

PRACTICE QUESTIONS

NVIDIA

NCA-AIIO

AI Infrastructure and Operations Associate

SHARED BY [u/Jazzy-Heron-2326](#) on aiexamcerts

EDITION Free Set 1

QUESTIONS 10

SHARED October 15, 2026

EXAM NVIDIA NCA-AIIO, AI Infrastructure and Operations Associate

Shared by [u/Jazzy-Heron-2326](#) on aiexamcerts

10 practice questions for the NVIDIA NCA-AII0 exam, AI Infrastructure and Operations Associate, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/nvidia/nca-aiio. Spotted a mistake? Email support@aiexamcerts.com with the exam code NCA-AII0 and the question number.

1. Which numeric precision format is commonly used in modern GPU training to reduce memory and increase throughput while keeping a wide dynamic range?
- A Fixed-point 8.8 format
 - B FP64 double precision
 - C INT4 integer weights
 - D BF16 (bfloat16)

Answer: D

BF16 keeps the 8-bit exponent of FP32, so it has a wide dynamic range while halving memory per value compared with FP32. FP64 is precise but slow and memory-hungry for deep learning. INT4 is an inference-oriented quantization format rather than a common training format.

2. The NVIDIA interconnect that links GPUs inside a server with high bandwidth is called _____.

Answer: NVLink

NVLink provides high-bandwidth GPU-to-GPU connections within a node, and NVSwitch extends this across GPUs. InfiniBand is the common fabric between nodes, which is a tempting but different answer.

3. In a neural network, what is the main purpose of an activation function such as ReLU?
- A To normalize the input data to zero mean and unit variance
 - B To introduce non-linearity so that stacked layers can model complex functions
 - C To store the training labels inside the model file for later reuse
 - D To split the dataset into training and validation sets

Answer: B

Without a non-linear activation, a stack of linear layers collapses into a single linear transformation, so ReLU lets the network learn complex patterns. Normalization is tempting because it also transforms values, but it is a separate preprocessing or layer step. Splitting datasets is a data management task, not a neuron function.

4. Overfitting is indicated when training performance is high while validation performance is markedly lower.
- A True
 - B False

Answer: A (True)

A widening gap between training and validation results shows the model has memorized training specifics. This pattern is the usual definition of overfitting, so the statement is true.

5. Which two statements about training a neural network with gradient descent are correct? (Choose two.)
- A Backpropagation computes gradients of the loss with respect to the weights
 - B Gradient descent always finds the global minimum for deep networks
 - C A larger learning rate always guarantees faster convergence
 - D The loss function is computed only at inference time
 - E Mini-batches approximate the full-dataset gradient with less memory

Answer: A, E

Backpropagation applies the chain rule to produce weight gradients, and mini-batches give noisy but cheaper gradient estimates. The claim that deep networks always reach the global minimum is false because the loss landscape is non-convex. A learning rate that is too large can diverge rather than converge faster.

6. When a model is too large for one GPU, engineers split the network's layers across several devices so each device holds only part of the model. What is this strategy called?
- A Model parallelism
 - B Data parallelism
 - C Speculative decoding
 - D Quantization-aware training

Answer: A

Model parallelism distributes parts of the model itself, such as layers or tensor shards, across devices. Data parallelism replicates the full model and splits the batch, which is tempting but does not solve a memory limit on one model copy. Speculative decoding and quantization-aware training are unrelated techniques.

7. An operator is comparing two GPU servers for training large models. Which interconnect is designed for high-bandwidth, low-latency communication between GPUs inside the same server?
- A BMC out-of-band management link
 - B Gigabit Ethernet management port
 - C SATA storage controller
 - D NVLink with NVSwitch

Answer: D

NVLink and NVSwitch provide high-bandwidth GPU-to-GPU links inside a node, which is critical for collective operations during training. The Gigabit Ethernet management port carries administrative traffic and is far too slow for tensor exchange. A BMC link and a SATA controller are not training interconnects.

8. The NVIDIA command-line tool that reports GPU utilization, memory and temperature is _____.

Answer: nvidia-smi (also accepted: nvidia smi)

nvidia-smi is the standard utility for quick GPU status checks on Linux hosts. nvcc is the compiler, which is an easy confusion, but it does not report device state.

9. Which tool is the standard command-line utility for querying NVIDIA GPU status such as utilization, memory use and temperature on a Linux host?
- A nvcc
 - B nvidia-smi
 - C ncu
 - D tensorrt-builder

Answer: B

nvidia-smi reports device state including utilization, memory and temperature, and is the usual first check on a host. nvcc is the CUDA compiler, which is easy to confuse with a monitoring tool. ncu is the Nsight Compute profiler for kernel analysis, and tensorrt-builder is not a standard monitoring command.

10. Which NVIDIA library provides optimized collective communication primitives such as all-reduce across multiple GPUs?
- A Triton Inference Server
 - B cuDNN
 - C RAPIDS cuDF
 - D NCCL

Answer: D

NCCL implements collective operations such as all-reduce, all-gather and broadcast across GPUs and nodes. cuDNN accelerates deep learning primitives like convolutions, which is tempting because both are CUDA libraries, but it does not handle multi-GPU collectives. RAPIDS cuDF is a dataframe library, and Triton serves models for inference.