

PRACTICE QUESTIONS

NVIDIA

NCP-AII

AI Infrastructure Professional

SHARED BY [u/Rustic-Router-7338](#) on aiexamcerts

EDITION Free Set 2

QUESTIONS 10

SHARED October 15, 2026

EXAM NVIDIA NCP-AII, AI Infrastructure Professional

Shared by [u/Rustic-Router-7338](#) 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-aii. Spotted a mistake? Email support@aiexamcerts.com with the exam code NCP-AII and the question number.

1. What is the main purpose of cable labeling in an AI cluster build?

- A To increase the bandwidth of each cable run
- B To replace the need for link testing after installation
- C To speed up fault isolation and maintenance across many links
- D To change the protocol that the port negotiates with its peer

Answer: C

Consistent labels make it quick to trace a failing link to its switch port and server port during maintenance. Labels do not change bandwidth, they do not replace link tests, and they do not alter negotiated protocols.

2. A redundant power supply setup protects against a single feed failure even when both supplies share one PDU circuit.

- A True
- B False

Answer: B (False)

Redundancy protects only when the supplies have independent paths. If both draw from the same circuit, one fault can remove both supplies. The statement is false because it treats redundant hardware as redundant power.

3. Why would an administrator enable persistence mode on NVIDIA GPUs in a compute cluster?

- A To raise the GPU's maximum clock speed permanently
- B To disable ECC memory checking for extra speed
- C To let the GPU drive a display output for the console
- D To keep the driver loaded between jobs

Answer: D

Persistence mode keeps the driver loaded between jobs, which avoids the initialization delay at each job start. It does not change clock limits, ECC settings or display output, which are configured separately.

4. The nvidia-smi option that prints the GPU topology matrix, including NVLink and PCIe paths between devices, is used with the topo command as topo _____.

Answer: -m (also accepted: --matrix, -matrix)

The topology matrix shows the link type between each pair of GPUs and CPUs, which is the first place to look when peer performance seems low. The option is combined with the topo subcommand, so it is written as nvidia-smi topo -m.

5. Which two steps belong in a standard post-installation validation of a GPU node? (Choose two.)
- A Run nvidia-smi and confirm that every GPU is listed and healthy
 - B Reformat the boot drive to confirm that it is empty
 - C Run a multi-GPU communication test such as nccl-tests
 - D Delete the driver logs so that the disk stays clean
 - E Disable all GPUs until the first job is scheduled

Answer: A, C

Confirming every GPU is listed and healthy, then testing communication across GPUs, verifies the node before it takes work. Reformatting the boot drive, deleting logs and disabling GPUs do not validate anything.

6. A RoCE fabric drops packets under heavy load, and training slows down. Which mechanism is typically configured to keep the Ethernet fabric lossless?
- A Packet fragmentation at the application layer only
 - B Sending all traffic through one VLAN without any QoS policy
 - C Disabling jumbo frames on every switch port in the fabric
 - D Priority flow control with ECN

Answer: D

Priority flow control prevents drops for RDMA traffic, and ECN signals congestion early so senders slow down before buffers overflow. Application-layer fragmentation, a single VLAN without QoS and disabling jumbo frames do not give RoCE a lossless behavior.

7. Which tool runs a point-to-point bandwidth test between two InfiniBand hosts?
- A ib_write_bw from the perftest package
 - B A ping test sent with 64-byte packets only
 - C A disk benchmark that writes sequential zeros to a local drive
 - D A DNS lookup timing test run from the login node

Answer: A

The perftest tools, such as `ib_write_bw`, measure RDMA bandwidth between two endpoints, which isolates the fabric from application behavior. Ping measures latency with small packets, a disk benchmark measures storage, and a DNS timing test measures name resolution.

8. The CUDA toolkit and the NVIDIA driver are separate components, and the driver must support the CUDA version being used.
- A True
 - B False

Answer: A (True)

The driver is installed on the host and the toolkit provides the development and runtime libraries. Each CUDA release has a minimum driver version, so the two must be compatible. The statement describes the normal relationship between them.

9. A training job is slow during data loading while the GPUs sit idle. Which first step is most useful?
- A Increase the GPU clock to finish the job sooner
 - B Measure storage throughput and read patterns
 - C Remove the GPU driver to free up some memory
 - D Turn off all monitoring so the job uses less CPU

Answer: B

Idle GPUs during loading point to the data path, so measuring storage throughput and the read pattern shows where the bottleneck is. Raising clocks, removing drivers and disabling monitoring do not address a data-loading bottleneck.

10. Which practice keeps shared storage healthy for multi-node training?
- A Store every dataset on one node's local disk
 - B Disable all quotas so that no job ever hits a limit
 - C Monitor capacity, latency and error rates
 - D Use a different file system for each job submitted to the cluster

Answer: C

Monitoring capacity, latency and errors detects problems before they stall training, and it is the basis for planning growth. Local-only storage, disabling quotas and per-job file systems each create failures or waste.