

PRACTICE QUESTIONS

NVIDIA

NCP-AII

AI Infrastructure Professional

SHARED BY [u/Tidy-Turtle-9809](#) on aiexamcerts

EDITION Free Set 1

QUESTIONS 10

SHARED October 15, 2026

EXAM NVIDIA NCP-AII, AI Infrastructure Professional

Shared by [u/Tidy-Turtle-9809](#) on aiexamcerts

10 practice questions for the NVIDIA NCP-AII exam, AI Infrastructure Professional, each followed by its answer and a short explanation. More free sets and the full practice exam: www.aiexamcerts.com/nvidia/ncp-aai. Spotted a mistake? Email support@aiexamcerts.com with the exam code NCP-AII and the question number.

1. Which factor most directly decides whether airflow in a GPU rack works as designed?

- A Whether the PDUs are mounted on the left or right rail
- B Hot and cold aisles with blanking panels
- C Whether the rack uses cage nuts or tool-less rails
- D The brand of the network switch placed in the top slot

Answer: B

Consistent front-to-back airflow and blanking panels that stop hot air recirculating are the physical conditions that keep intake temperatures in range. PDU side, rail style and switch brand do not change how air moves through the rack.

2. Cable _____ with source, destination and port on both ends makes fault isolation faster in large clusters.

Answer: labeling (also accepted: labelling, labels)

Labels on both ends let a technician confirm the physical path without tracing cables by hand, which shortens outages. Good labeling is part of the build documentation, not an optional extra.

3. An administrator runs `nvidia-smi topo -m` and sees that two GPUs are connected through SYS. What does this indicate?

- A The two GPUs share a single NVLink bridge between them
- B The two GPUs are inside the same MIG instance
- C The path crosses the CPU socket interconnect
- D The two GPUs are disabled in the BIOS setup screen

Answer: C

SYS in the topology matrix means the path traverses the system interconnect between CPU sockets, which is slower than a direct NVLink or PCIe switch path. A shared bridge would show NV links, a MIG instance is a partition of one GPU, and disabled GPUs would not appear as peers.

4. Multi-Instance GPU (MIG) shares one set of memory and compute slices between instances, so each instance can exceed its own memory limit.
- A True
 - B False

Answer: B (False)

MIG divides a GPU into isolated instances with dedicated memory and compute slices, so an instance cannot use memory beyond its allocation. The statement is false because it describes sharing across instances, which is what MIG is designed to prevent.

5. Which two items should be planned for a GPU cluster's storage tier? (Choose two.)
- A Capacity and throughput sized for the expected dataset and checkpoint volume
 - B Per-user home directories sized to equal the cluster's total GPU memory
 - C Backup or replication for data that cannot be regenerated
 - D Unlimited local disks installed in each GPU server
 - E Automatic deletion of all checkpoints after each training epoch

Answer: A, C

Storage has to hold the datasets and checkpoints at the throughput the GPUs need, and data that cannot be recreated needs protection. Oversized home directories, local disks in every server and deleting all checkpoints each epoch do not meet those needs.

6. What is the role of the subnet manager in an InfiniBand fabric?
- A It encrypts the DHCP leases issued to every host
 - B It assigns LIDs and computes routes across the fabric
 - C It stores the training datasets used by the cluster jobs
 - D It replaces the need for switch firmware updates over time

Answer: B

The subnet manager discovers the fabric, assigns local identifiers and computes forwarding routes, so ports cannot communicate without it. DHCP encryption, dataset storage and firmware replacement are not its job.

7. Which command shows the state and rate of InfiniBand ports on a host?

- A lsblk, which lists block devices and partition sizes
- B ss -tulpn, which lists listening sockets on the host
- C ibstat
- D uptime, which shows system load averages over time

Answer: C

ibstat reports each port's state, such as Active or Down, along with its rate and LID, which is the first check on a fabric link. The other commands report block devices, listening sockets and load, and they do not describe InfiniBand ports.

8. GPUDirect Storage moves data between storage and GPU memory with less involvement from the _____.

Answer: CPU (also accepted: host CPU, processor)

GPUDirect Storage lets data move directly between storage and GPU memory, which reduces copies through the CPU and can raise throughput. Encryption, driver replacement and weight compression are not its function.

9. Which command lets Docker run a container with all host GPUs visible?

- A docker run --privileged-gpu on
- B docker run --gpus all
- C docker start --attach-gpu
- D docker network create --gpu

Answer: B

The --gpus option of docker run selects which GPUs are exposed to the container, and all exposes every GPU on the host. The other commands are not valid Docker options for GPU access.

10. What is the purpose of the NVIDIA GPU Operator in Kubernetes?

- A To replace the Kubernetes scheduler entirely with a new one
- B To store container images in a central registry service
- C To deploy GPU drivers, device plugin and monitoring
- D To format the root disk of each node at every boot

Answer: C

The GPU Operator automates the GPU software components on nodes so that they are consistently deployed and monitored. It works with the scheduler rather than replacing it, and registries and disk formatting are different functions.