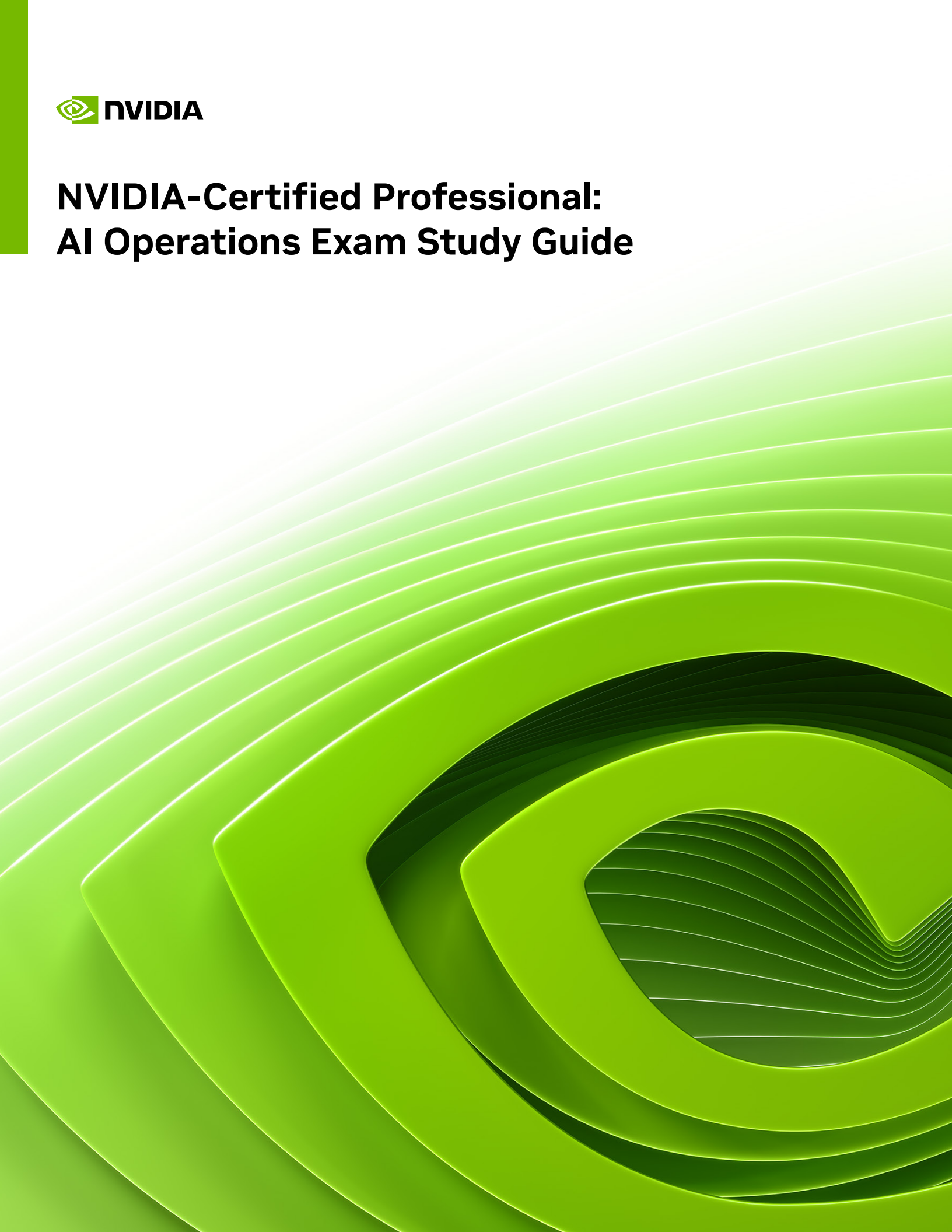
NVIDIA-Certified Professional: AI Operations Exam Study Guide



NVIDIA-Certified Professional: AI Operations Exam Study Guide

Contents

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

Information about NVIDIA certifications can be found here.

Job Description

The AI operations professional is responsible for managing and maintaining the AI infrastructure within a data center. Their role includes overseeing AI compute platforms, managing software and containers, configuring networking for AI workloads, and ensuring efficient storage and resource utilization. They need expertise in AI hardware and software, as well as data center technologies, to optimize performance and support AI-driven applications. Additionally, they might be involved in cluster management, workload scheduling, and implementing virtualization technologies to maximize the efficiency of the AI environment.

Job Responsibilities

1. Managing and maintaining AI compute platforms, including GPUs and other specialized hardware
2. Installing and configuring GPU drivers and software
3. Overseeing AI software stack and tools, including deep learning frameworks and data science libraries
4. Implementing and managing containerization technologies like Docker and NVIDIA NGC™
5. Configuring and optimizing networking infrastructure for AI workloads, including InfiniBand and Ethernet
6. Managing storage solutions for AI data, considering performance and capacity requirements
7. Deploying and managing data processing units (DPUs) to accelerate data center workloads
8. Monitoring and managing AI cluster health and resource utilization
9. Implementing workload management and scheduling tools like Slurm and Kubernetes
10. Ensuring efficient power and cooling for AI infrastructure to maintain optimal operating conditions

Recommended Qualifications and Experience

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

Certification Topics and References

Administration: Exam Weight 28%

Administration tasks for this job include overseeing NVIDIA Fleet Command™ and Slurm clusters, designing data center architecture specifically for AI workloads, managing NVIDIA Run:ai, and configuring NVIDIA Multi-Instance GPU (MIG) for both AI and high-performance computing (HPC) applications.

1.1	Administer Fleet Command.
1.2	Administer Slurm clusters.
1.3	Design data center architecture for AI workloads.
1.4	Administer Run:ai.
1.5	Configure MIG for AI and HPC.

Recommended Training (Optional)

Course reference: AI Operations

- Module 6, AI Data Center Management, Unit 2: Data Center Architecture for AI Workloads
- Module 6, AI Data Center Management, Unit 3: Scheduling AI Workloads With Slurm
- Module 6, AI Data Center Management, Unit 5: Managing AI at the Edge With Fleet Command
- Module 7, Virtualizing GPU Resources

Suggested Readings

- New Features for NVIDIA Fleet Command Deliver All-in-One Edge AI Management
- Accelerated Edge AI With Metropolis and Fleet Command | NVIDIA Technical Blog
- New Features and Applications Make Deploying Edge AI Easy With NVIDIA Fleet Command
- NVIDIA Base Command™ Manager
- Troubleshooting—NVIDIA Container Toolkit 1.17.3 documentation
- NVIDIA Multi-Instance GPU User Guide r560
- Product Overview—NVIDIA DGX™ Cloud Run:ai Documentation
- Developer Guide—NVIDIA Docs
- Open an Interactive Remote Console—NVIDIA Docs
- Install Administrator CLI—Run:ai Documentation Library
- NVIDIA Fleet Command Scales Edge AI Services for Enterprises
- Slurm Documentation
- NVIDIA Fleet Command
- How to Build Your GPU Cluster: Process and Hardware Options

Workload Management: Exam Weight 20%

Workload management tasks include administering Kubernetes clusters, troubleshooting issues using system management tools like NVIDIA Data Center GPU Manager (DCGM), NVIDIA System Management (NVSM), and nvidia-smi, and administering NVIDIA Base Command Manager (BCM) and cluster provisioning. These tasks are crucial for ensuring efficient resource allocation and monitoring and maintenance of AI infrastructure.

2.1 Administer Kubernetes cluster.
2.2 Use system management tools such as DCGM, NVSM, and nvidia-smi to troubleshoot issues.
2.3 Administer BCM and cluster provisioning.

Recommended Training (Optional)

Course reference: AI Operations

- Module 2, Compute, Unit 1: AI Compute Platforms Overview
- Module 2, Compute, Unit 3: Managing Compute Systems With NVSM
- Module 2, Compute, Unit 4: Operating AI Compute Platforms With DCGM
- Module 6, AI Data Center Management, Unit 2: Infrastructure Provisioning and Management With Base Command Manager
- Module 6, AI Data Center Management, Unit 4: AI Cluster Orchestration With Kubernetes

Suggested Readings

- NVIDIA System Management Interface
- Manage Builders | Docker Docs
- NVIDIA Bright Cluster Manager
- NVIDIA Container Toolkit
- DCGM Diagnostics
- Provisioning Nodes—NVIDIA DGX SuperPOD™
- DGX Superpod User Guide
- Split-Brain Recovery in HA Installation
- Concepts | Kubernetes
- kubectl Quick Reference

Install and Deploy: Exam Weight 32%

This section focuses on both software and hardware installation and configuration. A core responsibility is installing and configuring BCM and subsequently using BCM to install Kubernetes on NVIDIA hosts. The role also requires deploying containers from NGC and cloud virtual machine image (VMI) containers. Additionally, understanding the storage requirements of an AI data center and deploying NVIDIA DOCA™ services on DPU-Arm are key aspects of this position.

3.1	Install and configure BCM.
3.2	Install and initialize Kubernetes on NVIDIA hosts using BCM.
3.3	Deploy containers from NGC.
3.4	Deploy cloud VMI containers.
3.5	Understand storage requirements for AI data centers.
3.6	Deploy DOCA services on DPU-Arm.

Recommended Training (Optional)

Course reference: AI Operations

- Module 2: Compute, Unit 5: NVIDIA GPU Containers
- Module 4: Storage for AI, Unit 1: Storage for AI Overview
- Module 5: BlueField DPUs, Unit 3: Running DOCA Services on BlueField DPU
- Module 6: AI Data Center Management, Unit 2: Infrastructure Provisioning and Management With Base Command Manager

Suggested Reading List

- NVIDIA DOCA Software Framework
- Base Command Platform User Guide
- NVIDIA Cloud Native Technologies
- Kubernetes Deployment—NVIDIA DGX BasePOD™
- DOCA Documentation
- NVIDIA DOCA BlueMan Service Guide
- NVIDIA DOCA Installation Guide for Linux
- NGC User Guide
- NGC-Certified Public Clouds—NVIDIA Docs
- Docker Swarm vs Kubernetes: Which Is Better in 2025?

Troubleshooting and Optimization: Exam Weight 20%

Troubleshooting is a crucial aspect of maintaining optimal performance. Responsibilities include resolving issues with Docker, fabric manager service for NVIDIA NVLink™ and NVSwitch™ systems, BCM, Magnum IO™ components, and storage performance. Ensuring that these elements function seamlessly is essential for efficient, stable operation of AI workloads.

4.1 Troubleshoot Docker.
4.2 Troubleshoot the fabric manager service for NVLink and NVSwitch systems.
4.3 Troubleshoot Base Command Manager.
4.4 Troubleshoot Magnum IO components.
4.5 Troubleshoot storage performance.

Recommended Training (Optional)

Course reference: AI Operations

- Module 2: Compute, Unit 5: NVIDIA GPU Containers
- Module 3: Networking for AI, Unit 2: Operating InfiniBand Fabrics
- Module 3: Networking for AI, Unit 4: Monitoring AI Data Centers With NVIDIA UFM®
- Module 4: Storage for AI, Unit 1: Storage for AI Overview
- Module 6: AI Data Center Management, Unit 2: Infrastructure Provisioning and Management With Base Command Manager

Suggested Reading List

- Top Five Most Common Issues With Docker (and How to Solve Them) | Packagecloud Blog
- Optimizing Data Movement in GPU Applications With the NVIDIA Magnum IO Developer Environment
- Docker Container Logs
- Installing Docker and the Docker Utility Engine for NVIDIA GPUs—NVIDIA AI Enterprise: VMware Deployment Guide
- NVIDIA Base Command Manager
- Troubleshooting—NVIDIA Container Toolkit 1.17.3 Documentation
- Getting Started—NVIDIA DCGM Documentation
- Fabric Manager for NVIDIA NVSwitch Systems
- Containers for Deep Learning Frameworks User Guide—NVIDIA Docs
- NVIDIA Magnum IO GPUDirect® Storage
- Upgrading UFM Software—NVIDIA Docs
- NVIDIA UFM Enterprise User Manual
- NVIDIA UFM High-Availability User Guide v5.5.0
- NVIDIA Magnum IO

Questions?

Contact us here.

3770900. MAR25