NVIDIA-Certified Professional: AI Infrastructure Exam Study Guide

NVIDIA-Certified Professional: AI Infrastructure Exam Study Guide

Contents

This study guide provides an overview of each topic covered on the NVIDIA AI Infrastructure certification exam, as well as recommended training and suggested reading to help prepare for the exam.

Information about NVIDIA certifications can be found here.

Job Description

A professional who passes the NVIDIA NCP™-AII exam is skilled in deploying, configuring, and validating advanced NVIDIA AI infrastructure. They manage end-to-end system bring-up, physical installation, and networking, handle NVIDIA BlueField® and MIG configurations, install and maintain control plane and drivers, perform comprehensive cluster testing, storage validation, troubleshooting, and ongoing optimization for complex, GPU-powered environments.

Job Responsibilities

1. Lead deployment and validation of servers and systems for AI factories.
2. Configure and manage network topologies, BMC, OOB, TPM, power, and cooling.
3. Install, upgrade, and validate GPU-based servers, BlueField DPUs, cables, and transceivers.
4. Perform firmware upgrades, hardware validation, and storage setup.
5. Configure and administer physical and logical resources, including MIG partitioning and BlueField platforms.
6. Install and configure operating systems, cluster software, drivers, containers (Docker), and NGC CLI.
7. Manage and orchestrate clusters using NVIDIA Base Command™ Manager, Slurm, Pyxis, Enroot, and Run:ai.
8. Perform stress, benchmarking, and burn-in tests using HPL, NCCL, NVIDIA Nemo™, and ClusterKit.
9. Verify cabling, firmware/software versions, and network signal quality.
10. Troubleshoot and resolve hardware, software, storage, and performance faults.
11. Replace faulty components and optimize systems for AMD/Intel platforms.
12. Monitor, document, and report on cluster health, resource usage, and job performance.
13. Ensure secure, efficient, and scalable operation of NVIDIA AI infrastructure, including user access and workload management.

Recommended Qualifications and Experience

1. Bachelor’s degree in computer science, software engineering, AI, or a related field
2. Expertise in NVIDIA GPU/DPU technologies, AI software stacks, and data center management for high-performance AI workloads.

Certification Topics and References

Systems and Server Bring-Up Exam Weight 31%

These tasks involve end-to-end hardware deployment: rack, power, BMC, security, firmware, physical installation, initial server and network setup, cable validation, and hardware verification for AI workloads. Ensures all components are correctly deployed, configured, and operational before scaling AI operations.

1.1	Describe the sequence of events for deployment and validation
1.2	Describe network topologies for AI Factories
1.3	Perform initial configuration of BMC, OOB, and TPM
1.4	Perform firmware upgrades (including on NVIDIA HGX™ systems) and fault detection
1.5	Validate power and cooling parameter
1.6	Install GPU-based servers (SMI)
1.7	Validate installed hardware
1.8	Describe and validate cable types and transceivers
1.9	Install physical GPUs
1.10	Validate hardware operation for workloads
1.11	Configure initial parameters for third party storage

Recommended Training (Optional)

Course reference: AI Infrastructure Professional

Covered by these course sections:

- AI in the Data Center Overview
- Compute Platforms for AI
- BlueField Networking Platform – Overview, Bring-Up, Firmware, Management
- AI Data Center Management
- Practice: Bringing Up an AI cluster With BCM

Suggested Readings

- NVIDIA System Management Interface
- NVIDIA System Management User Guide
- NVIDIA CUDA® Compiler Driver NVCC
- NVIDIA Virtual GPU Software
- InfiniBand Fabric Utilities
- Getting Started With the NGC Command-Line Interface (CLI)
- NVIDIA LinkX® Cables and Transceivers
- DGX Basepod Deployment Guide
- NVIDIA DGX H100/H200 User Guide: Quickstart and Basic Operation
- AI Factory Whitepaper
- DGX Superpod Deployment Guide
- DGX Superpod Administration Guide
- DGX CentOS Install Guide
- Choosing the Right Storage: Blog
- Protecting Sensitive Data and AI Models: Blog
- Choosing the Right Storage for Enterprise AI Workloads: Blog

- DGX Superpod Data Center Design
- Datacenter Efficiency Metrics
- DGXOS User Guide
- CUDACuda Compiler Driver
- Using NVSM
- NVIDIA Networking
- DGX Superpod Design Guide
- Cabling Data Centers
- DGX H100 User Guide
- DGX A100 Service Manual

Physical Layer Management: Exam Weight 5%

This exam topic focuses on configuring and maintaining physical resources, including GPU/BlueField DPU networks, cable and transceiver management, and GPU partitioning with MIG. Ensures all physical components support secure, scalable, high-performance AI data center operations.

2.1 Configure and manage a BlueField Network Platform.

2.2 Configure MIG (AI and HPC).

Recommended Training (Optional)

Course reference: AI Infrastructure Professional

Covered by these course sections:

- BlueField Networking Platform
- Compute Platforms for AI/Virtualizing GPU Resources
- Practice: BlueField bring-up

Suggested Readings

- NVIDIA System Management User Guide
- NVIDIA RAPIDS™ cuDF Accelerates pandas Nearly 150x With Zero Code Changes | NVIDIA Technical Blog
- Choosing the Right Storage for Enterprise AI Workloads
- NVIDIA ConnectX® Card Replacement
- NVIDIA DCGM
- Data Center GPU Manager Guide
- Introduction to NVIDIA DGX™ H100/H200 Systems
- Spotlight: NVIDIA BlueField DPUs Power the VAST Data Platform for AI Workload Optimization
- NVIDIA-Certified Systems™
- Deploying DPU OS Using BFB From BMC
- DPU Modes of Operation
- Using mlxconfig - NVIDIA Docs
- Installing NVIDIA DOCA™ on a DPU
- Advanced GPU Configuration (Optional) — NVIDIA AI Enterprise: VMware Deployment Guide
- MIG User Guide — NVIDIA Multi-Instance GPU User Guide
- User Guide: NVIDIA AI Enterprise Documentation

Control Plane Installation and Configuration: Exam Weight 19%

These topics cover installation and configuration of operating systems, cluster managers, drivers, container tools, and management software. Enables orchestrated deployment, resource grouping, secure access, and reliable system software integration for NVIDIA AI clusters.

3.1 Install BCM, configure, and verify HA
3.2 Install OS
3.3 Install Cluster (configure category, configure interfaces, install Slurm/Enroot/Pyxis)
3.4 Install/update/remove NVIDIA GPU and DOCA drivers
3.5 Install the NVIDIA container toolkit
3.6 Demonstrate how to use NVIDIA GPUs with Docker
3.7 Install NGC CLI on hosts

Recommended Training (Optional)

Course reference: AI Infrastructure Professional

Covered by these course sections:

- AI Data Center Management
- Virtualizing GPU Resources
- NVIDIA AI Software
- Compute Platforms for AI

Suggested Reading List

- BCM Administrator Manual
- Initial Cluster Setup — NVIDIA DGX SuperPOD™
- DOCA-Host Installation and Upgrade
- NVIDIA-Certified Systems Configuration Guide
- Virtual GPU Client Licensing User Guide
- Installing the NVIDIA Container Toolkit
- Specialized Configurations With Docker — NVIDIA Container Toolkit
- Running a Sample Workload — NVIDIA Container Toolkit
- Getting Started With the NGC CLI

Cluster Test and Verification: Exam Weight 33%

The section of the exam focuses on validating cluster health and readiness through stress testing, benchmarking, cable integrity checks, firmware validation, and bandwidth verification. Includes end-to-end diagnostics and burn-in to ensure high reliability and optimal AI workload performance.

4.1 Perform single-nod stress test
4.2 Execute HPL (High-Performance Linpack)
4.3 Perform single-node NCCL (including verify NVIDIA NVLink™ Switch)
4.4 Validate cables by verifying signal quality
4.5 Confirm cabling is correct
4.6 Confirm FW/SW on switches
4.7 Confirm FW/SW on BlueField 3
4.8 Confirm FW on transceivers
4.9 Run ClusterKit to perform a multifaceted node assessment
4.10 Run NCCL to verify E/W fabric bandwidth
4.11 Perform NCCL burn-in
4.12 Perform HPL burn-in
4.13 Perform Nemo burn-in
4.14 Test storage

Recommended Training (Optional)

Course reference: AI Infrastructure

Covered by these course sections:

- Practice activities in Compute, Networking, Storage, and BlueField sections
- AI Data Center Management
- BlueField and Networking for AI

Suggested Reading List

- ib_write_lat
- AI Fabric Resiliency and Why Network Convergence Matters | NVIDIA Technical Blog
- Overview of NCCL — NCCL 2.27.5 Documentation
- NVIDIA Collective Communications Library (NCCL).
- Overview — NVIDIA NeMo Framework User Guide
- Train a Reasoning-Capable LLM in One Weekend With NVIDIA NeMo
- Storage - DGX
- Best Practices for DGX
- InfiniBand Port Counters - NVIDIA Docs
- UFM Supported Counter and Events
- Cabling Data Centers
- NVIDIA Cable Management Guidelines and FAQ
- Cable Validation - NVIDIA Docs
- NVIDIA DGX B200 Firmware Update Guide
- System Management Interface SMI | NVIDIA Developer

Troubleshooting and Optimization Exam Weight 12%

This section of the exam focuses on detecting, analyzing, and resolving hardware faults and performance bottlenecks. It encompasses root-cause analysis, component replacement, server/storage tuning, and ongoing optimization for multivendor hardware in NVIDIA-powered AI factories.

5.1	Identify and troubleshoot hardware faults (e.g., GPU/fan/network card)
5.2	Identify faulty cards/GPUs/power supplies
5.3	Replace faulty cards/GPUs/power supplies
5.4	Execute performance optimization for AMD and Intel Servers
5.5	Storage Optimization

Recommended Training (Optional)

Course reference: AI Infrastructure

Covered by these course sections:

- Compute Platforms for AI / Storage for AI
- BlueField Networking Platform
- Practices throughout the course
- AI Data Center Management

Suggested Reading List

- NVIDIA SMI
- DGX-2 Service Manual: DGX Systems Documentation
- NVIDIA DCGM
- Debugging and Troubleshooting — NVIDIA DCGM Documentation
- NVIDIA System Management (NVSM)
- NVIDIA-Certified Systems
- Overview — NVIDIA DCGM Documentation
- NVIDIA System Management User Guide
- Using the NVSM CLI — NVIDIA System Management User Guide
- Introduction to the NVIDIA DGX A100 System
- RAPIDS cuDF Accelerates pandas Nearly 150x With Zero Code Changes | NVIDIA Technical Blog
- Choose the Right Storage for Enterprise AI Workloads | NVIDIA Technical Blog
- Spotlight: NVIDIA BlueField DPUs Power the VAST Data Platform for AI Workload Optimization

Questions?

Contact us here.

NVIDIA