Reference Architecture: Enterprise AI/ML Workloads on Lenovo ThinkSystem and ThinkAgile Platforms

Last update: 26 December 2023
Version 1.0

Reference Architecture for AI/ML workloads on Lenovo ThinkSystem, ThinkAgile VX and HX with/without GPU

Contains performance data and sizing recommendations for AI/ML use cases

Describes deployment models for data lake and model training and inference

Contains detailed bill of materials for ThinkSystem and ThinkAgile Hyperconverged systems

Chandrakandh Mouleeswaran

LENOVO PRESS

Table of Contents

1 Introduction 4
1.1 Enterprise AI 4
2 Business problem and business value 5
2.1 Business problem 5
2.2 Business value 5
3 Requirements 6
3.1 Functional requirements 6
3.2 Non-functional requirements 7
4 Architectural Overview 8
5 Component Model 9
5.1 Compute Model 11
5.1.1 CPU 12
5.1.2 GPU 12
5.2 Distributed Computing and Orchestration Frameworks 12
5.3 Lenovo Intelligent Computing Orchestration (LiCO) 12
5.4 NVIDIA AI Enterprise 14
5.5 VMware Private AI Platform 15
5.5.1 VMware Cloud Foundation 15
5.5.2 VMware Private AI Foundation 16
5.6 Nutanix Enterprise AI Platform 16
5.6.1 Nutanix Cloud Platform 16
5.6.2 Nutanix Enterprise AI 17
5.6.3 Nutanix GPT in a Box 18
5.7 Storage Model 18
5.7.1 Lenovo Shared Storage Systems 19
5.7.2 VMware vSAN 19
5.7.3 Nutanix AOS 20
5.7.4 Lenovo Object Storage powered by Cloudian 20
5.8 MLOps Framework 20
6 AI/ML System Design 21
6.1 Selection of Use case and AI Technology 21
6.2 Define Hybrid Data Access 22
6.3 Feature Engineering and Feature Store 22
6.4 Selection of AI Platform and Framework 23
6.5 GPU Accelerated Analytics 25
6.6 ML Model Development 25
6.7 ML Algorithms and Model selection 25
6.8 Batch and Online Models 27
6.9 Training 28
6.10 Distributed Training 28
6.11 Scalable Inference 29

6.12 ML Pipeline 29
6.13 Security 29
6.14 AI Observability and Responsible AI 30
7 Sample AI/ML Applications Solution Design 31
7.1 Recommender Systems 31
7.2 Timeseries Forecasting 33
7.3 Conversational AI 35
7.4 Generative AI (Large Language Models) 37
7.5 Cognitive Services 39
8 Operational model 41
8.1 Operational model scenarios 41
8.2 Example Enterprise AI Infrastructure Solution 42
8.2.1 Enterprise Operational Model 44
8.3 OS, Hypervisor and Container Platform Support 44
8.4 MLOps and Orchestration Platform 45
8.5 Management Server 46
8.6 ThinkSystem Servers Bare Metal for HPC and AI 46
8.6.1 Management Cluster 46
8.6.2 HPC and AI/ML Compute Cluster 46
8.7 ThinkSystem Servers with Kubernetes for AI 47
8.7.1 Management Cluster 47
8.7.2 AI/ML Compute Cluster 48
8.8 ThinkAgile VX Servers with VMware vSAN for AI 48
8.8.1 VX Servers 49
8.8.2 Management Cluster 49
8.8.3 AI/ML Compute Cluster 52
8.9 ThinkAgile HX Servers with Nutanix for AI 52
8.9.1 HX Servers 52
8.9.2 Management Cluster 53
8.9.3 AI/ML Compute Cluster 54
8.10 ThinkSystem Servers with Cloudian for Data Lake 54
8.11 ThinkEdge Servers for Edge AI 55
8.12 Neptune Direct Water Cooling Solutions 55
8.13 System Management 56
8.14 Lenovo XClarity Orchestrator(LXCO) 57
8.15 Lenovo Intelligent Computing Orchestration (LiCO) 57
9 Conclusion 58
10 Appendix: Bill of materials 59
10.1 BOM for AI on ThinkSystem Server 59
10.2 BOM for AI on ThinkAgile VX Server 61
10.3 BOM for AI on ThinkAgile HX Server 63
10.4 BOM for AI Storage 65
Resources 66

1 Introduction

The intended audience for this document is CXOs, technical IT architects, data scientists, data engineers and system administrators who are interested to design and deployment of Artificial Intelligence (AI) and Machine learning (ML) solutions for their enterprise wide use cases with Lenovo ThinkSystem servers and storage and ThinkAgile hyperconverged infrastructure platforms.

The value of artificial intelligence technologies span across different functions and applications in an enterprise and it drives operational efficiency and productivity. The democratization of artificial intelligence in the enterprise business brings development of relevant applications or software programs to address tasks or use cases across industries and sectors.

Enterprise AI solutions implementation across the business does require different hardware platforms and tools for data preparation and analysis, model development, training, inference, application development and integrations with existing systems and workflows. There is no one size fit model to suit different use cases and the rise of deep learning technologies and evolution of generative AI capabilities bring operational challenges and it necessitates solutions to address performance and scalability, data security and privacy concerns, custom and complex model development, and responsible AI.

Lenovo systems and platforms streamline creating data foundation to manage data for analytics and AI/ML, build pipelines and train AI model and simplify deployment and scale applications in on-premises infrastructure. The on-premises AI deployment provide more control on data privacy and security, governance, customization, ownership and replicate and reuse easily to other functions in the business.

The document provides the operational model for AI/ML workloads and provides deployment options with Lenovo platforms, performance measurements, considerations and sizing guidance, and some example deployment models. The last section contains detailed bill of material configurations.

1.1 Enterprise AI

Enterprise AI comprises of infrastructure, technologies, platforms, and tools to develop and deploy AI models and AI applications and services to meet the data driven business objectives. The impact of Artificial Intelligence is converged across business operations, support functions, industries, technologies, manufacturing and security and AI is mainstream for any business and its functions. Most of the enterprise applications rely on getting value out of AI to improve product quality, operations, customer experience and attain time to value. Enterprise applications are integrated and developed with different technology stack, and they are dependent for data generation and consumption across different sources and implementing AI enablement does require seamless infrastructure and technology options to scale existing applications and develop and train AI/ML models and integrate with enterprise systems. Democratization of AI brings governance and efficiency in data sharing, model development, training and experimentation and measure AI performance and impact. Enterprise AI strategy objectives are solving complex business wide problems, scale infrastructure and AI/ML models, adapt technology convergence, self-service and federated learning , responsible AI and explainable AI.

2 Business problem and business value

2.1 Business problem

- Get insights out of massive volumes of data across business and its functions.
- Adapt evolving AI technology landscape and complex models.
- Accelerate AI development and production deployment.
- Build inhouse capabilities and develop dedicated AI team to address customer and partner demand.
- Optimize cost and agile in adopting advanced AI technologies.
- Scale and secure AI infrastructure and data
- Data privacy, model accuracy, bias and ethics

2.2 Business value

- Scalable infrastructure options for edge and datacenter deployment
- Unified Infrastructure for analytics, AI/ML and enterprise applications.
- Secure on-premises data management capabilities and automation.
- Distributed computing and infrastructure (edge-core-cloud)
- Seamless enterprise AI/ML platforms and attain highest level of data privacy.
- Manage and isolate enterprise applications and AI/ML applications in common infrastructure.
- Heterogeneous architecture and platforms and compatibility
- Accelerate queries and model response and performance.
- Accelerate model development and production deployment.
- Integration with Hybrid cloud to deploy, integrate and scale fast.
- Compatibility and integration with existing systems and applications

3 Requirements

3.1 Functional requirements

Functional requirements define the system's capabilities for end-to-end AI/ML model development and deployment and interactions with users, and integration with existing systems.

Data privacy and security	Support to store data on-premises and no data is shared outside. Provide capabilities to manage access and control over data and infrastructure across AI/ML lifecycle.
Data Sources	Support for different data sources platforms and applications RDBMS, NoSQL, Object storage and Parquet
Data Storage	Support for protocols NFS/Block/Object storage
DevOps	Support for develop, deploy and manage enterprise applications
ML Frameworks	Support deploying different ML frameworks and runtimes like PyTorch, Tensorflow, Keras, SparkML and others
MLOPs	Support to develop, manage, deploy and maintenance of machine learning models and applications. Manage and prepare data for model training and analytics.
ML Algorithms	Support for develop, train and optimize machine learning and deep learning algorithms with different complexity levels
Experiment tracking	Pre trained models and custom Model development
Modeling Languages	Support for Python, R, Spark, and other languages
Manage multiple projects	Features to provision and manage multiple machine learning projects
AI Applications	Support for develop and deploy AI/ML applications
AI Governance	Full Visibility and transparency
Virtualization and containerization	Support for virtual machines and containers for different operating systems and hypervisors
Customization	Fully customizable to support different machine learning frameworks, tools and applications
Integration	Support integration with existing infrastructure, applications and tools
Access and Availability	Control over end-to-end infrastructure and manage downtime
Ownership	Internal
Replicability	Reuse and replicate models across different functions and groups in a business
Edge	Hardware designed to adapt edge and run data processing and inference efficiently
API Support	Support APIs to access infrastructure resources, models, and application integration

3.2 Non-functional requirements

Requirement	Description
Scalability	Solution components scale for growth of massive dataset and models of different complexity level
Load balancing	AI Workload is distributed evenly across servers and clusters
Fault tolerance	Single component error will not lead to whole system unavailability
Ease of installation	Reduced complexity for solution deployment
Ease of management and operations	Reduced complexity for solution management
Flexibility	Solution supports variable deployment methodologies and frameworks
Security	Solution provides means to secure customer data and infrastructure
High performance	Solution components are designed to support high-performance compute and parallel processing
Portability	Solution support for migrating data, models and applications developed from cloud and similar hardware platforms
Storage Efficiency	Support for compression, deduplication, in-memory, cache to reduce storage footprint
Power Efficiency	Achieve higher performance and energy efficiency.
Accelerated Analytics	Improve query and data processing performance for enormous volume of data

4 Architectural Overview

Figure 1 shows common AI/ML architecture for enterprise AI and application deployment with Lenovo ThinkSystem, ThinkAgile and ThinkEdge platforms. Lenovo servers and storages provide flexible and scalable deployment options spanning high performance computing, hyperconverged infrastructure and shared storage system to consolidate data science and AI/ML development and production deployment environments and AI applications across an enterprise.

Lenovo next generation infrastructure platforms with support for latest generation processors from Intel and AMD and accelerator from NVIDIA provide more gains in reliability, scalability, efficiency and manageability. With multiple options for enterprise AI infrastructure, it drives to adopt flexible design and architecture choices to accelerate AI/ML life cycle without compromising performance, security and compliance.

Lenovo Enterprise AI solutions with combined software stack for system management, virtualization, software defined solutions, private and hybrid cloud management, integration and support for AI/ML frameworks and tools eases infrastructure management and AI/ML life cycle operations.

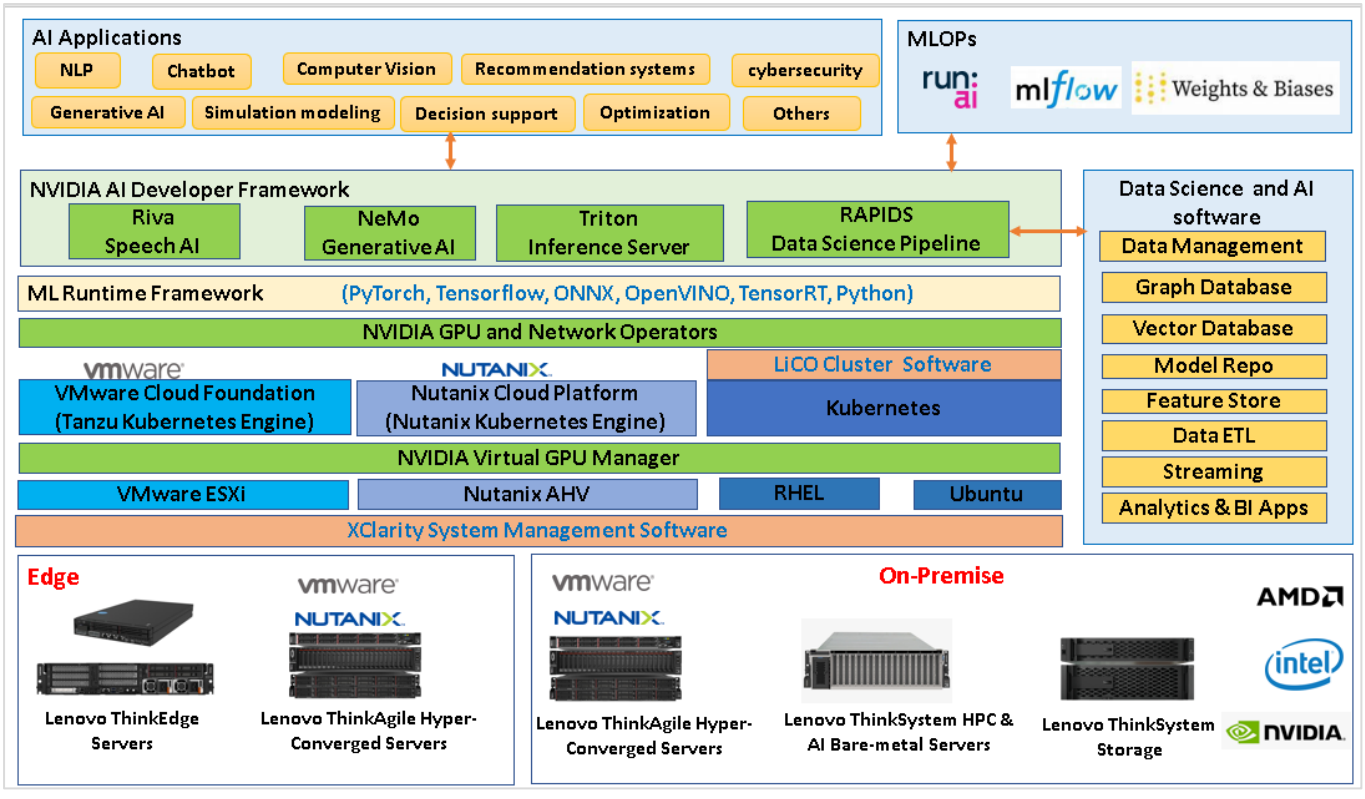


Figure 1: Lenovo Enterprise AI Infrastructure Solutions

5 Component Model

Figure 2 is a layered view of the Enterprise AI solution components comprises of hardware, software and applications.

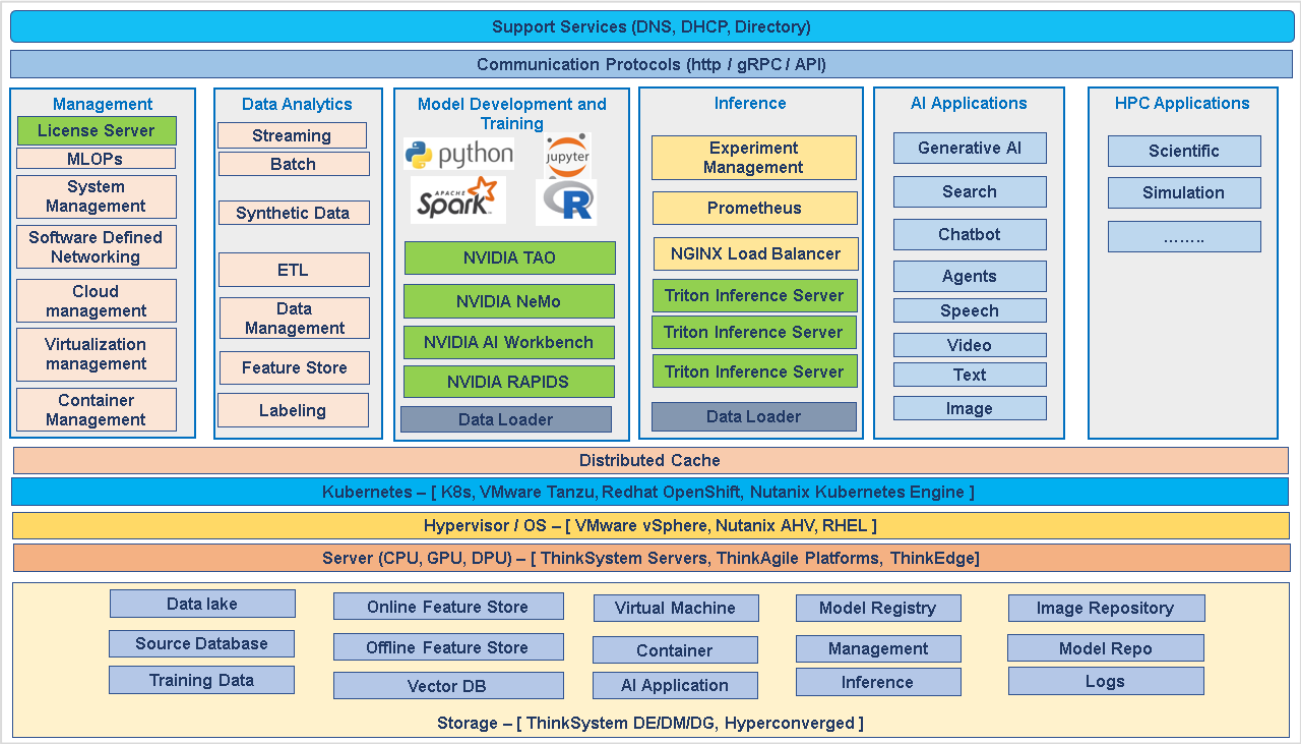


Figure 2: Component model for Enterprise AI Infrastructure

The following components sets foundation to build enterprise AI solution.

Compute Servers	Compute resources required to deploy and run virtual machines and containers for analytics and AI applications, models and enterprise applications. Lenovo provides the following options for servers. • ThinkSystem Servers with Intel and AMD CPU • ThinkAgile Hyperconverged Servers with Intel and AMD CPU - ThinkAgile VX (VMware vSAN) and ThinkAgile HX (Nutanix AHV) • ThinkEdge servers with Intel/AMD CPU • NVIDIA GPU • NVIDIA DPU • QUALCOMM Inference GPU

Storage	Storage is used to store data, virtual machines, containers and model repository. Lenovo storage solutions used for AI/ML are, • Shared Storage - ThinkSystem DE/DM/DG series • Hyperconverged Storage - VMware vSAN, Nutanix • Object Storage – ThinkSystem with Cloudian
Networking	Networking is used to connect across servers in the rack or datacenter. Lenovo servers supports 10GbE/25GbE/100GbE uplinks.
Virtualization and cloud platforms	Virtualization software runs on bare-metal servers and provide abstraction for server resources. Cloud management platform provides features to provision and operate virtual machines and contains. • VMware vCloud Foundation - ESXi • Nutanix Cloud Platform - Nutanix AHV • Lenovo Open Cloud Automation (LOC-A) – Bare-metal Linux
Container Platforms	Container platforms provides deploying and managing containers. Lenovo servers supports the following container platforms. • Kubernetes (Bare-metal Linux) • VMware Tanzu Kubernetes Engine (VMWare ESXi) • Nutanix Kubernetes Engine (Nutanix AHV) • Redhat OpenShift Container Engine (Bare-metal Linux)
Machine Learning Runtime	Machine learning runtime provides libraries and execution environment to run ML/DL models on a compute cluster with CPU and GPU. The major runtimes supported are PyTroch, TorchScript, TensorFlow, ONNX
AI Development Framework	AI development framework provides tools and libraries to develop, train and test machine learning and deep learning algorithms for different use cases. • Machine Learning Algorithms • Deep Learning • Natural Language Processing • Machine Translation • Automation • Computer Vision • Image Processing Speech Processing

MLOPs	Machine Learning operations (MLOPs) focuses on engineering and process to develop and production deployment and operationalize machine learning models. The following components are required for MLOPs and different software options are available, and a single component (software) can address multiple features. • Data Labeling • Data Visualization • Development IDE • Code Versioning • Feature Engineering • Model Training • Model Debugging and Validation • Model Tuning and HPO • Experiment Tracking • Model Packaging • Model Serving • Workflow Orchestration • Data Versioning • Model Registry • Feature Store • Model Monitoring

5.1 Compute Model

Lenovo servers, storage and networking portfolio provides diverse options to provide computing need for development, training, inference and deployment of machine learning and deep learning workloads.

Lenovo ThinkSystem servers are traditional systems which connect to other enterprise infrastructure, network switches and storage over network. These compute systems come with different form factor, local storage options and high-density GPU to accelerate and scale AI/ML workloads. These servers support bare-metal deployment and virtualization with VMware ESXi and Linux KVM hypervisors. ThinkSystem servers are suitable for wide range of HPC and/or AI applications which require more compute and large-scale AI deployment.

Lenovo ThinkAgile systems are hyperconverged servers comes with compute and software defined storage with high performance local SSD/NVMe drives, and this architecture provides simplified management, scalability, lower costs, flexibility, and cloud readiness. Lenovo ThinkAgile VX platforms are powered by VMware vSAN and vSphere and Lenovo ThinkAgile HX platforms are enabled with Nutanix AHV hypervisor. ThinkAgile servers are suitable for consolidating AI workloads, analytics and enterprise applications. These platforms come with private cloud and container management platforms to seamlessly support variety of workloads.

Lenovo ThinkEdge systems are designed to withstand harsh environmental requirement at edge locations and comes with accelerators for training and inference. These servers built with a smaller number of compute cores and local storage, accelerators, different network connectivity options and security hardening features for edge deployment.

5.1.1 CPU

Many data science and AI/ML workloads and software frameworks can be deployed on CPU and the latest generation processors with higher number of cores from Intel and AMD can address the performance requirement. Model serving and inference can run on CPU without compromising performance and latency. All AI/ML applications does require CPU for preprocessing and postprocessing in the training phase which utilizes GPU. ML training on CPU is not efficient and it takes longer duration for completion. CPUs are efficient for data preparation, feature extraction and small-scale models with minimal dataset and inference.

5.1.2 GPU

Machine Learning training and inference does require GPU to compute model efficiently. The complexity of model varies case to case and classical models require less GPU resources and deep learning and neural network models require large number of GPUs. Lenovo server platforms provide support to install GPUs and achieve low and high density.

5.2 Distributed Computing and Orchestration Frameworks

The Hight Performance Computing (HPC) and deep learning workloads requires massive compute and sharing and scaling compute for different AI/ML project need right workload management solution and automation to drive efficiency. The framework should support job scheduling, parallel processing, autoscaling, container support and upward integration. Enterprise AI adoption does require running and distributing AI/ML workloads on heterogeneous hardware and platforms. The framework should be interoperable across different ML frameworks as different functions in a business can adopt different technology stack for their AI/ML requirements. The scaling is applicable for data science applications, ingest and preprocessing, ML training, hyperparameter tuning, model serving and resource orchestration from the underlying hardware layer improve operational efficiency.

5.3 Lenovo Intelligent Computing Orchestration (LiCO)

The Lenovo HPC & AI Software Stack combines open-source with proprietary best-of-breed Supercomputing software to provide the most consumable open-source HPC software stack.

Lenovo Intelligent Computing Orchestration (LiCO) is a software solution that simplifies the use of clustered computing resources for Artificial Intelligence (AI) model development and training, and HPC workloads. LiCO interfaces with an open-source software orchestration stack, enabling the convergence of AI onto an HPC or Kubernetes-based cluster. LiCO enables a single cluster to be used for multiple AI workloads simultaneously, with multiple users accessing the available cluster resources at the same time. It provides web-based portal to

deploy, monitor and manage AI development and training jobs on a distributed cluster. LiCO comes with Lenovo Accelerated AI pre-defined training and inference templates for many common AI use cases.

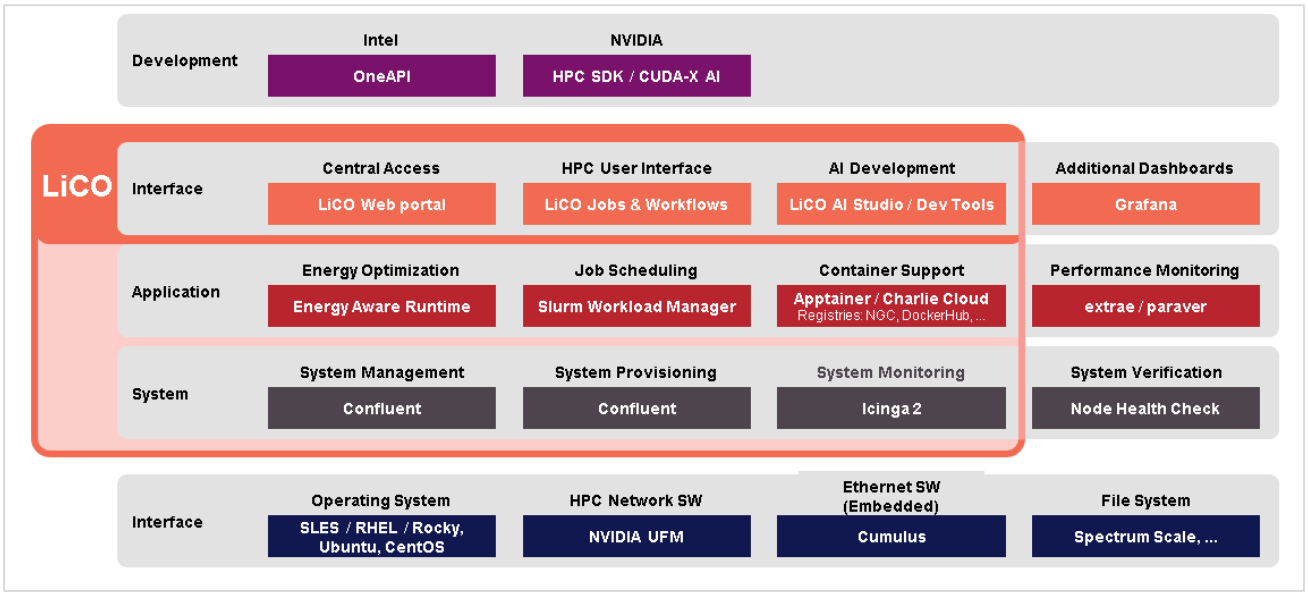


Figure 3: Lenovo Intelligent Computing Orchestration(LiCO) components

LiCO K8S/AI version provides docker-based containers and managed outside LiCO in the customer's docker repository and LiCO HPC/AI version containers are managed inside the LiCO interface and custom job submission templates defined as batch scripts (for SLURM, LSF, PBS) and also include HPC job templates.

It provides a fully tested and supported, complete but customizable HPC software stack to enable the administrators and users in optimally and environmentally sustainable utilizing their Lenovo Supercomputers.

The Lenovo HPC & AI Software Stack is built on the most widely adopted and maintained HPC community software for orchestration and management. It integrates third party components especially around programming environments and performance optimization to complement and enhance the capabilities, creating the organic umbrella in software and service to add value for our customers.

The key open-source components of the software stack are:

Lenovo Confluent Management

Confluent is Lenovo-developed open-source software designed to discover, provision, and manage HPC clusters and the nodes that comprise them. Confluent provides powerful tooling to deploy and update software and firmware to multiple nodes simultaneously, with simple and readable modern software syntax.

SLURM Orchestration

Slurm is integrated as an open source, flexible, and modern choice to manage complex workloads for faster processing and optimal utilization of the large-scale and specialized high-performance and AI resource capabilities needed per workload provided by Lenovo systems. Lenovo provides support in partnership with SchedMD.

LiCO Web portal

Lenovo Intelligent Computing Orchestration (LiCO) is a Lenovo-developed consolidated Graphical User Interface (GUI) for monitoring, managing, and using cluster resources. The web portal provides workflows for both AI and HPC, and supports multiple AI frameworks, including TensorFlow, Caffe, Neon, and MXNet, allowing you to leverage a single cluster for diverse workload requirements.

Energy Aware Runtime

EAR is a powerful European open-source energy management suite supporting anything from monitoring over power capping to live-optimization during the application runtime. Lenovo is collaborating with Barcelona Supercomputing Centre (BSC) and EAS4DC on the continuous development and support and offers three versions with differentiating capabilities.

5.4 NVIDIA AI Enterprise

NVIDIA AI Enterprise is an end-to-end, cloud native software platform that accelerates the data science pipeline and streamlines development and deployment of production-grade AI applications, including generative AI, computer vision, speech AI, and more. Enterprises that run their businesses on AI rely on the security, support, and stability provided by NVIDIA AI Enterprise to improve productivity of AI teams, reduce total cost of AI infrastructure, and ensure a smooth transition from pilot to production.

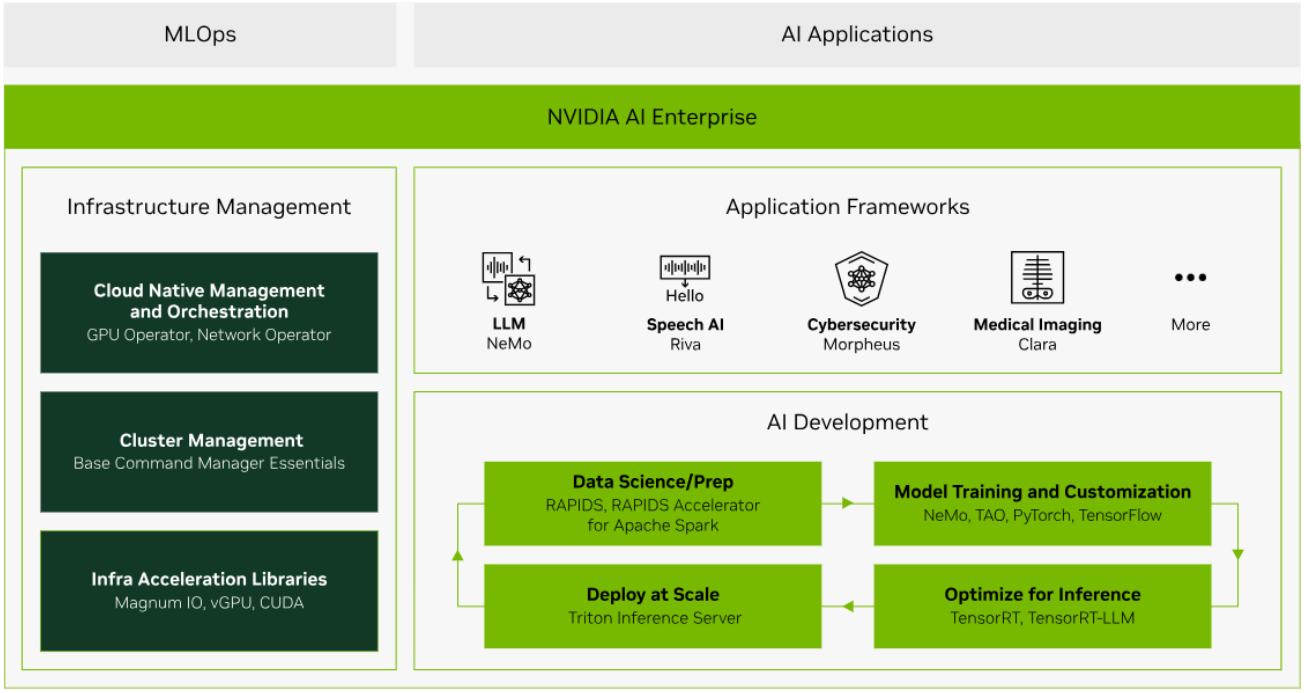


Figure 4: NVIDIA AI Enterprise Components

NVIDIA AI Enterprise includes the following components.

NVIDIA NeMo an end-to-end framework for building, customizing, and deploying enterprise-grade generative AI models; NeMo lets organizations easily customize pretrained foundation models from NVIDIA and select community models for domain-specific use cases.

NVIDIA RAPIDS is an open-source suite of GPU-accelerated data science and AI libraries with APIs that match the most popular open-source data tools. It accelerates performance by orders of magnitude at scale across data pipelines.

NVIDIA TAO Toolkit simplifies model creation, training, and optimization with TensorFlow and PyTorch and it enables creating custom, production-ready AI models by fine-tuning NVIDIA pretrained models and large training datasets.

NVIDIA TensorRT, an SDK for high-performance deep learning inference, includes a deep learning inference optimizer and runtime that delivers low latency and high throughput for inference applications. TensorRT is built on the NVIDIA CUDA parallel programming model, enables you to optimize inference using techniques such as quantization, layer and tensor fusion, kernel tuning, and others on NVIDIA GPUs.
https://developer.nvidia.com/tensorrt-getting-started

NVIDIA TensorRT-LLM is an open-source library that accelerates and optimizes inference performance of the latest large language models (LLMs). TensorRT-LLM wraps TensorRT's deep learning compiler and includes optimized kernels from FasterTransformer, pre- and post-processing, and multi-GPU and multi-node communication.
https://developer.nvidia.com/tensorrt

NVIDIA Triton Inference Server optimizes the deployment of AI models at scale and in production for both neural networks and tree-based models on GPUs.

NVIDIA AI Enterprise is certified on private cloud platforms VMware Cloud Foundation, Nutanix AHV, Red Hat Enterprise Linux and Ubuntu KVM and works seamlessly with Container orchestration VMware Tanzu, Red Hat OpenShift, Google Kubernetes Engine (GKE), Amazon Elastic Kubernetes, Service (EKS), Azure Kubernetes, Service (AKS), and upstream Kubernetes. It can be integrated with MLOps platforms such as ClearML, Domino Data Lab, Run:ai, UbiOps, and Weights & Biases.

5.5 VMware Private AI Platform

5.5.1 VMware Cloud Foundation

VMware Cloud Foundation (VCF) is a software defined solution to virtualize and manage compute, storage, and networking. Lenovo ThinkAgile VX servers with vSphere, vSAN, and NSX provide solid foundation to run any data science and AI workloads and applications. VMware Tanzu on top for vCloud Foundation provide container support and hybrid cloud deployment. Refer more details about the solution here Reference Design: VMware Cloud Foundation on Lenovo ThinkAgile VX

5.5.2 VMware Private AI Foundation

Lenovo ThinkAgile VX platform with VMware Cloud Foundation, Tanzu and NVIDIA AI Enterprise together provides complete solution with hardware and software stack to run enterprise AI workloads and Generative AI workloads with simplified management and build scalable deployment on on-premises and hybrid environment.

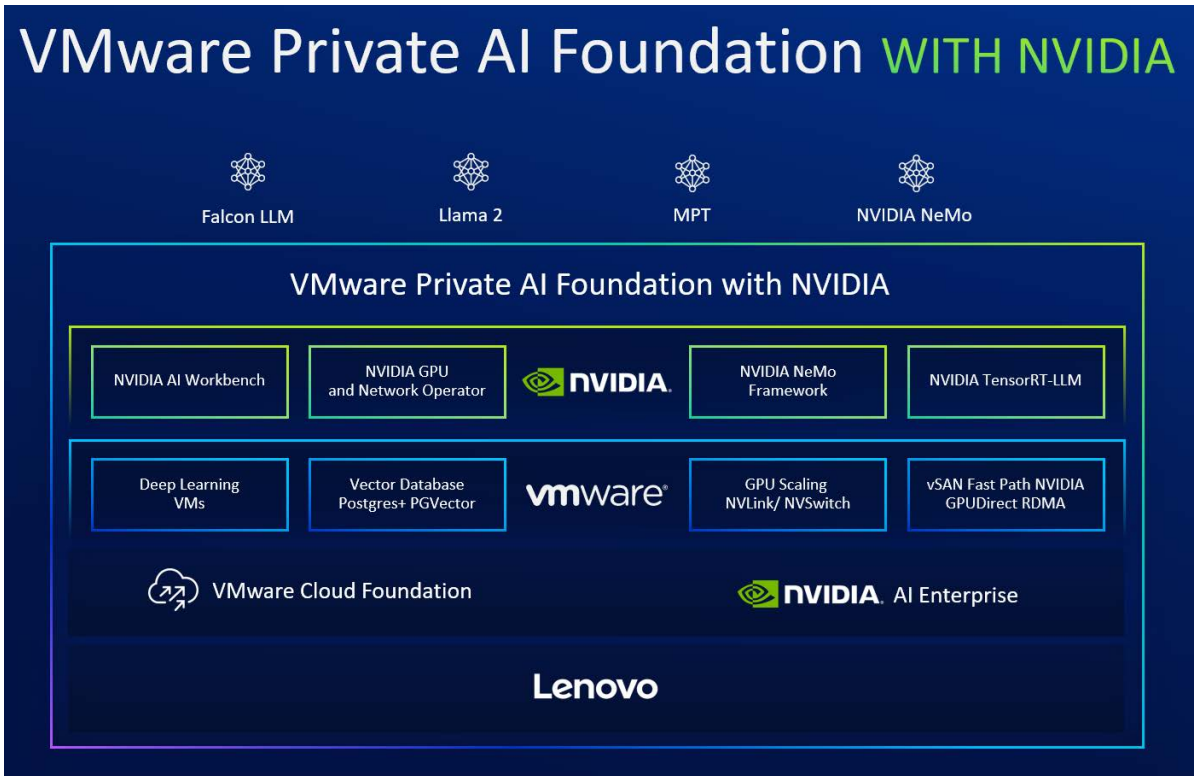


Figure 5: VMware Private AI Foundation

5.6 Nutanix Enterprise AI Platform

5.6.1 Nutanix Cloud Platform

Nutanix Cloud Platform (NCP) is a unified solution built on top of hyperconverged infrastructure to address hybrid cloud requirements and simplify operations for different workload scenarios. Lenovo ThinkAgile HX systems support NCP to provide reliable infrastructure as a foundation for a variety of use cases in hybrid cloud deployments. The platform comprises of the following key components which can be chosen based on the solution requirements:

Nutanix Cloud Infrastructure is a software defined infrastructure solution with compute, storage and networking for virtual machines and containers that can be deployed in private data centers on the hardware of your choice or in public clouds. The core platform is designed with AOS Storage, AHV, Karbon (Kubernetes Engine), Leap (Disaster Recovery) and Flow Network Security and Virtual Networking.

Nutanix Cloud Manager provides infrastructure management and operational support to build, manage and monitor deployments of virtual machines, containers and applications. It also delivers insights and automated remediation. The management stack includes Prism (Operations and Management), Calm (Self Service),

Beam (Cost Governance) and Security Central. Refer more information here
https://lenovopress.lenovo.com/lp0665-thinkagile-hx-series-reference-architecture

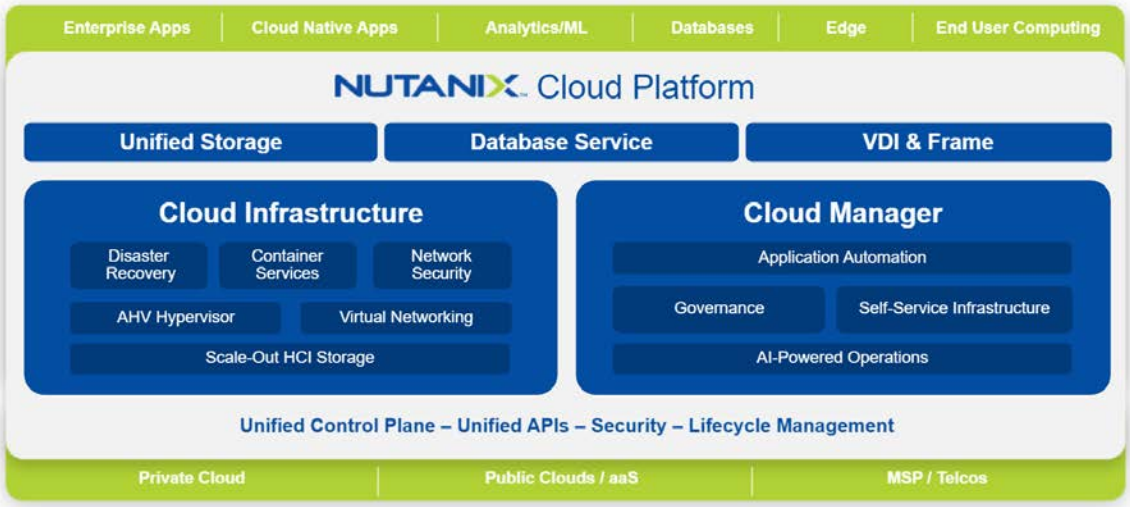


Figure 6: Nutanix Cloud Platform

5.6.2 Nutanix Enterprise AI

Lenovo ThinkAgile HX platform with Nutanix Cloud Platform and NVIDIA Enterprise AI provides unified solution for managing lifecycle of data science and AI/ML applications span across edge, core and public cloud infrastructure.

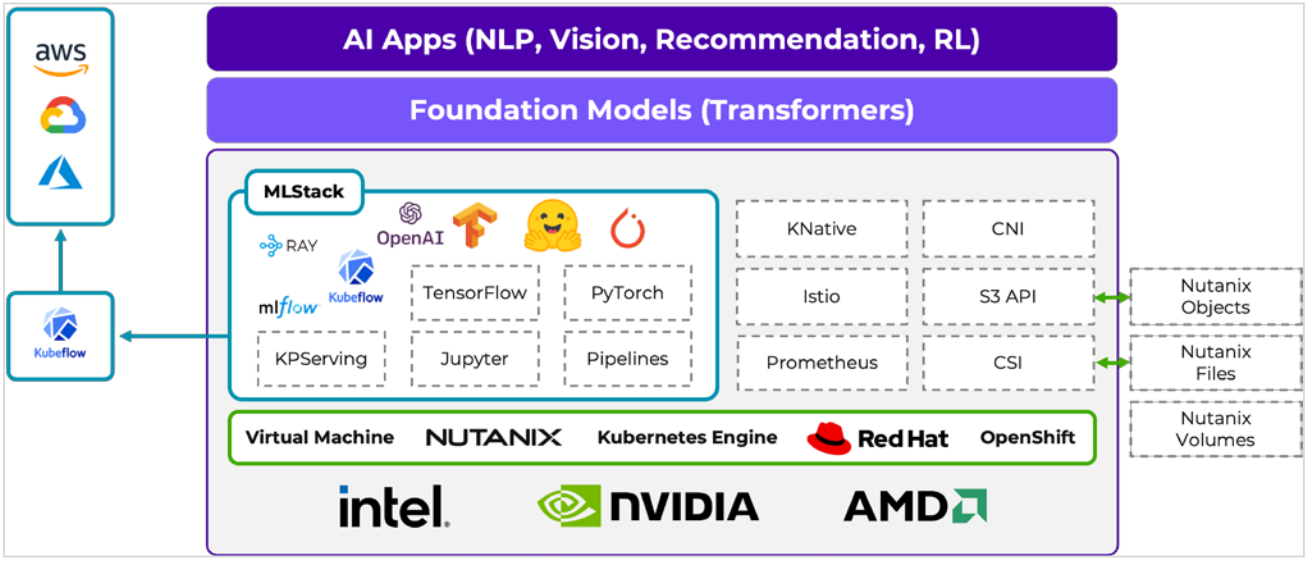


Figure 7: Nutanix Enterprise AI Solution

5.6.3 Nutanix GPT in a Box

Nutanix GPT-in-a-Box is an AI solution for large language models(LLM) and Generative AI in on-premises infrastructure with capabilities to control of their data and applications. With Lenovo ThinkAgile HX platforms, it provides everything needed to build AI-ready infrastructure, including,

- Nutanix Cloud Platform infrastructure on GPU-enabled server nodes
- Nutanix Files and Object storage for running and fine-tuning GPT models
- Open source software to deploy and run AI workloads, including PyTorch and Kubeflow
- Support for a curated set of LLMs (including Llama2, Falcon, and MPT)

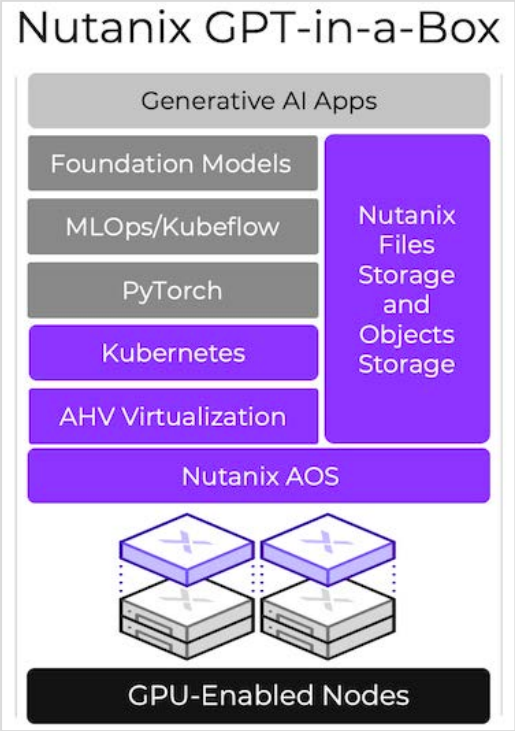


Figure 8: Nutanix Generative AI Platform

5.7 Storage Model

The AI applications and models uses different format and does require different storage class requirements and it can leverage block, NFS and object storages.

Table 1: AI/ML Storage requirements

Type	Description	Examples
Virtual machines	Virtual machines and containers for different applications does require storage for operating systems and store analytic data for processing locally	vmdk, vDisk
Images	virtual machine images and templates required to deploy different virtual machines and applications	OVA, snapshots, backups

Raw data	The data and documents used to train ML models. The data can be from any resources and can be in any format. The data are usually stored in data lake system to server as single source of truth.	documents, json, csv, web contents, audio, video, parquet, avro
Database	The source data for analytics and training can be from any of the RDBMS, NoSQL, Graph and VectorDB databases and feature stores require any database. All databases are stored within virtual disk	
Models	The trained models are stored in model repository. Pretrained models are available from public repositories	safetensors, pickle, bin, ONNX, HDF5
Notebooks	Development IDE and notebooks are stored in a common location to be used for MLOPs and sharing with other members in a team	.ipynb

5.7.1 Lenovo Shared Storage Systems

Lenovo ThinkSystem DE/DM/DG storages are scalable and performant to address enterprise-wide storage need for experimenting with different datasets for different AI/ML model architectures. ThinkSystem Storage can be scaled out or up to handle growing capacity needed to load new data and provide real time analytics and inference for AI applications. ThinkSystem storage is resilient and well suited for distributed training and provide optimal performance for various model tuning conditions. ThinkSystem storage support network-attached storage (NAS), storage-area networks (SAN) and enterprise can choose optimal configuration for different machine learning applications.

5.7.2 VMware vSAN

VMware vSAN is a hyperconverged storage and it supports two types of storage architectures leveraging latest hardware technologies and to provide scalable solution with different drive options to meet performance and capacity requirements.

vSAN Original Storage Architecture (OSA)

vSAN Original Storage Architecture comprises dedicated cache and capacity tiers and disk groups. All Flash configuration uses flash for both cache and capacity tier and hybrid configuration uses flash or NVMe drives for cache and HDDs for capacity tier. This architecture supports maximum 5 disk groups and maximum of 7 capacity drives per disk group. vSAN OSA does support different SSDs and HDDs type and sizes to create flexible configurations.

vSAN Express Storage Architecture (ESA)

vSAN Express Storage Architecture is a single tier storage solution and it is supported from vSAN 8 onwards. vSAN ESA requires high performance NVMe drives and leverages 25GbE/100GbE ethernet links to provide superior performance. It uses improved erasure coding which further reduces performance overhead and enables customers to achieve RAID 5/6 at the performance of RAID 1.

5.7.3 Nutanix AOS

AOS Storage simplifies storage for virtual and container environments by pooling the storage devices that are directly attached to a cluster of servers and presenting them to applications across a variety of storage protocols. Each node in a Nutanix cluster runs a VM called the Controller Virtual Machine (CVM) that runs the distributed storage services as well as other services necessary for a cluster environment.

AOS Storage absorbs incoming writes onto fast, low latency SSD tier. Data is then written to back-end storage resources asynchronously. This process ensures data exists in at least two independent locations in the cluster, making it fault-tolerant. AOS Storage is designed to automatically self-heal in the event of SSD, HDD/NVMe device, or node failures. It can also fully recover the management stack without requiring any user intervention or causing delays.

Intelligent Tiering provides automatic performance optimization on systems with multiple storage tiers (NVMe, SSD, HDD). Data is assigned to different tiers of storage within the storage pool using information lifecycle management (ILM) algorithms. Hot data is always preferentially targeted to the fastest storage tier.

5.7.4 Lenovo Object Storage powered by Cloudian

Lenovo Object Storage powered by Cloudian software defined storage platform provides a cost-effective, on-premises solution for large scale enterprise AI deployment to meet the need for capacity and performance requirements for data lakes and data analytics. Cloudian's native S3 API implementation offers the industry's highest level of S3 interoperability, letting you capitalize on the rapidly growing ecosystem of S3-enabled applications. With geo distribution architecture, data can be replicated across sites for better resilience and availability.

Lenovo Object Storage powered by Cloudian, is validated, and certified with leading data analytics platforms, including Greenplum, Teradata, Vertica, Apache Druid, Microsoft SQL Server 2022, Dremio, Splunk, Elastic, and Cribl, amongst others. The solution provides the flexibility, elasticity and scalability of the cloud infrastructure while allowing for separation of compute and storage scalability to modernize data analytics architectures on-premises. With a robust Data Lake solution within the security of your own firewall, customers can create data analytics solutions that comply with data privacy, residency and/or sovereignty regulations, all at a fraction of the total cost compared to traditional storage solutions. Refer more information here https://lenovopress.lenovo.com/lp1824-lenovo-object-storage-powered-by-cloudian

5.8 MLOps Framework

Machine learning operations include data preparation, model development, training, tuning, production deployment. MLOps framework is a collection of tools and process to achieve various aspects of continuous integration (CI)/continuous delivery (CD) capabilities and provide features such as experimental tracking, observability, and integration with software and infrastructure in the AI/ML environment. Kubeflow, ClearML and mlflow are referenced in this reference architecture and any compatible alternate solution can be easily fit in the solution.

6 AI/ML System Design

This section describes modular design approach for modern machine learning solution in on-premises to develop, integrate, maintain, and replicate across multiple teams in an enterprise. The design objective is to reuse most part of the ML pipeline components and infrastructure for across different projects in a business. This modular architecture works across Lenovo ThinkSystem and ThinkAgile infrastructure and partner solutions. Lenovo platforms provide adequate design assurance for verity of AI/ML applications developed on different technologies, platforms, and frameworks.

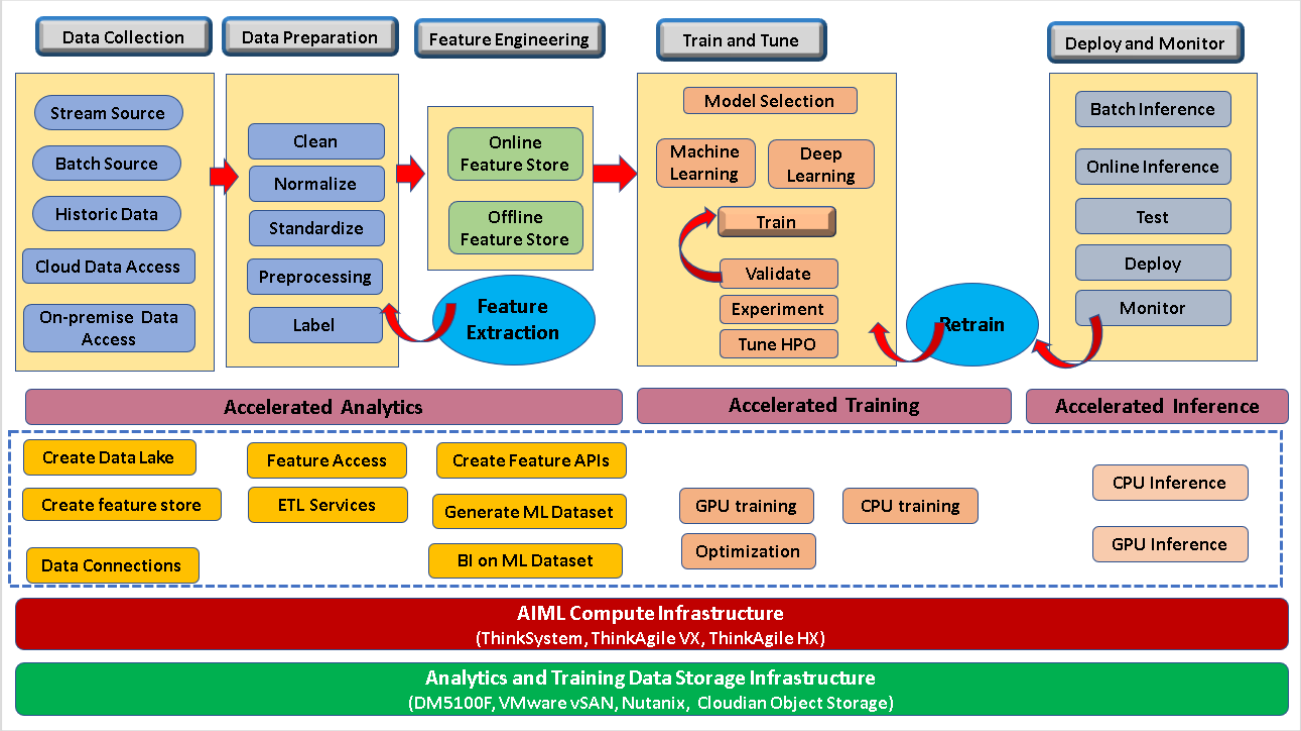


Figure 9: AI/ML System Design

6.1 Selection of Use case and AI Technology

Artificial intelligence is a computer capability to sense, learn, reason and engage from data and performing relevant tasks without being explicitly stated and build intelligence to act rationally and humanely. Artificial intelligence is applied in variety of technologies and domains and machine learning is core of it focused on developing algorithms leveraging data to learn and predict accurate results. Machine learning models caters to different use cases and enterprise AI and applications would leverage multiple models based on the requirements. The AI/ML models are applied in many technologies and applications such as data science, computer vision, natural language processing, robotics and many more. Artificial intelligence (AI) can be categorized into Artificial Narrow Intelligence (ANI), Artificial General intelligence (AGI) and Artificial Super Intelligence(ASI) based on the complexity of models and maturity of prediction and cognitive intelligence.

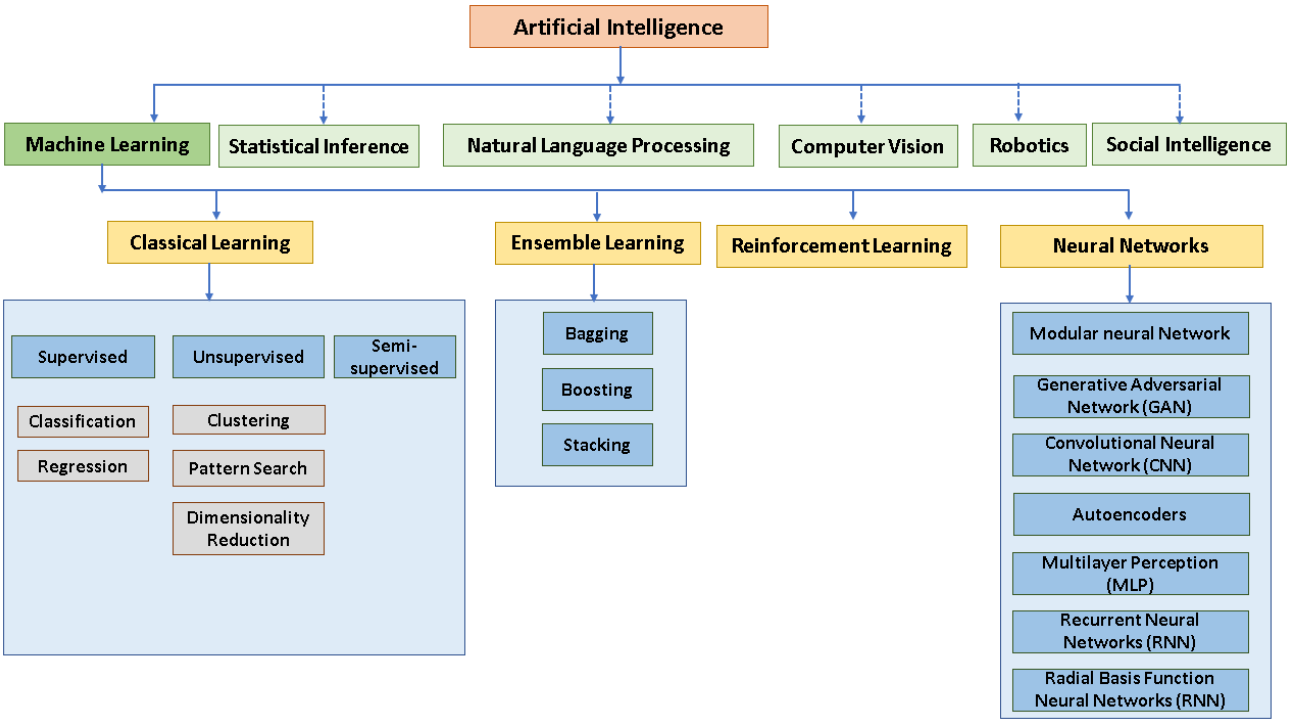


Figure 10: AI/ML Models Categories

6.2 Define Hybrid Data Access

Designing ML workloads in on-premises deployment requires robust storage solutions for different requirements as well as hybrid data access from different public cloud systems. Lenovo ThinkSystem storage solutions and ThinkAgile hyperconverged systems set strong foundation to create single source of truth of ML datasets and addresses different performance requirements for dated, new, cold, hot, and archival data. Lenovo hybrid cloud platforms powered by VMware Cloud Foundation and Nutanix Cloud Platforms provide features for connectivity to public clouds and migrate and use data for ML requirements.

6.3 Feature Engineering and Feature Store

The data from various sources in on-premises along with data generated from cloud are curated and used together to derive features for ML model training and serving for different use cases. Feature store is a centralized data storage with datasets used for ML training and serving and it is maintained as a separate asset and pipelines are created to do feature extracting and generating datasets for historical, batch and live data. Feature store and versioning helps to improve data quality by doing analysis and comparison with Business Intelligence (BI) tools.

Datasets can be stored in storage efficient manner either with parquet format or leveraging compression, deduplication and RAID5/6 features from underlying Lenovo storage systems or combination of both. Lenovo ThinkSystem and ThinkAgile platform storage solutions provide different classes of drives (HDD, SSD, NVMe) to build strong foundation infrastructure.

6.4 Selection of AI Platform and Framework

The AI/ML training and deployment can be managed in large scale using different components. Figure below shows different frameworks which can be adopted for HPC and/or AI/ML deployment and most of the components are interoperable with each other. Below are the solution options which can be adopted for Enterprise AI deployment with Lenovo server and storage platforms.

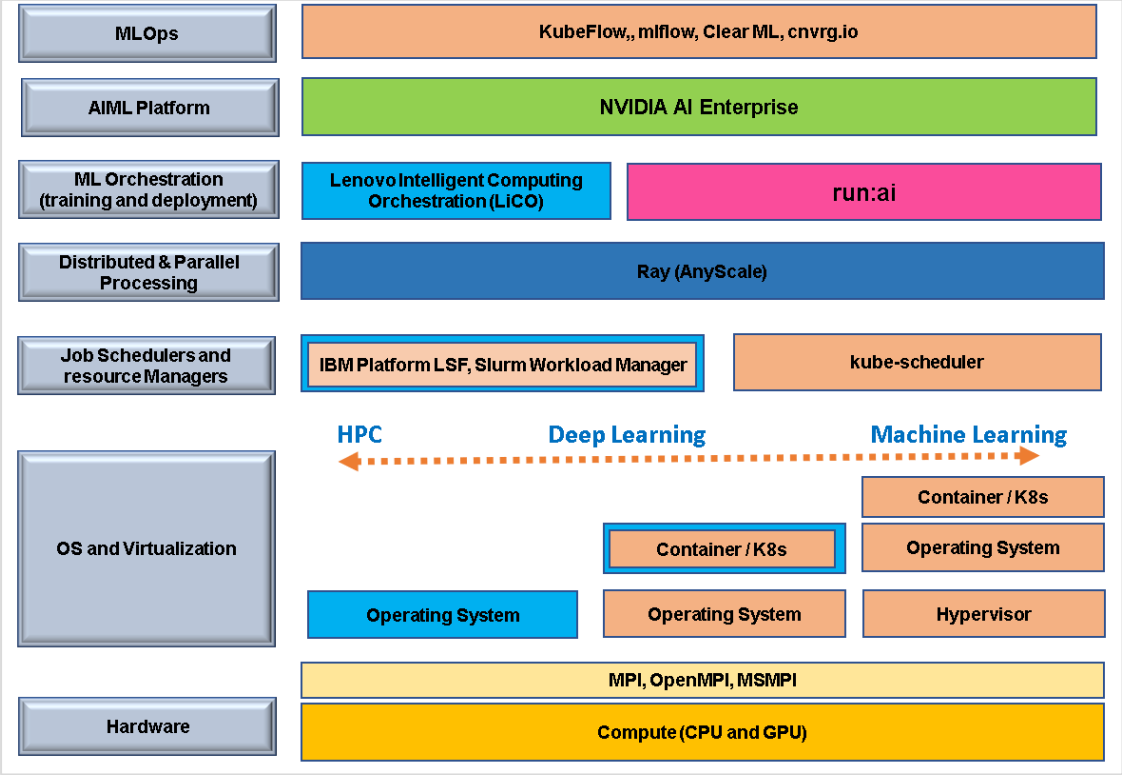


Figure 11: AI/ML Platform and framework

The deployment can be started with modular design using basic technology and components stack for individual AI use cases and then expanded with other software stack to get additional features for scaling and distributed computing. The AI infrastructure can be consolidated and managed across enterprise with combination of Lenovo platforms and software and Lenovo partner solutions and services. Table 1 shows different Lenovo platform choices and AI/ML frameworks can be used for enterprise AI deployment.

Table 2: Lenovo Enterprise AI Platform Options

Lenovo Platform	Standard Software Stack	Add on 1	Add on 2	Enterprise AI Consolidation
Bare metal AI/ML with Standard Kubernetes deployment	• Containers on Bare metal /Hypervisor (Lenovo ThinkSystem Servers) • K8s kube-scheduler	NVIDIA Enterprise AI	run:AI Ray	• HPC and AI/ML workloads together

	• Lenovo Intelligent Computing Platform (LiCO) for HPC/K8s • NVIDIA GPU libraries and Operator • MLOps platform			• Distributed computing across clusters • Heterogeneous platforms
Bare metal HPC & AI deployment	• Bare-metal or Bare-metal with container (Lenovo ThinkSystem Servers) • Resource managers (Slurm, LSF, PSB) • Lenovo Intelligent Computing Platform (LiCO) for HPC/AI • NVIDIA GPU libraries and Operator • MLOps platform	NVIDIA Enterprise AI	run:AI Ray	
VMware Hyperconverged AI Deployment	• ThinkAgile HX • VMware Tanzu container platform • VMware vCloud Foundation • VMware vSAN • NVIDIA GPU libraries and Operator • MLOps platform	NVIDIA Enterprise AI	run:AI Ray	
Nutanix Hyperconverged AI Deployment	• ThinkAgile HX • Nutanix Kubernetes Engine • Nutanix Cloud Platform • Nutanix AHV • NVIDIA GPU libraries and Operator • MLOps platform	NVIDIA Enterprise AI	Ray	

6.5 GPU Accelerated Analytics

The analytics workloads can be CPU, memory and IO intensive and data preparation for AI/ML takes more time in the life cycle. The larger the dataset, it requires more resources to clean and curate data for different ML project requirements. Analytics on GPU can improve data preparation speed and framework such as NVIDIA RAPIDS can leverage GPU capabilities to improve analytics performance and achieve real time AI objectives.

6.6 ML Model Development

ML models are developed on any of the programming languages python, C/C++ R, MATLAB, Java, Scala and it is required to evaluate libraries available for different algorithms and runtimes. ML runtime such as PyTorch, TensorFlow, Keras, Caffe provide tools, libraries and framework to train and inference ML/DL models across heterogeneous hardware. The performance of each language and framework varies for different algorithms and hardware, and it is recommended to consolidate different framework and use mostly used platform to avoid dual effort and model conversion challenges.

6.7 ML Algorithms and Model selection

Model developments are started with simple model and then expanded to complex models. The models experiment can be done with pretrained models and then based on the accuracy it can be retrained with custom data set. Pretrained models are trained with public datasets, and it may not work for all use cases. The custom model development takes considerable amount of time and training effort based on the selected model and training dataset and infrastructure. Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset. Table 1 shows the different ML methods and algorithms widely used for many use cases. The final model selection involves factors such as model accuracy, training time and cost and hyperparameter tuning efficiency. Ensuring the model performs best and can generalize well to new data, which is essential for real-world applications, and it has to be done case by case. Table 3 shows different ML model and algorithms commonly used and many projects require custom model development.

Table 3: AI/ML Models and Algorithms

Type	Method	Algorithm
Supervised (labeled data)	Classification	Decision Tree SVM kNN Logistic Regression Naive Bayes
	Regression	Linear Polynomial Lasso Ridge

Unsupervised (Unlabeled data)	Clustering	k-Means DBSCAN Fuzzy C-Means Mean Shift
	Pattern Search	Apriori ECLAT FP-Growth
	Dimensionality Reduction	PCA LDA SVD t-SNE QDA LLE LSA
Semi supervised (small labelled data)	Clustering	Clustering Anomaly Detection
Reinforcement Learning	Q-Learning	Deep Q-Network A3C Genetic Algorithm SARSA State-action-reward-state-action
Neural Networks	GAN	
	RNN	LSM LSTM GRU
	CNN	DCINN
	Transformers	
	Multilayer Perception (MLP)	
	Radial Basis Function Networks	
	Autoencoders	seq2seq
Ensemble models	Stacking	
	Boosting	
	Bagging	

NLP		NER (named entity recognizer) POS Tagged (a parts of speech tagger) Sentiment Analysis Information Extraction Information Abstraction Text Summarization Text Classification Spell Check language generation machine translation speech recognition Character recognition semantic parsing

Bias is an error resulted by simplifying the problem and leads to underfitting and variance is an error caused by excessive sensitivity to small delta in the training data, leading to overfitting. Low bias and low variance are the goals of a good model and finding a balance between bias and variance is known as the bias-variance trade-off and like regularization and ensemble methods helps to achieve it. A learning curve is a visual representation of how a model is performing and it can help identify if a model is overfitting, underfitting, high bias, or high variance.

The different types of model selection techniques in machine learning are -

- **Resampling Techniques** - Uses random sampling to generate multiple subsets of a given data set, which can be used to train and evaluate multiple models.
- **Probabilistic Techniques** - Used for selecting the finest model from a set of models by estimating the probability distributions of each model, given the data.

6.8 Batch and Online Models

The prediction and forecasting requirements vary from short term to long term and it is applicable for many use cases which does require real time prediction with live data. The training of a model can be done with historical data, batch data and live data and models tend to learn from new data. Also using larger dataset and parameters help models to identify more patterns and get more insights and learn to improve accuracy score.

Batch or offline learning is a traditional machine learning method, and it uses data accumulated over a period and the model performance is always based on historical data and model cannot learn incrementally. Training large batches of data requires also more computing resources and training frequency to push latest model to production which takes more effort and time.

Online machine learning systems receives data continuously and need to be able to adapt to rapidly changing conditions and learning takes place immediately as data become available.

6.9 Training

Training of ML model is a process of making ML algorithm to learn from training data. The model performance from the training will determine whether it can fit for use case. The quality of data and algorithm are important for the model performance and algorithm selection varies case by case.

Training data is written once and read bulk in nature and keeping history of training data is good practice for explainable and backtracking. Low-cost file, object or NAS storage is good fit for training dataset and however data loading and preprocessing across compute would benefit from high throughput and low latency drives.
Some algorithms require specialized data preparation in order to best expose the structure of the problem to the learning algorithm.

The table below shows different approach can be adopted to choose right model for the use case. The process involves validating model performance, optimization and algorithm efficiency and the objective to get better accuracy with less hardware. There are many methods available to compare models and it varies case by case. Table 4 provide general approach to choose right model and it varies based on scope and objectives like performance and infrastructure cost factors.

Table 4: AI/ML Model Selection methods

Task	Small Dataset	Large Dataset
Performance Estimation	• Repeated k-fold cross validation • confidence interval using 0.632 + Bootstrap	• Train/Test Split • Confidence interval using Normal approximation
Hyperparameter optimization	Repeated k-fold cross validation	3-way Train/Validation/Test Split
Model Selection	Repeated k-fold cross validation	3-way Train/Validation/Test Split
Model Algorithm comparison	• Nested Cross Validation • 5x2cv combined *F* test	• Multiple independent training and test datasets • McNemar Test • Cochran'sQ test • Friedman test

6.10 Distributed Training

Large datasets do require more resources for preprocessing and the distributed processing and training methods provide flexibility to choose heterogeneous hardware with different capacity and performance requirements. Many data processing tasks are CPU bound IO operations and large data can be split across compute instances with different CPU, memory and disk configurations to process and provide single output to training.

Data parallelism is more popular and frequently employed in large organizations for executing production-level deep learning algorithms.
Model parallelism, when a model does not fit on a single GPU, it has to be run on GPU across servers in the AI/ML infrastructure.

6.11 Scalable Inference

The end users and applications request inferencing to the deployed models through http, API and gRPC protocols. The inference server needs to scale based on the load from end users. The inference infrastructure should be reliable and performance to meet the varying load. The inference systems are optimized using the following techniques,

- TensorRT optimization
- F16 conversion
- INT8 Quantization
- Request Batch Size

The inference server can be scaled out to serve different models by distributing tasks on different CPU and GPU. Single inference server can server multiple models for small deployment and increasing batch size and grouping of requests drive maximum utilization of GPU and model performance.

6.12 ML Pipeline

Machine learning operations can be split into multiple pipelines to meet design and deployment requirements. It is recommended to have separate pipelines for training and deployment. The training part considers more iterations to find best fit and production deployment requires continuous monitoring, performance evaluation and feedback to training.

Each pipeline may take in the same raw training dataset and outputs a model that can be evaluated in the same manner but may require different or overlapping computational steps, such as:

- Data filtering.
- Data transformation.
- Feature selection.
- Feature engineering
- Model selection
- Model training
- Inference
- Post processing

6.13 Security

Appropriate security controls can be applied to protect models, data and infrastructure components in AI/ML lifecycle when multiple teams access and share common infrastructure. The cloud management and

Kubernetes platforms provide controls for resource isolation and sharing and access controls to manage multiple teams.

6.14 AI Observability and Responsible AI

AI value is measured by value for business implications and fairness, biases, accuracy and cost are factors commonly defines the success of AI adoption. The quality aspects should be considered from design to implementation and proper validation process should be followed from data collection, training dataset, algorithms and ML pipeline and monitoring model performance. AI/ML tasks and outputs should be explainable to meet use case objectives and validate properly.

- Right training data and right model
- Avoid data leakage
- Bias and errors
- Avoid and mitigate model drifts
- Fairness, model improvement & pipeline optimization

AI Observability and Explainable AI (XAI) tools and frameworks should be in place and ML pipeline should be optimal to push immediate changes or rollback.

7 Sample AI/ML Applications Solution Design

This section describes some of the AI use cases widely used and their corresponding system design to plan, and estimate required infrastructure. The system design can vary based on different components in the architecture and this section is focused on providing general reference to start with ML adoption.

7.1 Recommender Systems

Recommendation systems are prediction systems suggest relevant items and products for a user in e-commerce platforms based on various parameters and criteria such as search history, past purchases, rating and other factors. These platforms are designed to scale thousands of users and millions of products and big data systems to store data, events and the recommendation engine is a machine learning algorithm applies different filtering methods (content, user, collaborative, item) to personalize items and provide better search and user purchase experience.

Table 5: Recommender system functional and technical requirements

Category	Items	Example	Description
Capacity Estimation	Users	100k	Number of users accessing ecommerce platform
	Products	1000k	Number of products in the ecommerce platform
	Update Rate	5%	Daily change rate (price, description, stock,etc)
	Data Storage	200 TB	Storage requirements for database and services
	APIs	6+	Basic APIs required for recommendation systems (user, search, product, events, history, recommendation)
API Design	User ID	-	All APIs are required for an individual user who uses the platform.
Database	Products	-	NoSQL database for products
	Users	-	SQL/NoSQL database for users
	Events	-	NoSQL database for events
	Purchase History	-	NoSQL database for purchase history
ML Methods	Offline (Batch) Prediction	-	The recommendation engine predicts based on purchase history and runs as batch job daily for all users in the system to push recommendations.
	Online (Realtime)Prediction	-	The recommendation engine predicts online based on the current user activities in combination with history. It does run only when user logged in.
Architecture	Application Server	JBoss	Scale deployment based on number of users
	Load Balancer	HAProxy	Scale deployment based on number of users
	Database	Cassandra	Scale deployment based on number of users

	Recommendation Cache	Redis	Scale deployment based on number of users
	User Activity Queue	Kafka	Scale deployment based on number of users
	Recommendation Engine	Python	Scale deployment based on number of users

The recommender system can be implemented as wither batch or real time predictions. The figure below shows end-end deployment for recommender systems and different application and AI/ML services.

Batch(offline) recommendation system uses historical data of user activity and pushes predictions to the recommendation cache system on daily basis using the any of the filtering methods. In this approach, the ML model runs as batch for all the users and the predictions are pushed irrespective of user is logged in or not. It becomes expensive as the model runs for all the users and the recommendations may not be based on the live user search activities.

Realtime (online) recommendation system uses embedding models and retrieval and ranking algorithms to recommend most relevant matching items based on the current user session. The vector database contains embedding of user browsing history and recommendations from the batch system. When the user searches or clicks for any item, it gets searched in the vector database to retrieve hundreds or thousands of matches and then applied with ranking to return top relevant items. The search uses nearest neighbor indexing or geospatial indexing to find similarities. The advantage of the real time recommendation system is it runs only when the user is logged in and this reduces more compute but still it requires additional compute for search and other operations.

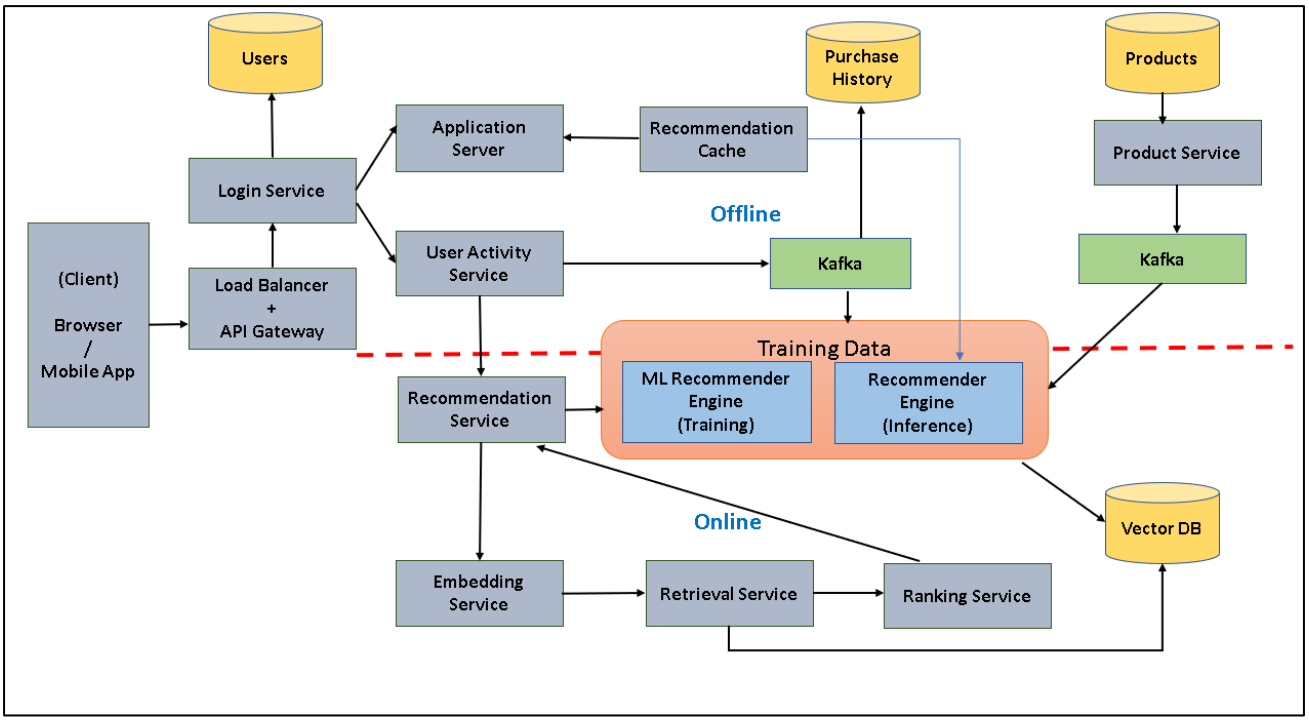


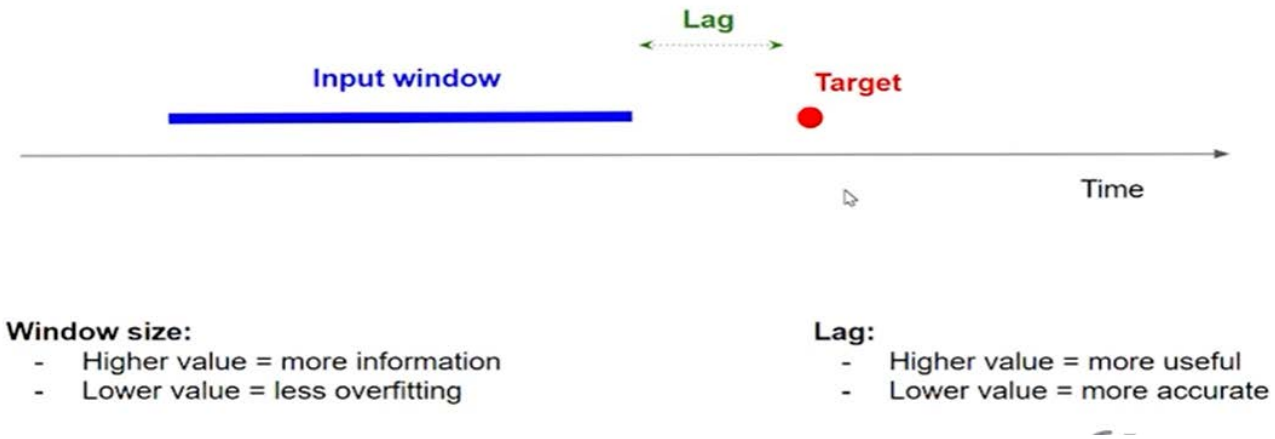
Figure 12: Recommender System Design

Refer documentation for NVIDIA Merlin, an open source framework to build and scale recommender systems https://developer.nvidia.com/merlin

7.2 Timeseries Forecasting

Forecasting systems predict future observations of a target based on information from past and it is classified under multi variant regression problem. This is common model used for weather forecasting, price forecasting, sales forecasting, energy forecasting, statistical and economical demand predictions. The forecasting models is based on time series data and one or more target variable to predict and it is combined with domain specific structured and unstructured data. The larger the number of target variables, the complexity of model increases.

Classification prediction is another time series problem and in which time is either discreate or continuous and there is no target variable input and prediction is always categorical.



There are many time series forecasting algorithms are available and the table below shows some of them. It is recommended to benchmark more than one models for the use case chosen.

Model Category	Example Model
Classical Time Series Model	ARIMA family
Supervised Learning Model	Linear Regression, Random Forest, XGBoost
Deep Learning	LSTM

https://neptune.ai/blog/select-model-for-time-series-prediction-task

Choosing right forecasting accuracy methods is complex and each method would not provide same value for different sets of data. The methods widely used are Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Some methods can work better for historical data and some for predicted data and data scientist should consider bias and outlier objectives to define KPIs for accuracy.

The following example shows energy demand forecasting for a power grid solution. The forecasting helps to predict demand which will help to plan power generation, tariffs and plan maintenance of power distribution systems. This is multi variant model and it considers weather, seasonal data (holiday, festival, etc) and the time series data from different substation is collected to through SCADA. This is sample approach, and it can vary case by case.

Table 6: Timeseries system functional and technical requirements

Category	Items	Example	Description
Capacity Estimation	Input Window Size	3/6/12/14 months	The amount/length of historical data used as base for forecasting
	Target		The future time which needs to be considered for prediction.
	Lag	Hourly, Daily, weekly, Monthly, yearly	The difference between last data point in the input window and future time
	Step Length	Hourly, daily, monthly	The interval to do continuous prediction and it is the distance between windows
API Design	Substation ID	-	All APIs are required for an individual Substation which requires demand forecasting.
Forecast Methods	Long Term	Monthly Quarterly Yearly	
	Short Term (Live)	hourly	
Database	Consumption	-	Time series database for power consumption metrics at substation level
	Weather	-	NoSQL database for weather
	Seasonal	-	NoSQL database for holidays and seasons
	Power Generation	-	Time series database for power source
ML Methods	ARIMA	-	
	Linear Regression		
	LSTM	-	
Architecture	Application Server (IOT)	JBoss	Scale deployment based on number of substations
	Database	Cassandra	Scale deployment based on number of users
	Analytics/BI Server	Tableau	Reports for different forecasting methods
	Energy Consumption Meter Queue	Kafka	Scale deployment based on number of users
	Forecasting services	Python	Scale deployment based on number of targets
	Plant Operation Services	Python/Java	Microservices for plant maintenance and operations

The figure below shows high level enterprise architecture components for power generation and distribution systems with AI forecasting services. Accurate energy prediction is primary objective on ESG and two methods are adopted in the industry.

Short-term load forecast (STLF) is the time-period of STLF lasts for a few minutes or hours to one-day ahead or a week. STLF aims at economic dispatch and optimal generator unit commitment while addressing real-time control and security assessment. The energy forecasting is done at substation level and then aggregated to predict total demand and power generation is planned accordingly. Due to growth of electric vehicles, solar power generation and it is necessary to predict demand and plan the distribution for residential and commercial operations.
Long-term load forecast (LTLF) is ranges from a few years up to 10 years ahead to plan for expansion of generation, transmission, and distribution.

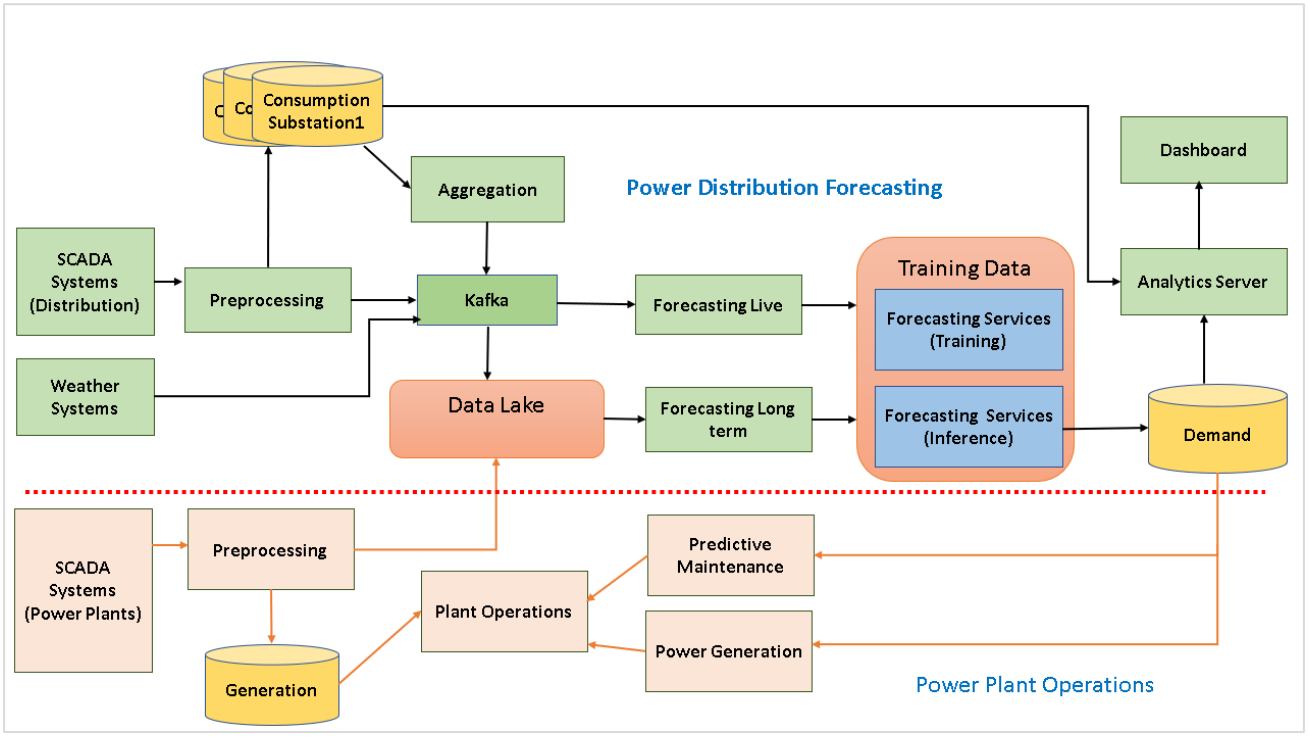


Figure 13: Timeseries forecasting system design

7.3 Conversational AI

Conversational AI combines NLP and machine learning to build systems to understand human language (text, audio, video) and process it and provide different services such as translation, spell check, content analysis, chatbot, question and answer, search, sentiment analysis, etc. The machine learning models processes large volumes of documents, web contents and videos and extract insights across different languages. Natural Language Processing involves preprocessing steps tokenization, stemming, lemmatization, normalization and vector embeddings.

The conversational AI applications provide semantic search, support for text and voice commands, context aware dialogue-based conversation than the classical chatbot which supports text conversation and predetermined rule-based workflows and navigations.

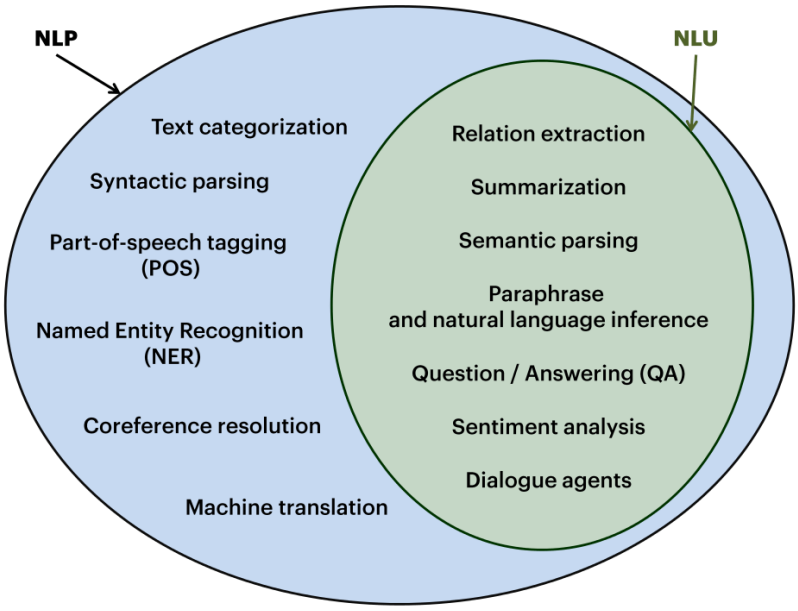


Figure 14: Natural Language Processing (NLP)

The example below describes developing conversational AI chatbot and semantic search. The chatbot can retrieve relevant answers from documents, knowledge bases, application and database systems in the enterprise. The solution includes search with features for summarization and classification. The number of chatbot agents can be developed separately for each use case. The chatbot system can connect with any other applications and services in the enterprise to answer relevant questions.

Table 7: Conversational AI functional and technical requirements

Category	Items	Example	Description
Capacity Estimation	Number of Chatbot Users	1000	
	Number of Intents	4000	
	No of Dialog Managers	2	
	Dialog Agents per bot engine	1	
	Semantic Search Volume	1000/day	
API Design	Search	-	
	Bot		
Database	Knowledge Base	MongoDB	
	Chat History	MongoDB	NoSQL database for weather

	Intents	MongoDB	
ML Methods	RASA	-	NLP framework
	Spacy	-	NLU services
	Flan-T5	-	Transformer model for summarization
Architecture	Bot Server	RASA	Scale deployment based on number of domains/users
	Database	MongoDB	Scale deployment based on volume
	Bot Agent		Bot agent to connect with chat clients
	Semantic Search	Python	Text summarization service with T5
	NLP models	Python	RASA NLP services

The figure below shows the services required for conversational AI and knowledge base with semantic search. The system uses RASA NLP framework and Spacy NLU model along with Flan-T5 transformer model for summarization and classification. The chat history is stored and model is retrained with chat history.

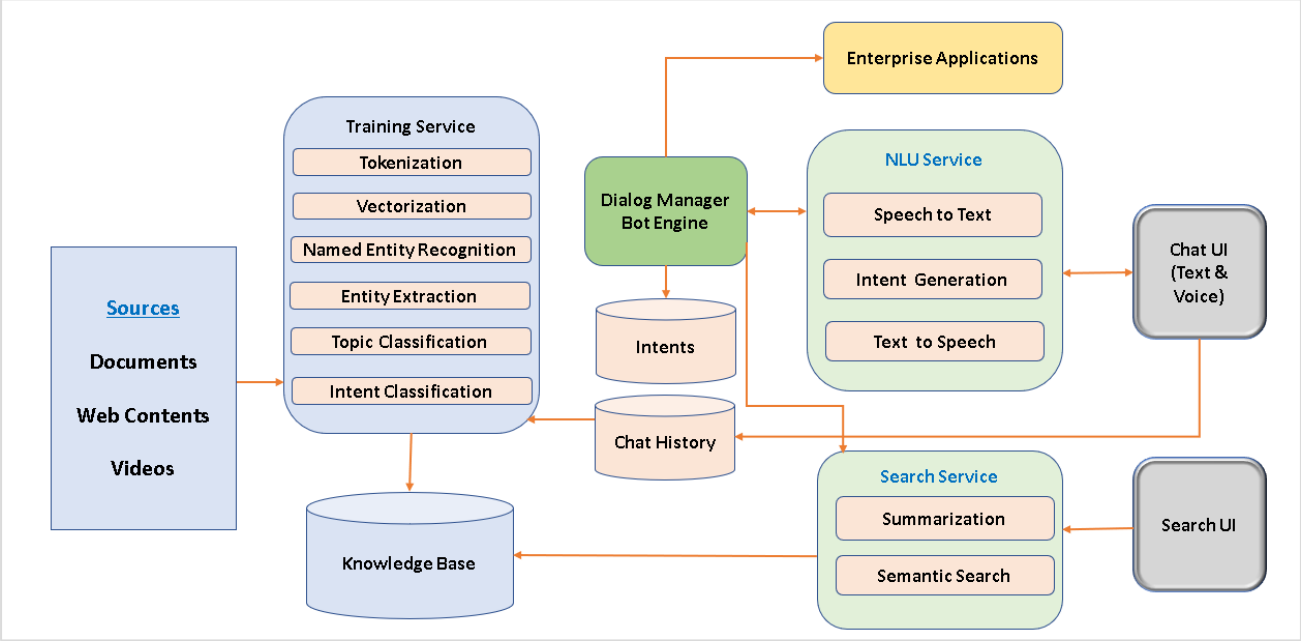


Figure 15: Conversational AI system design

7.4 Generative AI (Large Language Models)

The generative AI applications are powered with Foundation models (Large Language Models) to create text, images, code, audio, and videos. The LLMs are large models and trained with vast number of contents from web and synthetic datasets. The following are the use cases for generative AI.

- Content generation (text, image, audio, video)
- Summarization
- Question and answer

- Chatbots and virtual assistants
- Code (algorithm, deployment, testing)
- Data analytics (reports, statistical analysis)
- Design (3D models, prototype)

Refer https://lenovopress.lenovo.com/lp1798-reference-architecture-for-generative-ai-based-on-large-language-models for more details

The foundational models are alternate solution for classical NLP models and frameworks. LLM do not require intents and navigations and it is context aware and can provide relevant answers and summary and the models can be replicated to different domains with minimal effort. There are many pretrained opensource LLMs are available and training with custom dataset requires more infrastructure. The LLMs can be integrated with other AI/ML applications to provide better insights and analytics experience. The LLM models are categorized into text, image and video use cases and multimodal systems which combines different data formats and it is more complex to develop and train.

The LLMs use vector embedding and retrieval augmentation to convert source data and user intents. LLMs are context aware, and any user prompt is translated and provides relevant information.

Table 8: Generative AI functional and technical requirements

Category	Items	Example	Description
Capacity Estimation	Number of Chatbot Users	1000	
	No of Dialog Managers	2	
	Dialog Agents per bot engine	1	
	Semantic Search Volume	1000/day	
	Summarization	1000/day	
API Design	Search	-	
	Bot		
Database	Knowledge Base	MongoDB	
	Chat History	MongoDB	NoSQL database for weather
	Intents	MongoDB	
ML Methods		-	NLP framework
		-	NLU services
		-	Transformer model for summarization
Architecture	Bot Server	RASA	Scale deployment based on number of domains/users
	Vector Store	pgvector Postgress	Vector database to store embeddings

	Bot Agent	-	Bot agent to connect with chat clients
	Search API	python	Search service
	LLM models	Llama2	Appropriate models for summarization and chat
	Document Loaders	LangChain, LlamaIndex	To load, split and chunk documents

The figure shows conversation AI and search solution shown in section 7.4 designed using generative AI models. The user prompts are converted to embeddings and matching information is retrieved from the vector store and then passed to large language models to generate response for summarization or Q & A or chat conversations. The LLMs may not return same response for same query on every iteration and prompts should be given in clearly to get concise results without bias.

The LLMs are deep learning models with millions of parameters, and it does require high end infrastructure for training and inference. Private LLM project can be started with pretrained models and trained for domain specific data. The generative AI eliminates the need for intent creation and the models can be tuned with RLHF by using hundreds of prompts and full tuning require massive structures and unstructured data and labeled data provide more quality and performance than unlabeled data.

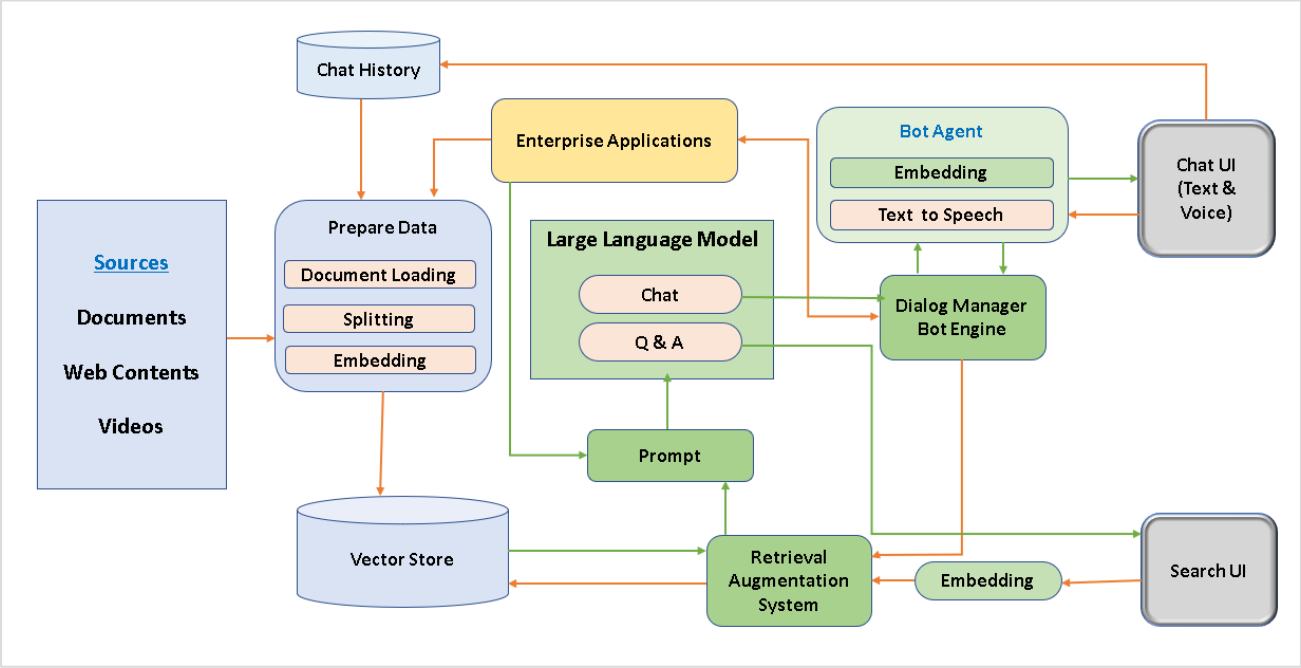


Figure 16: Generative AI system design

7.5 Cognitive Services

Artificial intelligence span across different verticals and industries and it does require machine learning models specific to different requirements. Developing and deploying new machine learning solution takes considerable time. Cognitive solutions are AI ready solutions tested on different data sets and real time use cases and it reduces complexity, expertise requirements and drives quicker deployment and implement AI solutions faster. Lenovo partnered with different cognitive solution providers to provide combined expertise on technology and

infrastructure to bring innovation and deliver scalable solution for any AI requirements. Table 9 lists some of Lenovo partners and their cognitive solutions for different AI use cases.

Table 9: Lenovo AI Innovation Partners and cognitive services

Lenovo AI Innovation Partner	Solution Details	Link
Cnvrg.io	Blueprints financial, media, social, medical and manufacturing industries and platform for data collection, inference and monitoring	https://cnvrg.io/blueprints/
byteLAKE	Provide pretrained models for Computational Fluid Dynamics and cognitive services for many use cases	https://www.bytelake.com/
Everseen	Visual AI platform is an end-end solution for data engineering, business process mapping and AI	https://everseen.com/
Infinia ML	Machine learning solution for document processing	https://infiniaml.com/
VISTRY	Machine learning solution for customer support	https://www.vistry.ai/
Deepbrain.ai	Conversational AI platforms	https://www.deepbrain.io/
GUISE AI	Predictive maintenance for oil & gas	https://guise.ai/
WaitTime	Real time AI powered crowd intelligence platform	https://www.thewaittimes.com/
AiFi	Autonomous retail solutions	https://aifi.com/
Signatrix	Visual intelligent platform for retail	https://www.signatrix.com/
Fingerprint	AI powered restaurant experience services	https://fingermark.ai/
GRAYMATICS	Cognitive video analytics platform	https://www.graymatics.com/
Pathr.ai	Spatial analytics platform and solutions	https://pathr.ai/
V7	Data labeling solution across industries	https://www.v7labs.com/
HU/EX	Cognitive solution for voice analytics	https://www.huex.ai/
AlwaysAI	Computer vision solutions and platform for data management, model training and deployment	https://alwaysai.co/
Chooch	Computer vision solutions for different industries	https://www.chooch.com/
InstaDeep	Decision making system for logistics, energy, biology and electronic design	https://www.instadeep.com/
PEAK technologies	Image recognition and package intelligence for logistics	https://www.peaktech.com/
VSBLTY	Computer vision and image recognition for retail and security	https://vsblty.net/

8 Operational model

This section describes the options for mapping the logical components of for enterprise AI solution onto hardware and software. The “Operational model scenarios” section gives an overview of the available mappings and has pointers into the other sections for the related hardware. Each subsection contains recommendations on how to size for that particular hardware, and a pointer to the BOM configurations that are described in section 10 on page 59. The last part of this section contains some deployment models for example customer scenarios.

8.1 Operational model scenarios

Figure 17 shows the operational models (solutions) in Lenovo Enterprise AI solution for enterprise and small-medium business (SMB).

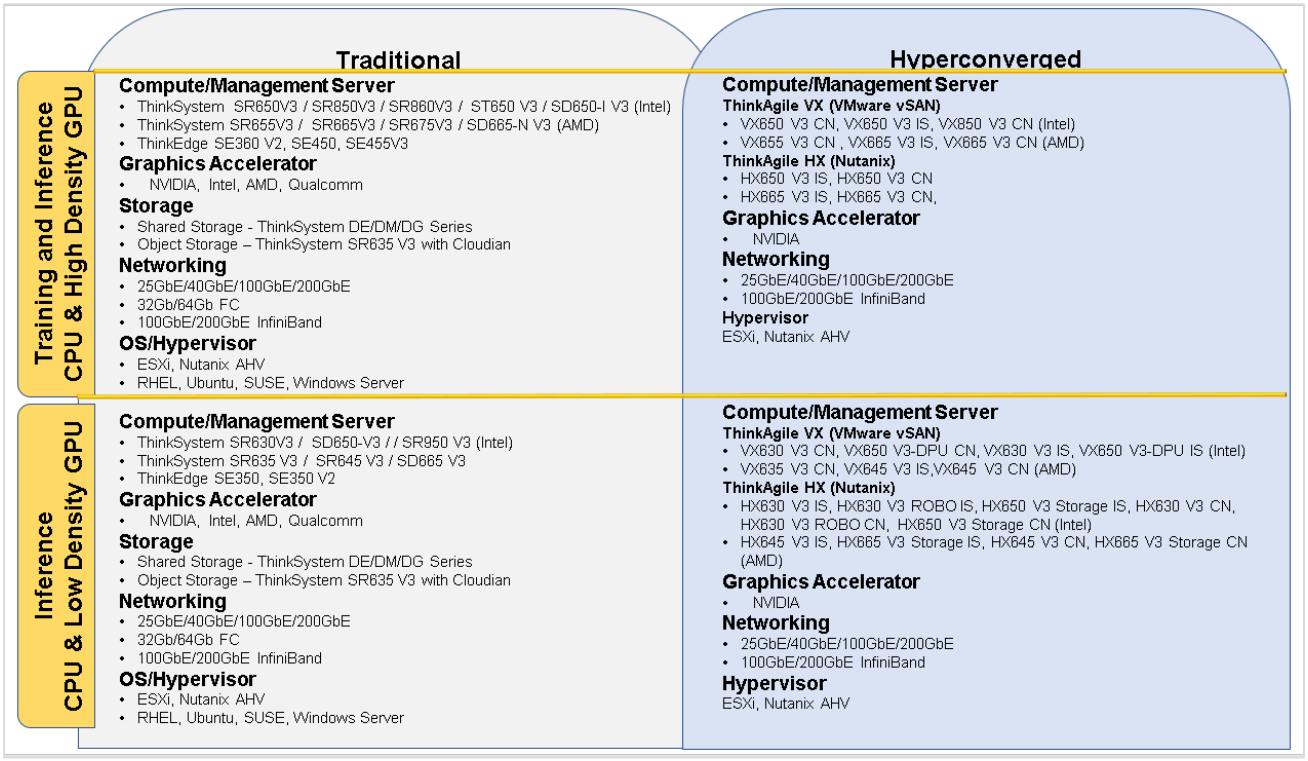


Figure 17: AI/ML solution operational model

The vertical axis is split into two categories of AI/ML workload compute requirements and many model inference can be performed on CPU and with low-end GPUs and accelerators. Machine learning training requires GPUs and inference for large model are addressed with GPU. CPU is required for both training and inference and the lower half can fit for SMB requirements and upper half for Enterprise requirements.
The horizontal axis is split into traditional and hyperconverged infrastructure solutions offered from Lenovo. The traditional systems are with rack-based systems, edge servers and shared storage and software defined object storage and the hyperconverged systems are with respective hypervisor and software defined storage software. Refer list of Lenovo systems with GPU support here https://lenovopress.lenovo.com/lp0768-thinksystem-thinkagile-gpu-summary

The enterprise AI architecture includes many analytics functions and support services for AI/ML deployment and hardware platforms can be mixed in such a way to consolidate and isolate applications appropriately without impacting performance. The number of models for different AI/ML requirements in a business varies from tens to hundreds range and each model requires its own data preparation, model development, training and deployment and this reference architecture covers different enterprise AI infrastructure solutions and software stack can be deployed on different Lenovo platforms and business can choose flexibly based on the AI/ML needs. Table 10 shows an approximate estimate to size required infrastructure for different business wide need for AI deployment.

Table 10: AI Models for different business sizes

Type of Business Size	Total models	Dataset Size (Total)	Number of Machine Learning models	Number of Deep Learning Models	Number of Generative AI Applications
Small	Up to 10	<100 TB	8	2	2
Medium	10-30	100-500TB	25	6	8
Enterprise	30+	1PB +	30+	10+	5+

8.2 Example Enterprise AI Infrastructure Solution

This section provides compute sizing for software and applications for a Enterprise AI requirements with 40 models of different sizes. This example is considered as base to provide sizing for different Lenovo ThinkSystem and ThinkAgile platform solutions described in the following sections.

Table 11 provides software components and resource requirements for data science and preparation phase in machine learning life cycle. Data preparation is critical task and the data sources can be any and this example considers streaming to process data coming from different data sources and process it to store features in different databases for different models.

Table 11: Data science and Machine Learning Framework System Configuration

Management service	Virtual processors	System memory	Storage	OS	GPU	HA needed	Workload Characteristics
Kafka Broker Cluster (2)	8 VCPU	32 GB	300 GB	Linux	N/A	Yes	Inventory events 1 million per day

Management service	Virtual processors	System memory	Storage	OS	GPU	HA needed	Workload Characteristics
Kafka Connector Cluster (2)	6 CPU	32 GB	200 GB	Linux	N/A	Yes	Inventory and product database
PostgreSQL	16 VCPU	128 GB	2000 GB	Linux	N/A	Yes	100k new product records per day
MongoDB	24 VCPU	128 GB	3000 GB	Linux	N/A	Yes	Training Data
Hopsworks Feature Store	16 VCPU	96 GB	300 GB	Linux	N/A	Yes	Feature store for model training and serving
KubeFlow	8 VCPU	24 GB	300 GB	Linux	N/A	Yes	MLOps

Compute for Machine Learning Training Cluster Workload

The number of AI/ML models in an enterprise varies for different use cases and each model also requires different CPU and GPU requirements based on the model algorithm and dataset complexity. Table 12 provides t-shirt sizing for ML training with NVIDIA multi-instance GPU (MIG) profile which partitions single H100 to 7 instances. The table below shows t-shirt sizing for training environment for 40 models of different sizes. The same cluster can be used for HPC workloads which is not sized in this reference architecture. The ML training can be done in distributed manner by running models across multiple nodes.

Table 12: Machine Learning Training Applications System Configuration

Model Type (count)	Virtual processors	System memory	Storage	GPU Profile	Workload Characteristics
Small (16)	8 VCPU	32 GB	100 GB	MIG 1g.10gb	Machine Learning
Medium (12)	16 VCPU	GB	200 GB	MIG 2g.20gb	Machine Learning
Large (8)	32 VCPU	128 GB	500 GB	MIG 3g.40gb	Machine Learning
Extra Large (3)	64 VCPU	256 GB	1000 GB	MIG 7g.80gb	Deep Learning
High End (1)	96 VCPU	1000 GB	2000 GB	8 x GPU (Direct)	Deep Learning

Compute for AI Inference Cluster Workload

AI Inference workloads can leverage both CPU and GPU. The number of inference instances can be scaled out to optimally use GPU memory. Inference can be performed with CPU only for many use cases. The inference scaling is based on the number of requests served and it must be scaled appropriately.

Table 13: Machine Learning Inference System Configuration

Model Type (count)	Virtual processors	System memory	Storage	GPU Profile	Workload Characteristics
Small (16)	12 VCPU	32 GB	100 GB	MIG 1g.10gb	Machine Learning
Medium (12)	32 VCPU	64 GB	200 GB	MIG 2g.20gb	Machine Learning
Large (8)	8 VCPU	128 GB	500 GB	MIG 3g.40gb	Machine Learning
Extra Large (3)	64 VCPU	512 GB	1000 GB	MIG 7g.80gb	Deep Learning
High End (1)	96 VCPU	1024 GB	2000 GB	4 x GPU (Direct)	Deep Learning

8.2.1 Enterprise Operational Model

For the enterprise operational model, see the following sections for more information about each component, its performance, and sizing guidance:

8.3 OS, Hypervisor and Container Platform Support
8.4 MLOps and Orchestration Platform
8.5 Management Server
8.6 ThinkSystem Servers Bare Metal for HPC and AI
8.7 ThinkSystem Servers with Kubernetes for AI
8.8 ThinkAgile VX Servers with VMware vSAN for AI
8.9 ThinkAgile HX Servers with Nutanix for AI
8.10 ThinkSystem Servers with Cloudian for Data Lake
8.11 ThinkEdge Servers for Edge AI
8.12 Neptune Direct Water Cooling Solutions
8.13 System Management

The SMB model is the same as the Enterprise model for all systems.

8.3 OS, Hypervisor and Container Platform Support

The following operating systems and hypervisors are supported on Lenovo ThinkSystem and ThinkAgile platforms. The enterprise AI software architecture is dependent on containers and relevant Kubernetes solution need to be chosen. If virtualization is used, then any supported Kubernetes version can be deployed on the operating system to be managed by the Enterprise AI stack.

Table 14: Lenovo server platforms and Kubernetes support

Server Platform	OS	Hypervisor	Kubernetes Platform
ThinkSystem	RHEL, Ubuntu, SUSE, Windows Server	VMware ESXi	Kubernetes
ThinkAgile VX	N/A	VMware ESXi	VMware Tanzu

ThinkAgile HX	N/A	Nutanix AHV	Nutanix Cloud Platform
ThinkEdge	RHEL, Ubuntu, SUSE, Windows Server	VMware ESXi	Kubernetes

8.4 MLOps and Orchestration Platform

The enterprise AI platform provides management, deployment and monitoring support for AI/ML deployment. The enterprise AI platform consists of many proprietary and open source components and it can be chosen based on the functional and operational requirements. The AI/ML model development and deployment can be done on any operating system and hypervisor as like any other workloads by not using any additional ML frameworks, but it may need additional manual effort and can result less operational efficiency. When more models, teams and infrastructure are involved, it is required to have robust compute resource management, optimization libraires and MLOps tools to achieve maximum performance and efficiency.

This reference architecture includes NVIDIA AI Enterprise platform which can be used across Lenovo ThinkSystem and ThinkAgile platforms and software ecosystem. Also, Lenovo platforms can work seamlessly with all open source frameworks which can operate with Kubernetes environment and NVIDIA GPU operator and this document can serve as a reference for many deployment options and frameworks.

Table 15: Lenovo Enterprise AI platforms and MLOps support

Server Platform	OS	Workload	Kubernetes Platform	Training and deployment	MLOPS	Enterprise AI Framework
ThinkSystem	Bare Metal (RHEL, Ubuntu, SUSE)	HPC/AI		LiCO	Kubeflow, mlflow	NVIDIA AI Enterpise
ThinkSystem	Bare Metal K8s(RHEL, Ubuntu, SUSE)	HPC/AI	Kubernetes	LiCO/ run:ai	Kubeflow, mlflow	NVIDIA AI Enterpise
ThinkSystem	Virtualized (VMware ESXi)	AI	Kubernetes	run:ai	Kubeflow, mlflow	NVIDIA AI Enterpise
ThinkAgile VX	VMware ESXi	AI	VMware Tanzu	run:ai	Kubeflow, mlflow	NVIDIA AI Enterpise
ThinkAgile HX	Nutanix AHV	Ai	Nutanix Cloud Platform	-		NVIDIA AI Enterpise
ThinkEdge	RHEL, Ubuntu, SUSE, Windows Server	AI	Kubernetes	run:ai	Kubeflow, mlflow	NVIDIA AI Enterpise

8.5 Management Server

Management servers should have the same hardware specification as compute servers so that they can be used interchangeably in a worst-case scenario. The management server components can co-exist with AI/ML workload components for small and medium deployment and for bare-metal deployments. It is recommended to have dedicated hardware for management servers. Management servers do not need GPU but if it is shared cluster deployment, then GPU can present. The management server requirement varies based on the Lenovo platform chosen, so refer the appropriate compute section.

8.6 ThinkSystem Servers Bare Metal for HPC and AI

This HPC and AI solution design uses bare-metal Lenovo ThinkSystem servers for running both HPC and AI workloads. The AI workloads can run as Singularity containers along with regular HPC applications. All the servers are managed by Lenovo LiCO HPC/AI edition.

8.6.1 Management Cluster

LiCO HPC/AI is a unified framework which is built with many components required for distributed computing, container management, HPC and AI/ML workload deployments and end to end infrastructure and applications monitoring. The table below provides minimum requirement, and the HPC login node performance requirement increases based on number of users and number of applications compiled.

Table 16: Management Clusters Configuration

Cluster Name	Number of Servers	Virtual processors	System memory	Storage	OS	HA needed
HPC Login Node	2 x ThinkSystem SR655 V3 2 x AMD EPYC 9124 16C 200W 3.0GHz, 128 GB	8 VPU	84 GB	500 GB	RHEL CentOS	Yes
LiCO/HPC Head Node	2 x ThinkSystem SR655 V3 2 x AMD EPYC 9124 16C 200W 3.0GHz. 128GB	8 VCPU	32 GB	1000 GB	RHEL CentOS	Yes

8.6.2 HPC and AI/ML Compute Cluster

The compute requirement considers end-end data science and AI/ML pipeline for model development, training, and production deployment. Data preparation is a critical task in any AI project, and it is recommended to have these applications and services isolated on separate cluster. The sample sizing given below uses Intel 4th Gen Xeon Scalable Processors for data science and AI applications and AMD 4th Gen EPYC processors with NVIDIA H100 GPU for HPC and AI training and inference use cases. The HPC and AI clustered are shared for both workloads.

Table 17: Compute Clusters Configuration for HPC and AI Applications

Cluster Name	Lenovo Platform	CPU	Memory	OS / Hypervisor	GPU
Data Science and AI Apps	4 x ThinkSystem SR650 V3	Intel Xeon Gold 6448H 32C 250W 2.4GHz	1000 GB	Linux (Singularity Container)	No
HPC and AI Compute Nodes – CPU Only	8 x ThinkSystem SR655 V3	2x AMD EPYC 9334 32C 210W 2.7GHz	512 GB	Linux	No
HPC and AI Training Nodes - GPU	3 x ThinkSystem SR675 V3	2x AMD EPYC 9454 48C 290W 2.75GHz	1000 GB	Linux	8xNVIDIA H100 80GB
HPC Storage Nodes	2 x ThinkSystem SR655 V3	2 x AMD EPYC 9124 16C 200W 3.0GHz	256 GB	Linux	No
AI Inference	2 x ThinkSystem SR675 V3	2 x AMD EPYC 9334 32C 210W 2.7GHz	1000 GB	Linux	4xNVIDIA H100 80GB

8.7 ThinkSystem Servers with Kubernetes for AI

This AI/ML solution design uses Lenovo ThinkSystem servers with Kubernetes standard or community edition for running both data science and AI workloads. All the servers are managed by LiCO K8s/AI and this deployment support non-Lenovo hardware and infrastructure compatible with Kubernetes. This solution is focused on running AI workloads and does not include HPC workloads.

8.7.1 Management Cluster

LiCO HPC/AI is a unified framework which is built with many components required for distributed computing, container management, AI/ML workload deployments and end to end infrastructure and applications monitoring. The table below provides minimum requirement, performance requirement increases based on number of containers managed. The management cluster hosts Kubernetes controller for all workload servers.

Table 18: Management Cluster Node Configuration

Cluster Name	Number of Servers	Virtual Machines
Management	3 x Intel Xeon Gold 6426Y 16C 185W 2.5GHz, 256 GB memory	Kuberneres Controller, Kubernetes API server Promotheus, Gibana, LiCO K8s/AI

Table 19: Management services configuration

Management service	Virtual processors	System memory	Storage	OS	HA needed
Kubernetes Controller VM	4 VCPU	16 GB	200 GB	Linux	3 Master Nodes
LiCO K8s/AI	8 VCPU	32 GB	200 GB	RHEL/CentOS	1 Node
XClarity Administrator	4 VCPU	16 GB	200 GB	Linux	No

8.7.2 AI/ML Compute Cluster

The compute requirement considers end-end data science and AI/ML pipeline for model development, training, and production deployment.

The compute requirement considers end-end data science and AI/ML pipeline for model development, training, and production deployment. Data preparation is a critical task in any AI project, and it is recommended to have these applications and services isolated on separate cluster. The sample sizing given below uses Lenovo ThinkSystem SR650 V3 server Intel 4th Gen Xeon Scalable Processors and NVIDIA H100 GPU for data science, AI/ML training AI training and inference use cases. ThinkSystem SR650 V3 systems support three NVIDIA H100 GPU per node and servers can be scaled out to meet workloads.

Table 20: Compute Clusters Configuration for AI Applications

Cluster Name	Lenovo Platform	CPU	Memory	OS / Hypervisor	GPU
Data Science and AI Apps	4 x ThinkSystem SR650 V3	Intel Xeon Gold 6448H 32C 250W 2.4GHz	2000 GB	Linux	No
AI Training Nodes	8 x ThinkSystem SR650 V3	Intel Xeon Platinum 8468 48C 350W 2.1GHz	1000 GB	Linux	3xNVIDIA H100 80GB
AI Inference	3 x ThinkSystem SR650 V3	Intel Xeon Platinum 8468 48C 350W 2.1GHz	1000 GB	Linux	3xNVIDIA H100 80GB

8.8 ThinkAgile VX Servers with VMware vSAN for AI

Lenovo ThinkAgile VX servers with VMware vSAN software defined storage solution provides flexible hardware options for CPU, memory, and IO intensive workloads. The VX systems with VMware Cloud Foundation and Tanzu enables end-end software stack and manage machine learning life cycle.

8.8.1 VX Servers

Lenovo ThinkAgile VX servers can be used for edge, management, or compute clusters with VMware vSAN and it supports All Flash and Hybrid configurations. With VMware vSphere 8 support, VX V3 servers can support original storage architecture(OSA) and express storage architecture(ESA) and the servers support Intel 4th Gen Xeon Scalable Processors and AMD 4th Gen EPYC processors which are perfect fit for AI workloads.

Table 21: Lenovo ThinkAgile VX system models

VX System	vSAN Support	Base System	CPU	Max drives	Max Possible Capacity
VX630 V3 IS	Hybrid All Flash	ThinkSystem SR630 V3	Intel Xeon 4th Gen	12	2 disk group 10 drives
VX630 V3 CN	Hybrid All Flash	ThinkSystem SR630 V3	Intel Xeon 4th Gen	12	2 disk group 10 drives
VX650 V3 IS	Hybrid All Flash	ThinkSystem SR650 V3	Intel Xeon 4th Gen	32	4 disk group 28 drives
VX650 V3 CN	Hybrid All Flash	ThinkSystem SR650 V3	Intel Xeon 4th Gen	32	4 disk group 28 drives
VX650V3 DPU IS	Hybrid All Flash	ThinkSystem SR650 V3	Intel Xeon 4th Gen	32	4 disk group 28 drives
VX650V3 DPU CN	Hybrid All Flash	ThinkSystem SR650 V3	Intel Xeon 4th Gen	32	4 disk group 28 drives
VX655 V3 Integrated System	Hybrid All Flash	ThinkSystem SR655 V3	AMD EPYC 4th Gen	40	5 disk group 35 drives
VX655 V3 Certified Node	Hybrid All Flash	ThinkSystem SR655 V3	AMD EPYC 4th Gen	40	5 disk group 35 drives
VX665 V3 Integrated System	Hybrid All Flash	ThinkSystem SR665 V3	AMD EPYC 4th Gen	40	4 disk group 28 drives
VX665 V3 Certified Node	Hybrid All Flash	ThinkSystem SR665 V3	AMD EPYC 4th Gen	40	4 disk group 28 drives

8.8.2 Management Cluster

The management cluster hosts virtual machines and containers for virtualization management, cloud management and Kubernetes services for managing complete infrastructure. The solution with VMware Cloud Foundation includes VMware vCenter for virtualization, vRealize Suite for private cloud, Tanzu Kubernetes

Engine and NSX-T for software defined networking. The software components for MLOps, AI framework and experiment tracking are deployed on the management cluster.

Table 22: Management Clusters Configuration

Cluster Name	Lenovo Platform	CPU	Memory	OS / Hypervisor	GPU
Management	4 x ThinkSystem VX650 V3	Intel Xeon Gold 6448H 32C 250W 2.4GHz	1000 GB	ESXi	No

Table 23: Management cluster VMs for Virtualization Management

VM description	CPU (vCPUs)	Memory (GB)	Storage (GB)	Network bandwidth	High availability
SDDC Manager	4	16	1000	1 GbE	vSphere HA
vCenter Server Appliance(1) Management Cluster	8	24	50	1 GbE	load balancer
vCenter Server Appliance(2) Edge and Compute Cluster	8	24	50	1 GbE	load balancer
vCenter Server Database (MS SQL)	4	8	200	1 GbE	SQL AlwaysOn Availability Group
vSphere Replication	2	4	20	1 GbE	not required
vSphere Data Protection	4	4	1600	1 GbE	not required
vRealize Orchestrator Appliance	2	3	12	1 GbE	Clustered

Table **24** lists each management cluster VM for vRealize Automation with its size in terms of virtual CPUs, RAM, storage, and networking.

Table 24: Management cluster VMs for vRealize Automation

VM description	CPU (vCPUs)	Memory (GB)	Storage (GB)	Network bandwidth	High availability
vRealize Suite Lifecycle Manager	4	16	135	1 GbE	N/A
vRealize Automation Appliance	4	16	30	1 GbE	load balancer
IaaS Database (MS SQL)	8	16	100	1 GbE	SQL AlwaysOn Availability Group

Infrastructure Web Server	2	4	40	1 GbE	load balancer
Infrastructure Manager Server	2	4	40	1 GbE	load balancer
Distributed Execution Manager (DEM)	2	6	40	1 GbE	load balancer
vSphere Proxy Agent	2	4	40	1 GbE	load balancer
vRealize Application Services	8	16	50	1 GbE	vSphere HA

Table 25 lists each management cluster VM for vRealize Operations Manager with its size in terms of virtual CPUs, RAM, storage, and networking.

Table 25: Management cluster VMs for vRealize Operations Manager

VM description	CPU (vCPUs)	Memory (GB)	Storage (GB)	Network bandwidth	High availability
vRealize Operations Manager – Master	4	16	500	1 GbE	clustered
vRealize Operations Manager – Data	4	16	500	1 GbE	not required
vRealize Configuration Manager – Collector	4	16	150	1 GbE	load balancer
vRealize Configuration Manager Database (MS SQL)	4	16	1000	1 GbE	SQL AlwaysOn Availability Group
vRealize Hyperic Server	8	12	16	1 GbE	load balancer
vRealize Hyperic Server - Postgres DB	8	12	75	1 GbE	load balancer
vRealize Infrastructure Navigator	2	4	24	1 GbE	not required

Table 26 lists the management VMs that are needed for NSX.

Table 26: NSX-T Management cluster VMs

VM description	CPU (vCPUs)	Memory (GB)	Storage (GB)	Network bandwidth	High availability
NSX-T Manager Management Cluster	4	12	300	1 GbE	vSphere HA
NSX-T Controller Management Cluster (odd # deployment; min 3)	4	4	20	1 GbE	Built-in/vSphere HA
NSX-T Manager Edge and Compute Cluster	4	12	60	1 GbE	vSphere HA

8.8.3 AI/ML Compute Cluster

The compute requirement considers end-end data science and AI/ML pipeline for model development, training, and production deployment.

The compute requirement considers end-end data science and AI/ML pipeline for model development, training, and production deployment. Data preparation is a critical task in any AI project, and it is recommended to have these applications and services isolated on separate cluster. The sample sizing given below uses Lenovo ThinkAgile VX650 V3 server Intel 4th Gen Xeon Scalable Processors and NVIDIA H100 GPU for data science, AI/ML training AI training and inference use cases. ThinkAgile VX650 V3 systems support three NVIDIA H100 GPU per node and servers can be scaled out to meet workloads.

Table 27: Compute Clusters Configuration for AI Applications

Cluster Name	Lenovo Platform	CPU	Memory	OS / Hypervisor	GPU
Data Science and AI Apps	4 x ThinkSystem SR650 V3	Intel Xeon Gold 6448H 32C 250W 2.4GHz	2000 GB	Linux	No
AI Training Nodes	8 x ThinkSystem SR650 V3	Intel Xeon Platinum 8468 48C 350W 2.1GHz	1000 GB	Linux	3xNVIDIA H100 80GB
AI Inference	4 x ThinkSystem SR650 V3	Intel Xeon Platinum 8468 48C 350W 2.1GHz	1000 GB	Linux	2xNVIDIA H100 80GB

8.9 ThinkAgile HX Servers with Nutanix for AI

Lenovo ThinkAgile HX servers with Nutanix software defined storage solution provides flexible hardware options for CPU, memory, and IO intensive workloads. The HX systems with Nutanix Cloud Platform and Nutanix Kubernetes Engine enable end-end software stack and manage machine learning life cycle.

8.9.1 HX Servers

Lenovo ThinkAgile HX Integrated Systems and Certified nodes are hyperconverged systems powered by Nutanix AHV offer a range of storage configurations: Hybrid nodes combine flash SSDs for performance and HDDs for capacity, All-flash nodes utilize traditional flash SSDs and NVMe nodes utilize NVMe SSDs. Different node types can be mixed in the same cluster. For data resiliency, Nutanix uses replication factor (RF), maintaining 2 or 3 data copies. This approach enables a Nutanix cluster to be self-healing in the event of a drive, node, block, or rack failure. Lenovo ThinkAgile HX systems with Intel 4th Gen Xeon Scalable Processors and AMD 4th Gen EPYC processors which are perfect fit for AI workloads.

Table 28: ThinkAgile HX Servers

HX System	Storage Support	Base System	CPU	Max drives
HX630 V3 Integrated System	Hybrid All Flash	ThinkSystem SR630 V3	Intel Xeon 4th Gen	18
HX630 V3 Certified Node	Hybrid All Flash	ThinkSystem SR630 V3	Intel Xeon 4th Gen	18
HX650 V3 Integrated System	Hybrid All Flash	ThinkSystem SR650 V3	Intel Xeon 4th Gen	24
HX650 V3 Certified Node	Hybrid All Flash	ThinkSystem SR650 V3	Intel Xeon 4th Gen	24
HX645 V3 Integrated System	Hybrid All Flash	ThinkSystem SR645 V3	AMD EPYC 4th Gen	12
HX645 V3 Certified Node	Hybrid All Flash	ThinkSystem SR645 V3	AMD EPYC 4th Gen	12
HX665 V3 Integrated System	Hybrid All Flash	ThinkSystem SR665 V3	AMD EPYC 4th Gen	32
HX665 V3 Certified Node	Hybrid All Flash	ThinkSystem SR665 V3	AMD EPYC 4th Gen	32

8.9.2 Management Cluster

Table 29: Management Clusters Configuration

Cluster Name	Lenovo Platform	CPU	Memory	OS / Hypervisor	GPU
Management	4 x ThinkSystem HX650 V3 IS	Intel Xeon Gold 6448H 32C 250W 2.4GHz	1000 GB	AHV	No

Table 30: Management cluster VMs for Virtualization Management

VM description	CPU (vCPUs)	Memory (GB)	Storage (GB)	Network bandwidth	High availability
Prism	4	16 GB	100 GB	1 GbE	Yes
Nutanix Prism Central	4	32 GB	500 GB	1 GbE	Yes

VM description	CPU (vCPUs)	Memory (GB)	Storage (GB)	Network bandwidth	High availability
Prism	4	16 GB	100 GB	1 GbE	Yes
Life Cycle Manager (LCM)	4	8 GB	200 GB	10 GbE	
MLOps	4	32 GB	500 GB	10 GbE	Yes
Nutanix Cloud Manager	4	32 GB	200 GB	10 GbE	Yes

8.9.3 AI/ML Compute Cluster

The compute requirement considers end-end data science and AI/ML pipeline for model development, training, and production deployment.

The compute requirement considers end-end data science and AI/ML pipeline for model development, training, and production deployment. Data preparation is a critical task in any AI project, and it is recommended to have these applications and services isolated on separate cluster. The sample sizing given below uses Lenovo ThinkSystem HX650 V3 IS server Intel 4th Gen Xeon Scalable Processors and NVIDIA H100 GPU for data science, AI/ML training AI training and inference use cases. ThinkSystem HX650 V3 IS systems support three NVIDIA H100 GPU per node and servers can be scaled out to meet workloads.

Table 31: Compute Clusters Configuration for AI Applications

Cluster Name	Lenovo Platform	CPU	Memory	OS / Hypervisor	GPU
Data Science and AI Apps	4 x ThinkSystem HX650 V3 IS	Intel Xeon Gold 6448H 32C 250W 2.4GHz	2000 GB	Linux	No
AI Training Nodes	8 x ThinkSystem HX650 V3 IS	Intel Xeon Platinum 8468 48C 350W 2.1GHz	1000 GB	Linux	3xNVIDIA H100 80GB
AI Inference	4 x ThinkSystem HX650 V3 IS	Intel Xeon Platinum 8468 48C 350W 2.1GHz	1000 GB	Linux	2xNVIDIA H100 80GB

8.10 ThinkSystem Servers with Cloudian for Data Lake

The configuration below provides minimum sizing to build 1 PB data lake for AI requirements. When sizing and configuring hardware systems for a VM or appliance-based Lenovo Object Storage powered by Cloudian, it is part of best practices (and a requirement) to engage in a sizing exercise to assess the workflow and workload

and determine what resources between the minimum and recommended values (or even above them) may be required. The configuration below uses 3 nodes for high availability. The data science, AI training and inference clusters can be connected to S3 compatible cloud storage over the network.

Table 32: Compute Node configuration for Cloudian

VM Configuration	Lenovo Platform	CPU	Memory	Metadata Tier	Raw Object Storage
3 Hyperstore Virtual Machines (16 VCPU+ 128 GB memory)	3 x ThinkSystem SR650 V2 Ready Node	Intel Xeon Gold 5318Y 24C 165W 2.1GHz, 384 GB	384 GB	2x 7.68TB NVMe	18x 22TB HDD (396TB)

8.11 ThinkEdge Servers for Edge AI

Table 33 lists Lenovo ThinkEdge servers designed for Edge AI use cases. The edge servers support high cores and GPUs to support doing inference and training at the edge.

Table 33: ThinkEdge Servers

Edge System	Storage Support	CPU	GPU
ThinkEdge SE455 V3	61.44TB	AMD EPYC 8004 Series Up to 64 cores, 576GB memory	6x single-wide GPUs 2x double-wide GPUs
ThinkEdge SE350 V2	30.72 TB	Intel Xeon D-2700 Series, (up to 16 cores, 256 GB memory	No
ThinkEdge SE360 V2	30.72 TB	Intel Xeon D-2700 Series, (up to 16 cores, 256 GB memory	NVIDIA A2 , NVIDIA L4 or Qualcomm Cloud AI 100
ThinkEdge SE350	16 TB	Intel Xeon D-2700 Series, (up to 16 cores, 256 GB memory	NVIDIA A2 , NVIDIA L4 , NVIDIA T4
ThinkEdge SE450	30.72 TB	1 x Intel Xeon 3rd Gen Up to 36 cores, 1 TB memory	4x single-wide GPUs 2x double-wide GPUs

8.12 Neptune Direct Water Cooling Solutions

Table 34 below lists Lenovo ThinkSystem servers support Lenovo Neptune Direct Water Cooling solution.

Table 34: ThinkSystem Servers with Neptune Water Cooling Support

Edge System	Storage Support	CPU	GPU
ThinkSystem SD665-N V3	30.72 TB	2 x AMD EPYC 9004 processors Up to 128 cores	4 x NVIDIA H100 GPUs

		3 TB memory	
ThinkSystem SD665- V3	61.44 TB	2 x AMD EPYC 9004 processors Up to 128 cores 3 TB memory	No
SD650-I V3	30.72 TB	2 x Intel Xeon 4th Gen Up to 60 cores Up to 2 TB memory	4x Intel Xeon Max Series GPUs
SD650- V3	30.72 TB	2 x Intel Xeon 4th Gen Up to 60 cores Up to 2 TB memory	No

8.13 System Management

Lenovo XClarity™ Administrator is a centralized resource management solution that reduces complexity, speeds up response, and enhances the availability of Lenovo® server systems and solutions.

The Lenovo XClarity Administrator provides agent-free hardware management for Lenovo's ThinkSystem, System x® rack servers and Flex System™ compute nodes and components, including the Chassis Management Module (CMM) and Flex System I/O modules. Figure 18 shows the Lenovo XClarity administrator interface, where Flex System components and rack servers are managed and are seen on the dashboard. Lenovo XClarity Administrator is a virtual appliance that is quickly imported into a virtualized environment server configuration.

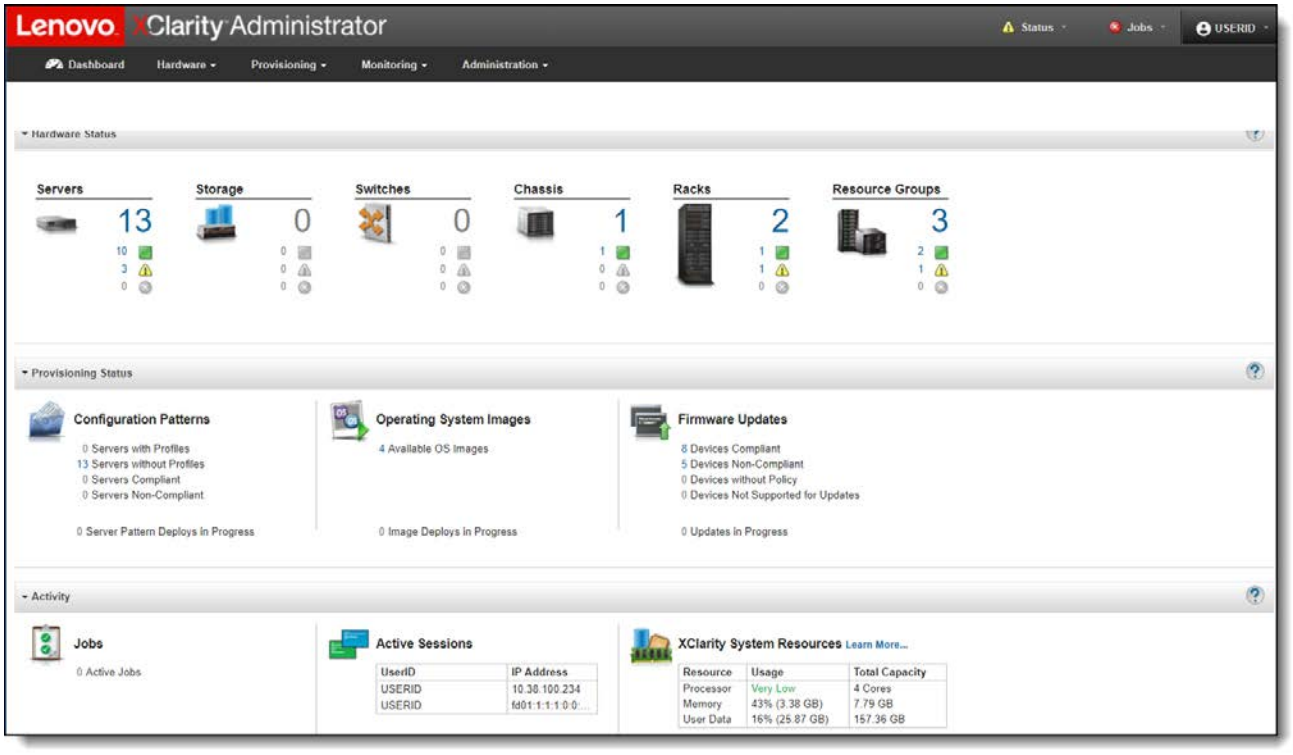


Figure 18: XClarity Administrator interface

8.14 Lenovo XClarity Orchestrator(LXCO)

XClarity Orchestrator provides a single interface to monitor and manage multiple Lenovo XClarity Administrators and the devices managed by them. LXCO supports deploying updates to Lenovo Lenovo XClarity Administrator and firmware updates to devices that are managed. LXCO can connect to third-party services (such as Splunk) for business intelligence machine learning and predictive analytics to collect resource utilization data and uses metric data to predict failures, create reports and custom alert rules that, when enabled, raise alerts when specific conditions exist in your environment.

8.15 Lenovo Intelligent Computing Orchestration (LiCO)

Lenovo Intelligent Computing Orchestration (LiCO) is a software solution that simplifies the use of clustered computing resources for Artificial Intelligence (AI) model development and training, and HPC workloads. LiCO interfaces with an open-source software orchestration stack, enabling the convergence of AI onto an HPC or Kubernetes-based cluster. Refer more information here https://lenovopress.lenovo.com/lp0858-lenovo-intelligent-computing-orchestration-lico

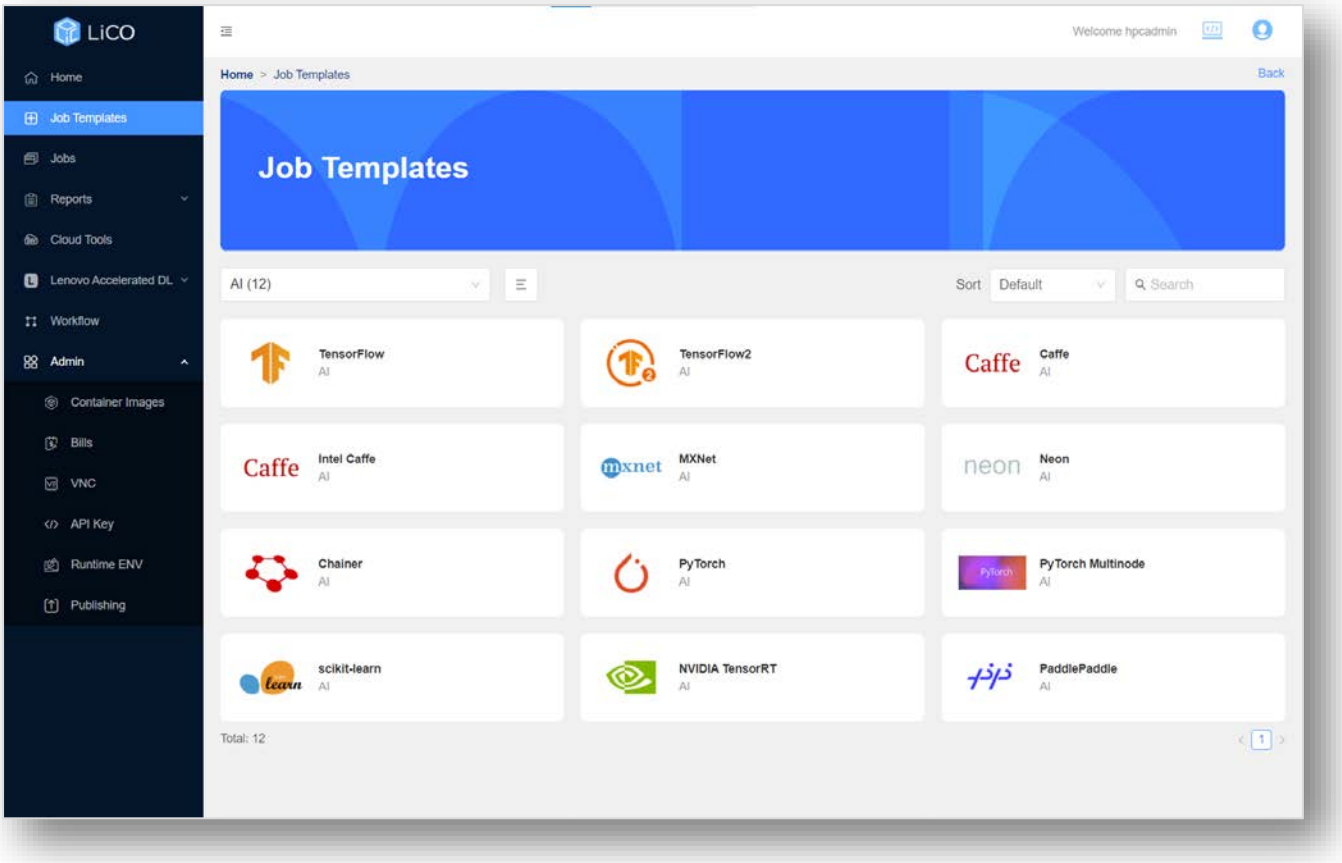


Figure 19: Lenovo Intelligent Computing Orchestration(LiCO)

9 Conclusion

Lenovo servers, storage and networking platforms options and solutions provides an ideal foundation infrastructure to build robust and resilient enterprise AI infrastructure. Lenovo partner ecosystem for virtualization, software defined solutions, container platforms and machine learning framework and cognitive services enable customer to start their AI/ML journey and consolidate enterprise wide AI use cases.

10 Appendix: Bill of materials

This appendix contains the bill of materials (BOMs) for different configurations of hardware. There are sections for user servers, management servers, storage, networking switches, chassis, and racks that are orderable from Lenovo. The last section is for hardware orderable from an OEM.

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.

10.1 BOM for AI on ThinkSystem Server

Table 35 shows bill of materials for ThinkSystem servers.

Table 35: ThinkSystem SR675 V3 for HPC & AI

Part number	Product Description	Qty
7D9RCTOLWW	Server: ThinkSystem SR675 V3 - 3yr Warranty - HPC&AI	4
BR7F	ThinkSystem SR675 V3 8DW PCIe GPU Base	4
BFYA	Operating mode selection for: "Maximum Efficiency Mode"	4
BPVJ	ThinkSystem AMD EPYC 9554 64C 360W 3.1GHz Processor	8
BQ3D	ThinkSystem 64GB TruDDR5 4800MHz (2Rx4) 10x4 RDIMM-A	96
5977	Select Storage devices - no configured RAID required	4
BFTQ	ThinkSystem 1x6 E1.S EDSFF Backplane Option Kit	4
B8P9	ThinkSystem M.2 NVMe 2-Bay RAID Enablement Kit	4
BKSR	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	8
BG3F	ThinkSystem NVIDIA Ampere NVLink 2-Slot Bridge	48
BQBN	ThinkSystem NVIDIA ConnectX-7 NDR200/HDR QSFP112 2-Port PCIe Gen5 x16 InfiniBand Adapter	20
B93E	ThinkSystem Intel I350 1GbE RJ45 4-port OCP Ethernet Adapter	4
BR9U	ThinkSystem NVIDIA H100 80GB PCIe Gen5 Passive GPU	32
BR7L	ThinkSystem SR675 V3 x16/x16 PCIe Riser Option Kit	8
BR7H	ThinkSystem SR675 V3 2x16 PCIe Front IO Riser	4
BK1E	ThinkSystem SR670 V2/ SR675 V3 OCP Enablement Kit	4
BR7S	ThinkSystem SR675 V3 Switched 4x16 PCIe DW GPU Direct RDMA Riser	8
B962	ThinkSystem 2400W 230V Platinum Hot-Swap Gen2 Power Supply	16
B4L2	2.0m, 16A/100-250V, C19 to IEC 320-C20 Rack Power Cable	16
BFTL	ThinkSystem SR670 V2/ SR675 V3 Toolless Slide Rail	4
3803	3m Blue Cat5e Cable	4
3793	3m Yellow Cat5e Cable	4
B7XZ	Disable IPMI-over-LAN	4
BR7V	ThinkSystem SR675 V3 System Board	4
BK15	High voltage (200V+)	4
BR7W	ThinkSystem SR670 V2/ SR675 V3 System Documentation	4

BE0D	N+1 Redundancy with Over-Subscription	4
BR82	ThinkSystem SR670 V2/ SR675 V3 WW Packaging	4
BR80	ThinkSystem SR675 V3 Agency Labels	4
BR85	ThinkSystem SR670 V2/ SR675 V3 Branding Label	4
B993	ThinkSystem V2 EDSFF Filler	24
BFD6	ThinkSystem SR670 V2/ SR675 V3 Power Mezzanine Board	4
BFTH	ThinkSystem SR670 V2/ SR675 V3 Front Operator Panel ASM	4
BRUC	ThinkSystem SR675 V3 CPU Heatsink	8
BFNU	ThinkSystem SR670 V2/ SR675 V3 Intrusion Cable	4
BR88	ThinkSystem SR670 V2/ SR675 V3 Service Label	4
BR7U	ThinkSystem SR675 V3 Root of Trust Module	4
BS03	ThinkSystem SR675 V3 2400W Power Supply Caution Label	4
BSD2	ThinkSystem SR675 V3 GPU Supplemental Power Cable 4	32
BS6Y	ThinkSystem 2U V3 M.2 Signal & Power Cable, SLx4 with 2X10/1X6 Sideband, 330/267/267mm	4
BR8G	ThinkSystem SR675 V3 Rear PCIe Riser Cable 4	4
BU22	ThinkSystem SR675 V3 Rear PCIe Riser Cable 6	4
BU23	ThinkSystem SR675 V3 Front OCP Cable 2	4
BR8Q	ThinkSystem SR675 V3 Front PCIe Riser Cable 6	4
BR8V	ThinkSystem SR675 V3 Front PCIe Riser Cable 2	4
BFTM	ThinkSystem SR670 V2/ SR675 V3 EDSFF Cage	4
BRUL	ThinkSystem SR675 V3 EDSFF Drive Sequence Label	4
BFGZ	ThinkSystem SR670 V2/ SR675 V3 Backplane Power Cable 4	4
BRUQ	ThinkSystem SR675 V3 EDSFF to Riser Cables	4
BF94	AI & HPC - ThinkSystem Hardware	4
5PS7B09635	Premier Essential - 3Yr 24x7 4Hr Resp + YDYD SR675 V3	4
5AS7A82992	Hardware Installation (Business Hours) for SR67x	4
5641PX3	XClarity Pro, Per Endpoint w/3 Yrs. SW S&S	4
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yrs. SW S&S	4
3444	Registration only	4
0724HEC	Switch: NVIDIA QM9700 64-Port Managed Quantum NDR InfiniBand Switch (PSE)	1
BP63	NVIDIA QM9700 64-Port Managed Quantum NDR InfiniBand Switch (PSE)	1
3791	0.6m Yellow Cat5e Cable	1
BQK3	Lenovo 3m NVIDIA NDRx2 OSFP800 to 4x NDR200 QSFP112 Passive Copper Splitter Cable	8
BRQ6	2.8m, 10A/100-250V, C15 to C14 Jumper Cord	2
BRET	NVIDIA QM97xx Enterprise RMK w/Air Duct	1
BF94	AI & HPC - ThinkSystem Hardware	1
5WS7B14266	Premier Essential - 3Yr 24x7 4Hr Resp NVID QM9700 PSE	1

7D5FCTO1WW-HPC	Switch: Mellanox AS4610-54T 1GbE Managed Switch with Cumulus (PSE)	1
BE2J	Mellanox AS4610-54T 1GbE Managed Switch with Cumulus (PSE)	1
3792	1.5m Yellow Cat5e Cable	1
6311	2.8m, 10A/100-250V, C13 to IEC 320-C14 Rack Power Cable	2
BEGG	Mellanox AS46xx Enterprise RMK w/Air Duct	1
BF94	AI & HPC - ThinkSystem Hardware	1

10.2 BOM for AI on ThinkAgile VX Server

Table 36 shows bill of materials for ThinkAgile VX650 V3 servers all flash ESA configuration.

Table 36: ThinkAgile VX650 V3 for AI

Part number	Product Description	Qty
7D6WCTOBWW	Server : Lenovo ThinkAgile VX650 V3 Certified Node with Controlled GPU	1
BRY9	ThinkAgile VX V3 2U 24x2.5" Chassis	1
BVGL	Data Center Environment 30 Degree Celsius / 86 Degree Fahrenheit	1
B0W3	XClarity Pro	1
B3XQ	3 Year	1
BN8K	ThinkAgile VX Remote Deployment	1
BQ6K	Intel Xeon Gold 6438M 32C 205W 2.2GHz Processor	2
BKTM	ThinkSystem 32GB TruDDR5 4800MHz (2Rx8) RDIMM	16
5977	Select Storage devices - no configured RAID required	1
B8P1	ThinkSystem 440-16i SAS/SATA PCIe Gen4 12Gb Internal HBA	1
B5MC	vSAN All Flash Config	1
B8LU	ThinkSystem 2U 8x2.5" SAS/SATA Backplane	2
B8P9	ThinkSystem M.2 NVMe 2-Bay RAID Enablement Kit	1
BTTY	M.2 NVMe	1
BKSR	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2
BQ8S	VMware ESXi 8.0 U1 (Factory Installed)	1
BPPW	ThinkSystem Broadcom 57504 10/25GbE SFP28 4-Port OCP Ethernet Adapter	1
B4RA	ThinkSystem Mellanox ConnectX-6 HDR100/100GbE QSFP56 2-port PCIe VPI Adapter	1
BR9U	ThinkSystem NVIDIA H100 80GB PCIe Gen5 Passive GPU	1
BPQV	ThinkSystem V3 2U x16/x16/E PCIe Gen5 Riser1 or 2	1
BLKN	ThinkSystem V3 2U G4 E/x16/x16 PCIe Riser1 or 2	1
BPKG	ThinkSystem V3 2U x16/x16 PCIe Gen4 Riser3 Kit with Cage	1
BPK9	ThinkSystem 1800W 230V Titanium Hot-Swap Gen2 Power Supply	2
BLL6	ThinkSystem 2U V3 Performance Fan Module	6
BK7W	ThinkSystem Toolless Friction Rail v2	1

BQQ2	ThinkSystem 2U V3 EIA Latch Standard	1
BPKR	TPM 2.0	1
BRPJ	XCC Platinum	1
B7XZ	Disable IPMI-over-LAN	1
BLL0	ThinkSystem SR650 V3 MB	1
BVMC	Trigger MFG to scan the SN from the CPU Board via this MI	1
BTSM	ThinkAgile VX650 V3 CN	1
AVEQ	ThinkSystem 8x1 2.5" HDD Filler	1
BQQ6	ThinkSystem 2U V3 EIA right with FIO	1
BXGY	Right EIA with FIO assembly	1
BQBP	ThinkSystem MCC CPU Clip	2
BP4Z	ThinkSystem SR650 V3 Performance Heatsink	2
BLL5	ThinkSystem SR650 V3 GPU Airduct	1
B8MB	ThinkSystem 2U MS GPU Air Duct Filler	2
B986	ThinkSystem HV 2U WW General PKG BOM	1
BLLD	ThinkSystem 2U MS 3FH Riser 1&2 Cage	1
BM8T	ThinkSystem SR650 V3 Firmware and Root of Trust Security Module	1
BMPF	ThinkSystem V3 2U Power Cable from MB to Front 2.5" BP v2	2
BACB	ThinkSystem V3 2U SAS/SATA Y Cable from CFF C0,C1/ C2,C3 to Front 8x2.5" BP	2
BPED	ThinkSystem SR650 V3 MCIO8x to SL8x CBL, PCIe4, CFF RAID INPUT, 250mm	1
BMP2	ThinkSystem V3 2U Power Cable from MB to CFF / Exp v2	1
BPES	ThinkSystem SR650 V3 MCIO8x Cable from PCIe 10/9 (MB) to MCIO 2/3(Riser3), 250mm	2
BRWK	ThinkSystem 400mm 2x6+4 GPU Power Cable	1
BS6Y	ThinkSystem 2U V3 M.2 Signal & Power Cable, SLx4 with 2X10/1X6 Sideband, 330/267/267mm	1
BPER	ThinkSystem SR650 V3 MCIO8x CBL from PCIe 1/2 (MB) to MCIO 1/4 (Riser3), 980mm	2
BPEZ	ThinkSystem SR650 V3 Riser 3 PWR CBL, 8P-10P, 150mm	1
BPEW	ThinkSystem V3 2U WH 24Pin Cable Riser3 SB, 120mm	1
BPK3	ThinkSystem WW Lenovo LPK	1
B265	ThinkAgile VX Pubkit	1
BE0E	N+N Redundancy With Over-Subscription	1
BK15	High voltage (200V+)	1
BQ16	G5 x16/x16/E PCIe Riser BPQV for Riser 1 Placement	1
BQ1A	G4 E/x16/x16 PCIe Riser BLKN for Riser 2 Placement	1
BNW9	ThinkSystem 2.5" PM1655 1.6TB Mixed Use SAS 24Gb HS SSD	4
BNWF	ThinkSystem 2.5" PM1653 3.84TB Read Intensive SAS 24Gb HS SSD	12
5641PX3	XClarity Pro, Per Endpoint w/3 Yr SW S&S	1
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yr SW S&S	1

B8Q8	ThinkSystem 440-16i SAS/SATA PCIe Gen4 12Gb Internal HBA Placement	1
5AS7B15971	Hardware Installation (Business Hours) for VX650 V3	1
5MS7A87711	ThinkAgile VX Remote Deployment (up to 4 node cluster)	1
7S0XCTO5WW	XClarity Controller	1
SBCV	Lenovo XClarity XCC2 Platinum Upgrade (FOD)	1

10.3 BOM for AI on ThinkAgile HX Server

Table 37 shows bill of materials for ThinkAgile HX650 V3 servers all flash Nutanix configuration.

Table 37: ThinkAgile HX650 V3 for AI

Part number	Product Description	Qty
7D6NCTODWW	Server : ThinkAgile HX650 V3 Certified Node with Controlled GPU	1
BRP4	ThinkAgile HX650 V3 Base	1
BVGL	Data Center Environment 30 Degree Celsius / 86 Degree Fahrenheit	1
B0W3	XClarity Pro	1
B15S	Nutanix SW Stack on Nutanix AHV	1
B0W1	3 Years	1
BM84	ThinkAgile HX Remote Deployment	1
BQ6K	Intel Xeon Gold 6438M 32C 205W 2.2GHz Processor	2
BKTM	ThinkSystem 32GB TruDDR5 4800MHz (2Rx8) RDIMM	16
B8P1	ThinkSystem 440-16i SAS/SATA PCIe Gen4 12Gb Internal HBA	1
B8LU	ThinkSystem 2U 8x2.5" SAS/SATA Backplane	2
B0SW	Nutanix Flash Node Config	1
BK7L	ThinkSystem 2.5" S4620 3.84TB Mixed Use SATA 6Gb HS SSD	6
B8P9	ThinkSystem M.2 NVMe 2-Bay RAID Enablement Kit	1
BTTY	M.2 NVMe	1
BKSR	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2
BN2T	ThinkSystem Broadcom 57414 10/25GbE SFP28 2-Port OCP Ethernet Adapter	1
BYFH	ThinkSystem NVIDIA L40S 48GB PCIe Gen4 Passive GPU	2
BPKG	ThinkSystem V3 2U x16/x16 PCIe Gen4 Riser3 Kit with Cage	1
BMUF	ThinkSystem 1800W 230V Platinum Hot-Swap Gen2 Power Supply	2
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2
BK7W	ThinkSystem Toolless Friction Rail v2	1
BQQ2	ThinkSystem 2U V3 EIA Latch Standard	1
BLL6	ThinkSystem 2U V3 Performance Fan Module	6
BPKR	TPM 2.0	1
B7Y0	Enable IPMI-over-LAN	1
BLL0	ThinkSystem SR650 V3 MB	1
BVMC	Trigger MFG to scan the SN from the CPU Board via this MI	1

B6C1	Node Cores	64
B6C2	Node Tebibytes	21
BHSS	MI for PXE with RJ45 Network port	1
9220	Preload by Hardware Feature Specify	1
5977	Select Storage devices - no configured RAID required	1
BRPJ	XCC Platinum	1
BK15	High voltage (200V+)	1
BPK3	ThinkSystem WW Lenovo LPK	1
BMPF	ThinkSystem V3 2U Power Cable from MB to Front 2.5" BP v2	2
BACB	ThinkSystem V3 2U SAS/SATA Y Cable from CFF C0,C1/ C2,C3 to Front 8x2.5" BP	2
BPED	ThinkSystem SR650 V3 MCIO8x to SL8x CBL, PCIe4, CFF RAID INPUT, 250mm	1
BMP2	ThinkSystem V3 2U Power Cable from MB to CFF / Exp v2	1
BPES	ThinkSystem SR650 V3 MCIO8x Cable from PCIe 10/9 (MB) to MCIO 2/3(Riser3), 250mm	2
BRWK	ThinkSystem 400mm 2x6+4 GPU Power Cable	2
BS6Y	ThinkSystem 2U V3 M.2 Signal & Power Cable, SLx4 with 2X10/1X6 Sideband, 330/267/267mm	1
BPER	ThinkSystem SR650 V3 MCIO8x CBL from PCIe 1/2 (MB) to MCIO 1/4 (Riser3), 980mm	2
BPEZ	ThinkSystem SR650 V3 Riser 3 PWR CBL, 8P-10P, 150mm	1
BPEW	ThinkSystem V3 2U WH 24Pin Cable Riser3 SB, 120mm	1
BE0E	N+N Redundancy With Over-Subscription	1
BQ11	G4 x16/x8/x8 PCIe Riser BLKL for Riser 1 Placement	1
BQ1A	G4 E/x16/x16 PCIe Riser BLKN for Riser 2 Placement	1
ATSB	Nutanix Solution Code MFG Instruction	1
BTSE	ThinkAgile HX650 V3 CN	1
AVEN	ThinkSystem 1x1 2.5" HDD Filler	6
AVEP	ThinkSystem 4x1 2.5" HDD Filler	1
AVEQ	ThinkSystem 8x1 2.5" HDD Filler	1
BQQ6	ThinkSystem 2U V3 EIA right with FIO	1
BXGY	Right EIA with FIO assembly	1
BQBP	ThinkSystem MCC CPU Clip	2
BP4Z	ThinkSystem SR650 V3 Performance Heatsink	2
BLL5	ThinkSystem SR650 V3 GPU Airduct	1
B8MB	ThinkSystem 2U MS GPU Air Duct Filler	1
AURS	Lenovo ThinkSystem Memory Dummy	16
B986	ThinkSystem HV 2U WW General PKG BOM	1
BLLD	ThinkSystem 2U MS 3FH Riser 1&2 Cage	1
BM8T	ThinkSystem SR650 V3 Firmware and Root of Trust Security Module	1
BU8S	ThinkAgile HX650 V3 - Nutanix IP	1
BLKN	ThinkSystem V3 2U G4 E/x16/x16 PCIe Riser1 or 2	1

BLKL	ThinkSystem V3 2U x16/x8/x8 PCIe Gen4 Riser1 or 2	1
5641PX3	XClarity Pro, Per Endpoint w/3 Yr SW S&S	1
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yr SW S&S	1
B8Q8	ThinkSystem 440-16i SAS/SATA PCIe Gen4 12Gb Internal HBA Placement	1
5AS7B15949	Hardware Installation (Business Hours) for HX650	1
5MS7B00045	ThinkAgile HX Remote Deployment (up to 3 node cluster)	1
7S0XCTO5WW	XClarity Controller	1
SBCV	Lenovo XClarity XCC2 Platinum Upgrade (FOD)	1

10.4 BOM for AI Storage

This section contains the bill of materials for shared storage Table 38 list the BOM for AI storage with ThinkSystem DG storage systems.

Table 38: ThinkSystem DG5000 - 300TB Storage

Part number	Product Description	Qty
7DE4CTO1WW	Controller : ThinkSystem DG5000 QLC All Flash Array	1
BF3C	Lenovo ThinkSystem Storage 2U NVMe Chassis	1
BQHN	Lenovo ThinkSystem 2U NVMe Controller with Titanium PSU	2
BXG8	Lenovo ThinkSystem 30.7TB (2x 15.36TB QLC NVMe SSD) Drive Pack for DG5000 - Unified Complete Bundle	10
BEVQ	Lenovo ThinkSystem Storage HIC, 10/25Gb iSCSI,4-ports	2
AVG0	3m Green Cat6 Cable	2
AV1W	Lenovo 1m Passive 25G SFP28 DAC Cable	2
B4BP	Lenovo ThinkSystem Storage USB Cable, Micro-USB	1
BX90	Lenovo ThinkSystem Storage ONTAP 9.12 Software Encryption - IPAv2	1
BY3B	SnapMirror License Bundle	1
5PS7B27398	Premier Essential 3Y 24x7x4+YDYD ThinkSystem DG5000 Unified Complete	1
5WS7B27496	Premier Essential 3Y 24x7x4 DG5000 307TB (20x15.36TB QLC NVMe)Pack Unified Complete	1
7S0L002EWW	Veeam Backup & Replication Universal License. Includes Enterprise Plus Edition features. - 1 Year Subscription Upfront Billing & Production (24/7) Support	2

Resources

For more information, see the following resources:

Lenovo ThinkSystem SR675 V3 server

https://lenovopress.lenovo.com/lp1611.pdf

ThinkAgile VX Series

https://lenovopress.lenovo.com/ds0104

ThinkAgile HX Series

https://lenovopress.lenovo.com/ds0019.pdf

VMware vSAN

https://www.vmware.com/in/products/vsan.html

Nutanix AOS Storage

https://www.nutanixbible.com/4c-book-of-aos-storage.html

ThinkSystem DM Series Hybrid Flash

https://lenovopress.lenovo.com/datasheet/ds0048-lenovo-thinksystem-dm-series-hybrid-flash

ThinkSystem DG5000 and DG7000

https://lenovopress.lenovo.com/lp1754-thinksystem-dg5000

https://lenovopress.lenovo.com/lp1755-thinksystem-dg7000

Document History

Version 1.0	26 December 2023 2023	• Initial version

Trademarks and special notices

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®
Flex System
Lenovo Neptune®
System x®
ThinkAgile®
ThinkEdge®
ThinkSystem®
XClarity®

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®, Azure®, SQL Server®, Windows Server®, and Windows® are trademarks 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.

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

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.