



- Expert Verified, Online, **Free**.



CERTIFICATION TEST

- CertificationTest.net - Cheap & Quality Resources With Best Support

A Slurm user needs to submit a batch job script for execution tomorrow.
Which command should be used to complete this task?

- A. sbatch -begin=tomorrow
- B. submit -begin=tomorrow
- C. salloc -begin=tomorrow
- D. srun -begin=tomorrow

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!

You are configuring networking for a new AI cluster in your data center. The cluster will handle large-scale distributed training jobs that require fast communication between servers.

What type of networking architecture can maximize performance for these AI workloads?

- A. Implement a leaf-spine network topology using standard Ethernet switches to ensure scalability as more nodes are added.
- B. Prioritize out-of-band management networks over compute networks to ensure efficient job scheduling across nodes.
- C. Use standard Ethernet networking with a focus on increasing bandwidth through multiple connections per server.
- D. Use InfiniBand networking to provide low-latency, high-throughput communication between servers in the cluster.

Suggested Answer: *D*

Currently there are no comments in this discussion, be the first to comment!

A system administrator needs to optimize the delivery of their AI applications to the edge.
What NVIDIA platform should be used?

- A. Base Command Platform
- B. Base Command Manager
- C. Fleet Command
- D. NetQ

Suggested Answer: *C*

Currently there are no comments in this discussion, be the first to comment!

A Slurm user is experiencing a frequent issue where a Slurm job is getting stuck in the "PENDING" state and unable to progress to the "RUNNING" state.

Which Slurm command can help the user identify the reason for the job's pending status?

- A. `sinfo -R`
- B. `scontrol show job <jobid></jobid>`
- C. `sacct -j <job[.step]></job[.step]>`
- D. `squeue -u <user_list></user_list>`

Suggested Answer: *B*

Currently there are no comments in this discussion, be the first to comment!

You are a Solutions Architect designing a data center infrastructure for a cloud-based AI application that requires high-performance networking, storage, and security. You need to choose a software framework to program the NVIDIA BlueField DPUs that will be used in the infrastructure. The framework must support the development of custom applications and services, as well as enable tailored solutions for specific workloads. Additionally, the framework should allow for the integration of storage services such as NVMe over Fabrics (NVMe-oF) and elastic block storage. Which framework should you choose?

- A. NVIDIA TensorRT
- B. NVIDIA CUDA
- C. NVIDIA NSight
- D. NVIDIA DOCA

Suggested Answer: *D*

Currently there are no comments in this discussion, be the first to comment!

You are managing a Slurm cluster with multiple GPU nodes, each equipped with different types of GPUs. Some jobs are being allocated GPUs that should be reserved for other purposes, such as display rendering.

How would you ensure that only the intended GPUs are allocated to jobs?

- A. Verify that the GPUs are correctly listed in both gres.conf and slurm.conf, and ensure that unconfigured GPUs are excluded.
- B. Use nvidia-smi to manually assign GPUs to each job before submission.
- C. Reinstall the NVIDIA drivers to ensure proper GPU detection by Slurm.
- D. Increase the number of GPUs requested in the job script to avoid using unconfigured GPUs.

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!

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

Suggested Answer: *B*

Currently there are no comments in this discussion, be the first to comment!

You have noticed that users can access all GPUs on a node even when they request only one GPU in their job script using `--gres=gpu:1`. This is causing resource contention and inefficient GPU usage.

What configuration change would you make to restrict users' access to only their allocated GPUs?

- A. Increase the memory allocation per job to limit access to other resources on the node.
- B. Enable cgroup enforcement in `cgroup.conf` by setting `ConstrainDevices=yes`.
- C. Set a higher priority for Jobs requesting fewer GPUs, so they finish faster and free up resources sooner.
- D. Modify the job script to include additional resource requests for CPU cores alongside GPUs.

Suggested Answer: *B*

Currently there are no comments in this discussion, be the first to comment!

A new researcher needs access to GPU resources but should not have permission to modify cluster settings or manage other users. What role should you assign them in Run:ai?

- A. L1 Researcher
- B. Department Administrator
- C. Application Administrator
- D. Research Manager

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!

When troubleshooting Slurm job scheduling issues, a common source of problems is jobs getting stuck in a pending state indefinitely. Which Slurm command can be used to view detailed information about all pending jobs and identify the cause of the delay?

- A. scontrol
- B. sacct
- C. sinfo

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!

What must be done before installing new versions of DOCA drivers on a BlueField DPU?

- A. Uninstall any previous versions of DOCA drivers.
- B. Re-flash the firmware every time.
- C. Disable network interfaces during installation.
- D. Reboot the host system.

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!

A Slurm user needs to display real-time information about the running processes and resource usage of a Slurm job. Which command should be used?

- A. `smap -j <jobid></jobid>`
- B. `scontrol show job <jobid></jobid>`
- C. `sstat -j <job(.step)></job(.step)>`
- D. `sinfo -j <jobid></jobid>`

Suggested Answer: C

Currently there are no comments in this discussion, be the first to comment!

Which two (2) ways does the pre-configured GPU Operator in NVIDIA Enterprise Catalog differ from the GPU Operator In the public NGC catalog? (Choose two.)

- A. It is configured to use a prebuilt vGPU driver image.
- B. It supports Mixed Strategies for Kubernetes deployments.
- C. It automatically installs the NVIDIA Datacenter driver.
- D. It is configured to use the NVIDIA License System (NLS).
- E. It additionally installs Network Operator.

Suggested Answer: AD

Currently there are no comments in this discussion, be the first to comment!

You are managing multiple edge AI deployments using NVIDIA Fleet Command. You need to ensure that each AI application running on the same GPU is isolated from others to prevent Interference.

Which feature of Fleet Command should you use to achieve this?

- A. Remote Console
- B. Secure NFS support
- C. Multi-Instance GPU (MIG) support
- D. Over-the-air updates

Suggested Answer: *C*

Currently there are no comments in this discussion, be the first to comment!

You are deploying AI applications at the edge and want to ensure they continue running even if one of the servers at an edge location fails. How can you configure NVIDIA Fleet Command to achieve this?

- A. Use Secure NFS support for data redundancy.
- B. Set up over-the-air updates to automatically restart failed applications.
- C. Enable high availability for edge clusters.
- D. Configure Fleet Command's multi-instance GPU (MIG) to handle failover.

Suggested Answer: *C*

Currently there are no comments in this discussion, be the first to comment!

You are an administrator managing a large-scale Kubernetes-based GPU cluster using Run:AI.

To automate repetitive administrative tasks and efficiently manage resources across multiple nodes, which of the following is essential when using the Run:AI Administrator CLI for environments where automation or scripting is required?

- A. Use the `runai-adm` command to directly update Kubernetes nodes without requiring `kubectl`.
- B. Use the CLI to manually allocate specific GPUs to individual jobs for better resource management.
- C. Ensure that the Kubernetes configuration file is set up with cluster administrative rights before using the CLI.
- D. Install the CLI on Windows machines to take advantage of its scripting capabilities.

Suggested Answer: *c*

Currently there are no comments in this discussion, be the first to comment!

A Fleet Command system administrator wants to create an organization user that will have the following rights:

For locations - read only -

For Applications - read/write/admin

For Deployments - read/write/admin

For Dashboards - read only -

What role should the system administrator assign to this user?

- A. Fleet Command Operator
- B. Fleet Command Admin
- C. Fleet Command Supporter
- D. Fleet Command Viewer

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!


An organization only needs basic network monitoring and validation tools.
Which UFM platform should they use?

- A. UFM Enterprise
- B. UFM Telemetry
- C. UFM Cyber-AI
- D. UFM Pro

Suggested Answer: B

Community vote distribution

D (100%)

  **cybe001** 2 months, 3 weeks ago

Selected Answer: D

D is the correct answer.

B. UFM Telemetry is not correct because, while it does provide network validation and basic network monitoring, its primary focus is on capturing and streaming rich, real-time telemetry data and workload information for analysis rather than on direct operational tools or easy validation suited for straightforward monitoring-only use cases. UFM Pro, on the other hand, is specifically designed to meet the needs of organizations requiring only basic network monitoring and validation without additional data streaming, advanced analytics, or integration features found in higher-tier platforms like UFM Telemetry, Enterprise, or Cyber-AI.

upvoted 2 times

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.

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!

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.

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!

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.

Suggested Answer: *D*

Currently there are no comments in this discussion, be the first to comment!

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.

Suggested Answer: C

Currently there are no comments in this discussion, be the first to comment!

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

Suggested Answer: C

Currently there are no comments in this discussion, be the first to comment!

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.

Suggested Answer: C

Currently there are no comments in this discussion, be the first to comment!

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

Suggested Answer: *C*

Currently there are no comments in this discussion, be the first to comment!

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.

Suggested Answer: *B*

Currently there are no comments in this discussion, be the first to comment!

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

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!

You are managing an on-premises cluster using NVIDIA Base Command Manager (BCM) and need to extend your computational resources into AWS when your local infrastructure reaches peak capacity.

What is the most effective way to configure cloudbursting in this scenario?

- A. Use BCM's built-in load balancer to distribute workloads evenly between on-premises and cloud resources without any pre-configuration.
- B. Manually provision additional cloud nodes in AWS when the on-premises cluster reaches its limit.
- C. Set up a standby deployment in AWS and manually switch workloads to the cloud during peak times.
- D. Use BCM's Cluster Extension feature to automatically provision AWS resources when local resources are exhausted.

Suggested Answer: *D*

Currently there are no comments in this discussion, be the first to comment!

In a high availability (HA) cluster, you need to ensure that split-brain scenarios are avoided.

What is a common technique used to prevent split-brain in an HA cluster?

- A. Configuring manual failover procedures for each node.
- B. Using multiple load balancers to distribute traffic evenly across nodes.
- C. Implementing a heartbeat network between cluster nodes to monitor their health.
- D. Replicating data across all nodes in real time.

Suggested Answer: C

Currently there are no comments in this discussion, be the first to comment!

You are configuring cloudbursting for your on-premises cluster using BCM, and you plan to extend the cluster into both AWS and Azure. What is a key requirement for enabling cloudbursting across multiple cloud providers?

- A. You only need to configure credentials for one cloud provider, as BCM will automatically replicate them across other providers.
- B. You need to set up a single set of credentials that works across both AWS and Azure for seamless integration.
- C. You must configure separate credentials for each cloud provider in BCM to enable their use in the cluster extension process.
- D. BCM automatically detects and configures credentials for all supported cloud providers without requiring admin input.

Suggested Answer: C

Currently there are no comments in this discussion, be the first to comment!

Which of the following correctly identifies the key components of a Kubernetes cluster and their roles?

- A. The control plane consists of the kube-apiserver, etcd, kube-scheduler, and kube-controller-manager, while worker nodes run kubelet and kube-proxy.
- B. Worker nodes manage the kube-apiserver and etcd, while the control plane handles all container runtimes.
- C. The control plane is responsible for running all application containers, while worker nodes manage network traffic through etcd.
- D. The control plane includes the kubelet and kube-proxy, and worker nodes are responsible for running etcd and the scheduler.

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!

A DGX H100 system in a cluster is showing performance issues when running jobs.
Which command should be run to generate system logs related to the health report?

- A. `nvsm show logs --save`
- B. `nvsm get logs`
- C. `nvsm dump health`
- D. `nvsm health --dump-log`

Suggested Answer: *C*

Currently there are no comments in this discussion, be the first to comment!

You are managing a Kubernetes cluster running AI training jobs using TensorFlow. The jobs require access to multiple GPUs across different nodes, but inter-node communication seems slow, impacting performance.

What is a potential networking configuration would you implement to optimize inter-node communication for distributed training?

- A. Increase the number of replicas for each job to reduce the load on individual nodes.
- B. Use standard Ethernet networking with jumbo frames enabled to reduce packet overhead during communication.
- C. Configure a dedicated storage network to handle data transfer between nodes during training.
- D. Use InfiniBand networking between nodes to reduce latency and increase throughput for distributed training jobs.

Suggested Answer: *D*

Currently there are no comments in this discussion, be the first to comment!