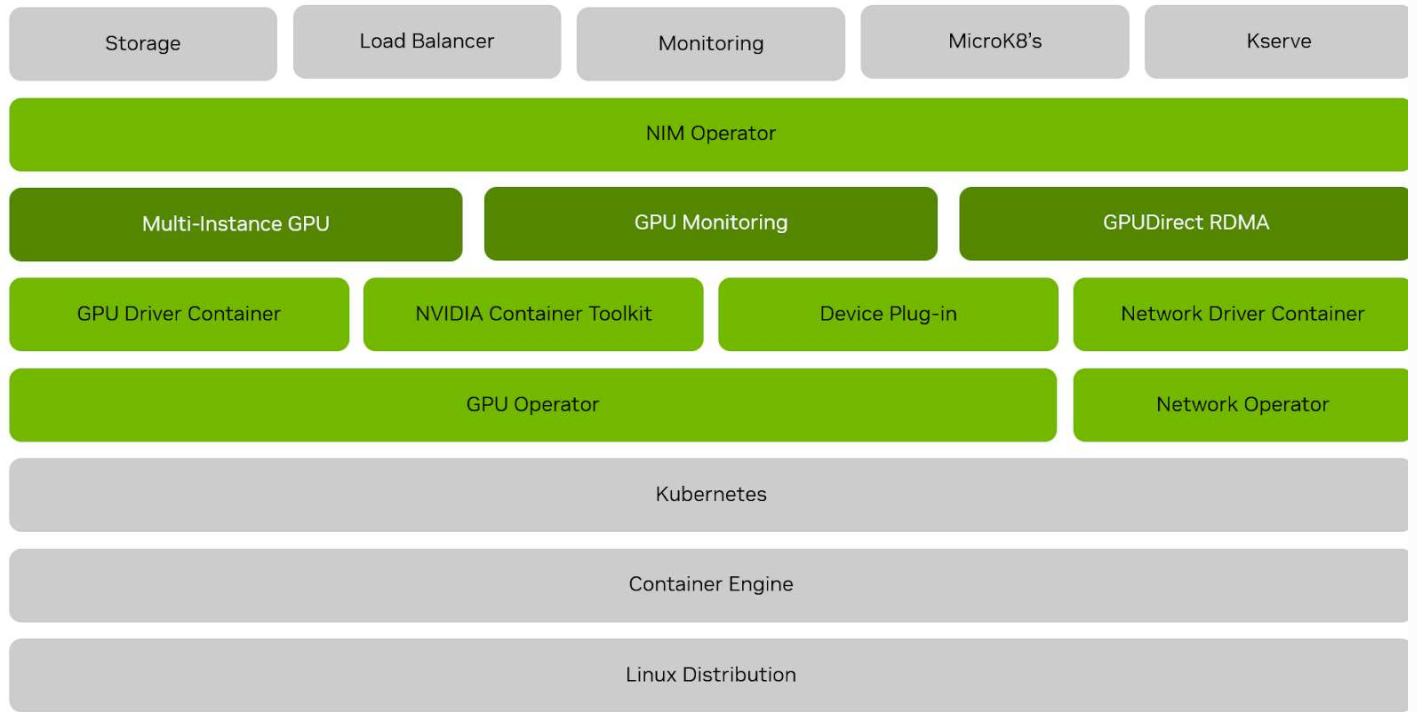


Software Stack

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 Platform	Upstream Kubernetes
Container Runtime	Containerd

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

The Role of Kubernetes

This architecture leverages Kubernetes, which is an open-source container orchestration platform that automates the deployment, scaling, and management of containerized applications across on-premises, cloud, and hybrid environments. Containers package applications with all required dependencies, enabling lightweight, portable, and consistent execution. As organizations started to scale many containers at once, Kubernetes emerged as the standard to allow a single administrator to manage large clusters efficiently. Kubernetes namespaces provide multi-tenancy within clusters, while NVIDIA GPU support and device plugins facilitate accelerated workloads in machine learning and data processing.

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, networking, or reference architecture provided by NVIDIA for enterprise deployments.

- **The NVIDIA Datacenter Driver, Container Toolkit, and 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 that is intended to run on top of the NVIDIA AI Enterprise Infrastructure software. A complete list of NVIDIA AI Enterprise-supported software can be found on [NGC](#).

Deployment Software

The NVIDIA [Cloud Native Stack](#) (CNS) provides the tooling necessary for the rapid deployment of NVIDIA AI Enterprise Infrastructure Software, such as NVIDIA Operators running on upstream Kubernetes.

Cloud-Native Stack enables developers to build, test, and run GPU-accelerated containerized applications. These applications can work seamlessly in production on enterprise Kubernetes-based platforms, such as NVIDIA Fleet Command, Red Hat OpenShift, and VMware vSphere.



[Privacy Policy](#) | [Your Privacy Choices](#) | [Terms of Service](#) | [Accessibility](#) | [Corporate Policies](#) | [Product Security](#) | [Contact](#)

Copyright © 2021-2026, NVIDIA Corporation.

Last updated on Mar 02, 2026.