KaozhengPro

IT認證考試題庫

專業平臺

考證寶提供最新考古題與模擬試題
協助您高效通過認證考試

www.kaozhengpro.com

Exam : **NCP-AIO**

Title : NVIDIA Certified Professional AI Operations

Version : DEMO

1.A data scientist is training a deep learning model and notices slower than expected training times. The data scientist alerts a system administrator to inspect the issue. The system administrator suspects the disk IO is the issue.
What command should be used?
A. tcpdump
B. iostat
C. nvidia-smi
D. htop
Answer: B
Explanation:
Comprehensive and Detailed Explanation From Exact Extract:
To diagnose disk IO performance issues, the system administrator should use theiostatcommand, which reports CPU statistics and input/output statistics for devices and partitions. It helps identify bottlenecks in disk throughput or latency affecting application performance.
tcpdumpis used for network traffic analysis, not disk IO.
nvidia-smimonitors NVIDIA GPU status but not disk IO.
htopshows CPU, memory, and process usage but provides limited disk IO details.
Therefore,iostatis the appropriate tool to assess disk IO performance and diagnose bottlenecks impacting training times.

2.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:
Comprehensive and Detailed Explanation From Exact Extract:
To verify the optimal communication path and diagnose issues between two nodes in an InfiniBand fabric, theibtracertcommand 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.
ibstatusprovides status information about local InfiniBand devices and ports.
ibpingtests connectivity and latency between nodes.
ibnetdiscoverdiscovers and prints the topology of the InfiniBand fabric but does not trace specific paths.
Therefore,ibtracertis the appropriate tool for path optimization verification between two compute nodes.

3.You are tasked with deploying a DOCA service on an NVIDIA BlueField DPU in an air-gapped data center environment. The DPU has the required BlueField OS version (3.9.0 or higher) installed, and you have access to the necessary container image from NVIDIA's NGC catalog. However, you need to ensure that the deployment process is successful without an internet connection.

Which of the following steps should you take to deploy the DOCA service on the DPU?
A. Install Docker on the DPU, pull the container directly from NGC, and run it using ‘docker run’ with appropriate environment variables.
B. Pull the container image from NGC using Docker and modify the YAML file before deployment.
C. Manually download the container image and YAML file beforehand, transfer them to the DPU, and deploy using Kubernetes with standalone Kubelet.
D. Use the host system’s Docker engine to pull the container image and deploy it on the DPU via SSH.
Answer: C
Explanation:
Comprehensive and Detailed Explanation From Exact Extract:
In an air-gapped environment where the DPU has no internet connectivity, direct pulling of container images from NVIDIA’s NGC catalog is not possible. The recommended approach is tomanually download the required container image and YAML deployment filesfrom a connected system, then transfer these files to the DPU. Deployment is then performed using Kubernetes with a standalone Kubelet on the DPU, which can deploy the preloaded container image offline. This ensures the deployment proceeds successfully without internet access.

4.A system administrator needs to scale a Kubernetes Job to 4 replicas.
What command should be used?
A. kubectl stretch job --replicas=4
B. kubectl autoscale deployment job --min=1 --max=10
C. kubectl scale job --replicas=4
D. kubectl scale job -r 4
Answer: C
Explanation:
Comprehensive and Detailed Explanation From Exact Extract:
The correct command to scale a Kubernetes Job to a specific number of replicas iskubectl scale job --replicas=4. This explicitly sets the number of desired pod instances for the Job resource. The other commands are either invalid (stretch), apply to Deployments rather than Jobs (autoscale deployment), or use incorrect syntax (-r).

5.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:
Comprehensive and Detailed Explanation From Exact Extract:
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.
Disk I/O operations per second (IOPS) measures random read/write operations and is less relevant for large sequential data streams in AI workloads.
Disk free space indicates available storage capacity but does not impact data feed rate.
Disk utilization in performance manager shows overall usage but does not specify the speed or consistency of data feed.
Therefore, focusing on sequential read speed (option C) is critical for ensuring consistent, high-throughput data feeding to GPUs, minimizing bottlenecks in deep learning runtime environments.
This is consistent with NVIDIA AI Operations best practices for system performance optimization and troubleshooting storage-related issues in AI infrastructure.