

PRACTICE QUESTIONS

NVIDIA

NCP-AIO

NVIDIA-Certified Professional: AI Operations

SHARED BY [u/witty_jaguar_60](#) on aiexamcerts

EDITION Free Set 2

QUESTIONS 10

SHARED October 18, 2026

EXAM NVIDIA NCP-AIO, NVIDIA-Certified Professional: AI Operations

Shared by [u/witty_jaguar_60](#) on aiexamcerts

10 practice questions for the NVIDIA NCP-AIO exam, NVIDIA-Certified Professional: AI Operations, each followed by its answer and a short explanation. More free sets and the full practice exam:

aiexamcerts.com/nvidia/ncp-aio. Spotted a mistake? Email support@aiexamcerts.com with the exam code NCP-AIO and the question number.

1. A new GPU node joins the cluster, but its GPUs do not appear in nvidia-smi. Which first check is most useful?
 - A Whether the NVIDIA kernel module loads and matches the running kernel
 - B Whether the Kubernetes dashboard add-on is installed
 - C Whether the Slurm job time limit is long enough
 - D Whether the DNS search suffix list is correct

Answer: A

If nvidia-smi cannot see a GPU, the driver module is usually missing or was built for another kernel, so checking that it loads comes first. Job time limits and DNS settings cannot hide a GPU from the driver, which makes them distractors.

2. What does the NVIDIA GPU Operator automate on a Kubernetes GPU cluster?
 - A Slurm job accounting and fairshare calculation
 - B Deployment of the driver, container toolkit and device plugin on GPU nodes
 - C Switch firmware upgrades on the leaf tier
 - D Installation of BlueField OS images on DPUs

Answer: B

The GPU Operator manages the software stack that makes GPUs usable in Kubernetes, including the driver, toolkit and device plugin. Job accounting, switch firmware and DPU imaging are separate concerns that the operator does not handle.

3. A BlueField DPU runs its own Arm operating system, separate from the x86 operating system of the host server.
 - A True
 - B False

Answer: A (True)

The DPU has its own Arm cores and OS image, which is why DOCA services and firmware are managed on the DPU itself. Confusing the DPU's OS with the host OS is a common mistake when troubleshooting offloaded networking.

4. Which pod spec field requests one NVIDIA GPU for a container?

- A spec.hostNetwork set to gpu
- B an annotation named gpu with the value 1 only
- C resources.limits with nvidia.com/gpu set to 1
- D a Slurm --gpus flag inside the pod spec

Answer: C

GPUs are extended resources in Kubernetes, so the request goes under resources.limits with nvidia.com/gpu. Annotations do not allocate devices, and Slurm flags belong to batch scripts, which makes them distractors in a pod spec.

5. A common way to keep ordinary pods off dedicated GPU nodes is to apply a _____ taint to those nodes.

Answer: NoSchedule (also accepted: no schedule)

A NoSchedule taint stops pods without a matching toleration from being placed on the node, which reserves the GPU nodes for GPU workloads. NoExecute also evicts running pods, so it is a stricter option than the one the question asks about.

6. Which storage requirement is most specific to saving AI model checkpoints during training?

- A Read-only archive retention for compliance
- B Tape rotation on a weekly schedule
- C Single-user home directory quotas
- D Burst write throughput and enough capacity for periodic saves

Answer: D

Checkpoints are written in bursts from many GPUs and must complete quickly to avoid stalling training, so write throughput and capacity drive the design. Archive retention and quotas are real storage concerns, but they do not describe the checkpoint write pattern.

7. Which NVIDIA service manages remote deployment of software and updates to many systems across multiple sites?

- A NVIDIA Fleet Command
- B NVIDIA Magnum IO communication library for collective operations
- C NVIDIA DOCA SDK
- D NVIDIA NetQ

Answer: A

Fleet Command handles remote deployment and lifecycle management for systems spread across sites and edge locations. Magnum IO is a communication library, DOCA is a developer SDK for DPUs, and NetQ monitors the network fabric.

8. Which two statements about NGC containers are correct? (Choose two.)

- A NGC images are OCI-compatible container images
- B NGC images replace the GPU driver on the host
- C NGC images are pulled from the nvcr.io registry
- D NGC images run only on Windows Server hosts
- E NGC pulls require a Slurm account for each user

Answer: A, C

NGC containers are served from nvcr.io and follow the OCI image format, so standard container runtimes can run them on Linux GPU hosts. The Windows-only claim and driver replacement are false, and Slurm accounts are unrelated to image pulls.

9. How does a Kubernetes pod get credentials for pulling an image from an authenticated registry?

- A It reads a plaintext password from a ConfigMap mounted into the pod
- B It references a docker-registry type Secret in the pod spec
- C It mounts a PersistentVolume that holds the password
- D It uses an Ingress rule that the registry checks

Answer: B

An `imagePullSecrets` reference to a docker-registry Secret supplies the credentials the kubelet needs to pull the image. Storing the password in a ConfigMap would expose it in plaintext, which is the tempting but insecure approach.

10. The command `dcgmi discovery -l` lists the GPUs that DCGM can see on the node.

- A True
- B False

Answer: A (True)

`dcgmi discovery -l` enumerates the GPUs DCGM manages on the host, which is the first check when a GPU seems missing from monitoring. Other `dcgmi` subcommands are used for diagnostics and policies.