Reference Architecture for Generative AI Compute and Storage with Lenovo ThinkSystem

Last update: **21 February 2024**

Version 1.0

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

Discusses design considerations for WEKA storage

Provides considerations for data strategy

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

Rob Reviere

Steve Eiland

LENOVO PRESS

Table of Contents

1 Introduction

Generative AI (GenAI) and its subset of large language models (LLMs) have exploded on the scene to offer capabilities never seen. From model based creational of original art and music to the ability to distil actionable information from the massive content on the internet, these models have become powerful tools for discovery and innovation. Lenovo's initiatives on this front are designed to mitigate a company's risks that they encountered when developing infrastructure supporting Generative AI. 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 extremely 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 that leverage accelerators, parallelized operators and maximize core usage. These challenges and approaches to address them are further detailed in the compute layer GenAI reference architecture, which also covers solution components recommended for a GenAI infrastructure.

This document focuses more the role external storage plays in a comprehensive GenAI solution. The sections that follow address the challenges above, provide an optimized reference architecture, and discuss the appropriate solution hardware. Although this document focuses on GenAI and LLMs, the content presented support equally as well any deep learning modelling methodology, like computer vision and natural language processing.

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 foundational design for the compute, storage and network layers needed for Generative AI use cases. The class of models that constitute Generative AI pushes the boundaries on 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 analysed carefully and resolved. This document will address these considerations and it will provide the building blocks for both small-scale and large-scale deployments. The intent of this document is to provide the planning, design considerations, and best practices for implementing a Generative AI solution. The actual use case(s) will dictate the final design; consequently, design workshops are highly recommended at the start of one's Generative AI journey.

2 Analytics Discussion

2.1 Architecture's Analytics Overview

Generative AI (GenAI) and large language models (LLM) involve creating algorithms that can generate new content based on patterns learned from existing data. Common models for GenAI are generative adversarial networks (GANs) to generate multi-media and realistic speech, variational autoencoders (VAEs) for signal processing, autoregressive models, and diffusion models for waveform signal processing, multi-modal modeling, molecular graph generation, time series modeling, and adversarial purification. A subset of Generative AI models is large language models that are trained on a broad set of unlabeled data that can be used for different tasks and fine-tuned for purposes across many verticals**.** These models have billions of parameters, 65 billion to 1 trillion for latest models, which require large numbers of accelerators, like GPUs, and extended time to train. The number of parameters is increasing at exponential rates; consequently, putting enormous strain on training resources. As such, the tendency of most organization is to use models pre-trained by others. These foundation models, either open-source or commercial, only require fine tuning for the specific use case and can be deployed relatively quickly for inferencing. Further details on this process can be found here. Although the focus of this document is on generative AI, the architecture and topics discussed are applicable for machine learning models and all deep learning models, including computer vision.

2.2 Common Use Cases

The types of GenAI use cases are as varied as the industry verticals to which they pertain. Across industry examples include:

- 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

Industry vertical examples are provided below. This by no means comprehensive; however, it provides a landscape of possibilities.

Manufacturing	Life Sciences	Retail and Finance
– Predictive Maintenance / Digital Twin – "Read" repair manuals, service bulletins, and warranty claims for new	– Medical imaging – Predicting protein structures for new gene therapies – Drug discovery	– AI concierge/branch banking – Personalized marketing campaigns

insights and quicker problem resolution – Design new parts and processes – Explore how a component can be manufactured faster – Detect patterns and trends for supplier risk analysis – Production line, robotics & process optimization	– Generate images of biological structures and processes for enhanced understanding – Protein engineering – Nefarious pattern identification and fraud detection – Regulatory intelligence	– Customer engagement / sales automation – Scenario simulation using market conditions, and macroeconomic factors for insights into risks and opportunities – Compliance and document analysis – Employee and customer - facing chatbots – Claims automation

3 Architectural Overview

The architecture developed represents components optimized to meet the computationally intensive work loads of Generative AI and large language models. The core components for the compute layer are built upon the Lenovo ThinkSystem SR675 V3 appliance using NVIDIA L40S GPUs and AMD 4th generation EPYC™ processors.

The NVIDIA L40S is based on the NVIDIA's Ada Lovelace architecture. These GPUs have fourth-generation tensor cores, third-generation RT cores, and CUDA cores. The L40S transformer engine improves AI performance by optimizing memory utilization and by automatically recasting between FP8 and FP16 precision as needed by the neural network architecture.

The AMD x86 server processors have industry leading performance with up to 96 "Zen 4" cores and 1152 MB of L3 cache per socket. This series of processors delivery leading 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.

For ultra-low latencies, all architectural components are connected to a NDR 400 InfiniBand fabric for ultra-low latencies. Included in the solution is the NVIDIA's AI Enterprise NeMo framework from which foundational models can be extracted, retrained to a customer's data and deployed. Further details of this configuration can be found here.

The storage layer is built upon on the ThinkSystem SR635 V3. The details of this appliance are supplied in the below hardware section and the architecture design is provided below.

3.1 Architecture Diagram

The below figures are of the same architecture and show two separate views: a technology view and the rack & stack view. The technology view illustrates the networking and switch layout in a foundational context serving as a repeatable unit that can be scaled to any level of need. The rack & stack view illustrates how this solution would be configured in a data center and is based on the BOM presented in the appendix.

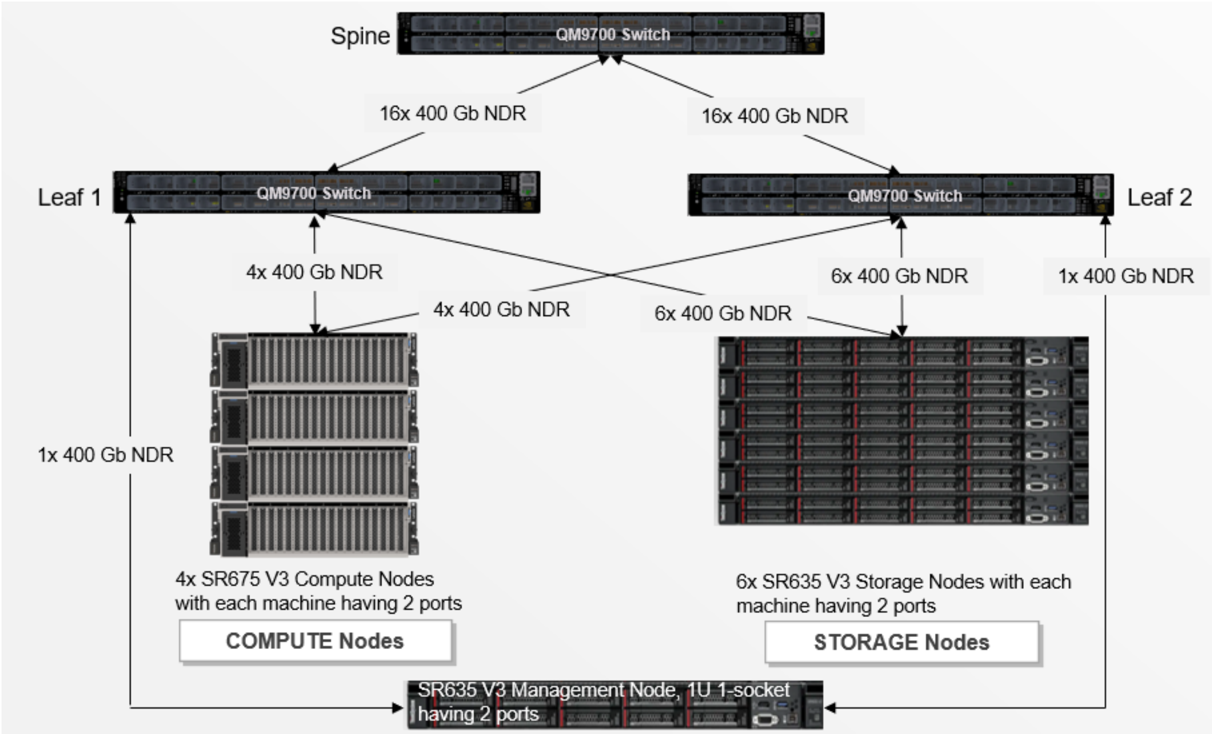

Figure 1. Technology view of the reference architecture

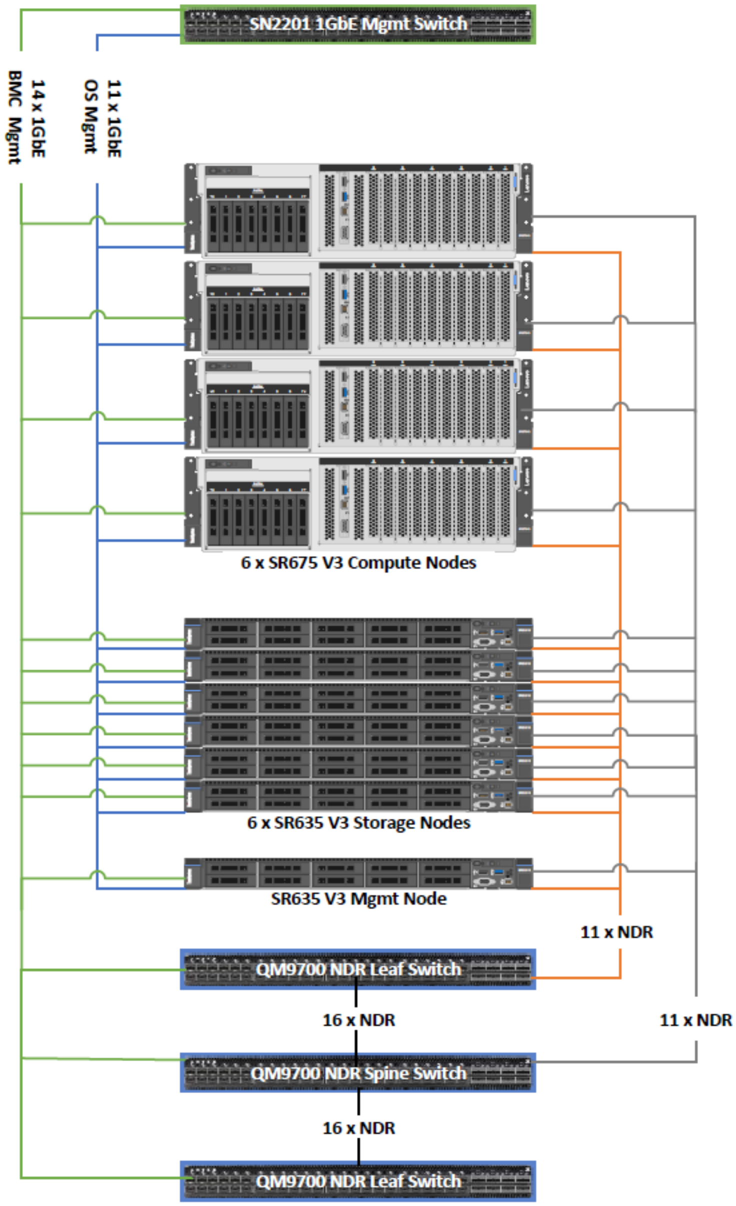

Figure 2. Rack and stack view of the reference architecture

The actual configuration implemented depends on the use case, the organization's existing infrastructure, the existing technical staffing, and the budget. Given these considerations, an all ethernet network topology could be used in place of InfiniBand.

3.2 Solution Design

The reference architecture design has a Fat Tree Topology with a unified fabric using a 2 leaf 1 spine configuration. This design has high availability networking (North/South and East/West bound traffic) with support of fail-over redundancy. This approach to architecture facilitates growth from small scale to massive scale operations in an economical way.

The compute layer of this solution consists of four Lenovo ThinkSystem SR675 V3 servers connected to two NVIDIA QM9700 InfiniBand switches (system user procedure) . Each server has 2 single-port ConnectX-7 400 Gb/s NDR network connections for inter-GPU system communication.

For the storage layer, six Lenovo ThinkSystem SR635 V3 servers are utilized. Each server is configured with ConnectX-7 400 Gb/s NDR InfiniBand ports and PCIe 5.0 x16 slots. Each SR635 V3 consists of single ports and contains 10x – 15.36TB NVMe SSDs. These support GPU direct functionality to maximize CPU performance utilization.

For GPU direct storage designs such as this one, a direct path exists between the GPU memory and the external storage enabling data to be moved into and out of GPU memory without taxing the CPU. This approach avoids extra copies of data through a bounce buffer in the CPU's memory. Without GPU direct, a bottleneck is created resulting in adverse effects on application performance because of increased time required to load data. This is especially prevalent as dataset sizes increase. Ref. https://developer.nvidia.com/gpudirect-storage

In this architecture, the Subnet Manager on the IB network is being run on the switch as opposed to on a server. The managed spine switches are QM9700's.

To round out the architectural design, a deployment/management server is also included. This management node is a SR635 V3 appliance having a 1U 1-socket configuration that is connected to the 400 Gb/s NDR network.

The network switch is configured with a message transmission unit (MTU) size of 4096 for InfiniBand (9000 for Ethernet).The WEKA AI RA does not require RoCE to be configured; however, if an Ethernet network is desired this can be a functional alternative to IB. This does not require priority flow control (PFC) to be configured on the network switch, greatly simplifying the network deployment.

4 Storage Platforms

4.1 WEKA storage platform

The WEKA system overview can be found here. Key highlights as they pertain to this architecture are provided below.

WEKA is a software solution that enables the implementation of a shareable, scalable, distributed filesystem storage.

The WEKA filesystem (WekaFS™) redefines storage solutions with its software-only approach, compatible with standard AMD or Intel x86-based servers and NVMe SSDs. WekaFS addresses common storage challenges by removing performance bottlenecks, making it suitable for environments requiring low latency, high performance, and cloud scalability.

WekaFS is a fully distributed parallel filesystem leveraging NVMe Flash for file services. Integrated tiering seamlessly expands the namespace to and from HDD object storage, simplifying data management. The intuitive GUI allows easy administration of exabytes of data without specialized storage training.

WekaFS stands out with its unique architecture, overcoming legacy systems' scaling and file-sharing limitations. Supporting POSIX, NFS, SMB, S3, and GPUDirect Storage, it offers a rich enterprise feature set, including snapshots, clones, tiering, cloud-bursting, and more.

Benefits include high performance across all IO profiles, scalable capacity, robust security, hybrid cloud support, private/public cloud backup, and cost-effective flash-disk combination. WekaFS ensures a cloud-like experience, seamlessly transitioning between on-premises and cloud environments.

WekaFS is particularly beneficial when integrated into Generative AI environments. Each AI pipeline step usually has a completely different profile for what the data looks like. When different IO demands are introduced to work on the data throughout the pipeline, this can cause issues with traditional storage. Some steps need low latency and random small IO. Others need massive streaming throughput. Still others need a concurrent mix of the two because of sub-steps within the process. In many cases, there can be multiple overlapping pipelines running against the same system at any given time.

A single pipeline makes it difficult to optimize a storage system for any stage of the pipeline, but overlapping pipelines in a consolidated environment brings out the dreaded "IO blender" problem. While originally coined for Virtual Machine consolidation, this same concept also applies to GenAI Data Pipelines.

Multi-pipeline impact on IO profiles

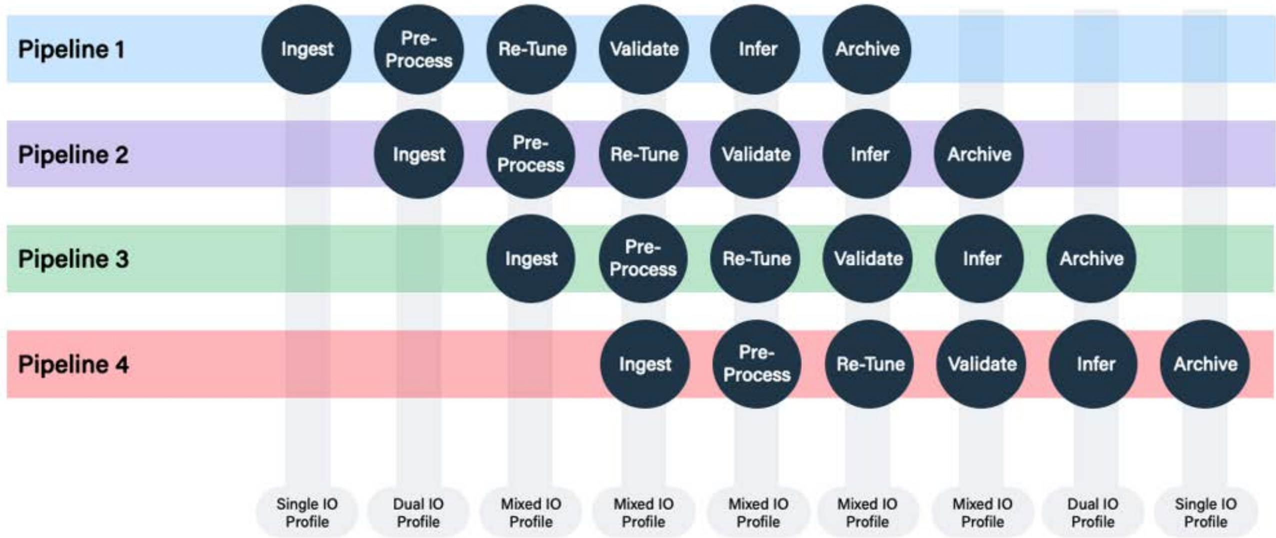

Figure 3. IO profile data pipelines as function of process function

When the Generative AI pipelines are overlapped, a storage system no longer must deal with just discrete IO for every step in a pipeline, but now must handle the mixed IO from various stages of every pipeline simultaneously as illustrated in Figure 3.

Consider a department of various researchers or developers launching training and/or re-tuning jobs at various times: the overlap blurs the IO patterns to the point where the storage system is dealing with a mixed IO profile that tends to be random in nature. This mixed IO has become quite typical in GenAI environments, and stresses most traditional storage and filesystems that were originally developed to handle one or two types of siloed IO. Ultimately, mixed IO can slow down legacy storage environments to the point that the trade-off for consolidation may not have been worthwhile.

The Lenovo High Performance File System Storage Solution with the WEKA® Data Platform is built on Lenovo ThinkSystem SR635 V3 server. The flexibility provided by using WEKA-certified Lenovo Ready Nodes includes enhancements from current 15.36TB PCIe Gen5 NVMe to larger future capacities such as 30.72TB. Both AMD and Intel processor versions are supported, as well a range of high-speed InfiniBand and Ethernet network adapters.   This solution provides a modern, high-performance solution to the mixed IO pipelines observed in GenAI and other GPU-focused workloads.

Additional details on WEKA systems on Lenovo hardware can be found below.

[Lenovo EveryScale Design Architecture for WEKA Storage]

5 Hardware and Network Topology

The compute and data-intensive applications that run GenAI workloads demand maximum performance. Both the SR635 V3 and the SR675 V3 have been optimized to run specific layers of the GenAI solution stack, one for the storage layer and one for the compute layer, respectively. The SR635 V3 supports GPU direct with the benefits described and allows for a smaller footprint system not requiring any major data center changes. The SR675 V3 is a GPU dense box designed for AI workloads. Both types of hardware utilize the AMD platform that provides enough PCIe Gen5 lanes to future proof the initial investment.

5.1 Storage layer: SR635 V3 product highlights

As shown in the below interconnect diagram, the ThinkSystem SR635 V3 is a 1-socket 1U single CPU server that features the fourth generation AMD EPYC 9004 "Genoa" family of processors. With up to 128 processor cores and support for PCIe 5.0 standard for I/O, the SR635 V3 is designed for performance in a slim design. The SR635 amplifies speed with support for the latest TruDD5 memory and storage capacity that is needed for these complex workloads.

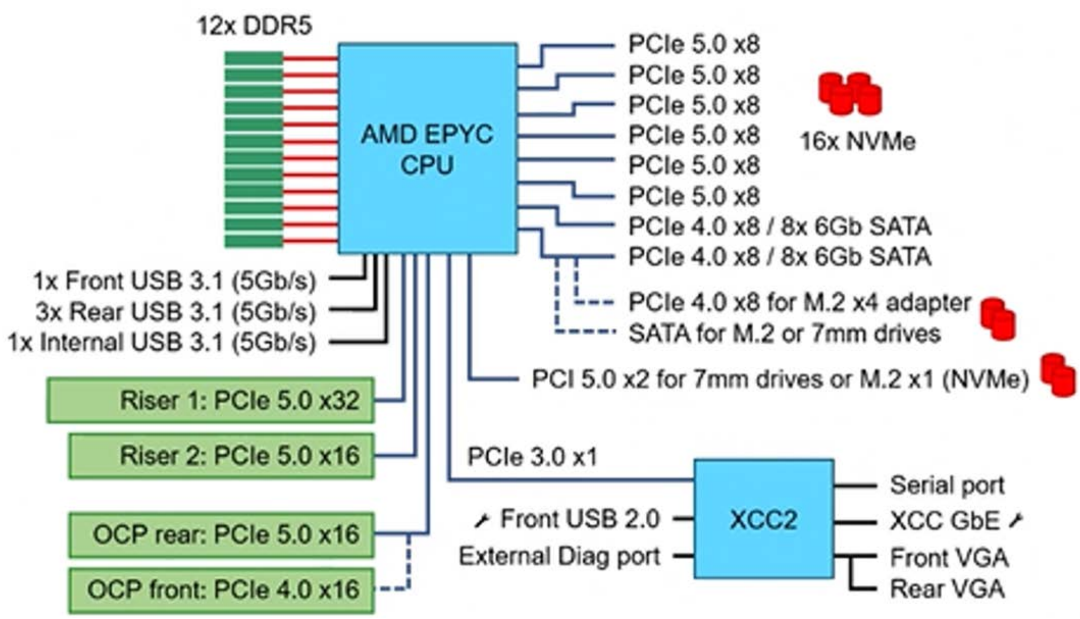

Figure 4. SR635 V3 Server Interconnect Diagram, showing major components and their connections.

The SR635 is designed to be a storage ready node whose features and bus speeds are optimized to provide the best disk performance for AI workloads.

This server is ideal for dense workloads, like Generative AI, that can take advantage of GPU direct processing and high-performance NVMe drives.

The SR635 V3 server has been designed to take advantage of the features of the 4th generation AMD EPYC processors, such as the full performance of a 360W 128-core processor, support for 4800 MHz memory and PCIe Gen 5.0 support. The server also offers onboard NVMe PCIe ports that allow direct connections to 16x NVMe SSDs, which results in faster access to store and access data. Included with the SR635 V3 is Lenovo's XClarity software that provides easy management from deployment to decommission along with added security features like monitoring for unexpected adds or changes.

Further details regarding this server are provided in its product guide.

5.2 Compute layer: SR675 V3 product highlights

For Generative AI it is important to find a balance between performance, power consumption and scalability. The ThinkSystem SR675 V3 AI ready node meets this balance. This appliance is a versatile GPU-rich 3U rack server that supports four or eight double-wide GPUs including the new NVIDIA H100 NVL and L40S Tensor Core GPUs. As shown in the below interconnect diagram, the ThinkSystem SR675 V3 server is based on 2x AMD EPYC 9004 Series processors with up to 3TB DDR memory, up to 6x EDSFF E1.S NVMe SSDs, and up to 160 PCIe lanes.

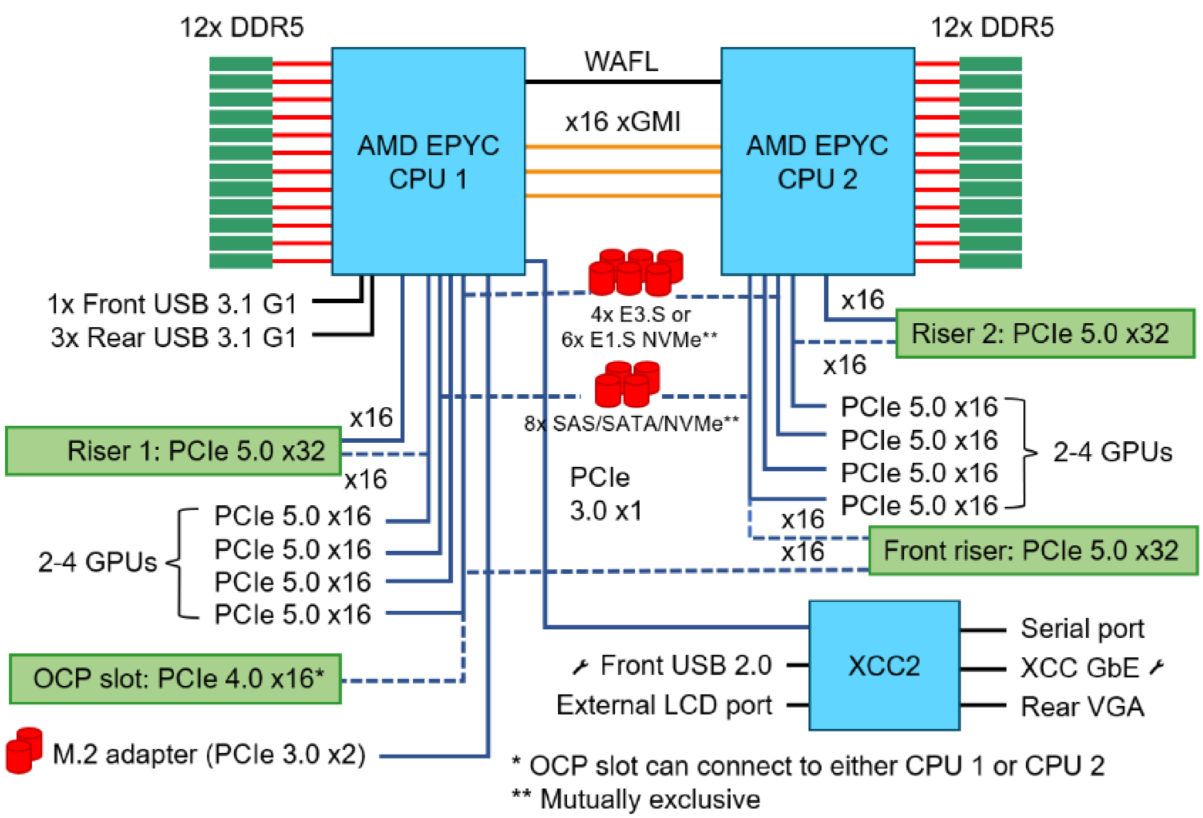

Figure 5. SR675 V3 Server Interconnect Diagram, showing major components and their connections.

This appliance has a high core count compute engine tailor made for data center applications with up to 128 high performance Zen 4c cores. Each SR675 V3 has high core counts per PCI lanes per socket to provide maximum configuration flexibility. Its large capacity cache is tailor made for simulations and computational fluid dynamics that play a key part in some digital twin applications that can complement AI workloads. This appliance has been designed for scale-out computing with drop-in capability for next generation GPUs and sufficient thermal headroom for future power requirements. It is configurable with a broad range of accelerators from low to high-end and has the versatility of supporting multiple PCIe configurations.

Further details regarding this server are provided in its product guide.

5.3 Network Topology

The networking chosen for this architecture is NVIDIA's InfiniBand, which is a high-speed, low latency, low CPU overhead, highly efficient, and scalable server, and storage interconnect technology. The switches employed provide ultra-high data throughput and density required of highly parallelized algorithms common for GenAI models.

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.

Ultimately, the decision should be based on the customer's specific requirements, existing infrastructure, budget, and the expertise available within the organization to manage and optimize the chosen networking technology.

Further details on the networking and switches used can be found in the compute layer GenAI reference architecture.

6 Other Considerations

In this section, a high-level survey of considerations is provided to help an organization on its journey to create and deploy an AI environment that supports GenAI initiatives. The various components comprising a GenAI ecosystem include computing, networking, storage, analytics data strategy, data lifecycle management, data ingestion pipelines, model management, and security. Building a system so that all its components act in synergy requires a balancing of performance, cost, and complexity. The approaches taken are dependent upon whether the system is a brownfield or greenfield implementation. For both types of implementations, building on the company's established cybersecurity, governance, and policies, while the solution is under development, facilitates the creation of a responsible AI offering that is secure and flexible for scaling.

An emerging trend that needs to be considered is more and more enterprise applications have GenAI capabilities core to their functionality. The convergence of technologies, like ML and computer vision, with GenAI and the integration into IoT networks makes the need for a holistic development process a must. All these challenges require a cross functional team representing personas of IT, OT, data engineering, and data science to participate in the solution development. The sections that follow are by no means exhaustive; however, they provide a starting point for the GenAI journey that results in a well-crafted GenAI solution.

6.1 Data strategy and lifecycle

A comprehensive view of what happens to data from creation to modeling to archival needs to be part of any sound data strategy. Artificial intelligence is not the old business intelligence story where data is stored in a transactional database that has been optimized for standard reports. In the artificial intelligence paradigm, all the following need to be considered: data velocity (hot, cold, warm), data refresh rates, data munging, data retention policy, and understanding of what happens with the inference results. The days of pushing data to a database and forgetting it can no longer be the approach.

Streaming data, hot data, can be produced and refreshed at deep sub-seconds. In some cases, AI models must act upon the data inflight. In other cases, acting upon the data in a few seconds achieves the objectives. At the other extreme, cold data, traditional batch processing is used. In most cases, the analytic model creators may not know about its velocity or refresh rate and as such treat it as a static dataset, which leads to answers not based upon ground truth.

All data has varying degrees of fidelity. Data from different sources could have values that should be the same but because of how it was captured/measured may be off and need reconciliation. Data can be captured at different time stamps, not standardized to UTC and/or measured at different velocities, which if not corrected would lead to incorrect model results. Data can be measured in different units, imperial vs. metric, this too much be reconciled to produce an accurate outcome. These are just a few data quality issues that must be addressed prior to modeling. The data munging required to form data with sufficient fidelity for modeling must be pushed upstream as much as practical. Having each model developer do munging in isolation can lead to inaccuracies and conflicting outcomes.

Data retention policies, which are often dictated by taxing authorities and regulators, must be considered in any design. A common rule of thumb in US is to retain financial data for 7 years. In some cases, only metadata needs to be retained. The company's own policies drive these requirements. In the context of AI modeling, often a requirement exists to be able to reproduce the original model, which is dependent upon the data retention policy.

Once a model has been pushed to production and inference results are being produced by new data, what happens next needs to be considered. Does the inference result kick-off a process? Does the inference result act as an input to another model? Do the inference results require human intervention? Do the inference results need to be saved and if so, what is the governing policy? All these questions need to be considered and mapped to appropriate responses. The data infrastructure will need to support this need.

In the below figure, an example analytics data lifecycle is provided. The analytic data pipelines and the AI enabled Lenovo infrastructure solutions are mapped to the data flow and its associated modeling activities.

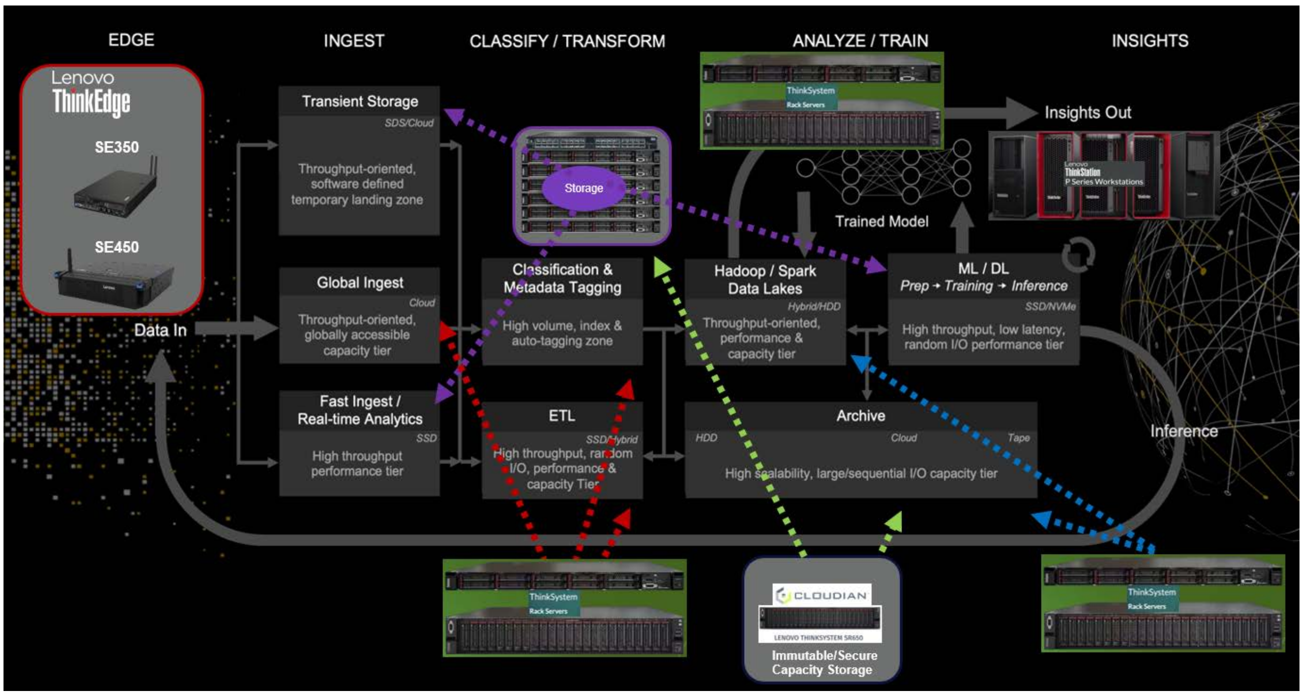

Figure 6. Analytics data pipelines enabled by Lenovo Infrastructure Solutions and leading software partners

The Lenovo AI-Ready portfolio consists of ThinkSystem servers, ThinkEdge devices, and ThinkStations. These play a role in an overall AI strategy. Below provides a description of each machine category and the most common AI personas that would use.

- ThinkSystem Servers – the SR675 V3 and the soon to be released SR685a V3 are the go-to appliances for GenAI. ThinkSystem servers are built to deliver the performance, security, and uptime needed in AI deployments.
  - Persona: data scientist, who is training a new model from scratch, and/or heavily re-training foundational models.
  - Persona: data engineers converting a data scientist's model into production grade code for deployment and inferencing.
- ThinkEdge – compact, ruggedized systems built for AI inferencing where the data is created: in remote locations in the field, at the point of contact. ThinkEdge systems are purpose-built to deliver AI at the edge in the form of light weight inference models.
  - Persona: Data engineer deploying a production grade inference model.

- ThinkStation – powerful workstations designed for complex ML and AI algorithm development. These systems are built to reduce the labor-intensive tasks of pre-processing and data preparation, where data scientists spend nearly 80% of their time.
    - Persona: data scientist who is developing or prototyping a new lighter weight model, e.g., 7 billion parameter LLMs, or determining which foundational model is best suited for a use case.

6.2 Data lake house and source systems integration

The below figure represents the data pipeline, data integration, data lakehouse and modelling steps for a typical artificial intelligence solution. This represents the analytics architecture showing the flow of source data, edge data acquisition and operations data acquisition, into the data lakehouse where it is stored, processed, and managed for analytics consumption. Once a model is created, trained, to address a use case, it is converted into production grade code for deployment through a process, DevOps etc. That model's performance is monitored and archived when it decays due to data drift. The archived model along with the original data used in its creation is stored in the lakehouse for traceability. The model output, inference results, are feed to downstream processes, not depicted, via a case management system.

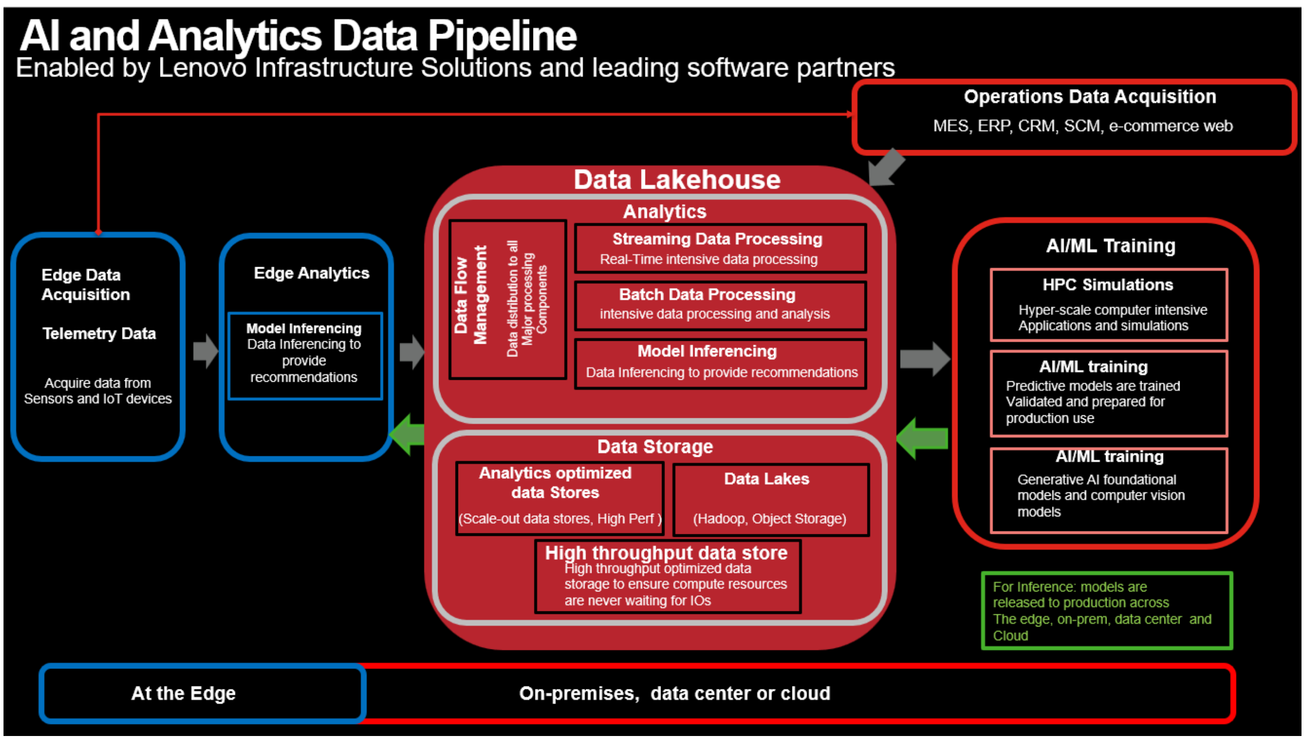

Figure 7, Representation of common data and analytics patterns relating to modelling and Generative AI

In the 1980's, data warehouses were the choice for large, centralized repositories of integrated data collected from various sources. As data sizes grew, warehouse technology also evolved. This was in a time where most organization only had to be concerned about cleaned structured data for BI reporting. Data warehouses are ideal for this situation. Now, organizations doing data science are concerned with structured, unstructured, and semi-structured data that has a range of sizes and data velocities. Data warehouses are not suited for these needs and are inefficient for these use cases. A more modern construct, data lakes are designed to store raw unstructured data in any format. Lakes store data, but do not support transactional queries, nor enforce data quality. Lakes do not have isolation/consistency making mixing of appends and reads or batch and streaming difficult. To address this challenge, organization were forced to create a multitude of systems containing both data lakes and warehouses; however, this approach introduced complexity and latencies for the modelling community.

Companies now require systems for diverse data applications including real-time monitoring, SQL querying, machine learning and data science. As such, architects have gravitated to an open architecture single system design that can accommodate this variety of data needs leading to the creation of data lakehouses. Lakehouses are a new design combining the benefits of data warehousing with data lakes. The aim is to store raw data of any form while also enabling structured queries common to warehouses. Lakehouses allow data management features and data structures of a warehouse to be built on low-cost storage in open formats. They support object stores that can store a broader range of data types: structured, semi-structured and unstructured (images, video, audio, semi-structured data, and text). They support openness in the form of open and standardized formats, like parquet, and have APIs that can be used to integrate python and R libraries directly with the data.

Lakehouses can also support ad-hoc querying and complex analytics. They allow schemas to evolve over time unlike warehouses that require a pre-defined schema. A single platform can now integrate data storage, data processing and analytics. In addition, lakehouses support ACID transactions which ensure consistency across concurrent reads or writes. They support schema enforcement and governance as well as BI reporting tools applied directly to the source data.

Data lakehouses can include vector databases that contain massive amounts of high-dimensional vector data. These types of databases are needed for development of large language models and chatbots that rely upon them. The primary feature of vector databases is that they can efficiently index, search, and store high-dimensional data points, which are clustered based on similarity and not criteria based exact matching as would be the case for traditional databases. In more traditional databases, SQL and NoSQL, individual data points are represented by rows and columns in a structured tabular format. Vector databases contain data points that are represented by vectors with a fixed number of dimensions. Vectors, in this context, are numerical features, embeddings of images or text, speech/audio, or other diverse ranges of information. Vector databases are one part of the data and AI ecosystem. Knowledge graphs are often confused with vector databases. These graphs represent relationships between objects or events and are a fit-for-purpose databases storing the graph structure and the knowledge graph information. Knowledge graphs represent known entities and complex relationships between them while vectors databases are high-dimensional vectors. Knowledge graphs are preferred for processing complex relationships whereas vector databases are more suited to handle images, videos and LLM text.

The final components of an enterprise grade data and analytics system are security, access control and data governance capabilities, which includes lineage, retention, and auditing. A means to monitor and respond, as appropriate, to each of these areas should be incorporated into the design. There are many tools in the market that address these needs.

6.3 API capabilities for foundational GenAI models

A foundation GenAI model is often found in library repositories, like those of Hugging Face, OpenAI, or LangChain. To utilizes these models on one’s own corpus of data, APIs are required. API advances are making this more seamless. As an example, GPT-4 Turbo allows API access to OpenAI's image generation tool (Dall-E 3) and its advanced data analysis tool. These API enable easy input of enormous amounts of data for customized graphics or low-level data analysis. The API call for GPT-4 Turbo now allows for multiple functions to be called in a single API message leading to lower latency with increased speed and less computational overhead.

Additional information can be found at the below links.

https://python.langchain.com/docs/integrations/platforms/openai

https://www.linkedin.com/pulse/generative-ai-made-another-big-leap-forward-business-week-greenstein-snyce/

6.4 Model lifecycle management considerations

Models like data must be managed throughout their lifecycle. All models decay over time where there predictive power lessons because of data drift. In many cases, organizations require the ability to recreate a model. As such, only maintaining model revisioning, with the original supporting data and metadata can achieve this. To understand the complexities of this task an understanding of the modelling process is needed. The CRISP-DM methodology is an industry standard that is used not only for data mining, for which it was originally developed, but for modelling in general. This is a structured approach when combined with agile methodologies leads to the best results. In this process, there are several key steps: business understanding, data understanding, data preparation, modelling, evaluation, and model deployment typically through CI/CD and DevOps processes. To monitor model performance model ops is employed. Systems that perform model ops look at metrics that pertain to the model being used. For example, a common metric for regression models is ROC or lift curves. The analyst, who creates the model, decides upon the metric and its threshold used to signal that the model has decayed to a sufficient level to warrant review. In most model ops platforms, the chosen metric is automatically calculated and monitored, appropriate alerts are generated and supporting validating and auditing reports created. All these capabilities should be considered in a well-designed AI platform.

7 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

7.1 BOM for Weka design

PN	Description	Quantity
	Server_Compute - 4DW L40S	4
7D9RCTOLWW	ThinkSystem SR675 V3 3yr Warranty - HPC&AI with Controlled GPU	4
BR7N	ThinkSystem SR675 V3 x16 PCIe Gen5 Rear IO Riser	4
BR7H	ThinkSystem SR675 V3 2x16 PCIe Front IO Riser	4
BR7Q	ThinkSystem SR675 V3 Direct 4x16 PCIe DW GPU Riser	4
BQ1N	ThinkSystem NVIDIA ConnectX-7 NDR OSFP400 1-Port PCIe Gen5 x16 InfiniBand Adapter	8
BLL2	ThinkSystem V3 2U 8x2.5" AnyBay Gen5 Backplane	4
B8P9	ThinkSystem M.2 NVMe 2-Bay RAID Enablement Kit	4
BE0E	N+N Redundancy With Over-Subscription	4
BFYA	Operating mode selection for:"Maximum Efficiency Mode"	4
BKTJ	ThinkSystem 2600W 230V Titanium Hot-Swap Gen2 Power Supply	16
B7Y0	Enable IPMI-over-LAN	4
5977	Select Storage devices - no configured RAID required	4
BYFH	ThinkSystem NVIDIA L40s 48GB PCIe Gen4 Passive GPU	16
BFTL	ThinkSystem SR670 V2/ SR675 V3 Toolless Slide Rail	4
B93E	ThinkSystem Intel I350 1GbE RJ45 4-port OCP Ethernet Adapter	4
BR7G	ThinkSystem SR675 V3 4DW PCIe GPU Base	4
BPVJ	ThinkSystem AMD EPYC 9554 64C 360W 3.1GHz Processor	8
BKSR	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	8

PN	Description	Quantity
BQ3D	ThinkSystem 64GB TruDDR5 4800MHz (2Rx4) 10x4 RDIMM-A	96
BK1E	ThinkSystem SR670 V2/ SR675 V3 OCP Enablement Kit	4
BRUD	ThinkSystem SR675 V3 Front Video/USB/Diagnostic for 4-DW GPU model	4
6252	2.5m, 16A/100-250V, C19 to IEC 320-C20 Rack Power Cable	16
5PS7B09635	Premier Essential - 3Yr 24x7 4Hr Resp + YDYD SR675 V3	4
2306	Integration >1U Component	4
BABV	ThinkSystem Screw for fix M.2 Adapter	4
BFNU	ThinkSystem SR670 V2/ SR675 V3 Intrusion Cable	4
BFGY	ThinkSystem SR670 V2/ SR675 V3 Backplane Power Cable 3	4
BFD6	ThinkSystem SR670 V2/ SR675 V3 Power Mezzanine Board	4
BFCZ	ThinkSystem SR670 V2/ SR675 V3 PCIe Rear Riser Bracket Filler	4
BFTH	ThinkSystem SR670 V2/ SR675 V3 Front Operator Panel ASM	4
BK15	High voltage (200V+)	4
BR7U	ThinkSystem SR675 V3 Root of Trust Module	4
BR7V	ThinkSystem SR675 V3 System Board	4
BR80	ThinkSystem SR675 V3 Agency Labels	4
BR88	ThinkSystem SR670 V2/ SR675 V3 Service Label	4
BR85	ThinkSystem SR670 V2/ SR675 V3 Branding Label	4
BR8W	ThinkSystem SR675 V3 Front PCIe Riser Cable 3	4
BR8R	ThinkSystem SR675 V3 Front PCIe Riser Cable 4	4
BR8E	ThinkSystem SR675 V3 Backplane to MB Cable 3	4
BR8S	ThinkSystem SR675 V3 Direct DW/SW GPU Riser Cables 1	4
BR8F	ThinkSystem SR675 V3 Backplane to MB Cable 4	4
BRNM	ThinkSystem SR670 V2/SR675 V3 2600W Power Supply Caution Label	4
BRUC	ThinkSystem SR675 V3 CPU Heatsink	8
BRUS	ThinkSystem SR675 V3 Rear OCP Cable	4
BR7W	ThinkSystem SR670 V2/ SR675 V3 System Documentation	4
BS6Y	ThinkSystem 2U V3 M.2 Signal & Power Cable, SLx4 with 2X10/1X6 Sideband, 330/267/267mm	4
BSD2	ThinkSystem SR675 V3 GPU Supplemental Power Cable 4	16

PN	Description	Quantity
AVEN	ThinkSystem 1x1 2.5" HDD Filler	32
BF93	AI & HPC - LeSI Solutions	4
A102	Advanced Grouping	4
8971	Integrate in manufacturing	4
A193	Integrated Solutions	4
8034	LeSI Solution Component	4
	Server_Mgmt	1
7D9GCTOLWW	ThinkSystem SR635 V3 - 3yr Warranty - HPC&AI	1
BLK7	ThinkSystem V3 1U x16 PCIe Gen5 Riser1	1
BVGL	Data Center Environment 30 Degree Celsius / 86 Degree Fahrenheit	1
BPKR	TPM 2.0	1
BLKD	ThinkSystem 1U V3 10x2.5" Media Bay w/ Ext. Diagnostics Port	1
BQ1N	ThinkSystem NVIDIA ConnectX-7 NDR OSFP400 1-Port PCIe Gen5 x16 InfiniBand Adapter	2
B8N0	ThinkSystem 1U 8x2.5" SAS/SATA Backplane	1
BNWC	ThinkSystem 2.5" PM1653 960GB Read Intensive SAS 24Gb HS SSD	2
BE0E	N+N Redundancy With Over-Subscription	1
BFYA	Operating mode selection for:"Maximum Efficiency Mode"	1
BLKH	ThinkSystem 1100W 230V Titanium Hot-Swap Gen2 Power Supply	2
B7Y0	Enable IPMI-over-LAN	1
5977	Select Storage devices - no configured RAID required	1
B8LA	ThinkSystem Toolless Slide Rail Kit v2	1
BQ26	ThinkSystem SR645 V3/SR635 V3 1U High Performance Heatsink	1
B93E	ThinkSystem Intel I350 1GbE RJ45 4-port OCP Ethernet Adapter	1
BLK9	ThinkSystem V3 1U MS LP+LP BF Riser Cage	1
BRQV	ThinkSystem RAID 5350-8i PCIe 12Gb Internal Adapter	1
BLK4	ThinkSystem V3 1U 10x2.5" Chassis	1
BRED	ThinkSystem AMD EPYC 9254 24C 200W 2.9GHz Processor	1
BQ3C	ThinkSystem 16GB TruDDR5 4800MHz (1Rx8) RDIMM-A	12

PN	Description	Quantity
BH9M	ThinkSystem V3 1U Performance Fan Option Kit v2	6
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2
BLKA	ThinkSystem V3 1U x16 PCIe Gen5 Riser2	1
5PS7B08712	Premier Essential - 3Yr 24x7 4Hr Resp + YDYD SR635 V3	1
2305	Integration 1U Component	1
AUTQ	ThinkSystem small Lenovo Label for 24x2.5"/12x3.5"/10x2.5"	1
AUWG	Lenovo ThinkSystem 1U VGA Filler	1
AWF9	ThinkSystem Response time Service Label LI	1
B5X6	ThinkSystem 1U 2x2.5" Fixed Filler	1
B97B	XCC Label	1
B8NJ	ThinkSystem 1U MS Fan Dummy	2
B8NK	ThinkSystem 1U Super Cap Holder Dummy	1
B984	ThinkSystem 1U PLV Top Cover Sponge	1
B8KV	ThinkSystem 1U 8x2.5" SAS/SATA HDD Type Label	1
B8GR	ThinkSystem V3 1U Power Cable to CFF RAID/HBA	1
BA1Z	ThinkSystem 1U CFF RAID to 10x2.5" Backplane SAS/SATA G4 Cable 2	1
BHSS	MI for PXE with RJ45 Network port	1
BPK3	ThinkSystem WW Lenovo LPK	1
BK15	High voltage (200V+)	1
BMJC	ThinkSystem 8x2.5" BP and 6+4 x2.5" BP Power Cable v2	1
BPDF	ThinkSystem 1100W Ti Power rating Label WW	1
BQPS	ThinkSystem logo Label	1
BLK6	ThinkSystem V3 1U LP Riser Cage	1
BQ7L	ThinkSystem SR635 V3 MB	1
BQ7Q	ThinkSystem SR635 V3 Model Name Label	1
BQ7R	ThinkSystem SR635 V3 Agency Label	1
BQ7V	ThinkSystem SR635 V3 Service Label for WW	1
BQQW	ThinkSystem CFF Input Swift 7 (MB) CFF RAID 480mm	1
BSR6	ThinkSystem SR635 V3/SR655 V3 RoT Module LV-RoW	1

PN	Description	Quantity
AVEN	ThinkSystem 1x1 2.5" HDD Filler	6
BF93	AI & HPC - LeSI Solutions	1
A2HV	Configuration ID 07	1
BZM8	DSS-G 5.0 SW Stack	1
AVW7	DSSG Solution	1
8971	Integrate in manufacturing	1
A193	Integrated Solutions	1
8034	LeSI Solution Component	1
BRPJ	XCC Platinum	1
	Storage_Weka FS	6
7D9GCTO8WW	-SB- ThinkSystem SR635 V3 for WEKA	6
BTTY	M.2 NVMe	6
BLK7	ThinkSystem V3 1U x16 PCIe Gen5 Riser1	6
BVGL	Data Center Environment 30 Degree Celsius / 86 Degree Fahrenheit	6
BPKR	TPM 2.0	6
BLKD	ThinkSystem 1U V3 10x2.5" Media Bay w/ Ext. Diagnostics Port	6
BQ1N	ThinkSystem NVIDIA ConnectX-7 NDR OSFP400 1-Port PCIe Gen5 x16 InfiniBand Adapter	12
BRQX	ThinkSystem 1U 2.5" 10 NVMe Gen5 Backplane	6
BM8X	ThinkSystem M.2 SATA/x4 NVMe 2-Bay Enablement Kit	6
BNEQ	ThinkSystem 2.5" U.2 P5520 15.36TB Read Intensive NVMe PCIe 4.0 x4 HS SSD	60
BE0E	N+N Redundancy With Over-Subscription	6
BFYB	-SB- Operating mode selection for: "Maximum Performance Mode"	6
BLKH	ThinkSystem 1100W 230V Titanium Hot-Swap Gen2 Power Supply	12
B7Y0	Enable IPMI-over-LAN	6
5977	Select Storage devices - no configured RAID required	6
B8LA	ThinkSystem Toolless Slide Rail Kit v2	6
BQ26	ThinkSystem SR645 V3/SR635 V3 1U High Performance Heatsink	6

PN	Description	Quantity
B93E	ThinkSystem Intel I350 1GbE RJ45 4-port OCP Ethernet Adapter	6
BLK9	ThinkSystem V3 1U MS LP+LP BF Riser Cage	6
BC4V	Non RAID NVMe	6
BLK4	ThinkSystem V3 1U 10x2.5" Chassis	6
BREG	ThinkSystem AMD EPYC 9354P 32C 280W 3.25GHz Processor	6
BKSR	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	12
BQ39	ThinkSystem 32GB TruDDR5 4800MHz (1Rx4) 10x4 RDIMM-A	72
BH9M	ThinkSystem V3 1U Performance Fan Option Kit v2	36
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	12
BLKA	ThinkSystem V3 1U x16 PCIe Gen5 Riser2	6
5PS7B25064	Premier Essential - 3Yr 24x7 4Hr Resp + YDYD SR635 V3 SDS WEKA	6
2305	Integration 1U Component	6
AUTQ	ThinkSystem small Lenovo Label for 24x2.5"/12x3.5"/10x2.5"	6
AUWG	Lenovo ThinkSystem 1U VGA Filler	6
AWF9	ThinkSystem Response time Service Label LI	6
B97B	XCC Label	6
B8NJ	ThinkSystem 1U MS Fan Dummy	12
B8NK	ThinkSystem 1U Super Cap Holder Dummy	6
B984	ThinkSystem 1U PLV Top Cover Sponge	6
B8KJ	ThinkSystem 1U 10x2.5" NVMe HDD Type Label	6
BHSS	MI for PXE with RJ45 Network port	6
BPK3	ThinkSystem WW Lenovo LPK	6
BK15	High voltage (200V+)	6
BPDF	ThinkSystem 1100W Ti Power rating Label WW	6
BQPS	ThinkSystem logo Label	6
BLK6	ThinkSystem V3 1U LP Riser Cage	6
BQ7L	ThinkSystem SR635 V3 MB	6
BQ7Q	ThinkSystem SR635 V3 Model Name Label	6
BQ7R	ThinkSystem SR635 V3 Agency Label	6

PN	Description	Quantity
BQ7V	ThinkSystem SR635 V3 Service Label for WW	6
BQRU	ThinkSystem PCIe Cbl MB_Swift 5/Swift 6 10x2.5" AnyBay 390/390mm	6
BQRV	ThinkSystem PCIe Cbl MB_Swift 2 10x2.5" AnyBay(Havana) 520mm	6
BQRW	ThinkSystem PCIe Cbl MB_Swift 4/Swift3 10x2.5" AnyBay470/470mm	6
BRF3	ThinkSystem Cable132	6
BRQ3	ThinkSystem V3 2U WH CBL, 20Pin, 320mm,Tin-plated	6
BSR6	ThinkSystem SR635 V3/SR655 V3 RoT Module LV-RoW	6
BSEA	M.2 Module,MB to M.2 Signal Cable,275mm,(SLMx4)x2/SB to Module	6
BF93	AI & HPC - LeSI Solutions	6
A2HP	Configuration ID 01	1
A2HQ	Configuration ID 02	1
A2HR	Configuration ID 03	1
A2HS	Configuration ID 04	1
A2HT	Configuration ID 05	1
A2HU	Configuration ID 06	1
8971	Integrate in manufacturing	6
A193	Integrated Solutions	6
BRPJ	XCC Platinum	6
	Ethernet Management	1
7D5FCTOFWW	Nvidia SN2201 1GbE Managed Switch with Cumulus (PSE)	1
BPC7	Nvidia SN2201 1GbE Managed Switch with Cumulus (PSE)	1
6201	1.5m, 10A/100-250V, C13 to IEC 320-C14 Rack Power Cable	2
BSNB	NVIDIA SN2201 Enterprise Rack Mount Kit w/Air Duct	1
5WS7B14386	Premier Essential - 3Yr 24x7 4Hr Resp NVID SN2201 PSE	1
2305	Integration 1U Component	1
BF93	AI & HPC - LeSI Solutions	1
8971	Integrate in manufacturing	1
A193	Integrated Solutions	1

PN	Description	Quantity
8034	LeSI Solution Component	1
	Highspeed IB NDR Leaf b	1
0724HEC	NVIDIA QM9700 64-Port Managed Quantum NDR InfiniBand Switch (PSE)	1
BP63	NVIDIA QM9700 64-Port Managed Quantum NDR InfiniBand Switch (PSE)	1
BRQ6	2.8m, 10A/100-250V, C15 to C14 Jumper Cord	2
BRET	NVIDIA QM97xx Enterprise RMK w/Air Duct	1
5WS7B14266	Premier Essential - 3Yr 24x7 4Hr Resp NVID QM9700 PSE	1
2305	Integration 1U Component	1
BTH8	Add dust caps on all unused NDR switch ports	1
BF93	AI & HPC - LeSI Solutions	1
8971	Integrate in manufacturing	1
A193	Integrated Solutions	1
8034	LeSI Solution Component	1
	Highspeed IB NDR Leaf a	1
0724HEC	NVIDIA QM9700 64-Port Managed Quantum NDR InfiniBand Switch (PSE)	1
BP63	NVIDIA QM9700 64-Port Managed Quantum NDR InfiniBand Switch (PSE)	1
BRQ6	2.8m, 10A/100-250V, C15 to C14 Jumper Cord	2
BRET	NVIDIA QM97xx Enterprise RMK w/Air Duct	1
5WS7B14266	Premier Essential - 3Yr 24x7 4Hr Resp NVID QM9700 PSE	1
2305	Integration 1U Component	1
BTH8	Add dust caps on all unused NDR switch ports	1
BF93	AI & HPC - LeSI Solutions	1
8971	Integrate in manufacturing	1
A193	Integrated Solutions	1
8034	LeSI Solution Component	1

PN	Description	Quantity
	Highspeed IB NDR Spine	1
0724HEC	NVIDIA QM9700 64-Port Managed Quantum NDR InfiniBand Switch (PSE)	1
BP63	NVIDIA QM9700 64-Port Managed Quantum NDR InfiniBand Switch (PSE)	1
BRQ6	2.8m, 10A/100-250V, C15 to C14 Jumper Cord	2
BRET	NVIDIA QM97xx Enterprise RMK w/Air Duct	1
5WS7B14266	Premier Essential - 3Yr 24x7 4Hr Resp NVID QM9700 PSE	1
2305	Integration 1U Component	1
BTH8	Add dust caps on all unused NDR switch ports	1
BF93	AI & HPC - LeSI Solutions	1
8971	Integrate in manufacturing	1
A193	Integrated Solutions	1
8034	LeSI Solution Component	1
	Rack_1	1
1410O42	Lenovo EveryScale 42U Onyx Heavy Duty Rack Cabinet	1
BHC4	Lenovo EveryScale 42U Onyx Heavy Duty Rack Cabinet	1
BJPD	21U Front Cable Management Bracket	2
BJPC	Side Panel Right Installation	1
6012	DPI Single-phase 30A/208V C13 Enterprise PDU (US)	4
BJ2N	Front Installation of 180mm Extension Kit	1
BJPB	Side Panel Left Installation	1
BNAQ	Enable selection of latest options and features where available. Previous versions are disabled. To be able to select the previous versions, select None.	1
BHC7	ThinkSystem 42U Onyx Heavy Duty Rack Side Panel	2
BHC9	ThinkSystem 42U Onyx Rack Extension,180mm Depth	1
BJPA	ThinkSystem 42U Onyx Heavy Duty Rack Rear Door	1
2304	Integration Prep	1
2310	Solution Specific Test	1
AU8K	LeROM Validation	1
B1EQ	Network Verification	1

PN	Description	Quantity
5WS7A92829	Premier Essential - 3Yr 24x7 4Hr Resp_EveryScale 42U Rack	1
4275	5U black plastic filler panel	3
4273	3U black plastic filler panel	1
4271	1U black plastic filler panel	1
5AS7B07693	Lenovo EveryScale Rack Setup Services	1
9123	2-bay arrangement	1
BF93	AI & HPC - LeSI Solutions	1
8971	Integrate in manufacturing	1
A193	Integrated Solutions	1
BHXP	LeSI Service and Support Documentation	1
8034	LeSI Solution Component	1
9134	Use 200V (high voltage)	1
	Software	
S6Z3	NVIDIA AI Enterprise Subscription License and Support per GPU Socket, 3 Years	16
S9W2	WEKA Extreme Performance and Scale Edition - XPS Per TB up to 1PB 3Yr License w/WEKA support	93
SBCV	Lenovo XClarity XCC2 Platinum Upgrade (FOD)	7
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yr SW S&S	11
	Cables	
AVFX	1.0m Green Cat6 Cable	5
AVFZ	1.5m Green Cat6 Cable	2
AVG0	3m Green Cat6 Cable	12
BQJF	Lenovo 0.5m NVIDIA NDRx2 OSFP800 to NDRx2 OSFP800 Passive Copper Cable	32
BQJW	Lenovo 1.5m NVIDIA NDRx2 OSFP800 to 2x NDR OSFP400 Passive Copper Splitter Cable	6
BQJV	Lenovo 1m NVIDIA NDRx2 OSFP800 to 2x NDR OSFP400 Passive Copper Splitter Cable	2

PN	Description	Quantity
BQJY	Lenovo 3m NVIDIA NDRx2 OSFP800 to 2x NDR OSFP400 Passive Copper Splitter Cable	2
BQJX	Lenovo 2m NVIDIA NDRx2 OSFP800 to 2x NDR OSFP400 Passive Copper Splitter Cable	2
	Extras	
	Note: Parts added as Extras ship directly to the customer separately from the rack order.	
	Other parts	
7S1G002PWW	WEKA Installation (up to 30 nodes, remote)	1

8 Appendix:

Resources

Document history

Version 1.0	February 2024	First version includes both compute and storage layers

Trademarks and special notices

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.