



Vendor: NVIDIA

Exam Code: NCP-AIO

Exam Name: AI Operations

Version: DEMO

QUESTION 1

Your organization is running multiple AI models on a single A100 GPU using MIG in a multi-tenant environment. One of the tenants reports a performance issue, but you notice that other tenants are unaffected.

What feature of MIG ensures that one tenant's workload does not impact others?

- A. Hardware-level isolation of memory, cache, and compute resources for each instance.
- B. Dynamic resource allocation based on workload demand.
- C. Shared memory access across all instances.
- D. Automatic scaling of instances based on workload size.

Answer: A

Explanation:

NVIDIA's Multi-Instance GPU (MIG) technology provides hardware-level isolation of critical GPU resources such as memory, cache, and compute units for each GPU instance. This ensures that workloads running in one instance are fully isolated and cannot interfere with the performance of workloads in other instances, supporting multi-tenancy without contention.

QUESTION 2

You are deploying an AI workload on a Kubernetes cluster that requires access to GPUs for training deep learning models. However, the pods are not able to detect the GPUs on the nodes.

What would be the first step to troubleshoot this issue?

- A. Verify that the NVIDIA GPU Operator is installed and running on the cluster.
- B. Ensure that all pods are using the latest version of TensorFlow or PyTorch.
- C. Check if the nodes have sufficient memory allocated for AI workloads.
- D. Increase the number of CPU cores allocated to each pod to ensure better resource utilization.

Answer: A

Explanation:

The first step in troubleshooting Kubernetes pods that cannot detect GPUs is to verify whether the NVIDIA GPU Operator is properly installed and running. The GPU Operator manages the installation and configuration of all NVIDIA GPU components in the cluster, including drivers, device plugins, and monitoring tools. Without it, pods will not have access to GPU resources. Ensuring correct installation and operational status of the GPU Operator is essential before checking application-level versions or resource allocations.

QUESTION 3

Your Kubernetes cluster is running a mixture of AI training and inference workloads. You want to ensure that inference services have higher priority over training jobs during peak resource usage times.

How would you configure Kubernetes to prioritize inference workloads?

- A. Increase the number of replicas for inference services so they always have more resources than training jobs.
- B. Set up a separate namespace for inference services and limit resource usage in other namespaces.
- C. Use Horizontal Pod Autoscaling (HPA) based on memory usage to scale up inference services during peak times.

- D. Implement ResourceQuotas and PriorityClasses to assign higher priority and resource guarantees to inference workloads over training jobs.

Answer: D

Explanation:

To prioritize inference workloads over training jobs in Kubernetes, administrators should configure PriorityClasses and ResourceQuotas. PriorityClasses allow assigning different priority levels to pods, ensuring that during resource contention, higher-priority pods (inference services) receive resources first. ResourceQuotas limit the resource consumption per namespace or user, controlling overall usage and reserving capacity for critical workloads. This setup effectively manages resource allocation and guarantees performance for inference jobs during peak times.

QUESTION 4

You are managing a high availability (HA) cluster that hosts mission-critical applications. One of the nodes in the cluster has failed, but the application remains available to users.

What mechanism is responsible for ensuring that the workload continues to run without interruption?

- A. Load balancing across all nodes in the cluster.
- B. Manual intervention by the system administrator to restart services.
- C. The failover mechanism that automatically transfers workloads to a standby node.
- D. Data replication between nodes to ensure data integrity.

Answer: C

Explanation:

In an HA cluster, the failover mechanism is responsible for detecting node failures and automatically transferring workloads to a standby or redundant node to maintain service availability. This process ensures mission-critical applications continue running without interruption. Load balancing helps distribute traffic but does not handle node failures. Manual intervention is not ideal for HA, and data replication ensures data integrity but does not itself manage workload continuity.

QUESTION 5

A system administrator needs to collect the information below:

```
GPU behavior monitoring
GPU configuration management
GPU policy oversight
GPU health and diagnostics
GPU accounting and process statistics
NVSwitch configuration and monitoring
```

What single tool should be used?

- A. nvidia-smi
- B. CUDA Toolkit
- C. DCGM
- D. Nsight Systems

Answer: C

Explanation:

The NVIDIA Data Center GPU Manager (DCGM) is the comprehensive management tool that provides all the requested functionalities: monitoring GPU behavior, managing configurations, enforcing policies, health diagnostics, process accounting, and NVSwitch monitoring. DCGM is designed for large-scale GPU management in data centers and AI clusters, providing detailed telemetry and control over NVIDIA GPUs and NVSwitches.

QUESTION 6

Your organization is deploying an AI workload that requires high-throughput access to shared storage across multiple servers. The workload involves both training and inference tasks that need fast read and write speeds.

Which storage architecture would best support this AI workload?

- A. Use local storage on each server to minimize network traffic between nodes.
- B. Prioritize write performance over read performance since training tasks dominate AI workflows.
- C. A high-performance shared storage system that supports both high read and write IO performance.
- D. Use SSD-based shared storage systems to save costs while scaling up storage capacity.

Answer: C

Explanation:

For AI workloads involving both training and inference across multiple servers, a high-performance shared storage system that supports both high read and write I/O performance is essential. This ensures fast data access and efficient coordination between distributed compute nodes, preventing bottlenecks in data throughput. Local storage may minimize network traffic but lacks the necessary data sharing and coordination. Prioritizing only write performance neglects inference workload needs, and cost-saving SSD options might not deliver the required performance at scale. Hence, option C is the best choice for balanced, high-throughput AI workloads.

QUESTION 7

What is the primary purpose of assigning a provisioning role to a node in NVIDIA Base Command Manager (BCM)?

- A. To configure the node as a container orchestration manager
- B. To enable the node to monitor GPU utilization across the cluster
- C. To allow the node to manage software images and provision other nodes
- D. To assign the node as a storage manager for certified storage

Answer: C

Explanation:

In NVIDIA Base Command Manager (BCM), assigning the provisioning role to a node enables that node to manage software images and perform provisioning tasks for other nodes in the cluster. This role allows automated deployment and configuration of cluster nodes, ensuring consistency and simplifying large-scale management. It is not primarily responsible for container orchestration, GPU monitoring, or storage management.

QUESTION 8

You are using BCM for configuring an active-passive high availability (HA) cluster for a firewall system. To ensure seamless failover, what is one best practice related to session synchronization between the active and passive nodes?

- A. Configure both nodes with different zone names to avoid conflicts during failover.
- B. Use heartbeat network for session synchronization between active and passive nodes.
- C. Ensure that both nodes use different firewall models for redundancy.
- D. Set up manual synchronization procedures to transfer session data when needed.

Answer: B

Explanation:

A best practice for active-passive HA clusters, such as for firewall systems managed via BCM, is to use a heartbeat network to synchronize session state data between active and passive nodes. This real-time synchronization allows the passive node to take over seamlessly in case the active node fails, maintaining session continuity and minimizing downtime. Configuring different zone names or firewall models can cause incompatibility, and manual synchronization is prone to errors and delays.

QUESTION 9

A system administrator of a high-performance computing (HPC) cluster that uses an InfiniBand fabric for high-speed interconnects between nodes received reports from researchers that they are experiencing unusually slow data transfer rates between two specific compute nodes. The system administrator needs to ensure the path between these two nodes is optimal. What command should be used?

- A. ibtracert
- B. ibstatus
- C. ibping
- D. ibnetdiscover

Answer: A

Explanation:

To verify the optimal communication path and diagnose issues between two nodes in an InfiniBand fabric, the ibtracert command is used. It traces the route that InfiniBand packets take through the fabric, identifying each hop and any potential bottlenecks or faulty links along the path.

QUESTION 10

An administrator is troubleshooting a bottleneck in a deep learning run time and needs consistent data feed rates to GPUs. Which storage metric should be used?

- A. Disk I/O operations per second (IOPS)
- B. Disk free space
- C. Sequential read speed
- D. Disk utilization in performance manager

Answer: C

Explanation:

When troubleshooting performance bottlenecks related to feeding data consistently to GPUs during deep learning workloads, the key storage metric to consider is sequential read speed. Deep learning training typically involves streaming large datasets sequentially from storage to GPUs. The sequential read speed measures how fast data can be read in a continuous stream, directly impacting the ability to keep GPUs fed without stalls.

Thank You for Trying Our Product

Lead2pass Certification Exam Features:

- ★ More than **99,900** Satisfied Customers Worldwide.
- ★ Average **99.9%** Success Rate.
- ★ **Free Update** to match latest and real exam scenarios.
- ★ **Instant Download** Access! No Setup required.
- ★ Questions & Answers are downloadable in **PDF** format and **VCE** test engine format.
- ★ Multi-Platform capabilities - **Windows, Laptop, Mac, Android, iPhone, iPod, iPad**.
- ★ **100%** Guaranteed Success or **100%** Money Back Guarantee.
- ★ **Fast**, helpful support **24x7**.



View list of all certification exams: <http://www.lead2pass.com/all-products.html>



10% Discount Coupon Code: ASTR14