Free Questions for NCA-AIIO

Shared by Stein on 15-05-2025

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

Which of the following statements correctly highlights a key difference between GPU and CPU architectures?

Options:

A- CPUs are optimized for parallel processing, making them better for AI workloads, while GPUs are designed for sequential tasks
B- GPUs typically have higher clock speeds than CPUs, allowing them to process individual tasks faster
C- CPUs are specialized for graphical computations, whereas GPUs handle general-purpose computing
D- GPUs are optimized for parallel processing, with thousands of smaller cores, while CPUs have fewer, more powerful cores for sequential tasks

Answer:

D

Explanation:

GPUs are optimized for parallel processing, with thousands of smaller cores, while CPUs have fewer, more powerful cores for sequential tasks, correctly highlighting a key architectural difference. NVIDIA GPUs (e.g., A100) excel at parallel computations (e.g., matrix operations for AI), leveraging thousands of cores, whereas CPUs focus on latency-sensitive, single-threaded tasks. This is detailed in NVIDIA's 'GPU Architecture Overview' and 'AI Infrastructure for Enterprise.'

Option (A) reverses the roles. GPUs don't have higher clock speeds (B); CPUs do. CPUs aren't for graphics (C); GPUs are. NVIDIA's documentation confirms (D) as the accurate distinction.

Question 2

Question Type: MultipleChoice

Which two software components are directly involved in the life cycle of AI development and deployment, particularly in model training and model serving? (Select two)

Options:

A- Prometheus
B- MLflow
C- Airflow
D- Apache Spark
E- Kubeflow

Answer:

B, E

Explanation:

MLflow (B) and Kubeflow (E) are directly involved in the AI development and deployment life cycle, particularly for model training and serving. MLflow is an open-source platform for managing the ML lifecycle, including experiment tracking, model training, and deployment, often used with NVIDIA GPUs. Kubeflow is a Kubernetes-native toolkit for orchestrating AI workflows, supporting training (e.g., via TFJob) and serving (e.g., with Triton), as noted in NVIDIA's 'DeepOps' and 'AI Infrastructure and Operations Fundamentals.'

Prometheus (A) is for monitoring, not AI lifecycle tasks. Airflow (C) manages workflows but isn't AI-specific. Apache Spark (D) processes data but isn't focused on model serving. NVIDIA's ecosystem integrates MLflow and Kubeflow for AI workflows.

Question 3

Question Type: MultipleChoice

A retail company is considering using AI to enhance its operations. They want to improve customer experience, optimize inventory management, and personalize marketing campaigns. Which AI use case would be most impactful in achieving these goals?

Options:

A- AI-powered recommendation systems, which personalize product suggestions for customers based on their behavior
B- Natural language processing for automated customer support chatbots
C- AI-driven fraud detection to prevent unauthorized transactions
D- Image recognition for automatic labeling of products in warehouses

Answer:

A

Explanation:

AI-powered recommendation systems are the most impactful use case for improving customer experience, optimizing inventory, and personalizing marketing in retail. These systems, accelerated by NVIDIA GPUs and deployed via Triton Inference Server, analyze customer behavior to deliver tailored suggestions, driving sales, reducing overstock, and enhancing campaigns. NVIDIA's 'State of AI in Retail and CPG' report highlights recommendation systems as a top retail AI application.

NLP chatbots (B) improve support but don't address inventory or marketing directly. Fraud detection (C) is security-focused, not operational. Image recognition (D) aids warehousing but lacks broad impact. NVIDIA prioritizes recommendations for retail goals.

Question 4

Question Type: MultipleChoice

Which of the following is a primary challenge when integrating AI into existing IT infrastructure?

Options:

A- Ensuring AI models have a user-friendly interface
B- Scalability of the AI workloads
C- Finding AI tools that are compatible with existing hardware
D- Selecting the right cloud service provider

Answer:

B

Explanation:

Scalability of AI workloads is a primary challenge when integrating AI into existing IT infrastructure. AI tasks, especially training and inference on NVIDIA GPUs, demand significant compute, memory, and networking resources, which legacy systems may not handle efficiently. Scaling these workloads across clusters or hybrid environments requires careful planning, as noted in NVIDIA's 'AI Infrastructure and Operations Fundamentals' and 'AI Adoption Guide.'

User-friendly interfaces (A) are secondary to technical integration. Hardware compatibility (C) is less challenging with NVIDIA's broad support. Cloud provider selection (D) is a decision, not a core challenge. NVIDIA identifies scalability as a key integration hurdle.

Question 5

Question Type: MultipleChoice

A large manufacturing company is implementing an AI-based predictive maintenance system to reduce downtime and increase the efficiency of its production lines. The AI system must analyze data from thousands of sensors in real-time to predict equipment failures before they occur. However, during initial testing, the system fails to process the incoming data quickly enough, leading to delayed predictions and occasional missed failures. What would be the most effective strategy to enhance the system's real-time processing capabilities?

Options:

A- Reduce the number of sensors to decrease the amount of data the AI system must process
B- Use a more complex AI model to enhance prediction accuracy
C- Implement edge computing to preprocess sensor data closer to the source before sending it to the central AI system
D- Increase the frequency of sensor data collection to provide more detailed inputs for the AI model

Answer:

C

Explanation:

Implementing edge computing to preprocess sensor data closer to the source is the most effective strategy to enhance real-time processing capabilities for a predictive maintenance system. Using NVIDIA Jetson devices at the edge, raw sensor data can be filtered, aggregated, or preprocessed (e.g., via DeepStream), reducing the volume sent to the central GPU cluster (e.g., DGX). This lowers latency and ensures timely predictions, as outlined in NVIDIA's 'Edge AI Solutions' and 'AI Infrastructure for Enterprise.'

Reducing sensors (A) risks missing critical data. A more complex model (B) increases processingdemands, worsening delays. Higher data frequency (D) exacerbates the bottleneck. Edge computing is NVIDIA's recommended solution for real-time IoT workloads.

Question 6

Question Type: MultipleChoice

In an effort to improve energy efficiency in your AI infrastructure using NVIDIA GPUs, you're considering several strategies. Which of the following would most effectively balance energy efficiency with maintaining performance?

Options:

A- Enabling deep sleep mode on all GPUs during processing times
B- Disabling all energy-saving features to ensure maximum performance
C- Running all GPUs at the lowest possible clock speeds
D- Employing NVIDIA GPU Boost technology to dynamically adjust clock speeds

Answer:

D

Explanation:

Employing NVIDIA GPU Boost technology to dynamically adjust clock speeds is the most effective strategy to balance energy efficiency and performance in an AI infrastructure. GPU Boost, available on NVIDIA GPUs like A100, adjusts clock speeds and voltage based on workload demands and thermal conditions, optimizing Performance Per Watt. This ensures high performance when needed while reducing power use during lighter loads, as detailed in NVIDIA's 'GPU Boost Documentation' and 'AI Infrastructure for Enterprise.'

Deep sleep mode (A) during processing disrupts performance. Disabling energy-saving features (B) wastes power. Lowest clock speeds (C) sacrifice performance unnecessarily. GPU Boost is NVIDIA's recommended approach for efficiency.

Question 7

Question Type: MultipleChoice

You are assisting a professional administrator in ensuring data integrity during AI model training in an AI data center. Which of the following strategies would best contribute to maintaining data integrity across distributed GPU nodes?

Options:

A- Use a single master node with GPUs to manage all data processing and then distribute the results to other nodes
B- Assign data verification tasks to DPUs, allowing GPUs to focus solely on model training
C- Utilize redundant GPU nodes to independently process data and compare results post-training
D- Implement a distributed file system with replication, ensuring that each GPU node has access to the same consistent dataset

Answer:

D

Explanation:

Implementing a distributed file system with replication (e.g., GPFS, Lustre) is the best strategy to maintain data integrity across distributed GPU nodes during AI model training. This ensures allnodes access a consistent, replicated dataset, preventing corruption or discrepancies that could skew training results. NVIDIA's 'DGX SuperPOD Reference Architecture' and 'AI Infrastructure and Operations Fundamentals' recommend distributed file systems for data consistency in multi-node GPU clusters, supporting scalability and fault tolerance.

A single master node (A) risks bottlenecks and single-point failures. DPUs for verification (B) offload networking, not data integrity tasks. Redundant processing (C) is inefficient and post-hoc. NVIDIA's guidance favors distributed file systems for integrity.

Question 8

Question Type: MultipleChoice

In a virtualized AI environment, you are responsible for managing GPU resources across several VMs running different AI workloads. Which approach would most effectively allocate GPU resources to maximize performance and flexibility?

Options:

A- Deploy all AI workloads in a single VM with multiple GPUs to centralize resource management
B- Assign a dedicated GPU to each VM to ensure consistent performance for each AI workload
C- Implement GPU virtualization to allow multiple VMs to share GPU resources dynamically based

on demand

D- Use GPU passthrough to allocate full GPU resources directly to one VM at a time, based on the highest priority workload

Answer:

C

Explanation:

Implementing GPU virtualization to allow multiple VMs to share GPU resources dynamically based on demand is the most effective approach for maximizing performance and flexibility in a virtualized AI environment. NVIDIA's GPU virtualization (e.g., via vGPU or GPU Operator in Kubernetes) enables time-slicing or partitioning (e.g., MIG on A100 GPUs), allowing workloads to access GPU resources as needed. This optimizes utilization and adapts to varying demands, as outlined in NVIDIA's 'GPU Virtualization Guide' and 'AI Infrastructure for Enterprise.'

A single VM (A) limits scalability. Dedicated GPUs per VM (B) wastes resources when idle. GPU passthrough (D) restricts sharing, reducing flexibility. NVIDIA recommends virtualization for efficient resource allocation in virtualized AI setups.