NVIDIA NCA-AIIO Practice Questions

Shared by Randolph on 21-08-2026

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

What is an advantage of InfiniBand over Ethernet?

Options:

A- InfiniBand always provides higher bandwidth than Ethernet.
B- InfiniBand supports RDMA while Ethernet does not.
C- InfiniBand offers lower latency than Ethernet.

Answer:

C

Explanation:

InfiniBand's advantage over Ethernet lies in its lower latency, achieved through a streamlined protocol and hardware offloads, delivering microsecond-scale communication critical for AI clusters. While InfiniBand often offers high bandwidth, Ethernet can match or exceed it (e.g., 400 GbE), and Ethernet supports RDMA via RoCE, making latency the standout differentiator.

(Reference: NVIDIA Networking Documentation, Section on InfiniBand vs. Ethernet)

Question 2

Question Type: MultipleChoice

For which workloads is NVIDIA Merlin typically used?

Options:

A- Recommender systems
B- Natural language processing
C- Data analytics

Answer:

A

Explanation:

NVIDIA Merlin is a specialized, end-to-end framework engineered for building and deploying large-scale recommender systems. It streamlines the entire pipeline, including data preprocessing (e.g., feature engineering, data transformation), model training (using GPU-accelerated frameworks), and inference optimizations tailored for recommendation tasks. Unlike general-purpose tools for natural language processing or data analytics, Merlin is optimized to handle the unique challenges of recommendation workloads, such as processing massive user-item interaction datasets and delivering personalized results efficiently.

(Reference: NVIDIA Merlin Documentation, Overview Section)

Question 3

Question Type: MultipleChoice

What NVIDIA tool should a data center administrator use to monitor NVIDIA GPUs?

Options:

A- NVIDIA System Monitor
B- NetQ
C- DCGM

Answer:

C

Explanation:

The NVIDIA Data Center GPU Manager (DCGM) is the recommended tool for data center administrators to monitor NVIDIA GPUs. It provides real-time health monitoring, telemetry (e.g., utilization, temperature), and diagnostics, tailored for large-scale deployments. NetQ focuses on network monitoring, and there's no ''NVIDIA System Monitor'' in this context, making DCGM the correct choice. (Note: The document incorrectly lists D; C is intended.)

(Reference: NVIDIA DCGM Documentation, Overview Section)

Question 4

Question Type: MultipleChoice

Which GPUs should be used when training a neural network for self-driving cars?

Options:

A- NVIDIA H100 GPUs
B- NVIDIA L4 GPUs
C- NVIDIA DRIVE Orin

Answer:

A

Explanation:

Training neural networks for self-driving cars requires immense computational power and high-bandwidth memory to process vast datasets (e.g., sensor data, video). NVIDIA H100 GPUs, with their cutting-edge architecture and massive throughput, are ideal for these demanding workloads. L4 GPUs are optimized for inference and efficiency, while DRIVE Orin targets in-vehicle inference, not training, making H100 the best choice.

(Reference: NVIDIA AI Infrastructure and Operations Study Guide, Section on GPU Selection for Training)

Question 5

Question Type: MultipleChoice

How many Mellanox ConnectX-6 Single Port VPI cards are in a DGX A100 system?

Options:

A- 8
B- 16
C- 4

Answer:

A

Explanation:

The DGX A100 system includes eight Mellanox ConnectX-6 Single Port VPI cards, providing high-speed connectivity (up to 200 Gb/s) for clustering and data transfer. These cards support versatile protocols (InfiniBand or Ethernet), enabling robust multi-node AI workloads, with eight being the standard configuration for this system.

(Reference: NVIDIA DGX A100 System Documentation, Networking Section)

Question 6

Question Type: MultipleChoice

How many 1 Gb Ethernet in-band network connections are in a DGX H100 system?

Options:

A- 1
B- 2
C- 0

Answer:

C

Explanation:

The DGX H100 system uses high-speed NVIDIA ConnectX-7 QSFP56 ports (supporting 10 GbE and above) for in-band management and storage traffic, with no 1 Gb Ethernet interfaces allocated to in-band networks. A single 1 GbE RJ45 port exists, but it's reserved for out-of-band Baseboard Management Controller (BMC) tasks, not in-band connectivity.

(Reference: NVIDIA DGX H100 System Documentation, Networking Section)

Question 7

Question Type: MultipleChoice

When training a neural network, What is the most common pattern of storage access?

Options:

A- Random write
B- Sequential read
C- Sequential write

Answer:

B

Explanation:

Training neural networks typically involves streaming large datasets from storage in a sequential read pattern. This ordered access maximizes throughput and minimizes seek overhead, as training pipelines ingest data in batches for processing across epochs. Writes (e.g., model checkpoints) are less frequent and typically sequential, while random writes are rare, making sequential reads the dominant pattern. (Note: The document incorrectly lists C as the answer; B aligns with NVIDIA's documentation.)

(Reference: NVIDIA AI Infrastructure and Operations Study Guide, Section on Storage Access Patterns)

Question 8

Question Type: MultipleChoice

Which aspect of computing uses large amounts of data to train complex neural networks?

Options:

A- Machine learning
B- Deep learning
C- Inferencing

Answer:

B

Explanation:

Deep learning, a subset of machine learning, relies on large datasets to train multi-layered neural networks, enabling them to learn hierarchical feature representations and complex patterns autonomously. While machine learning encompasses broader techniques (some requiring less data), deep learning's dependence on vast data volumes distinguishes it. Inferencing, the application of trained models, typically uses smaller, real-time inputs rather than extensive training data.

(Reference: NVIDIA AI Infrastructure and Operations Study Guide, Section on Deep Learning Fundamentals)

Question 9

Question Type: MultipleChoice

A customer is evaluating an AI cluster for training and is questioning why they should use a large number of nodes. Why would multi-node training be advantageous?

Options:

A- The model is too large to fit into GPU memory.
B- The model is being used by a large number of users.
C- The model is being used for large-scale inference workloads.

Answer:

A

Explanation:

Multi-node training is advantageous when a model's size---its parameters, activations, and gradients---exceeds the memory capacity of a single GPU. By sharding the model across multiple nodes (using techniques like data parallelism or model parallelism), training becomes feasible and efficient. User count and inference scale are unrelated to training architecture needs, which focus on compute and memory distribution.

(Reference: NVIDIA AI Infrastructure and Operations Study Guide, Section on Multi-Node

Training Benefits)

Question 10

Question Type: MultipleChoice

How is the architecture different in a GPU versus a CPU?

Options:

A- A GPU acts as a PCIe controller to maximize bandwidth.
B- A GPU is architected to support massively parallel execution of simple instructions.
C- A GPU is a single large and complex core to support massive compute operations.

Answer:

B

Explanation:

A GPU's architecture is designed for massive parallelism, featuring thousands of lightweight cores that execute simple instructions across vast data elements simultaneously---ideal for tasks like AI training. In contrast, a CPU has fewer, complex cores optimized for sequential execution and branching logic. GPUs don't function as PCIe controllers (a hardware role), nor are they single-core designs, making the parallel execution focus the key differentiator.

(Reference: NVIDIA GPU Architecture Whitepaper, Section on GPU Design Principles)

Question 11

Question Type: MultipleChoice

Which NVIDIA tool aids data center monitoring and management?

Options:

A- NVIDIA Mellanox Insight

B- NVIDIA Clara
C- NVIDIA TensorRT
D- NVIDIA DCGM

Answer:

D

Explanation:

NVIDIA Data Center GPU Manager (DCGM) aids data center monitoring and management by providing detailed GPU telemetry, health diagnostics, and performance tracking at scale. Clara targets healthcare, TensorRT optimizes inference, and Mellanox Insight isn't a standard NVIDIA tool, making DCGM the go-to solution.

(Reference: NVIDIA DCGM Documentation, Overview Section)

Question 12

Question Type: MultipleChoice

What is a significant benefit of using containers in an AI development environment?

Options:

A- They increase the base accuracy of AI models by optimizing their algorithms.
B- They ensure that AI applications run consistently across different computing environments.
C- They can automatically generate AI datasets for machine learning model training.
D- They directly increase the processing speed of GPUs used in AI computations.

Answer:

B

Explanation:

Containers (e.g., Docker) encapsulate AI applications with their dependencies, ensuring consistent execution across diverse environments---from development laptops to production clusters---without manual reconfiguration. They don't inherently improve model accuracy, generate datasets, or boost GPU speed, focusing instead on portability and reproducibility. (Note: The document incorrectly lists A; B is correct per NVIDIA standards.)

(Reference: NVIDIA AI Infrastructure and Operations Study Guide, Section on Containers in AI Development)

P2P exams