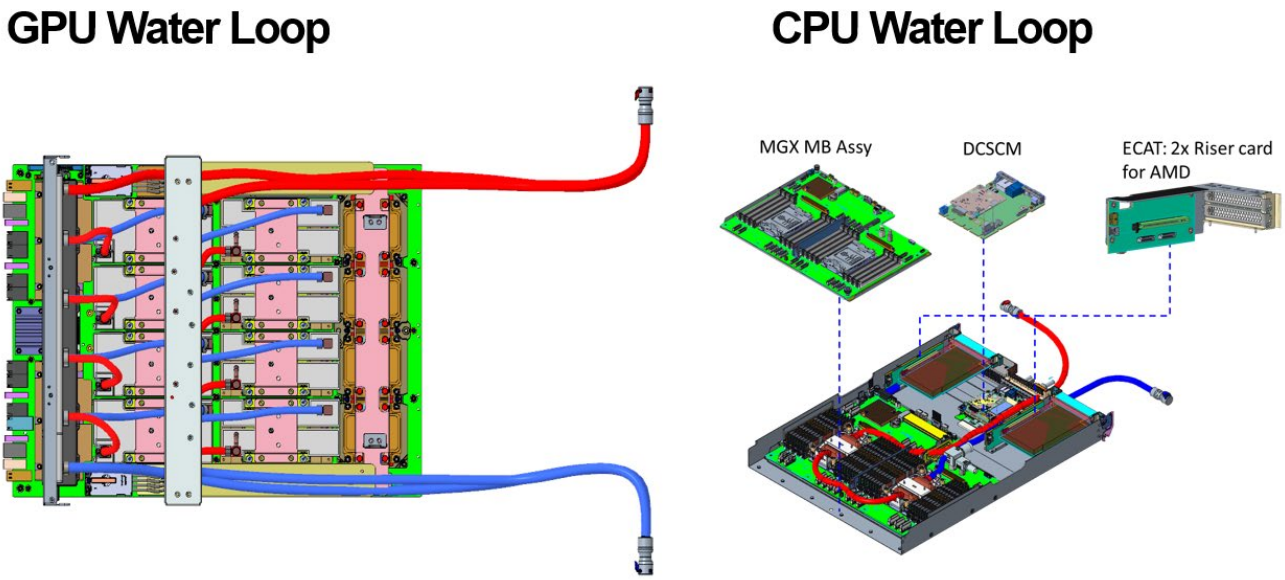

Figure 19: GPU and CPU Water Loop Configuration.

Neptune liquid cooling enables the powerful GPUs and CPUs in the SR780a V3 to sustain higher performance over longer time periods when compared to traditional air-cooled appliances. Liquid cooling more effectively removes heat allowing GPUs and CPUs to run in accelerated mode for extended periods producing lower chip temperatures, which enable higher clock speeds and reduced electricity consumption. This method of thermal transfer also improves component and overall system reliability because of less thermal fatigue. Because a reduced portion of the system requires air cooling, space utilization is improved and higher server density with flexible placement is achieved. Neptune's robust cooling is key to preventing thermal throttling and thus performance can be maintained.

Neptune customers have reported significant energy reductions and performance gains producing better energy efficiencies when compared to their traditional air-cooled approaches. This capability provides higher performance while being sustainable and has higher computing density in a lower footprint. As with everything, there are trade-offs. Specifically, liquid cooling implementations have more complicated infrastructure requirements, high maintenance, and risk of leaks. Lenovo addresses these challenges through direct warm water-cooling options that turns waste heat into value by allowing hot water reuse in a facility.

Lenovo Neptune offers 3 methods of liquid cooling: liquid assisted, rack water cooled, and direct water cooling. A description of these methods along with their benefits is provided in the below figure.

Cooling method	Description	Benefits
Liquid Assisted Cooling	With either a Thermal Transfer Module (TTM) or Liquid to Air Module (L2A), heat is drawn away from GPUs and CPUs with liquid cooled by a traditional heat sink, or existing system fans.	• Special heat sinks to transfer heat quickly • Enabling higher performing processors and GPUs • Allows for data centers that rely solely on air cooling to continue operating
Rack Water Cooling	Keep heat down in the rack with either a Rear-door Heat Exchanger (RDHX) or In-rack Cooling Distribution Units (CDU).	• Rack option takes heat out of data center in chilled water • 3.5X more efficient than air only • Saves enough power to run 4,000 LED bulbs
Direct Water Cooling	With Full Systems and Core Systems available, cooled water draws heat away from hot components – including power supplies for completely fanless operation.	• Warm water inside server • Higher perf/density than air cooled servers • Up to 40% reduction in power costs

Figure 20: Types of Neptune Water Cooling.

Results from a lab test, which compared two Lenovo compute nodes, one based on air cooling, and one based on Neptune cooling are provided in the below figure. The standardized test used was a High-Performance Linpack (HPL) benchmark test. The HPC benchmark is a compute intensive test measuring a system's floating-point computing power while solving a dense system of linear equations. In this instance of the test, a comparison of the A100 PCIe card was made to an A100 SXM card. Although this is an older test, directionally consistent results are expected for the SR780a V3 with H100/H200 GPUs.

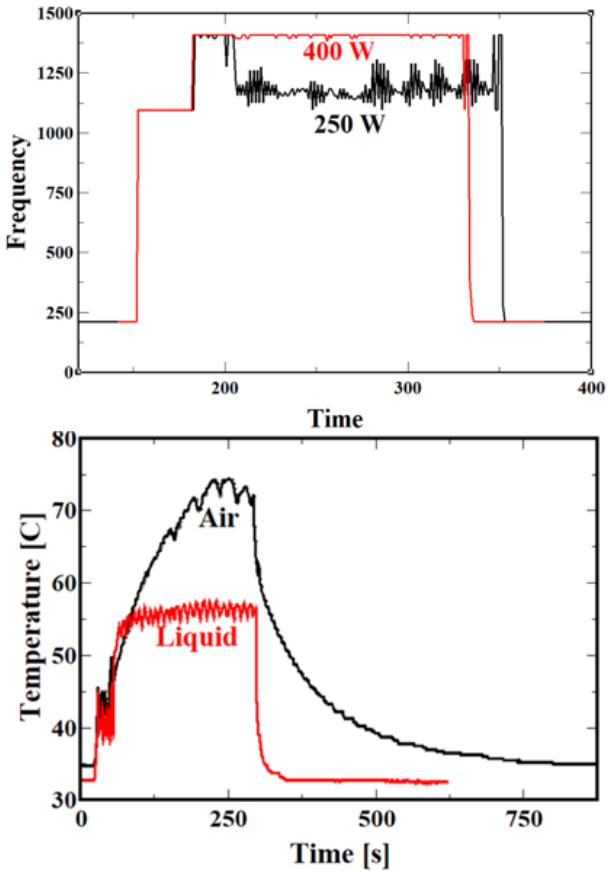

Figure 21: HPL Test Results for Air-cooled and Water-cooled (red line) Appliances.

The test shows an ~ 20C margin exists between liquid cooled plates compared to an air-cooled system. The Neptune cooled appliance provides the highest absolute performance and efficiency. The thermal efficiency is increased by the liquid cooled plates. The Neptune appliance provides the maximum stable frequency and the lowest operating temperatures. The net effect of water cooling is reduced training run times, increased packaging density, more reliable clock speeds and improved reliability.

There are several excellent articles regarding best practices and considerations for the development of water-cooled based data centers. Some of which are listed below.

Deploying liquid cooling in the data center, deploying-liquid-cooling-in-the-data-center-a-guide-to-high-density-cooling-white-paper.pdf (vertiv.com)

An introduction to liquid cooling in the data center, An introduction to liquid cooling in the data center - DCD (datacenterdynamics.com)

NVIDIA Adds Liquid-Cooled GPUs for Sustainable, Efficient Computing, Driving Efficiency, Liquid Cooled GPUs Debut at Computex | NVIDIA Blogs

NVIDIA to Build AI Factories and Data Centers for the Next Industrial Revolution, NVIDIA to build AI factories and data centres for the next industrial revolution (datacenternews.asia)

Liquid cooling of data centers: A necessity facing challenges, Liquid cooling of data centers: A necessity facing challenges - ScienceDirect

6 Appendix: Lenovo Bill of Materials

This appendix contains the bill of materials (BOMs) for different configurations of the solution. There are sections for user servers, management servers, storage, and networking.

The BOM lists in this appendix are not meant to be exhaustive and must always be double-checked with the configuration tools. Any discussion of pricing, support, and maintenance options is outside the scope of this document.

This solution is a racked, switched, and cabled solution that follows the LeSI framework for designing, manufacturing, integrating, and delivering data center solutions. More information about LeSI can be found at the following URL: https://lenovopress.lenovo.com/lp0900-lenovo-everyscale-lesi#benefits

6.1 BOM for SR680 V3 compute and DDN storage

6.2 BOM for SR780 V3 compute and DDN storage

6.3 BOM for SR685 V3 compute and DDN storage

7 Resources:

- Acknowledgement:

The below individuals advised and reviewed this document. Their help is greatly appreciated and was extremely valuable.

Michael Ridout

Lerone Latouche

Ted Vojnovich

Ranga Sankaranarayanan

Ajay Dholakia

Delia DeCourcy

Valerio Rizzo

- Lenovo, New 8-GPU AI Servers from Lenovo
- Google, tf.data: A Machine Learning Data Processing Framework. arXiv:2101.12127v2 [cs.LG] 23Feb2021
- Meta, How Meta trains large language models at scale
- Lenovo, Monaco Monza MarinaBay System Specification V1.0
- Memory Population Requirements for Lenovo ThinkSystem Servers
- ThinkSystem NVIDIA BlueField-3 QSFP112 2-Port 200Gb DPU Adapter
- NVDIA GH200 Grace Hopper Superchip Architecture
- NVIDIA Next-Generation Networking for the Next Wave of AI
- Lenovo adds new AI solutions, expands Neptune cooling range to enable heat reuse
- NVIDIA AI GPU Servers: PCIe vs. SXM
- Meta, Composable Data Management
- Meta's approach to machine learning prediction robustness

Document History

Version 1.0	July 2024	Initial version

Trademarks and special notices

References in this document to Lenovo products or services do not imply that Lenovo intends to make them available in every country.

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:

Lenovo®
Neptune®
ThinkSystem®

The following terms are trademarks of other companies:

AMD, AMD CDNA™, AMD EPYC™, and Infinity Fabric™ are trademarks of Advanced Micro Devices, Inc.

Intel® and Xeon® are trademarks of Intel Corporation or its subsidiaries.

Linux® is the trademark of Linus Torvalds in the U.S. and other countries.

Microsoft® is the trademark of Microsoft Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Information is provided "AS IS" without warranty of any kind.

All customer examples described are presented as illustrations of how those customers have used Lenovo products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

Information concerning non-Lenovo products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by Lenovo. Sources for non-Lenovo list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. Lenovo has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-Lenovo products. Questions on the capability of non-Lenovo products should be addressed to the supplier of those products.

All statements regarding Lenovo future direction and intent are subject to change or withdrawal without notice and represent goals and objectives only. Contact your local Lenovo office or Lenovo authorized reseller for the full text of the specific Statement of Direction.

Some information addresses anticipated future capabilities. Such information is not intended as a definitive statement of a commitment to specific levels of performance, function or delivery schedules with respect to any future products. Such commitments are only made in Lenovo product announcements. The information is presented here to communicate Lenovo’s current investment and development activities as a good faith effort to help with our customers' future planning.

Performance is based on measurements and projections using standard Lenovo benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

Photographs shown are of engineering prototypes. Changes may be incorporated in production models.

Any references in this information to non-Lenovo websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this Lenovo product and use of those websites is at your own risk.