NVIDIA AI Enterprise Software Reference Architecture

This software reference architecture provides an example deployment of NVIDIA AI Enterprise software suite. It showcases a bare metal deployment, and provides example workloads to showcase the platform's capabilities. Topics such as hardware, network, and workload software will be discussed.

NVIDIA AI Enterprise

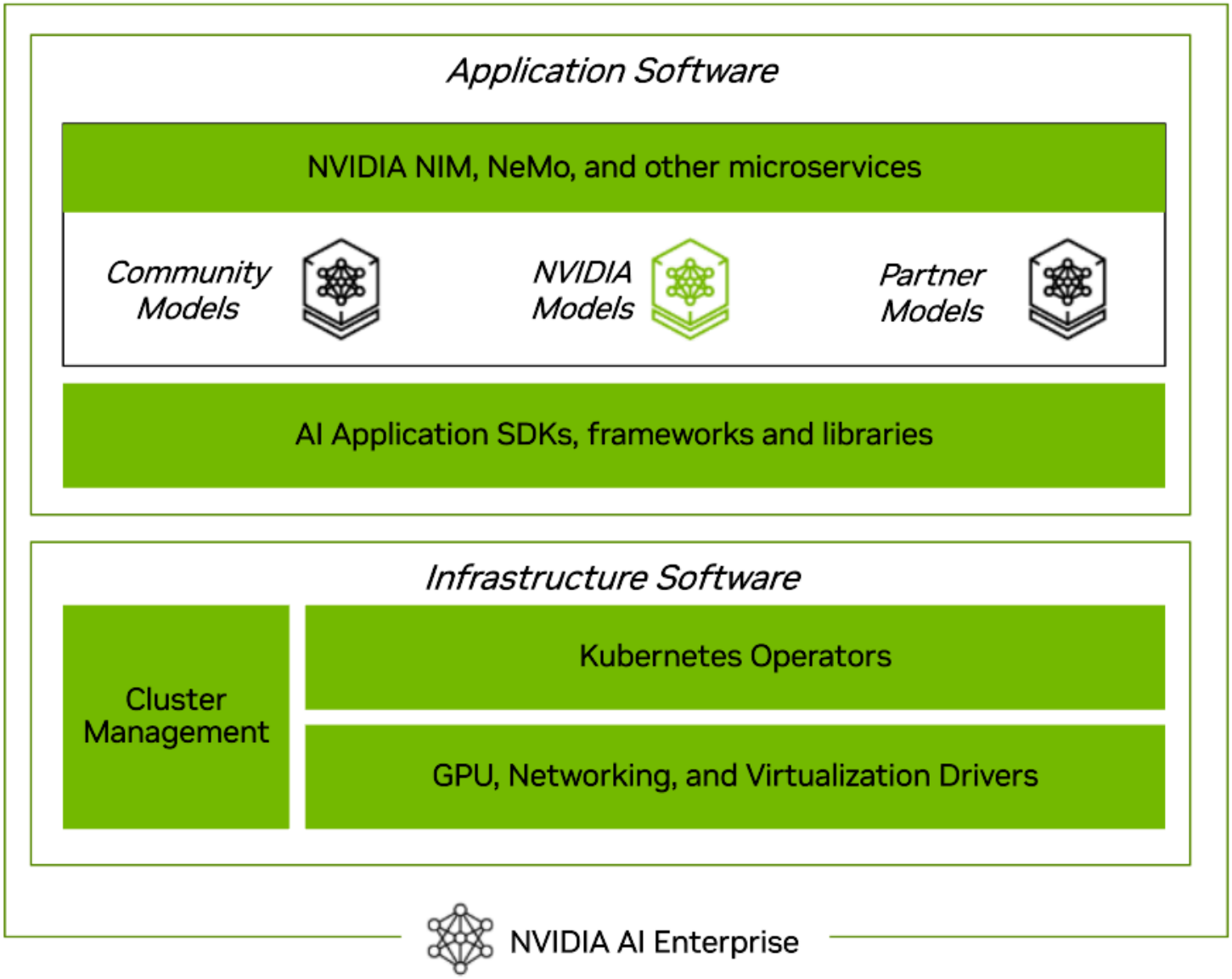
Software Reference Architecture

Support

Notices

Platform Overview

NVIDIA AI Enterprise is a cloud-native suite of software tools, libraries and frameworks designed to deliver optimized performance, robust security, and stability for production AI deployments. Easy-to-use microservices optimize model performance with enterprise-grade security, support, and stability, ensuring a streamlined transition from prototype to production for enterprises that run their businesses on AI. It consists of two primary layers: the infrastructure layer and the application layer.



Application Layer

The application layer provides specialized SDKs, frameworks and state of the art AI models for developing AI applications. It includes:

1. Optimized microservices that enhance AI model performance and speed time to deployment for a wide range of AI workflows.
2. Development and deployment tools: Support for popular AI tools like Triton, TensorFlow, PyTorch, and NVIDIA's own SDKs.
3. Optimized libraries for deep learning, data science, and machine learning.
4. Access to a repository of pre-trained models for various AI tasks.

By separating the infrastructure layer (which is versioned) from the application layer, NVIDIA AI Enterprise ensures that foundational updates and improvements do not disrupt the development and deployment of AI applications. This modular approach allows for flexibility and scalability in AI projects.

Application Layer Software

NVIDIA AI application frameworks, NVIDIA pretrained models and all other NVIDIA AI software available on NGC are supported with an NVIDIA AI Enterprise license. With 100+ AI frameworks and pretrained models including NeMo, Maxine, cuOpt and more, look for the **NVIDIA AI Enterprise Supported** label on NGC.

Organizations start their AI journey by using the open, freely available NGC libraries and frameworks to experiment and pilot. Now, when they're ready to move from pilot to production, enterprises can easily transition to a fully managed and secure AI platform with an NVIDIA AI Enterprise subscription. This gives enterprises deploying business critical AI, the assurance of business continuity with NVIDIA Enterprise Support and access to NVIDIA AI experts.

Table 1: Application Layer Software

Component	Description	Branch Type	NGC Catalog	Documentation
NVIDIA NIM	NVIDIA NIM provides microservices for accelerated AI model deployment.	Feature Branch (FB)	NIM Feature Branch (FB) on NGC Catalog	NVIDIA NIM Documentation
		Production Branch (PB)	NIM Production Branch (PB) on NGC Catalog	Production Branch (PB) Release Notes
Application Frameworks, AI Toolkits, SDKs and more	Building blocks and software tools to build AI workflows. Includes core AI and data science frameworks.	Feature Branch (FB)	Feature Branch (FB) on NGC Catalog	Feature Branch (FB) Release Notes Varies by framework/toolkit. Refer to the documentation links on the product pages on NGC Catalog.
		Production Branch (PB)	Production Branch (PB) on NGC Catalog	Production Branch (PB) Release Notes
		Long Term Supported Branch (LTSB)	Long Term Supported Branch (LTSB) on NGC Catalog	Long Term Supported Branch (LTSB) Release Notes
Production-Ready Pretrained Models	Pretrained AI models simplify and speed up development by eliminating the need to build from scratch.	Models available with NVIDIA AI Enterprise support.	Pretrained Models on NGC Catalog	Varies by model. Refer to the documentation links on the product pages on NGC Catalog.

Infrastructure Layer

The infrastructure layer includes various components that ensure efficient deployment, management, and scaling of AI applications. Key features include:

1. Versioning to maintain compatibility and stability across different deployments. Each version provides feature updates, security patches, and performance improvements.
2. Drivers to optimize utilization of NVIDIA GPUs and Networking in bare metal and virtualized environments.
3. Kubernetes operators for managing GPU and networking in containers and the lifecycle of microservices and AI pipelines.
4. Cluster Management software to provision and monitor servers at scale.

Infrastructure Layer Software

The NVIDIA AI Enterprise Infrastructure Release packages all software for managing and optimizing infrastructure and workloads.

Table 2: Infrastructure Layer Software

Component	Description	NGC Link	Documentation
NVIDIA Data Center Driver	Provides hardware support for NVIDIA GPUs. Consult the appropriate NVIDIA AI Enterprise Release Notes to see which GPUs and operating systems are supported by each version of the driver.	GPU Driver on NGC	NVIDIA Data Center Driver Documentation
NVIDIA vGPU (C-Series) Host	The NVIDIA driver is to be deployed in the hypervisor for virtualized environments.	NVIDIA vGPU (C-Series) Host	NVIDIA vGPU C-Series Documentation

Component	Description	NGC Link	Documentation
Driver		Driver on NGC	
NVIDIA vGPU (C-Series) Guest Driver	NVIDIA virtual GPU (vGPU) software driver is to be deployed in the VM or on a bare metal operating system to enable multiple virtual machines (VMs) to have simultaneous, direct access to a single physical GPU.	NVIDIA vGPU (C-Series) Guest Driver on NGC	NVIDIA vGPU C-Series Driver Documentation
NVIDIA DOCA Driver for Networking	Enables rapidly creating and managing applications and services on the BlueField networking platform, leveraging industry-standard APIs.	DOCA Driver on NGC	NVIDIA DOCA Drivers Documentation
GPU Operator	NVIDIA GPU Operator simplifies the deployment of NVIDIA AI Enterprise by automating the management of all NVIDIA software components needed to provision GPUs in Kubernetes.	GPU Operator on NGC	NVIDIA GPU Operator Documentation
Network Operator	NVIDIA Network Operator simplifies the provisioning and management of NVIDIA networking resources in a Kubernetes cluster.	Network Operator on NGC	NVIDIA Network Operator Documentation
NVIDIA NIM Operator	NVIDIA NIM Operator enables cluster administrators to operate the software components and services required to run LLM, embedding, and other models using NVIDIA NIM microservices in Kubernetes.	NIM Operator on NGC	NVIDIA NIM Operator Documentation
Base Command Manager	NVIDIA Base Command Manager streamlines cluster provisioning, workload management, and infrastructure monitoring across data centers and edge locations. It comprises the features of NVIDIA Base Command Manager that are certified for use with NVIDIA AI Enterprise.	Base Command Manager on NGC	NVIDIA Base Command Manager Documentation

Note

NVIDIA virtual GPU (vGPU) C-Series drivers allow virtual machines to utilize the full performance of GPUs and access advanced features such as sharing, live migration, and monitoring. The host driver is installed in the hypervisor of each physical host, while the guest driver is installed on each virtual machine.

NVIDIA AI Enterprise is certified to run across public cloud, data centers, workstations, DGX platform to edge. A complete list of supported configurations is listed in the NVIDIA AI Enterprise Infra Support Matrix.

Introduction

The NVIDIA AI Enterprise Software Reference Architecture (RA) provides an example infrastructure stack build that is geared towards OEMs and NVIDIA partners who intend to build systems that are ready for single-tenant production grade AI workloads. While hardware components of the infrastructure stack can be modular, the software components of the infrastructure stack are consistent for various workloads, e.g. Inference, Finetuning, & Retrieval Augmented Generation. There are many different ways to configure and optimize NVIDIA Systems, and this Software Reference Architecture is intended to be software & hardware agnostic while being updated for each NVIDIA AI Enterprise major release. This enables the NVIDIA partner ecosystem to provide additional value and enterprise customers to confidently deliver AI solutions faster, allowing them to focus on running their business rather than fighting deployments. Whether you choose to implement a full-fledged data center using our guidelines or adapt the node configurations with your own networking, the NVIDIA AI Enterprise Software Reference Architecture provides an invaluable starting point. The RA provides full-stack hardware and software recommendations for building high-performance, scalable, secure accelerated computing infrastructure and contains detailed guidance on optimal server, cluster, and network configurations for modern production AI workloads.

Last updated on Jul 01, 2025.

Compute Node Hardware

The Software Reference Architecture is comprised of individually optimized NVIDIA-Certified System servers that follow a prescriptive design pattern to ensure optimal performance when deployed in a cluster environment. There are currently three types of server configurations for which Enterprise RAs are designed: PCIe Optimized 2-4-3, PCIe Optimized 2-8-5, and HGX systems. For the PCIe Optimized configurations (for example, 2-8-5), the respective digits refer to the number of sockets (CPUs), the number of GPUs, and the number of network adapters. For Further details on Enterprise RA designs, refer to the NVIDIA Enterprise Reference Architecture Overview Whitepaper.

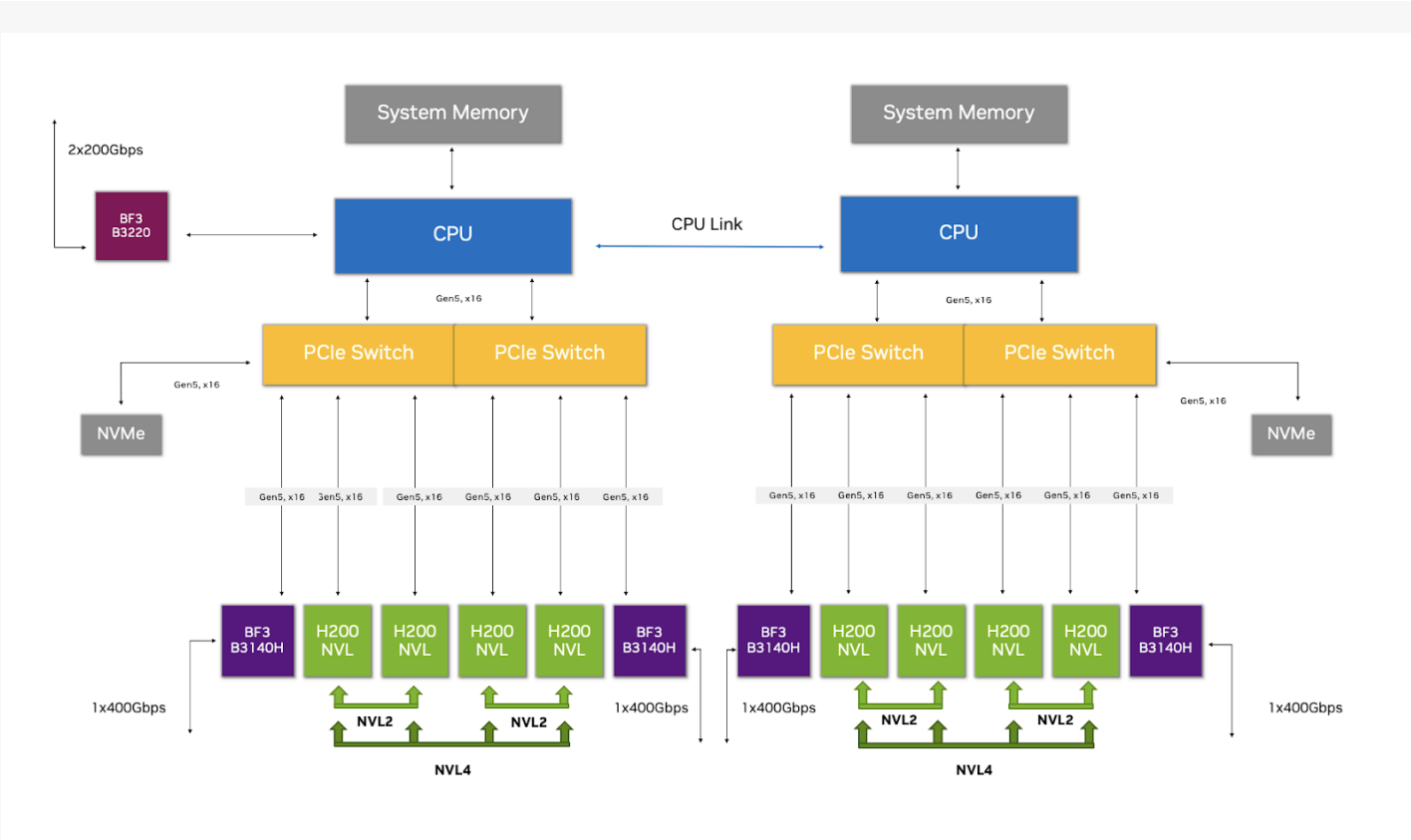
H200 NVL Systems

The NVIDIA AI Enterprise RA leverages an H200 NVL PCIe Optimized 2-8-5 reference configuration. Additional detailed reference configurations for 2-4-3 & 2-8-9 HGX systems can be made available via request. Additional types of systems with L40S, L40, L4, H100/H200 can also be used and will have different configurations

Note

The NVIDIA H200 NVL includes an NVIDIA AI Enterprise License and can be activated through NGC. Not all supported GPUs include an NVIDIA AI Enterprise License.

Diagram of PCIe Optimized 2-8-5 configuration with NVIDIA H200 NVL.



Components for the H200 NVL NVIDIA-Certified system are listed in the below table.

Parameter	System Configuration
GPU configuration	GPUs are balanced across CPU sockets and root ports. • Inference servers: 2x, 4x, and 8x GPUs per server • Training and DL servers: Minimum 8 GPUs per server

Parameter	System Configuration
	See the topology diagram above for details
NVLink Interconnect	H200 NVL supports NVL4 and NVL2 bridges. Pairing of GPU cards under the same CPU socket is best; pairing of GPU cards under different CPU sockets is acceptable but not recommended. See topology diagram above for NVLink bridging recommendations.
CPU	Intel Emerald Rapids, Intel Sapphire Rapids, Intel Granite Rapids and Intel Sierra Forest AMD Genoa and AMD Turin
CPU sockets	Two CPU sockets minimum
CPU speed	2.0 GHz minimum CPU clock
CPU cores	Minimum 7 physical CPU cores per GPU > • For configuration using MIG, 2 CPU cores required per MIG instance > • For OS kernel or virtualization, additional two cores per GPU
System memory (total across all CPU sockets)	Minimum 128 GB of system memory per GPU
DPU	One NVIDIA® BlueField®-3 DPU per serve

Parameter	System Configuration
PCI Express	One Gen5 x16 link per maximum two GPUs. Recommend one Gen5 x16 link per GPU
PCIe topology	Balanced PCIe topology with GPUs spread evenly across CPU sockets and PCIe root ports. NIC and NVMe drives should be under the same PCIe switch or PCIe root complex as the GPUs. Note that a PCIe switch may not be needed for low-cost inference servers; direct-attach to CPU is best if possible. See the topology diagram above for details
PCIe switches	Gen5 PCIe switches as needed (where additional link fanout is not required, direct attach is best).
Compute (E-W) NIC	Four NVIDIA® BlueField®-3 SuperNICs per server Up to 400 Gbps
Local storage	Local storage recommendations are as follows: • Inference Servers: Minimum 1 TB NVMe drive per CPU socket. • Training / DL Servers: Minimum 2 TB NVMe drive per CPU socket. • HPC Servers: Minimum 1 TB NVMe drive per CPU socket
Remote systems	SMBPBI over SMBus (OOB) protocol to BMC PLDM T5-enabled, SPDM-

Networking Physical Topologies

H200 NVL System Networking

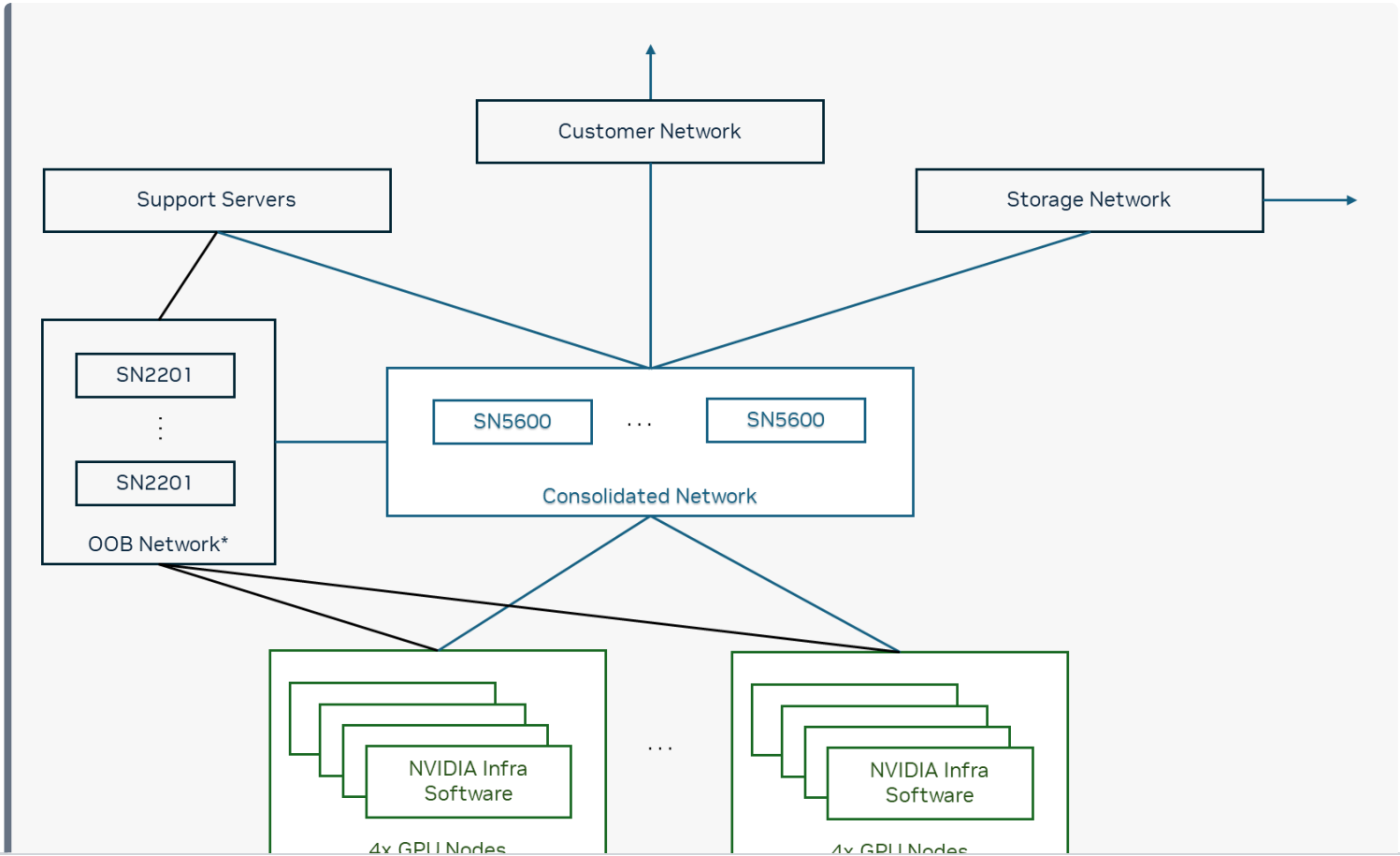
The NVIDIA platform networking configuration enables the highest AI performance and scale, while ensuring enterprise manageability and security. It leverages the NVIDIA expertise in AI data centers and optimizes network traffic flow:

- **East-West (Compute Network) traffic**: This refers to traffic between NVIDIA H200 NVL systems within the cluster, typically for multi-node AI training, HPC collective operations, and other workloads. These recommendations are critical for AI processing, handling internal data transfers that affect model training and scaling, requiring high bandwidth and low latency solutions. These are tailored for AI clusters to improve communication between GPUs and other components, ensuring seamless data flow within the data center. This is critical for scaling as data is processed and passed between various layers in AI models (across GPUs, CPUs, and storage). Poorly managed east-west traffic can lead to bottlenecks, slowing down training times and reducing the overall efficiency of the AI pipeline.
- **North-South (Customer and Storage Network) traffic**: This involves traffic between NVIDIA H200 NVL systems and any external resources including cloud management and orchestration systems, remote data storage nodes, and other parts of the data center or the Internet. Supports external communication and is especially important for storage connectivity for data ingestion and result delivery. Presently, NVIDIA recommends NVIDIA

BlueField Data Processing Units (DPUs) for all North-South traffic to offload and ensure secure, efficient handling of requests from outside the network.
- **Switching**: For all Enterprise RAs, NVIDIA provides configuration recommendations for Ethernet, which is the preferred switching for enterprise workloads.

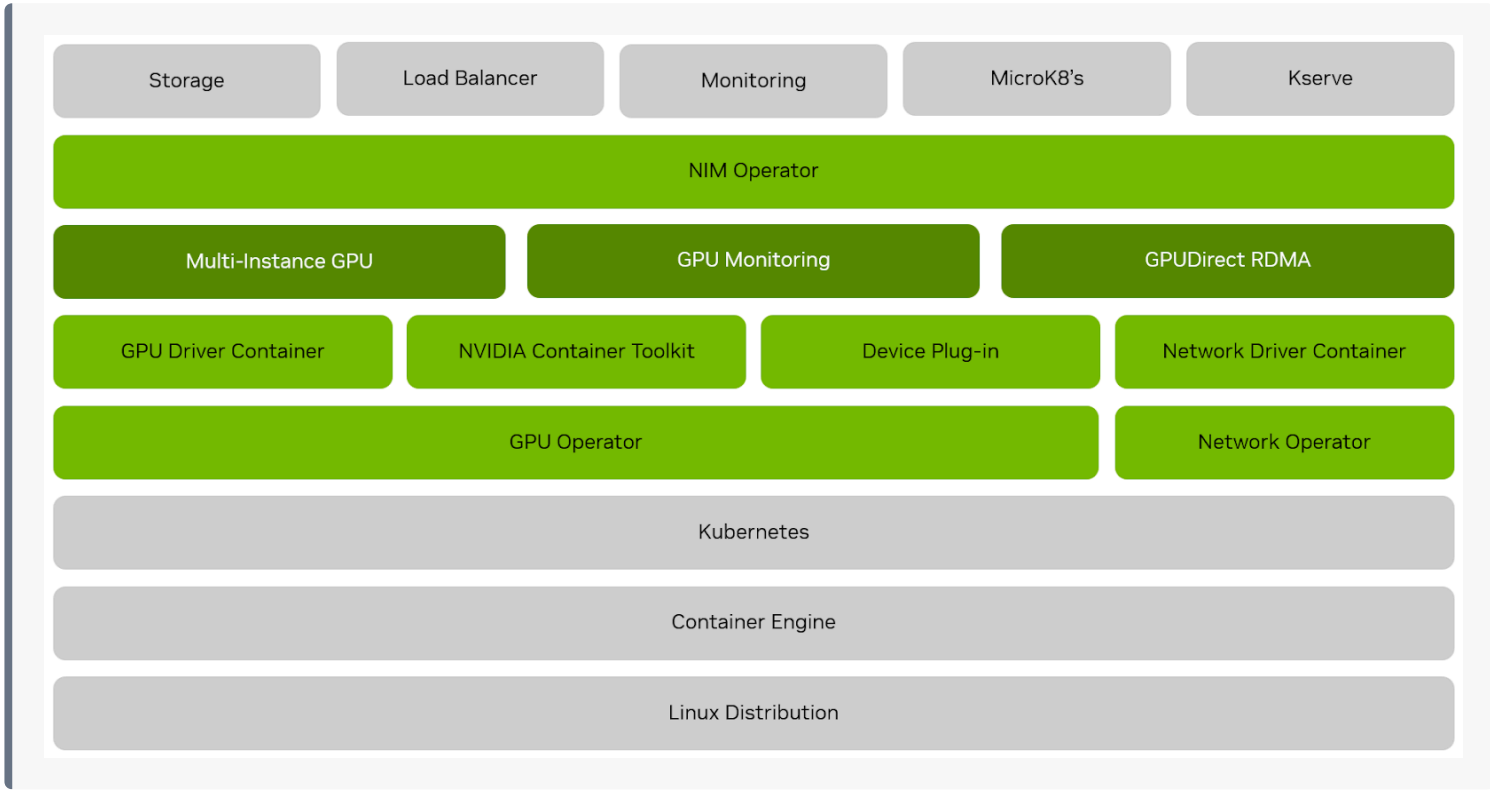
For optimal performance, it is recommended to use NVIDIA networking in conjunction with the NVIDIA H200 NVL platform. NVIDIA networking platforms provide end-to-end InfiniBand and Ethernet connectivity solutions. Combined with NVIDIA Spectrum-X Ethernet, NVIDIA H200 NVL platform delivers the highest performance for DL training and inference, data science, scientific simulation, and other modern workloads.

Network Topology Diagram



Software Stack

A sample software stack is provided as a fully supported agnostic starting point that is independent of the underlying NVIDIA Certified Systems (hardware). Other supported software can be found in the NVIDIA AI Enterprise Software Support Matrix. The example software stack provides examples for, Operating System, Orchestration Platform, Container Runtime, and NVIDIA Infrastructure Software.



Platform Software

The following platform software is used as an agnostic starting point for running NVIDIA AI Enterprise workloads.

Operating System	Ubuntu
Orchestration Platorm	Upstream Kubernetes
Compute Engine	Containerd

Supported versions of platform software for a given NVIDIA AI Enterprise Release can be found in the NVIDIA AI Enterprise Software Support Matrix.

NVIDIA Infrastructure Software

NVIDIA delivers infrastructure software for running workloads in development and production environments. The software used is hardware agnostic, i.e. the same software can be used regardless of the underlying hardware, netowrking, or reference architecture provided by NVIDIA for enterprise deplyoments.

- **The NVIDIA Driver, Container Toolkit, Kubernetes device plugin** are provisioned before GPU resources are available to the cluster. These software components allow the GPU to run on Kubernetes.
- **NVIDIA GPU Operator** automates the lifecycle management of the software required to use GPUs with Kubernetes. It takes care of the complexity that arises from managing the lifecycle of special resources like GPUs. It also handles all the configuration steps required to provision

NVIDIA GPUs, making them as easy to scale as other resources. Advanced features of GPU Operator allow for better performance, higher utilization, and access to GPU telemetry. Certified and validated for compatibility with industry leading Kubernetes solutions, GPU Operator allows organizations to focus on building applications, rather than managing Kubernetes infrastructure.

- **The NVIDIA Network Operator** simplifies the provisioning and management of NVIDIA networking resources in a Kubernetes cluster. The operator automatically installs the required host networking software - bringing together all the needed components to provide high-speed network connectivity. These components include the NVIDIA networking driver, Kubernetes device plugin, CNI plugins, IP address management (IPAM) plugin and others. The NVIDIA Network Operator works in conjunction with the NVIDIA GPU Operator to deliver high-throughput, low-latency networking for scale-out, GPU computing clusters. A Helm chart easily deploys the Network operator in a cluster to provision the host software on NVIDIA-enabled nodes.
- **The NVIDIA DOCA Driver for Networking** - is provisioned before network resources are available to the cluster. These software components allow the NIC, Smart NICs, & DPUs to run on Kubernetes.
- **GPUDirect® RDMA (GDR)** technology is a BlueField-3 feature that unlocks high-throughput, low-latency network connectivity to feed GPUs with data. GPUDirect RDMA allows efficient, zero-copy data transfers between GPUs using the hardware engines in the BlueField-3 ASIC.
- **GPUDirect Storage (GDS)** provides a direct path to local or remote storage (like NVMe or NVMe-oF) and GPU memory. BlueField-3 enables this direct communication within a distributed environment, when the GPU and storage media are not hosted in the same enclosure. BlueField-3 GDS provides increased bandwidth, lower latency, and increased capacity between storage and GPUs. This is especially important, as dataset sizes no longer fit into system

memory, and data IO to the GPUs becomes the runtime bottleneck. Enabling a direct path alleviates this bottleneck for scale-out AI and data science workloads.

- **NVIDIA NIM™**, part of NVIDIA AI Enterprise, is a set of easy-to-use microservices designed for secure, reliable deployment of high-performance AI model inferencing across workstations, data centers, and the cloud. Supporting a wide range of AI models, including open-source community and NVIDIA AI Foundation models, NVIDIA NIM ensures seamless, scalable AI inferencing, on-premises or in the cloud, leveraging industry-standard APIs.
- **NVIDIA NIM Operator**, automates the deployment and lifecycle management of generative AI applications built with NVIDIA NIM microservices on Kubernetes. NIM Operator delivers a better MLOps/LLMOps experience and improves performance by abstracting the deployment, configuration, and management of NIM microservices, allowing users to focus on the end to end application.

Supported versions of NVIDIA Infrastructure software for a given NVIDIA AI Enterprise Release can be found in the NVIDIA AI Enterprise Software Support Matrix.

> **Note**
>
> NVIDIA AI Enterprise includes additional software for building and running applications and is intended to run on top of the NVIDIA infrastructure software. A complete list of NVIDIA AI Enterprise supported software can be found on NGC.

Deployment Software

To install this software stack, NVIDIA provides the Cloud Native Stack (CNS) tooling to get started quickly. Additionally, enterprises can leverage Base Command Manager Essentials for software

Last updated on Jul 01, 2025.

Target Workloads

This RA provides optimal configuration for finetuning and inference of Large Language Model, as well as Traditional DL Inference Models. To run example workloads with NIM Operator refer to the Caching Models and NIM Services sections of the NVIDIA NIM Operator Documentation. Additionally, Sample RAG Application documentation is provided to extend the capabilities of NIM. Additionally, KServe may also be used to orchestrate and expose the APIs included in NIM. An example KServe implementation is provided on the CNS Github.

NVIDIA provides reference solutions for various AI use cases via NVIDIA Blueprints, some of which can leverage the K8's stack. NVIDIA also provides software for AI development with NVIDIA NeMo.

- **NVIDIA NeMo** is a set of microservices that help enterprise AI developers to easily curate data at scale, customize LLMs with the popular fine-tuning techniques, evaluate models on standard and custom benchmarks, and guardrail them for appropriate and grounded outputs.

  - **NeMo Curator:** A powerful microservice for enterprise developers to efficiently curate high-quality datasets for training LLMs, thereby enhancing model performance and accelerating the deployment of AI solutions.
  - **NeMo Customizer:** A high-performance, scalable microservice that simplifies the fine-tuning and alignment of LLMs with popular parameter efficient fine tuning techniques including LoRA, DPO.
  - **NeMo Evaluator:** An enterprise-grade microservice that provides industry-standard benchmarking of generative AI models, synthetic data generation, and end-to-end RAG pipelines.

- **NeMo Guardrails:** Microservice for developers to implement robust safety and security measures in LLM-based applications, ensuring that these applications remain reliable and aligned with organizational policies and guidelines.