Free Questions for NCA-AIIO

Shared by Green on 06-01-2026

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

You are responsible for managing an AI data center that handles large-scale deep learning workloads. The performance of your training jobs has recently degraded, and you've noticed that the GPUs are underutilized while CPU usage remains high. Which of the following actions would most likely resolve this issue?

Options:

A) Increase the GPU memory allocation.
B) Reduce the batch size during training.
C) Add more GPUs to the system.
D) Optimize the data pipeline for better I/O throughput.

Answer:

D

Explanation:

GPU underutilization with high CPU usage during training suggests a bottleneck in the data pipeline, where CPUs can't feed data to GPUs fast enough, starving them of work. Optimizing the data pipeline for better I/O throughput---using NVIDIA DALI for GPU-accelerated data loading or improving storage (e.g., NVMe SSDs)---ensures data reaches GPUs efficiently, maximizing utilization. This is a common issue in NVIDIA DGX systems, where pipeline optimization is critical for large-scale workloads.

Increasing GPU memory (Option A) doesn't address data delivery. Reducing batch size (Option B) might lower GPU demand but reduces throughput, not solving the root cause. Adding GPUs (Option C) exacerbates underutilization without fixing the bottleneck. NVIDIA's training optimization guides prioritize pipeline efficiency.

Question 2

Question Type: MultipleChoice

Which NVIDIA software component is specifically designed to accelerate the end-to-end data science workflow by leveraging GPU acceleration?

Options:

A) NVIDIA CUDA Toolkit
B) NVIDIA DeepStream SDK
C) NVIDIA TensorRT
D) NVIDIA RAPIDS

Answer:

D

Explanation:

NVIDIA RAPIDS is a suite of GPU-accelerated libraries (e.g., cuDF, cuML) designed to speed up the end-to-end data science workflow, from data preparation to machine learning, on NVIDIA GPUs. It integrates with tools like Pandas and Scikit-learn, providing dramatic performance boosts for tasks like ETL, feature engineering, and model training, as used in DGX systems and cloud environments.

The CUDA Toolkit (Option A) is a general-purpose GPU programming platform, not data science-specific. DeepStream SDK (Option B) targets video analytics, not broad data science. TensorRT (Option C) optimizes inference, not the full workflow. RAPIDS is NVIDIA's dedicated data science accelerator.

Question 3

Question Type: MultipleChoice

You manage a large-scale AI infrastructure where several AI workloads are executed concurrently across multiple NVIDIA GPUs. Recently, you observe that certain GPUs are underutilized while others are overburdened, leading to suboptimal performance and extended processing times. Which of the following strategies is most effective in resolving this imbalance?

Options:

A) Implementing dynamic GPU load balancing across the infrastructure
B) Reducing the batch size for all AI workloads
C) Increasing the power limit on underutilized GPUs
D) Disabling GPU overclocking to normalize performance

Answer:

A

Explanation:

Uneven GPU utilization in a multi-GPU infrastructure indicates poor workload distribution. Implementing dynamic GPU load balancing---using tools like NVIDIA Triton Inference Server or Kubernetes with GPU Operator---assigns tasks based on real-time GPU usage, ensuring balanced workloads and optimal performance. This strategy, common in DGX clusters, reduces processing times by preventing overburdening or idling.

Reducing batch size (Option B) lowers GPU demand uniformly but doesn't address imbalance and may reduce throughput. Increasing power limits (Option C) might boost underutilized GPUs slightly but doesn't fix distribution. Disabling overclocking (Option D) ensures consistency but not balance. Dynamic balancing is NVIDIA's recommended approach.

Question 4

Question Type: MultipleChoice

You are responsible for managing an AI infrastructure that runs a critical deep learning application. The application experiences intermittent performance drops, especially when processing large datasets. Upon investigation, you find that some of the GPUs are not being fully utilized while others are overloaded, causing the overall system to underperform. What would be the most effective solution to address the uneven GPU utilization and optimize the performance of the deep learning application?

Options:

A) Reduce the size of the datasets being processed.
B) Increase the clock speed of the GPUs.
C) Add more GPUs to the system.
D) Implement dynamic load balancing for the GPUs.

Answer:

D

Explanation:

Intermittent performance drops due to uneven GPU utilization stem from workload imbalance. Dynamic load balancing, enabled by NVIDIA tools like Triton Inference Server or Kubernetes with GPU Operator, redistributes tasks based on GPU utilization, ensuring even processing of large datasets. This optimizes performance in DGX or multi-GPU setups by preventing overload and underuse, directly addressing the root cause.

Reducing dataset size (Option A) compromises model quality and doesn't fix distribution. Increasing clock speed (Option B) may help overloaded GPUs but not underutilized ones. Adding GPUs (Option C) increases capacity but not balance. NVIDIA's infrastructure solutions favor dynamic balancing for critical applications.

Question 5

Question Type: MultipleChoice

Your company is planning to deploy a range of AI workloads, including training a large convolutional neural network (CNN) for image classification, running real-time video analytics, and performing batch processing of sensor dat

a. What type of infrastructure should be prioritized to support these diverse AI workloads effectively?

Options:

A) On-premise servers with large storage capacity
B) CPU-only servers with high memory capacity
C) A cloud-based infrastructure with serverless computing options
D) A hybrid cloud infrastructure combining on-premise servers and cloud resources

Answer:

D

Explanation:

Diverse AI workloads---training CNNs (compute-heavy), real-time video analytics (latency-sensitive), and batch sensor processing (data-intensive)---require flexible, scalable infrastructure. A hybrid cloud infrastructure, combining on-premise NVIDIA GPU servers (e.g., DGX) with cloud resources (e.g., DGX Cloud), provides the best of both: on-premise control for sensitive data or latency-critical tasks and cloud scalability for burst compute or storage needs. NVIDIA's hybrid solutions support this versatility across workload types.

On-premise alone (Option A) lacks scalability. CPU-only servers (Option B) can't handle GPU-accelerated AI efficiently. Serverless cloud (Option C) suits lightweight tasks, not heavy AI workloads. Hybrid cloud is NVIDIA's strategic fit for diverse AI.

Question 6

Question Type: MultipleChoice

You are working with a large healthcare dataset containing millions of patient records. Your goal is to identify patterns and extract actionable insights that could improve patient outcomes. The dataset is highly dimensional, with numerous variables, and requires significant processing power to analyze effectively. Which two techniques are most suitable for extracting meaningful insights from this large, complex dataset? (Select two)

Options:

A) SMOTE (Synthetic Minority Over-sampling Technique)
B) Data Augmentation
C) Batch Normalization
D) K-means Clustering
E) Dimensionality Reduction (e.g., PCA)

Answer:

D, E

Explanation:

A large, high-dimensional healthcare dataset requires techniques to uncover patterns and reduce complexity. K-means Clustering (Option D) groups similar patient records (e.g., by symptoms or outcomes), identifying actionable patterns using NVIDIA RAPIDS cuML for GPU acceleration. Dimensionality Reduction (Option E), like PCA, reduces variables to key components, simplifying analysis while preserving insights, also accelerated by RAPIDS on NVIDIA GPUs (e.g., DGX systems).

SMOTE (Option A) addresses class imbalance, not general pattern extraction. Data Augmentation (Option B) enhances training data, not insight extraction. Batch Normalization (Option C) is a training technique, not an analysis tool. NVIDIA's data science tools prioritize clustering and dimensionality reduction for such tasks.

Question 7

Question Type: MultipleChoice

You are working on a project that involves monitoring the performance of an AI model deployed in production. The model's accuracy and latency metrics are being tracked over time. Your task, under the guidance of a senior engineer, is to create visualizations that help the team understand trends in these metrics and identify any potential issues. Which visualization would be most effective for showing trends in both accuracy and latency metrics over time?

Options:

A) Pie chart showing the distribution of accuracy metrics.
B) Box plot comparing accuracy and latency.
C) Dual-axis line chart with accuracy on one axis and latency on the other.
D) Stacked area chart showing cumulative accuracy and latency.

Answer:

C

Explanation:

Tracking accuracy and latency trends over time requires a visualization that shows both metrics' evolution clearly. A dual-axis line chart, with accuracy on one axis and latency on the other, plots each as a line against time, revealing correlations (e.g., latency spikes reducing accuracy) and trends. NVIDIA RAPIDS supports such visualizations on GPUs, enhancing real-time monitoring in production environments like DGX or Triton deployments.

Pie charts (Option A) show distributions, not trends. Box plots (Option B) summarize static data, not time-based changes. Stacked area charts (Option C) imply cumulative values, confusing for independent metrics. Dual-axis is NVIDIA-aligned for performance analysis.

Question 8

Question Type: MultipleChoice

Your organization runs multiple AI workloads on a shared NVIDIA GPU cluster. Some workloads are more critical than others. Recently, you've noticed that less critical workloads are consuming more GPU resources, affecting the performance of critical workloads. What is the best approach to ensure that critical workloads have priority access to GPU resources?

Options:

A) Implement GPU Quotas with Kubernetes Resource Management
B) Use CPU-based Inference for Less Critical Workloads
C) Upgrade the GPUs in the Cluster to More Powerful Models
D) Implement Model Optimization Techniques

Answer:

A

Explanation:

Ensuring critical workloads have priority in a shared GPU cluster requires resource control. Implementing GPU Quotas with Kubernetes Resource Management, using NVIDIA GPU Operator, assigns resource limits and priorities, ensuring critical tasks (e.g., via pod priority classes) access GPUs first. This aligns with NVIDIA's cluster management in DGX or cloud setups, balancing utilization effectively.

CPU-based inference (Option B) reduces GPU load but sacrifices performance for non-critical tasks. Upgrading GPUs (Option C) increases capacity, not priority. Model optimization (Option D) improves efficiency but doesn't enforce priority. Quotas are NVIDIA's recommended strategy.

Question 9

Question Type: MultipleChoice

A financial services company is developing a machine learning model to detect fraudulent transactions in real-time. They need to manage the entire AI lifecycle, from data preprocessing to model deployment and monitoring. Which combination of NVIDIA software components should they integrate to ensure an efficient and scalable AI development and deployment process?

Options:

A) NVIDIA Clara for model training, TensorRT for data processing, and Jetson for deployment.
B) NVIDIA RAPIDS for data processing, TensorRT for model optimization, and Triton Inference Server for deployment.
C) NVIDIA DeepStream for data processing, CUDA for model training, and NGC for deployment.
D) NVIDIA Metropolis for data collection, DIGITS for training, and Triton Inference Server for

deployment.

Answer:

B

Explanation:

The AI lifecycle for real-time fraud detection needs efficient data preprocessing, model optimization, and deployment. NVIDIA RAPIDS accelerates data processing on GPUs, TensorRToptimizes models for low-latency inference, and Triton Inference Server scales deployment across platforms---perfect for financial use cases in NVIDIA DGX or cloud environments.

Clara (Option A) is healthcare-focused, not fraud. DeepStream (Option C) is video-centric, and CUDA isn't a full training solution. Metropolis (Option D) targets smart cities, and DIGITS is outdated. Option B aligns with NVIDIA's lifecycle strategy.