Free Questions for NCA-AIIO

Shared by Harrell on 16-04-2026

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

What is a key benefit of using NVIDIA GPUDirect RDMA in an AI environment?

Options:

A- It increases the power efficiency and thermal management of GPUs.
B- It reduces the latency and bandwidth overhead of remote memory access between GPUs.
C- It enables faster data transfers between GPUs and CPUs without involving the operating system.
D- It allows multiple GPUs to share the same memory space without any synchronization.

Answer:

C

Explanation:

NVIDIA GPUDirect RDMA allows network adapters to directly access GPU memory, bypassing the CPU and operating system kernel. This accelerates data transfers between GPUs and CPUs (or other devices), reducing latency and CPU overhead in AI workflows, such as multi-node training. It doesn't focus on power efficiency or unsynchronized memory sharing, making faster transfers its key benefit.

(Reference: NVIDIA GPUDirect RDMA Documentation, Overview Section)

Question 2

Question Type: MultipleChoice

In an AI cluster, what is the importance of using Slurm?

Options:

A- Slurm is used for data storage and retrieval in an AI cluster.
B- Slurm is responsible for AI model training and inference in an AI cluster.
C- Slurm is used for interconnecting nodes in an AI cluster.

D- Slurm helps with managing job scheduling and resource allocation in the cluster.

Answer:

D

Explanation:

Slurm (Simple Linux Utility for Resource Management) is a workload manager critical for AI clusters, handling job scheduling and resource allocation. It ensures tasks are assigned to available GPUs/CPUs efficiently, supporting scalable training and inference. It doesn't manage storage, perform training, or interconnect nodes---those are separate functions.

(Reference: NVIDIA AI Infrastructure and Operations Study Guide, Section on Slurm in AI Clusters)

Question 3

Question Type: MultipleChoice

What is a common tool for container orchestration in AI clusters?

Options:

A- Kubernetes
B- MLOps
C- Slurm
D- Apptainer

Answer:

A

Explanation:

Kubernetes is the industry-standard tool for container orchestration in AI clusters, automating deployment, scaling, and management of containerized workloads. Slurm manages job scheduling, Apptainer (formerly Singularity) runs containers, and MLOps is a practice, not a tool, making Kubernetes the clear leader in this domain.

(Reference: NVIDIA AI Infrastructure and Operations Study Guide, Section on Container Orchestration)

Question 4

Question Type: MultipleChoice

When deploying high-density workloads in a data center, what are the three main resource constraints that need to be considered?

Options:

A- Processing speed, storage capacity, and network connectivity.
B- Power, cooling, and physical space.
C- Bandwidth, security, and redundancy.

Answer:

B

Explanation:

High-density workloads (e.g., GPU clusters for AI) strain data center resources, primarily power (to supply dense servers), cooling (to dissipate heat from tightly packed hardware), and physical space (to house equipment). While processing speed, bandwidth, and other factors matter, power, cooling, and space are the physical constraints most critical to deployment feasibility.

(Reference: NVIDIA AI Infrastructure and Operations Study Guide, Section on Data Center Resource Constraints)

Question 5

Question Type: MultipleChoice

Which two components are included in GPU Operator? (Choose two.)

Options:

A- Drivers
B- PyTorch
C- DCGM

D- TensorFlow

Answer:

A, C

Explanation:

The NVIDIA GPU Operator is a tool for automating GPU resource management in Kubernetes environments. It includes two key components: GPU drivers, which provide the necessary software to interface with NVIDIA GPUs, and the NVIDIA Data Center GPU Manager (DCGM), which offers health monitoring, telemetry, and diagnostics for GPU clusters. Frameworks like PyTorch and TensorFlow are separate AI development tools, not part of the GPU Operator, which focuses on infrastructure rather than application layers.

(Reference: NVIDIA GPU Operator Documentation, Components Section)

Question 6

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 7

Question Type: MultipleChoice

When should RoCE be considered to enhance network performance in a multi-node AI computing environment?

Options:

A- A network that experiences a high packet loss rate (PLR).
B- A network with large amounts of storage traffic.
C- A network that cannot utilize the full available bandwidth due to high CPU utilization.

Answer:

C

Explanation:

RoCE (RDMA over Converged Ethernet) enhances network performance by offloading data transport to the NIC via RDMA, bypassing CPU involvement. It's particularly valuable when high CPU utilization limits bandwidth usage, as it reduces overhead and unlocks full link capacity. While RoCE can handle storage traffic, it's less effective with high packet loss (requiring reliable networks), making CPU-bound scenarios its prime use case.

(Reference: NVIDIA Networking Documentation, Section on RoCE Benefits)

Question 8

Question Type: MultipleChoice

Which NVIDIA software provides the capability to virtualize a GPU?

Options:

A- Horizon
B- vGPU
C- virtGPU

Answer:

B

Explanation:

NVIDIA vGPU (Virtual GPU) software enables GPU virtualization by partitioning a physical GPU into multiple virtual instances, assignable to virtual machines or containers for accelerated workloads. Horizon is a VMware product, and ''virtGPU'' isn't an NVIDIA offering, confirming vGPU as the correct solution.

(Reference: NVIDIA vGPU Documentation, Overview Section)

Question 9

Question Type: MultipleChoice

Which option best statements is true about GPUs and CPUs?

Options:

A- GPUs are optimized for parallel tasks, while CPUs are optimized for serial tasks.
B- GPUs have very low bandwidth main memory while CPUs have very high bandwidth main memory.
C- GPUs and CPUs have the same number of cores, but GPUs have higher clock speeds.
D- GPUs and CPUs have identical architectures and can be used interchangeably.

Answer:

A

Explanation:

GPUs and CPUs are architecturally distinct due to their optimization goals. GPUs feature thousands of simpler cores designed for massive parallelism, excelling at executing many lightweight threads concurrently---ideal for tasks like matrix operations in AI. CPUs, conversely, have fewer, more complex cores optimized for sequential processing and handling intricate control flows, making them suited for serial tasks. This divergence in design means GPUs outperform CPUs in parallel workloads, while CPUs excel in single-threaded performance, contradicting claims of identical architectures or interchangeable use.

(Reference: NVIDIA GPU Architecture Whitepaper, Section on GPU vs. CPU Design)

Question 10

Question Type: MultipleChoice

What is the name of NVIDIA's SDK that accelerates machine learning?

Options:

A- Clara
B- RAPIDS
C- cuDNN

Answer:

C

Explanation:

The CUDA Deep Neural Network library (cuDNN) is NVIDIA's SDK specifically designed to accelerate machine learning, particularly deep learning tasks. It provides highly optimized implementations of neural network primitives---such as convolutions, pooling, normalization, and activation functions---leveraging GPU parallelism. Clara focuses on healthcare applications, and RAPIDS accelerates data science workflows, but cuDNN is the core SDK for machine learning acceleration.

(Reference: NVIDIA cuDNN Documentation, Introduction)

Question 11

Question Type: MultipleChoice

Which of the following statements correctly differentiates between AI, Machine Learning, and Deep Learning?

Options:

A- Machine Learning is a subset of AI, and AI is a subset of Deep Learning.
B- AI and Deep Learning are the same, while Machine Learning is a separate concept.
C- AI is a subset of Machine Learning, and Machine Learning is a subset of Deep Learning.
D- Deep Learning is a subset of Machine Learning, and Machine Learning is a subset of AI.

Answer:

D

Explanation:

Artificial Intelligence (AI) is the overarching field encompassing techniques to mimic human intelligence. Machine Learning (ML), a subset of AI, involves algorithms that learn from data. Deep Learning (DL), a specialized subset of ML, uses neural networks with many layers to tackle complex tasks. This hierarchical relationship---DL within ML, ML within AI---is the correct differentiation, unlike the reversed or conflated options.

(Reference: NVIDIA AI Infrastructure and Operations Study Guide, Section on AI, ML, and DL Definitions)

Question 12

Question Type: MultipleChoice

In an AI cluster, what is the purpose of job scheduling?

Options:

A- To gather and analyze cluster data on a regular schedule.
B- To monitor and troubleshoot cluster performance.
C- To assign workloads to available compute resources.
D- To install, update, and configure cluster software.

Answer:

C

Explanation:

Job scheduling in an AI cluster assigns workloads (e.g., training, inference) to available compute resources (GPUs, CPUs), optimizing resource utilization and ensuring efficient execution. It's distinct from data analysis, monitoring, or software management, focusing solely on workload distribution.

(Reference: NVIDIA AI Infrastructure and Operations Study Guide, Section on Job Scheduling)