NVIDIA-Certified Systems Configuration Guide for PCIe Servers

Design Guide

DG-10105-001 _v07 | January 2024

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

This PCIe server system configuration guide provides the server topology and system configuration recommendations for server designs that integrate NVIDIA® PCIe form factor graphics processing units (GPUs) from the Ampere GPU architecture. (NVIDIA HGX™ system configurations are covered in the NVIDIA HGX-specific collateral.)

The optimal PCIe server configuration depends on the target workloads (or applications) for that server. While in some cases, the server design can be optimized for a particular use case or target workload, generally speaking, a GPU server can be configured to execute the following types of applications or target workloads:

- Large Language Models (LLM)
- Natural Language Recognition (NLR)
- Omniverse applications
- Inference and Intelligent Video Analytics (IVA)
- Deep learning (DL) training / AI
- High-performance computing (HPC)
- Aerial Edge AI and 5G vRAN
- Cloud gaming
- Rendering and virtual workstation
- Virtual Desktop Infrastructure (VDI)
- Virtual workstation (vWS)
- Transcoding

Note: This system configuration guide provides system design recommendations for inference, IVA, DL training, transcoding, virtual desktop/workstation, and HPC applications. The other target workloads listed in this chapter are discussed in separate documents.

1.1. NVIDIA-Certified Systems

Servers that meet the recommendations provided in this design guide should be submitted for testing as an NVIDIA-Certified System™. NVIDIA certification covers

both single and multi-node configurations. Consult the NVIDIA-Certified systems documentation for more details regarding that program.

1.2. References

The following public GPU documentation is available to end customers:

- H100 HGX
- H100
- L40s
- L40
- L4

1.3. Initial Discussion Points

These discussion points outline architectural considerations for a solution between sales and end customers. They are not required for NVIDIA-Certified Systems certification testing but to serve as conversation starters for specific configuration and design deep dives.

1.3.1. Application Considerations

1.3.1.1. Workload Types

Generally, you can configure a GPU server to execute many different workloads. Some examples would be:

- Artificial Intelligence
  - Inferencing
    - NVIDIA TensorRT-LLM (Large Language Models)
    - NVIDIA Triton
    - DeepStream Intelligent Video Analytics (IVA)
    - Natural Language Recognition (NLR)
  - Training
    - Large Language Model Training (LLM)
    - Natural Language Processing (NLP)
    - Recommender Training
    - Deep Learning (DL) training / AI
    - Computer Vision Training

    - High-performance computing (HPC)
    - RAPIDS
    - TensorFlow
    - ResNet
    - HPC-SDK
    - Spark
- Omniverse and Visualization
    - Blender
    - V-Ray
    - Octane
    - Redshift
    - OGL
    - Rendering
    - Simulation
    - 3d Design Collaboration
- Virtual Desktop Infrastructure

Many workloads can leverage NVIDIA NeMo™ as an end-to-end, cloud-native enterprise framework for developers to build, customize, and deploy generative AI models with billions of parameters. For more information, refer to Nemo Framework for Generative AI - Get Started | NVIDIA Developer.

1.3.1.2. Workload Sizing

The size of your application workload, datasets, models, and specific case use will impact your hardware selections and deployment considerations. This guide will only provide an overview of options as a starting point. Additional cluster sizing overviews are available in the NVIDIA AI Enterprise Sizing Guide. Please discuss specific requirements with your provider to ensure your solution will meet your business needs.

1.3.2. GPU Scalability Considerations

Enterprise hardware can be designed to fit your AI application case use.

1.3.2.1. Single GPU

An application or workload has access to the entire GPU.

1.3.2.2. Multi-Instance GPU Partitioning

Certain GPUs can be run as a single unit or partitioned into multiple virtual GPUs to support multiple parallel threads. The most straightforward example is during Virtual Desktop Infrastructure deployments, where a single server can run multiple virtualized workstations using a single GPU. Other workloads can also take advantage of this if required as long as the tradeoffs in performance are considered.

1.3.2.3. Multiple GPUs

Having multiple GPUs within a single server allows a greater range of hardware acceleration. Depending on your needs, these GPUs can be MIG capable, shared across multiple workloads, or dedicated to a high-performance computing workload within that server. An application or workload has access to the entire GPU.

1.3.2.4. Single Node Workloads

Single Node workloads are a deployment pattern designed around being able to allocate resources within a single server or workstation. This can mean training or inferencing on a single server, using the entire system, or partitioning the GPU to run multiple applications all within the same node. There may be options to upgrade resources by adding additional GPU, CPU, or memory within that server, but these solutions typically do not scale to the cluster level. Single node deployments do not require high-speed networking to connect multiple nodes for your AI workload.

1.3.2.5. Clustered Workloads

Workload clustering is an application deployment pattern designed around being able to allocate additional resources across multiple servers. This means multiple nodes are connected with high-speed networking (either InfiniBand or RoCE) or via NVLink and NVSwitch to allow the workload to spread resources across multiple nodes in a cluster. Much like the considerations of how your application workload processes threads on a single GPU, MIG partition, or multiple GPUs on a single server, your workload can also do processing across multiple GPUs on multiple servers, at multiple locations to run the most complex high-performance workloads.

Note: This system configuration guide provides high-level design recommendations. Other target workloads are discussed in separate documents. Your individual needs and constraints may need additional consideration.

- NVIDIA GPUDirect Storage Benchmarking and Configuration Guide
- Recommender Systems
- Riva Hardware

1.3.3. Deployment Considerations

AI workloads can be deployed in multiple locations depending on the requirement for the application and case use. The specific requirement for your case use will help guide your hardware needs. The following sections describe example locations.

1.3.3.1. Data Center

Data Centers (DC) encompass the standard IT infrastructure location. These typically include servers deployed into Racks with Top-Of-Rack switches (TOR) connecting multiple racks within a Row. Rows are laid out with hot and cold aisles to service the hardware.

1.3.3.2. Edge

Edge is a very broad term, but you can think of this as anywhere that is not a standard data center. It can be in the backroom at a retail location, a cellphone tower, or a manufacturing site. You can even have edge-based data centers. In general, edge refers to locations not connected to your organization's backbone networking infrastructure. There are two types of edge systems, each with its own distinct characteristics:

- Enterprise Edge
- Industrial Edge

These are discussed in the following sections.

1.3.3.3. Enterprise Edge

Enterprise edge locations cover non-standard data center locations and include Remote Management capabilities found in the data center. Often the same servers can be found in standard data centers or edge locations. These systems are usually based on traditional enterprise servers and have been adapted for use in edge applications. They are typically intended for use in temperature-controlled environments.

1.3.3.4. Industrial Edge

Industrial Edge locations would cover applications where non-standard DC management capabilities traditionally do not exist, such as factory floors or cell phone towers. Systems deployed to Industrial locations tend to have more rigorous thermal and shock and vibe testing to handle a range of applications that a standard server in a DC would not be able to tolerate. These systems are ruggedized industrial PCs or other specialized devices deployed on-premises or vehicles. They are specifically designed for the environment in which they are deployed.

1.3.3.5. Workstations

Workstations are your typical desktop and laptop computers.

1.3.3.6. Desktop Workstations

These are tower-based systems designed only to move around a few times.

1.3.3.7. Mobile Workstations

These are typically laptop-based systems designed around mobility.

1.3.4. Security Considerations

Security becomes paramount as your accelerated workloads scale beyond the traditional data center. Specific security recommendations are beyond the scope of this guide, but the following features are validated as part of the certification process.

1.3.4.1. Trusted Platform Module

NVIDIA-Certified systems are tested for TPM 2.0 modules. TPM is an international security standard that allows for Platform Integrity, Disk Encryption, and system identification and attestation.

1.3.4.2. Unified Extensible Firmware Interface

UEFI is a public specification that replaces the legacy Basic Input/Output System (BIOS) boot firmware. NVIDIA-Certified systems are tested for UEFI bootloader compatibility.

1.3.5. Thermal Considerations

NVIDIA Certified Systems are qualified and tested to run workloads within the OEM manufacturer's temperature and airflow specifications.

Industrial-certified systems are tested at the OEM's maximum supported temperature.

Component temperature can impact workload performance, which in turn is affected by environmental, airflow, and hardware selections. When building a solution to ensure optimal performance, consider these variables.

Please review the GPU Thermal section under Design Discussion for more information on specific model thresholds.

Chapter 2. Configurations

2.1. Inference System Configurations

Inference application performance is greatly accelerated with the use of NVIDIA GPUs and includes workloads such as:

- Large Language Model Inference
- Natural Language Recognition (NLR)
- Omniverse applications
- DeepStream – GPU-accelerated Intelligent Video Analytics (IVA)
- NVIDIA® TensorRT™, Triton – inference software with GPU acceleration

A GPU server designed for executing inference workloads can be deployed at the edge or in the data center. Each server location has its own set of environmental and compliance requirements. For example, an edge server may require NEBS compliance with more stringent thermal and mechanical requirements.

Table 1 provides the system configuration requirements for an inference server using NVIDIA GPUs. Large Language Models should target the higher-end specs. Omniverse and visualization application usage will need L40S/L40.

Table 1. Inference Server System Configuration

Parameter	Inference Server Configuration
GPU	L40S L40 L4 H100 H100 HGX
GPU Configuration	2x / 4x / 8x GPUs per server 4x is recommended to remove the need for a PCIe switch. GPUs should be balanced across CPU sockets and root ports.

Parameter	Inference Server Configuration
CPU	x86 PCIe Gen5 capable CPUs are recommended, such as Intel Xeon scalable processor (Sapphire Rapids) or AMD Genoa.
CPU Sockets	2 CPU sockets minimum
CPU Speed	2.1 GHz minimum base clock
CPU Cores	6x physical CPU cores per GPU
System Memory	Minimum 1.5x of total GPU memory / 2.0x is recommended. Evenly spread across all CPU sockets and memory channels.
DPU	One Bluefield®-3 DPU per server
PCI Express	Minimum of one Gen5 x16 link per Gen5 GPU is recommended. Minimum of one Gen4 x16 link per Gen4 GPU is recommended. Minimum of one Gen5 x16 link per 2x GPUs for PCIe Switch configurations.
PCIe Topology	For balanced PCIe architecture, GPUs should be evenly distributed across CPU sockets and PCIe root ports. NICs and NVMe drives should be placed within the same PCIe switch or root complex as the GPUs. It's important to note that a PCIe switch may be optional for cost-effective inference servers.
PCIe Switches	Direct CPU attach is preferred. ConnectX®-7 Gen5 PCIE Switches as needed.
Network Adapter (NIC)	ConnectX®-7 (up to 400 Gbps) BlueField®-3 DPU in NIC mode (up to 400 Gbps). See Section Network for details.
NIC Speed	Up to 400 Gbps per GPU Minimum 50 Gbps per GPU for single-node inference Minimum 200 Gbps for multi-node inference
Storage	One NVMe per CPU socket
Remote Systems Management	Redfish 1.0 (or greater) compatible
Security	TPM 2.0 module (secure boot)

2.2. Deep Learning Training System Configurations

Deep Learning (DL) training application performance is greatly accelerated by the use of NVIDIA GPUs and includes workloads such as:

- NVIDIA TensorRT-LLM (Large Language Model Training)
- Recommender Training

- Natural Language Processing Training
- Computer Vision Training

GPU servers optimized for training workloads are usually located in data centers. Each data center or Cloud Service Provider (CSP) may have their own environmental and compliance standards, but these tend to be less strict than the requirements for NEBS or edge servers.

Table 2 provides the system configuration requirements for a DL training server using NVIDIA GPUs.

Table 2. Training Server - System Configuration

Parameter	Deep Learning Server Configuration
NVIDIA GPU	L40S L40 H100 H100 NVL
GPU Configuration	2x / 4x / 8x GPUs per server GPUs are balanced across CPU sockets and root ports. See topology diagrams for details.
CPU	Gen5 CPUs are recommended.
CPU Sockets	2 CPU sockets minimum
CPU Speed	2.1 GHz minimum base clock
CPU Cores	6x physical CPU cores per GPU (minimum)
System Memory	Minimum 1.5x of total GPU memory / 2.0x is recommended. Evenly spread across all CPU sockets and memory channels.
DPU	One BlueField-3 DPU per server
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. See topology diagrams for details.
PCIe Switches	ConnectX®-7 Gen5 PCIE Switches as needed.
Network Adapter (NIC)	ConnectX®-7 (up to 400 Gbps) BlueField® -3 DPU in NIC mode recommended (includes 400 Gbps NIC and PCIe switch). See Section Network for details.
NIC Speed	Up to 400 Gbps
Storage	One NVMe drive per CPU socket

Parameter	Deep Learning Server Configuration
Remote Systems Management	Redfish 1.0 (or greater) compatible
Security	TPM 2.0 module (secure boot)

2.3. Virtual Desktop Infrastructure System Configuration

Virtual Desktop Infrastructure (VDI) delivers a true workstation experience served from a data center machine. With VDI, IT can enable users to transition seamlessly between the office, home, or elsewhere.

The NVIDIA L4 paired with NVIDIA Virtual PC (or vPC) software enables the flexible work environment of the future. IT can provide a user experience virtually indistinguishable from a physical PC, even with multimedia-rich applications like video conferencing. Plus, IT can quickly support new users by provisioning new virtual desktops within minutes.

The VDI workloads for each user may vary based on factors like the number and types of applications, file sizes, number of monitors, and their resolution. NVIDIA recommends testing specific workloads to determine the best NVIDIA virtual GPU (vGPU) solution for your needs. The most successful customer deployments begin with a proof of concept (POC) and are continuously optimized throughout their lifecycle.

A GPU server optimized for VDI is usually located in a data center. Each data center or Cloud Service Provider (CSP) may have their own environmental and compliance standards, but these are usually less strict than the requirements for NEBS or edge servers.

Table 3 provides the system configuration requirements for a VDI server using NVIDIA L4 PCIe server cards.

Table 3. VDI Server System Configuration

Parameter	VDI Server Configuration
NVIDIA GPU	A16
GPU Configuration	1x / 2x / 3x / 4x GPUs per server GPUs are balanced across CPU sockets and root ports. See topology diagrams for details.
CPU	Gen5 CPUs are recommended. Gen4 CPUs are acceptable
CPU Sockets	1P / 2P
CPU Speed	2.1 GHz minimum base clock
CPU Cores	2-4 virtual CPU cores per user (oversubscription expected)

Parameter	VDI Server Configuration
System Memory	16 GB RAM per user
PCI Express	Gen4 / Gen5
PCIe Topology	Balanced PCIe topology with GPUs spread evenly across CPU sockets and PCIe root ports is preferred.
Driver	Virtual GPU 13.0 or later (NVIDIA Virtual PC License)
PCIe Switches	Broadcom PEX88000 Gen4 and Microsemi PM40100 Gen4 PCIe Switches
Network Adapter (NIC)	ConnectX-7 BlueField-3 ConnectX-6 LX ConnectX-6 ConnectX-4 LX See Section Network for details.
NIC Speed	Up to 400 Gbps
Storage	35 GB per user
Remote Systems Management	Redfish 1.0 (or greater) compatible
Security	TPM 2.0 module (secure boot)

2.4. Virtual Workstation System Configurations

NVIDIA RTX™ Virtual Workstation (vWS) is specialized NVIDIA software that enables designers and engineers to complete the most demanding work faster. With vWS, IT departments can virtualize any data center application with an experience effectively indistinguishable from that of a physical workstation. This feature enables workstation performance from any device.

vWS users are classified into "light use," "medium use," and "heavy use" categories. Light users require less graphical content and typically work with smaller model sizes. In entry-level vWS circumstances, an IT department must support mostly virtual PC users. However, an NVIDIA A16 can also be an affordable solution for a few light virtual workstation users.

For medium to heavy virtual workstation users who require up to the highest performance and work with the largest models and datasets, NVIDIA recommends the NVIDIA A40 PCIe card.

A GPU server optimized for vWS service is usually located in the data center. Each data center or cloud service provider (CSP) may have their own environmental and compliance standards, but these are usually less strict compared to the requirements for NEBS or edge servers.

Table 4 provides the system configuration requirements for a vWS server using NVIDIA A16 PCIe server cards.

Table 4. Entry Level vWS Server System Configuration

Parameter	Entry Level vWS Server Configuration
NVIDIA GPU Product	A16
GPU Configuration	1x / 2x / 3x / 4x A16 cards per server Max recommended is two L4 cards per CPU socket See topology diagrams for details
CPU	Gen5 CPUs are recommended. Gen4 CPUs are acceptable
CPU Sockets	1P / 2P
CPU Speed	2.8 GHz minimum
CPU Cores	4-6 virtual CPU cores per user (oversubscription expected)
System Memory	16 GB RAM per user
PCI Express	Gen4 / Gen5
PCIe Topology	Balanced PCIe topology with GPUs spread evenly across CPU sockets and PCIe root ports is preferred.
Driver	Virtual GPU 13.0 or later (NVIDIA RTX vWS License)
PCIe Switches	Broadcom PEX88000 Gen4 and Microsemi PM40100 Gen4 PCIe Switches
Network Adapter (NIC)	ConnectX-7 BlueField-3 ConnectX-6 LX ConnectX-6 ConnectX-4 LX See Section Network for details.
NIC Speed	Up to 400 Gbps
Storage	40 GB per user
Remote Systems Management	Redfish 1.0 (or greater) compatible
Security	TPM 2.0 module (secure boot)

2.5. Transcode System Configurations

Video content distributed to viewers is often transcoded into several adaptive bit rate (ABR) profiles for delivery. Content in production may arrive in one of the large numbers

of CODEC formats that must be transcoded into another for distribution or archiving. The recommended system configuration is for compressed video formats.

High-quality video applications' processing demands have pushed limits for broadcast and telecommunication networks. Live streaming will drive overall video data traffic growth for both cellular and Wi-Fi as consumers move beyond watching the on-demand video to viewing live streams.

NVIDIA GPUs integrally include an on-chip hardware encoder and decoder unit (often called NVENC and NVDEC). Separate from NVIDIA® CUDA® cores, NVENC and NVDEC execute encoding or decoding workloads without slowing the execution of graphics or CUDA workloads simultaneously.

A GPU server designed for transcoding is typically deployed in the data center. Each data center or cloud service provider (CSP) may have their own environmental and compliance requirements. However, those requirements are typically less stringent than NEBS or edge server requirements.

Table 5 provides the system configuration requirements for a Transcode server using NVIDIA A16 PCIe server cards.

Table 5. Transcode Server System Configuration

Parameter	Transcode Server Configuration
NVIDIA GPU Product	A16
GPU Configuration	1x / 2x / 3x / 4x A16 cards per server Max recommended is two L4 cards per CPU socket See topology diagrams for details
CPU	Gen5 CPUs are recommended. Gen4 CPUs are acceptable
CPU Sockets	1P / 2P
CPU Speed	2.1 GHz minimum
CPU Cores	4 CPU cores per L4 PCIe card
System Memory	1.5x to 2x GPU memory
PCI Express	Gen4 / Gen5
PCIe Topology	Balanced PCIe topology with GPUs spread evenly across CPU sockets and PCIe root ports is preferred.
Driver	R470 TRD1 or later
PCIe Switches	Broadcom PEX88000 Gen4 and Microsemi PM40100 Gen4 PCIe Switches
Network Adapter (NIC)	ConnectX-7 BlueField-3 ConnectX-6 LX

Parameter	Transcode Server Configuration
	ConnectX-6 ConnectX-4 LX See Section Network for details.
NIC Speed	Up to 400 Gbps
Storage	One NVMe per socket
Remote Systems Management	Redfish 1.0 (or greater) compatible
Security	TPM 2.0 module (secure boot)

2.6. Inference and Deep Learning Training Topology Diagrams

This chapter shows the system configurations that correspond to those outlined in Table 1 and Table 2 for inference and DL training servers, starting from the simplest configuration to the most complex.

Note that some server configurations may not require a PCIe switch. For example, depending on the number of PCIe lanes available from the CPU, a server with one or two GPUs per socket may not require a PCIe switch..

Figure 1. 2P Server with Two GPUs

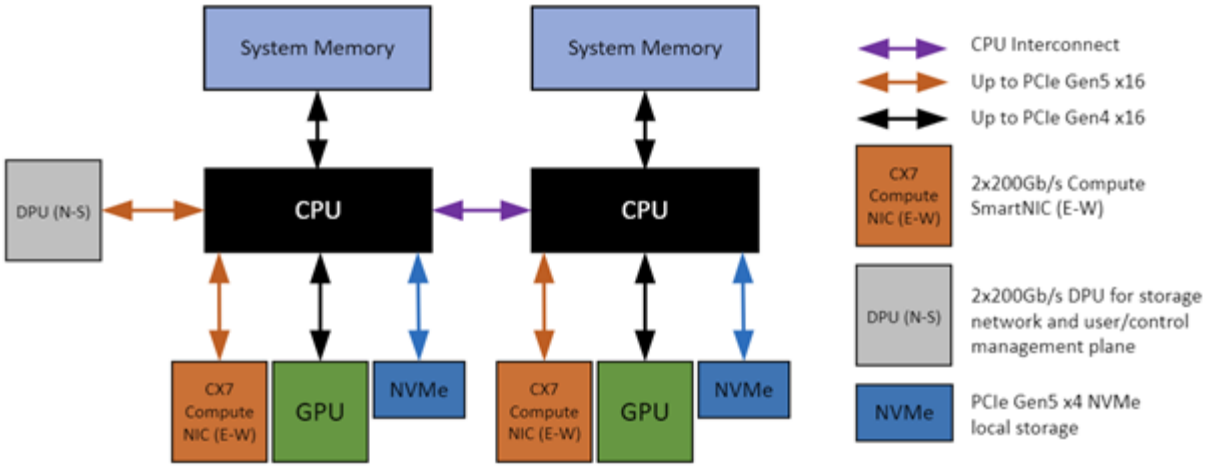


Figure 2. 2P Server with Four GPUs

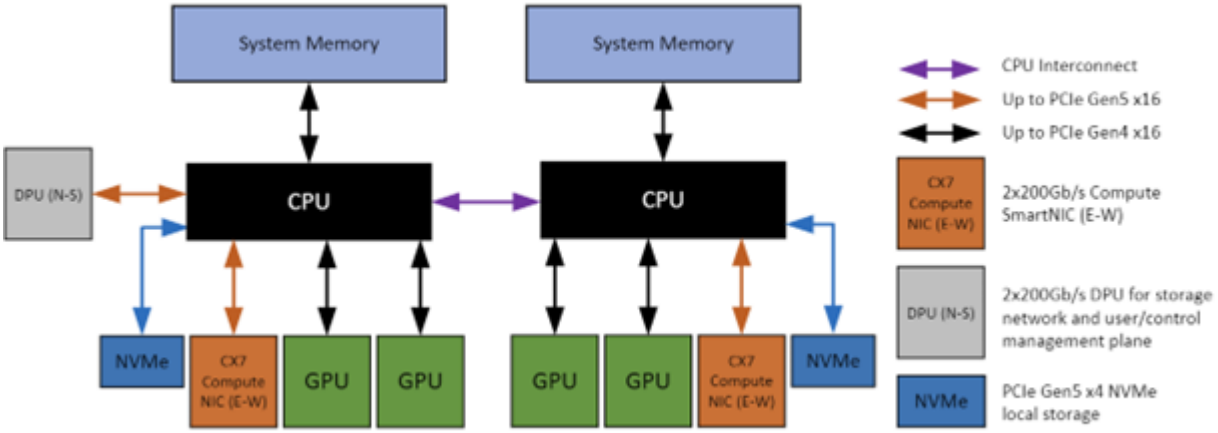
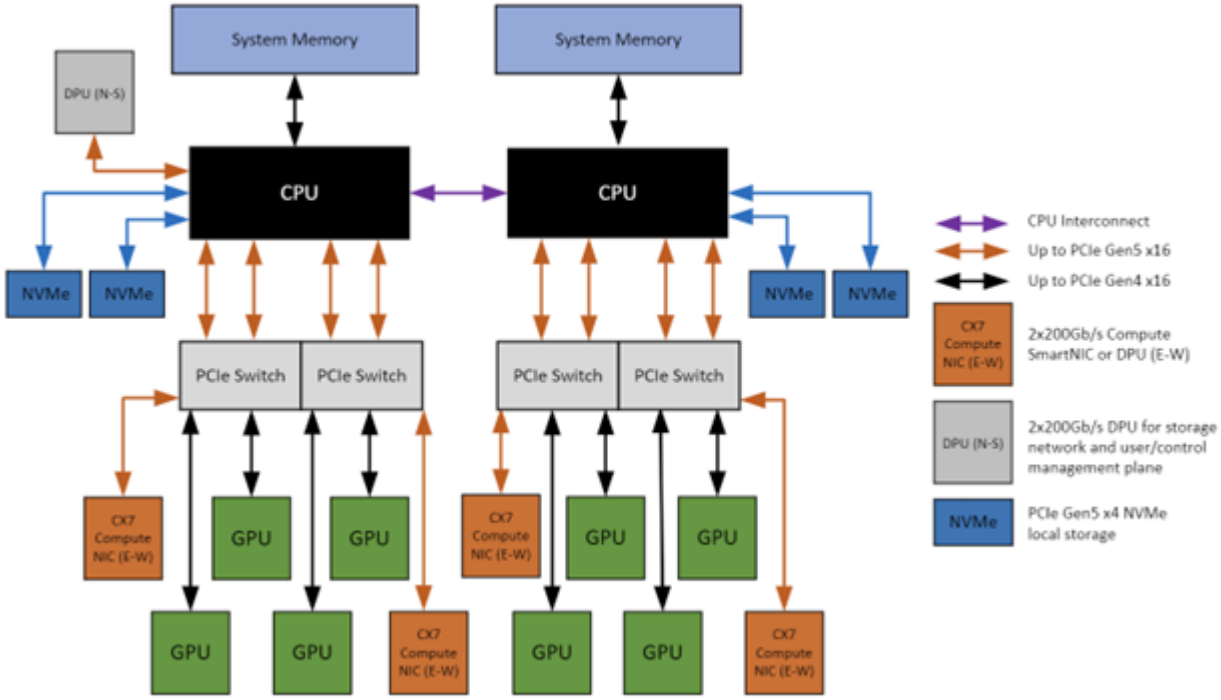


Figure 3. 2P Server with Eight GPUs and PCIe Switch



Chapter 3. Design Discussion

This chapter discusses design considerations for the NVIDIA GPU server cards identified in this design guide. When considering the NVIDIA A16 PCIe card, the term "GPU" in the following sections implies one NVIDIA A16 PCIe card.

3.1. GPU

GPUs are best deployed in servers in powers of two (1, 2, 4, or 8 GPUs), with the GPUs evenly distributed across CPU sockets and root ports. A balanced GPU configuration results in optimal GPU application performance and will avoid performance bottlenecks in critical interfaces such as PCIe and system memory.

3.2. GPU Power

GPU power requirements are dependent on the specific model. Refer to the table below for more information and links to model-specific details. Ensuring you have appropriate power helps ensure optimal performance of the GPU. Please work with your vendor for proper sizing of your PSU based on the GPU model and any other components in the system.

3.3. GPU Thermal

GPU thermal operating temperatures vary by model. Refer to the table below for more information. Note that GPU performance can be impacted by temperature. Restrictive airflow or environmental thermal issues can affect your workload. Please work with your vendor for appropriate thermal considerations based on your entire system requirements and design.

Table 6. GPU Power and Thermal

GPU	Typical Workload	Max Power Draw
L40S	Inference/Training	350 W
L40	Inference/Training	300 W

GPU	Typical Workload	Max Power Draw
L4	Inference	72 W
H100 HGX	Training	350 W per gpu
H100	Inference/Training	310-350 W[1]

Note:

1. Power wattage availability will vary depending on the OEM cabling implementation. Please check with your vendor for system-specific details.

3.4. CPU

Single-socket (1P) or dual-socket (2P) configurations are allowed. The primary design constraint for the CPU is to choose a configuration with sufficient CPU cores per GPU. At least six physical CPU cores per GPU is recommended.

The CPU configuration, 1P or 2P, depends on the number of GPUs used in the server and the number of cores available in either an AMD or Intel enterprise-class CPU. 1P server designs are typically used in edge server applications.

If you are interested in designing a 4P server with NVIDIA GPUs, you may contact NVIDIA for further assistance.

3.4.1. AMD PCIe Port Mapping

Servers using the AMD CPUs may suffer from IO performance issues if the correct PCIe ports are not used in the server configuration.

The general design guidelines for optimal IO performance and device-to-host bandwidth when using AMD Genoa CPUs are:

- 4 GPUs in 1P – use ports P0/P1/P2/P3
- 4 GPUs in 2P with PCIe Switch (4xGMI with 4G links)– use ports P0/P2
- 4 GPUs in 2P (4xGMI with 4G links)– use ports P0/P2 or P1/P3
- 4 GPUs in 2P (4xGMI with 2G+2P links – use ports P0/P2 or G1/G3
- 8 GPUs in 2P with inline GPUs – use ports P0/P1/P2/P3
- 8 GPUs in 2P with PCIE Switch (8x root ports) – use ports P0/P1 and P2/P3
- 8 GPUs in 2P with PCIE Switch (4xGMI with 4G links) – use ports P0/P2 or P1/P3
- 8 GPUs in 2P with PCIE Switch (4xGMI with 2G+2P links) – use ports P0/P2 or G1/G3

The port mappings for prior AMD CPU generations are different. The general design guidelines for optimal IO performance and device-to-host bandwidth when using AMD Rome and Milan CPUs are:

- 2 GPUs in 1P – use P0/P3 or P1/P2

- 4 GPUs in 1P – use ports P0/P1/P2/P3
- 4 GPUs in 2P – use P0/P3 or P1/P2 on both sockets
- 8 GPUs in 2P – use ports P0/P1/P2/P3

3.5. System Memory

Total system memory should be at least 1.5 times greater than the total amount of GPU frame buffer memory. At least 2.0 times greater is preferred for the best GPU application performance.

For example, in the server with four NVIDIA A100 80 GB GPUs – each with 80 GB of frame buffer memory – the total GPU memory is 320 GB. Thus, the total system memory should be at least 480 GB, and 640 GB is preferred.

The system memory should be evenly distributed across all CPU sockets and memory channels for optimal performance.

3.5.1. System Memory ECC

Note the following requirements and guidelines regarding system memory ECC:

- To achieve NVIDIA Certification, all datacenter and enterprise edge systems must have ECC capability on system memory.
- For NVIDIA Certification purposes, ECC capability on system memory is optional for Industrial Edge systems.
- On all systems capable of supporting system memory ECC, system memory ECC must be enabled during NVIDIA Certification testing. For the best user experience, it is recommended but not required that "system memory ECC enabled" be the default configuration for NVIDIA-Certified Systems.

3.6. PCIe Interface

This section details the PCIe interface, including speed and land width, topology, and port mapping.

3.6.1. PCIe Speed and Lane Width

PCIe Gen5 x16 to each GPU is recommended for maximum throughput.

- Different CPU models will support different PCIe generations. Depending on your CPU selection a Gen4 or Gen3 x16 connection to each GPU is acceptable.
- PCIe Gen5 or Gen4 x8 links to the GPU are also supported. However, GPU performance may be impacted for some applications.

Table 7. PCIe Throughput Comparison Table

Version	Introduced	Transfer Rate Per Lane	Throughput	
			x8	x16
3.0	2010	8.0 GT/s	7.877 GB/s	15.754 GB/s
4.0	2017	16.0 GT/s	15.754 GB/s	31.508 GB/s
5.0	2019	32.0 GT/s	31.508 GB/s	63.015 GB/s

Source: Wikipedia

3.6.2. PCIe Topology

For servers with one or two GPUs installed, it is likely that a PCIe switch is not needed to connect the PCIe devices (GPUs, NICs, and NVMe drives) to the host PCIe root ports. This is especially true with the AMD Rome and Milan CPUs and the upcoming Intel Ice Lake CPUs, which all provide sufficient PCIe Gen4 lanes for device connectivity.

- In general, it is strongly recommended to distribute the GPUs across the CPU sockets and root ports for a balanced configuration. One Gen5 or Gen4 x16 link per GPU is strongly recommended for best performance.
- As the number of GPUs and other PCIe devices increases for larger, higher-performance servers, PCIe switches may be needed to provide sufficient connectivity. In that case, the number of PCIe switch layers should be limited to one or two to minimize PCIe latency.
- If a PCIe switch is used, a maximum of two GPUs should share an upstream PCIe Gen 5 or Gen4 x16 link for optimal performance.

3.6.3. AMD PCIe Port Mapping

Refer to Section AMD PCIe Port Mapping for specific recommendations for using the PCIe ports on the AMD CPUs.

3.7. BIOS Settings and ECC

The following BIOS settings should be enabled on your system to ensure optimal performance for certification testing. CPU Virtualization may be required in environments where Virtual Machines are in use.

Description	Setting
ECC Memory (for systems with ECC memory)	Enabled
ACS	Disabled
IOMMU	Disabled[1] [1] Intel Sapphire Rapids requires IOMMU to be enabled
CPU Virtualization	Disabled

Description	Setting
PCIe ACS Enable (Access Control Services)	Enabled[1] [1] NVIDIA GPUDirect requires ACS to be disabled
PCIe Ten Bit Tag Support	Enabled
PCIe Relaxed Ordering (RO)	Enabled

3.8. PCIe Relaxed Ordering

PCIe relaxed ordering (RO) can improve performance by allowing the path between the Requestor and Completer to re-order some transactions received before other transactions which were previously enqueued. For data center applications, NVIDIA recommends the following PCIe RO settings for CPU, GPU, NVME SSDs, and PCIe Switches.

Description	Setting
CPU	Enabled
GPU	Default VBIOS Setting
NVMe SSD	Enabled if using GPUDirect Storage (GDS) Work with your NVMe SSD supplier to enable RO for PCIe writes when using GDS
Broadcom PCIe Switches	Enabled for PCIe read completions. For other PCIe switch vendors, please work with the vendor for optimal configuration and settings.

3.9. Network

NVIDIA recommends using a minimum of 200Gb/s InfiniBand and Ethernet CX7 and BlueField-3 data processing units (DPUs). The NVIDIA-Certified systems testing process requires a standardized testing environment that uses one of the following NVIDIA products:

- ConnectX-7
- ConnectX-7 Dx
- ConnectX-6
- ConnectX-6 Dx
- BlueField-3

East-West Network

- NVIDIA recommends the ConnectX-7 (CX7) smart network adapters to connect GPUs to the East-West (E-W) network, providing low-latency and high-bandwidth communication between GPUs within the AI cluster.

CX7 adapters also incorporate the NVIDIA® GPUDirect® RDMA that further optimizes data movement and allows direct memory access (DMA) between GPUs, bypassing the CPU and improving overall system performance.

- Recommended Compute Network Bandwidth per GPU
  - 200 GBps

North-South Network

This section describes the recommended BlueField-3 model for the North-South (N-S) network.

The NVIDIA BlueField-3 data processing unit (DPU) is a networking based infrastructure compute platform that enables organizations to securely deploy and operate NVIDIA GPU-enabled servers in AI cloud data centers at massive scales. BlueField-3 offers several essential capabilities and benefits within the NVIDIA GPU-enabled servers. For more information, refer to BlueField Data Processing Units (DPUs) | NVIDIA.

The table below lists the supported network and security protocols offloaded by NVIDIA networking cards.

Table 8. Supported Network Protocols and Offloaded Security Protocols

SmartNIC/DPU	Network Protocol	Security Offload Engine		
		IPsec	TLS	AES-XTS
BlueField-3	Ethernet/ InfiniBand	Yes	Yes	Yes
ConnectX-7	Ethernet/ InfiniBand	Yes	Yes	Yes
ConnectX-6	Ethernet/ InfiniBand	No	No	Yes
ConnectX-6 Dx	Ethernet	Yes	Yes	Yes

3.9.1. NIC Speeds

The following PCIe recommendations apply to the different NIC speeds used in the server design for best performance.

- Each 400 Gbps NIC needs Gen5 x16
- Each 200 Gbps NIC needs Gen4 x16
- Each 100 Gbps NIC needs Gen4 x8 or Gen3 x16

3.10. Storage

The size and scale of the application and underlying dataset will guide your storage requirements. Like networking, these are unique to the individual solution performance and constraints.

3.10.1. Certification Requirements

NVIDIA Certified Systems tests are run using internal NVME drives.

3.10.2. Internal Storage

Storage devices should be located under the same PCIE root complex as the GPU and NIC devices for optimal Non-Uniform Memory Access (NUMA) performance.

In server configurations that use PCIe switches, the NVMe drives should be located under the same PCIe switch as the GPU(s) and NICs for best NVIDIA® GPUDirect® Storage (GDS) performance.

3.10.3. External Storage

External Storage solutions are supported for production deployments but require additional consideration during your solution design. The same considerations for PCIe throughput, network cards, and switches that apply to a clustered GPU solution also apply for external/shared storage integration. Please consult with your server and storage solution partners to ensure your workload will perform as expected within your environment. External storage is not required for certification testing.

3.10.4. GPU Direct Storage

GDS is supported for NVMe drives located in the same server as the GPUs or in external storage servers. If internal NVMe drives are present in the server, they must be on the same PCIe switch or root complex to be used with GDS.

For more details on GPUDirect Storage, refer to the NVIDIA GPUDirect Storage Documentation.

3.11. Remote Management

While Remote Management is not a requirement for any particular AI application workload, there are scenarios where remote systems management capabilities are highly desirable.

3.11.1. In-Band

In-band management solutions refer to something happening within the Operating System. These will typically revolve around GPU-oriented telemetry and performance, where the OS can access the GPU via the GPU drivers.

3.11.1.1. NVIDIA System Management Interface

The NVIDIA System Management Interface (nvidia-smi) is a command line utility based on top of the NVIDIA Management Library (NVML), intended to aid in managing and monitoring NVIDIA GPU devices.

This utility allows administrators to query the GPU device state and permits administrators to modify the GPU device state with the appropriate privileges. It targets the Tesla™, GRID™, Quadro™, and Titan X products, though limited support is also available on other NVIDIA GPUs.

For more details, please refer to thenvidia-smi documentation.

3.11.1.2. NVIDIA Data Center GPU Manager

The NVIDIA Data Center GPU Manager (DCGM) is a suite of tools for managing and monitoring NVIDIA data center GPUs in cluster environments. The suite includes active health monitoring, comprehensive diagnostics, system alerts, and governance policies, including power and clock management. Infrastructure teams can use it as a standalone tool or easily integrate it with cluster management tools, resource scheduling, and monitoring products from NVIDIA's partners.

3.11.2. Out-of-Band

Baseboard Management Controllers (BMC) provide an external interface to manage servers from a central location over what is referred to as an Out Of Band channel. BMC software runs independently of the Operating System. Specific BMC features will vary across OEMs, but you can generally expect hardware-level management (power on/off, serial over LAN, and telemetry features to be available. While the OS may have external access over the internet, your BMC management port will typically not be exposed due to the significant security risks this presents. A BMC communicates to the underlying server hardware via the I2C or SMBUS interfaces through either IPMI or Redfish.

NVIDIA-Certified Systems are tested for both IPMI and Redfish management capabilities.

3.11.2.1. Intelligent Platform Management Interface

The Intelligent Platform Management Interface, or IPMI, is a standard protocol for remote systems management. This is typically considered a legacy approach to remote hardware management.

3.11.2.2. Redfish

The Redfish standard is a suite of specifications developed by the DMTF organization that delivers an industry-standard protocol providing a RESTful interface for out-of-band management of modern servers, storage, networking, and converged infrastructure.

NVIDIA participates in developing Redfish as a Leadership Partner within the DMTF organization, helping to drive GPU modeling and management standards across the industry.

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation ("NVIDIA") makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, BlueField, ConnectX, CUDA, GPUDirect, NVIDIA-Certified Systems, NVIDIA HGX, NVIDIA RTX, and TensorRT are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Arm

Arm, AMBA and Arm Powered are registered trademarks of Arm Limited. Cortex, MPCore and Mali are trademarks of Arm Limited. All other brands or product names are the property of their respective holders. "Arm" is used to represent Arm Holdings plc; its operating company Arm Limited; and the regional subsidiaries Arm Inc.; Arm KK; Arm Korea Limited.; Arm Taiwan Limited; Arm France SAS; Arm Consulting (Shanghai) Co. Ltd.; Arm Germany GmbH; Arm Embedded Technologies Pvt. Ltd.; Arm Norway, AS and Arm Sweden AB.

Copyright

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
www.nvidia.com