NVIDIA
NCA-AIIO
AI Infrastructure and Operations

Latest Version: 9.0

Question: 1

Which NVIDIA solution is specifically designed to accelerate data analytics and machine learning workloads, allowing data scientists to build and deploy models at scale using GPUs?

A. NVIDIA CUDA
B. NVIDIA JetPack
C. NVIDIA RAPIDS
D. NVIDIA DGX A100

Answer: C

Explanation:
NVIDIA RAPIDS is an open-source suite of GPU-accelerated libraries specifically designed to speed up data analytics and machine learning workflows. It enables data scientists to leverage GPU parallelism to process large datasets and build machine learning models at scale, significantly reducing computation time compared to traditional CPU-based approaches. RAPIDS includes libraries like cuDF (for dataframes), cuML (for machine learning), and cuGraph (for graph analytics), which integrate seamlessly with popular frameworks like pandas, scikit-learn, and Apache Spark.
In contrast:
NVIDIA CUDA(A) is a parallel computing platform and programming model that enables GPU acceleration but is not a specific solution for data analytics or machine learning—it's a foundational technology used by tools like RAPIDS.
NVIDIA JetPack(B) is a software development kit for edge AI applications, primarily targeting NVIDIA Jetson devices for robotics and IoT, not large-scale data analytics.
NVIDIA DGX A100(D) is a hardware platform (a powerful AI system with multiple GPUs) optimized for training and inference, but it's not a software solution for data analytics workflows—it's the infrastructure that could run RAPIDS.
Thus, RAPIDS (C) is the correct answer as it directly addresses the question's focus on accelerating data analytics and machine learning workloads using GPUs.
Reference:NVIDIA RAPIDS documentation on nvidia.com; NVIDIA AI Infrastructure overview.

Question: 2

Your team is running an AI inference workload on a Kubernetes cluster with multiple NVIDIA GPUs. You observe that some nodes with GPUs are underutilized, while others are overloaded, leading to inconsistent inference performance across the cluster. Which strategy would most effectively balance the GPU workload across the Kubernetes cluster?

A. Deploying a GPU-aware scheduler in Kubernetes
B. Implementing GPU resource quotas to limit GPU usage per pod

C. Using CPU-based autoscaling to balance the workload
D. Reducing the number of GPU nodes in the cluster

Answer: A

Explanation:
Deploying a GPU-aware scheduler in Kubernetes (A) is the most effective strategy to balance GPU workloads across a cluster. Kubernetes by default does not natively understand GPU resources beyond basic resource requests and limits. A GPU-aware scheduler, such as the NVIDIA GPU Operator with Kubernetes, enhances the orchestration by intelligently distributing workloads basedon GPU availability, utilization, and specific requirements of the inference tasks. This ensures that underutilized nodes are assigned work while preventing overloading of others, leading to consistent performance.
Implementing GPU resource quotas(B) can limit GPU usage per pod, but it doesn't dynamically balance workloads across nodes—it only caps resource consumption, potentially leaving some GPUs idle if quotas are too restrictive.
Using CPU-based autoscaling(C) focuses on CPU metrics and ignores GPU-specific utilization, making it ineffective for GPU workload balancing in this scenario.
Reducing the number of GPU nodes(D) might exacerbate the issue by reducing overall capacity, not addressing the imbalance.
The NVIDIA GPU Operator integrates with Kubernetes to provide GPU-aware scheduling, monitoring, and management, making (A) the optimal solution.
Reference:NVIDIA GPU Operator documentation; Kubernetes integration with NVIDIA GPUs on nvidia.com.

Question: 3

A large enterprise is deploying a high-performance AI infrastructure to accelerate its machine learning workflows. They are using multiple NVIDIA GPUs in a distributed environment. To optimize the workload distribution and maximize GPU utilization, which of the following tools or frameworks should be integrated into their system? (Select two)

A. NVIDIA CUDA
B. NVIDIA NGC (NVIDIA GPU Cloud)
C. TensorFlow Serving
D. NVIDIA NCCL (NVIDIA Collective Communications Library)
E. Keras

Answer: A,D

Explanation:
In a distributed environment with multiple NVIDIA GPUs, optimizing workload distribution and GPU utilization requires tools that enable efficient computation and communication:
NVIDIA CUDA(A) is a foundational parallel computing platform that allows developers to harness GPU power for general-purpose computing, including machine learning. It's essential for programming GPUs and optimizing workloads in a distributed setup.

NVIDIA NCCL(D) (NVIDIA Collective Communications Library) is designed for multi-GPU and multi-node communication, providing optimized primitives (e.g., all-reduce, broadcast) for collective operations in deep learning. It ensures efficient data exchange between GPUs, maximizing utilization in distributed training.
NVIDIA NGC(B) is a hub for GPU-optimized containers and models, useful for deployment but not directly responsible for workload distribution or GPU utilization optimization.
TensorFlow Serving(C) is a framework for deploying machine learning models for inference, not for optimizing distributed training or GPU utilization during model development.
Keras(E) is a high-level API for building neural networks, but it lacks the low-level control needed for distributed workload optimization—it relies on backends like TensorFlow or CUDA.
Thus, CUDA (A) and NCCL (D) are the best choices for this scenario.
Reference:NVIDIA CUDA Toolkit documentation; NVIDIA NCCL documentation on nvidia.com.

Question: 4

You are managing the deployment of an AI-driven security system that needs to process video streams from thousands of cameras across multiple locations in real time. The system must detectpotential threats and send alerts with minimal latency. Which NVIDIA solution would be most appropriate to handle this large-scale video analytics workload?

A. NVIDIA Clara Guardian
B. NVIDIA Jetson Nano
C. NVIDIA DeepStream
D. NVIDIA RAPIDS

Answer: C

Explanation:
NVIDIA DeepStream (C) is specifically designed for large-scale, real-time video analytics workloads. It provides a software development kit (SDK) that leverages NVIDIA GPUs to process multiple video streams simultaneously, enabling tasks like object detection, classification, and tracking with minimal latency. DeepStream integrates with deep learning frameworks (e.g., TensorRT) and supports scalable deployment across distributed systems, making it ideal for a security system processing thousands of camera feeds.
NVIDIA Clara Guardian(A) is focused on healthcare applications, such as smart hospitals and medical imaging, not general-purpose video analytics for security.
NVIDIA Jetson Nano(B) is an edge computing platform for small-scale AI tasks, unsuitable for handling thousands of streams due to its limited processing power.
NVIDIA RAPIDS(D) accelerates data analytics and machine learning, not real-time video processing.
DeepStream's ability to handle high-throughput video analytics with low latency makes it the best fit (C).
Reference:NVIDIA DeepStream SDK documentation on nvidia.com.

Question: 5

A data center is running a cluster of NVIDIA GPUs to support various AI workloads. The operations team needs to monitor GPU performance to ensure workloads are running efficiently and to prevent potential hardware failures. Which two key measures should they focus on to monitor the GPUs effectively? (Select two)

A. Disk I/O rates
B. CPU clock speed
C. GPU temperature and power consumption
D. GPU memory utilization
E. Network bandwidth usage

Answer: C,D

Explanation:
To monitor GPU performance effectively in an AI data center, the focus should be on metrics directly tied to GPU health and efficiency:
GPU temperature and power consumption(C) are critical to prevent overheating and power-related failures, which can disrupt workloads or damage hardware. High temperatures or excessive power draw indicate potential issues requiring intervention.
GPU memory utilization(D) reflects how much of the GPU's memory is being used by workloads. High utilization can lead to memory bottlenecks, while low utilization might indicate underuse, both affecting efficiency.
Disk I/O rates(A) relate to storage performance, not GPU operation directly.
CPU clock speed(B) is a CPU metric, irrelevant to GPU monitoring in this context.
Network bandwidth usage(E) is important for distributed systems but doesn't directly assess GPU performance or health.
NVIDIA tools like NVIDIA System Management Interface (nvidia-smi) provide these metrics (C and D), making them essential for monitoring.
Reference:NVIDIA Data Center GPU Management documentation; nvidia-smi usage guide on nvidia.com.

Thank You for Trying Our Product

For More Information – Visit link below:

https://www.examsboost.com/

15 USD Discount Coupon Code:

G74JA8UF

FEATURES

- ✓ 90 Days Free Updates
- ✓ Money Back Pass Guarantee
- ✓ Instant Download or Email Attachment
- ✓ 24/7 Live Chat Support
- ✓ PDF file could be used at any Platform
- ✓ 50,000 Happy Customer