NVIDIA NCP-AII Practice Questions

Shared by Travis on 21-08-2026

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

You are following the official steps to install the NVIDIA Container Toolkit using a package manager on Ubuntu. After importing the NVIDIA package repository and GPG key, what is the next action?

Options:

A- Reboot the host system to apply the repository changes and proceed.
B- Install the nvidia-container-toolkit package using your package manager.
C- Format the disk to clear any existing NVIDIA-related dependencies first.
D- Download the CUDA toolkit installer from NVIDIA'S official website.

Answer:

B

Explanation:

The NVIDIA Container Toolkit (formerly nvidia-docker2) is the essential middleware that allows Docker, Podman, or Containerd to 'see' and utilize the host's GPU hardware. The standard installation workflow on Debian-based systems like Ubuntu involves three core phases: repository configuration, package installation, and runtime configuration. Once the GPG key is added (to ensure package integrity) and the .list file is placed in /etc/apt/sources.list.d/ (to point to the NVIDIA production servers), the local package index must be refreshed via apt-get update. Immediately following this, the administrator must install the toolkit using the command sudo apt-get install -y nvidia-container-toolkit. Rebooting (Option A) is unnecessary at this stage because no kernel modules have been modified yet. Downloading the CUDA Toolkit (Option D) is a separate step; notably, the Container Toolkit allows containers to run CUDA applications even if the host only has the NVIDIA driver installed, making the driver---not the host CUDA toolkit---the primary prerequisite.

Question 2

Question Type: MultipleChoice

A team is validating a DGX BasePOD deployment. Using cmsh, they run a command to check GPU health across all nodes. What indicates that the system is ready for AI workloads?

Options:

A- The command output is ignored if the system powers on without errors.
B- At least half of the GPUs report Status_Health = OK.
C- All GPUs report Status_Health = OK and Health = OK for each device.
D- Only the head node's GPUs need to be healthy.

Answer:

C

Explanation:

In an NVIDIA DGX BasePOD or SuperPOD environment, 'Cluster Health' is a binary state: either the entire fabric and all compute resources are ready, or the cluster is considered degraded. Using the Bright Cluster Manager (BCM) shell (cmsh), administrators can aggregate telemetry from every node in the cluster. For a system to be considered 'Production Ready,' every single GPU across the multi-node deployment must report a status of Health = OK. This verification ensures that the hardware is communicating correctly over the PCIe bus, the NVLink fabric is initialized, and no ECC (Error Correction Code) memory errors are present. If even a single GPU in a 32-node cluster is unhealthy, collective communication libraries like NCCL may hang or experience significant performance penalties during 'All-Reduce' operations, as the entire job typically scales to the speed of the slowest/unhealthiest component. Therefore, seeing Status_Health = OK for every device is the mandatory exit criterion for the bring-up phase.

Question 3

Question Type: MultipleChoice

Your company is planning to expand its AI capabilities significantly over the next five years. To future-proof your storage infrastructure, you need a solution that can scale in both capacity and performance. Which of the following strategies best ensures that your storage infrastructure remains adaptable to future AI demands?

Options:

A- Deploy an all-flash array and remove data tiering to reduce latency.
B- Implement single-tier cloud storage solution to leverage cloud scalability.
C- Use a hybrid cloud model combining scalable cloud resources with on-premises infrastructure.
D- Implement on-premises block storage system with periodic hardware upgrades.

Answer:

C

Explanation:

Future-proofing AI storage requires balancing the massive throughput needs of training with the vast capacity needs of data lakes. A hybrid cloud model (Option C) is the most adaptable strategy because it allows organizations to maintain high-performance, low-latency 'hot' data on-premises (near the DGX clusters) while leveraging the elastic, near-infinite scalability of cloud storage for 'cold' archival data or sudden bursts in compute demand. This model supports data tiering, where active datasets live on NVMe-based local parallel file systems to saturate GPU memory, and older datasets migrate to lower-cost cloud object storage. While an all-flash array (Option A) is fast, it can become cost-prohibitive for multi-petabyte AI data lakes without a cloud-integrated tiering strategy. The hybrid approach provides the architectural flexibility to adopt new technologies over a five-year horizon without being locked into rigid on-premises hardware upgrade cycles.

Question 4

Question Type: MultipleChoice

You are validating the environment of an NVIDIA GPU-accelerated data center during post-deployment checks. Which one action is essential to confirm that power and cooling are sufficient for the stable operation of NVIDIA DGX H100 systems?

Options:

A- Confirm the system fans are running at 100% under all workloads to prevent overheating.
B- Review the system BIOS to ensure GPU overclocking is enabled for maximum performance.
C- Use NVSM to disable unused PCIe devices to reduce overall system heat output.
D- Verify that each DGX system is connected to redundant, properly rated PDUs and that all power supplies are reporting nominal input.

Answer:

D

Explanation:

Stable operation of high-density AI infrastructure like the DGX H100 requires strict adherence

to power and thermal specifications. A single DGX H100 system can draw up to 10.2kW under peak load. Therefore, the most essential validation step is ensuring the electrical 'infrastructure-to-server' handoff is healthy. This involves verifying that the system is connected to redundant PDUs (Power Distribution Units) capable of handling the amperage requirements without tripping breakers. Using NVSM (NVIDIA System Management), an administrator must check that all six power supplies (PSUs) are functional and receiving nominal input voltage (typically 200V-240V). If a PSU reports sub-optimal input or a 'Loss of Redundancy,' the system may throttle performance or shut down unexpectedly during a heavy training run. Fans running at 100% (Option A) at all times would actually indicate an inefficient or failed cooling policy, as fans should dynamically scale based on thermals. Overclocking (Option B) is not supported or recommended for enterprise DGX systems, as they are already factory-tuned for the highest stable performance.

Question 5

Question Type: MultipleChoice

During HPL execution on a DGX cluster, the benchmark fails with "not enough memory" errors despite sufficient physical RAM. Which HPL.dat parameter adjustment is most effective?

Options:

A- Reduce the problem size while maintaining the same block size.
B- Set PMAP to 1 to enable process mapping.
C- Increase block size to 6144 to maximize GPU utilization.
D- Disable double-buffering via BCAST parameter.

Answer:

A

Explanation:

High-Performance Linpack (HPL) is a memory-intensive benchmark that allocates a large portion of available GPU memory to store the matrix N. While a server may have 2TB of physical system RAM, the 'not enough memory' error usually refers to the HBM (High Bandwidth Memory) on the GPUs themselves. In a DGX H100 system, each GPU has 80GB of HBM3. If the problem size (N) specified in the HPL.dat file is too large, the required memory for the matrix will exceed the aggregate capacity of the GPU memory. Reducing the problem size (N) while maintaining the optimal block size (NB) ensures that the problem fits within the GPU memory limits while still pushing the computational units to their peak performance. Increasing the block size (Option C) would actually increase the memory footprint of certain internal buffers, potentially worsening the issue. Reducing N is the standard procedure to

stabilize the run during the initial tuning phase of an AI cluster bring-up.

Question 6

Question Type: MultipleChoice

ClusterKit's NCCL bandwidth test shows 350 GB/s on a 400G InfiniBand fabric. How should this result be interpreted?

Options:

A- Optimal performance, indicating healthy fabric and GPUDirect RDMA.
B- Suboptimal performance; requires FEC tuning to reach 380+ GB/s.
C- Critical failure; expected is >390 GB/s for HDR InfiniBand.
D- Inconclusive; rerun with --stress=cpu to validate.

Answer:

A

Explanation:

Interpreting NCCL (NVIDIA Collective Communications Library) results requires understanding the difference between theoretical link rate and effective bus bandwidth. A 400G (NDR) InfiniBand link has a theoretical unidirectional capacity of 50 GB/s. In an 8-GPU DGX system, where each GPU is typically mapped to its own 400G HCA, the aggregate theoretical bandwidth is 400 GB/s. However, NCCL performance benchmarks report Bus Bandwidth, which accounts for the mathematical overhead of collective operations like all_reduce. Achieving 350 GB/s (Option A) is widely considered optimal and representative of a healthy, well-configured fabric. This indicates that the system is effectively utilizing GPUDirect RDMA to bypass host memory and that the network switches are routing traffic without significant congestion or packet drops. Expecting >390 GB/s (Option C) is unrealistic once protocol headers and algorithmic overhead are subtracted from the raw bit rate. Rerunning with CPU stress (Option D) would not provide insights into the InfiniBand/GPU fabric health.

Question 7

Question Type: MultipleChoice

A system engineer needs to set the vGPU scheduling behavior for all GPUs to share the scheduling equally with the default time slice length. What command should be used?

Options:

A- esxcli system module parameters set -m nvidia -p 'NVreg_RegistryDwords=RmPVMRL=0x01'
B- esxcli graphics module parameters set -m nvidia -p 'NVreg_RegistryDwords=RmPVMRL=0x01'
C- esxcli system module parameters set -m nvidia -p 'NVreg_RegistryDwords=FRL=0x01'
D- esxcli system module parameters set -m nvidia -p 'NVreg_RegistryDwords=RmPVMRL=0x00'

Answer:

A

Explanation:

When deploying NVIDIA vGPU on VMware ESXi, the NVIDIA driver provides several scheduling policies to determine how GPU physical resources are shared among multiple virtual machines. The default behavior is often the 'Best Effort' scheduler, but for environments requiring predictable performance across all users, the 'Equal Share' scheduler is preferred. This scheduler gives each vGPU an equal 'time slice' of the physical GPU's engines. The configuration is managed via module parameters passed to the nvidia kernel driver during host boot. The specific registry key for this behavior is RmPVMRL. Setting RmPVMRL=0x01 enables the Equal Share scheduler (Option A). Conversely, 0x00 would revert to the default time-sliced behavior. It is critical to use system module parameters set to ensure the setting persists across reboots and is applied globally to the NVIDIA driver stack. This ensures that no single 'noisy neighbor' VM can monopolize the GPU cycles, which is a common requirement in shared AI research labs or virtual desktop infrastructures where consistency is more important than raw peak throughput of a single task.

Question 8

Question Type: MultipleChoice

What command is needed to measure BER (Bit Error Rate)?

Options:

A- mlxconfig -d <device> q
B- ethtool -S <device>
C- mlxlink -d <device> -c -e
D- mstflint -d <device> q full

Answer:

C

Explanation:

In NVIDIA networking environments, specifically those utilizing InfiniBand or high-speed Ethernet via ConnectX adapters, monitoring the physical link quality is critical for preventing packet loss and RDMA retransmissions. The mlxlink tool is part of the NVIDIA Firmware Tools (MFT) package and is the primary utility for checking the status and health of the physical link. Using the -d flag specifies the device (e.g., /dev/mst/mt4123_pciconf0), while the -c (counters) and -e (error counters/BER) flags provide a detailed readout of the link's performance. Bit Error Rate (BER) is a fundamental metric for signal integrity. NVIDIA systems typically distinguish between 'Raw BER' (errors before Forward Error Correction) and 'Effective BER' (errors remaining after FEC). A high BER often points to a failing transceiver, a dirty fiber connector, or a marginal DAC cable. While ethtool can show general statistics in Ethernet mode, mlxlink is the verified method for granular BER measurement across InfiniBand and high-speed fabrics, allowing engineers to determine if a link meets the 'Error-Free' operation standards required for large-scale AI collective communications like NCCL.

Question 9

Question Type: MultipleChoice

You are leading a project to enhance the energy efficiency of a data center that heavily relies on AI workloads. NVIDIA suggests moving beyond traditional metrics like Power Usage Effectiveness (PUE) to better capture the efficiency of modern data centers. Which strategy should you prioritize?

Options:

A- Use Power Usage Effectiveness as the primary metric while supplementing it with additional measures of useful work done per unit of energy.
B- Use watts used as the primary measure of efficiency, as it accurately reflects the power input at any given time.
C- Develop benchmarks tailored to specific workloads, such as MLPerf for AI applications, to better understand energy use in real-world scenarios.
D- Focus on integrating kilowatt-hours into existing metrics to better reflect the actual energy

used for productive work.

Answer:

C

Explanation:

Traditional data center metrics like PUE (Power Usage Effectiveness) only measure how much energy is 'wasted' by cooling and power delivery relative to the IT load; they say nothing about how efficiently that IT load is performing its task. In an AI Factory, 'Efficiency' is better defined by the amount of AI training or inference performed per watt. NVIDIA advocates for the use of workload-specific benchmarks, such as MLPerf, to quantify this. MLPerf measures the time and energy required to complete standardized AI tasks (like training a ResNet-50 model or an LLM). By prioritizing these benchmarks (Option C), an organization can compare the energy efficiency of different hardware architectures (e.g., A100 vs. H100) or different software optimizations (e.g., FP8 vs. FP16). For example, even if an H100 system draws more peak power than an older system, its ability to complete a training job 9x faster results in a significantly lower 'Total Energy Consumed per Job'. This shift from 'infrastructure efficiency' (PUE) to 'computing efficiency' (MLPerf-per-watt) is essential for modern AI data centers aiming for sustainability and cost-effective scaling.

Question 10

Question Type: MultipleChoice

After initial setup and health checks, the DGX H100 system administrator wants to verify that containers can access GPUs before running production workloads. Which method is recommended for this validation?

Options:

A- sudo docker run --gpus all --rm nvcr.io/nvidia/cuda:12.1.1-base-ubuntu22.04 systemctl
B- sudo docker run --gpus all --rm nvcr.io/nvidia/cuda:12.1.1-base-ubuntu22.04 ls -la
C- sudo docker run --rm nvcr.io/nvidia/cuda:12.1.1-base-ubuntu22.04 nvidia-smi
D- sudo docker run --gpus all --rm nvcr.io/nvidia/cuda:12.1.1-base-ubuntu22.04 nvidia-smi

Answer:

D

Explanation:

The definitive 'smoke test' for an NVIDIA-accelerated container environment is running nvidia-smi from within a container pulled from the NVIDIA GPU Cloud (NGC). Option D is the only 100% verified command because it includes all three necessary components:

--gpus all: This flag is required by the NVIDIA Container Toolkit to map the host's GPU device nodes and driver libraries into the container. Without it (as in Option C), the container will not 'see' any GPUs.

nvcr.io/nvidia/cuda: Using an official CUDA base image from NGC ensures that the container has the necessary user-space libraries to communicate with the NVIDIA driver on the host.

nvidia-smi: This command specifically queries the driver and hardware. If it successfully prints the GPU table (showing the H100/A100 modules, their temperatures, and memory), it proves that the entire stack---from the physical hardware and kernel driver to the container runtime and toolkit---is functioning correctly.

Running ls -la (Option B) or systemctl (Option A) only tests that the container can run, but it does nothing to verify GPU accessibility or driver communication.

Question 11

Question Type: MultipleChoice

For an NVIDIA Enterprise AI Factory with 256 GPUs, which storage solution characteristic is most critical to validate during scaling tests?

Options:

A- Consistent per-node throughput >8 GiB/s.
B- Single-node write performance during idle clusters.
C- RAID rebuild times under disk failure.
D- Maximum 4K random read IOPS exceeding 1 million.

Answer:

A

Explanation:

Scaling an AI cluster to 256 GPUs (32 nodes of DGX H100) creates a massive 'Incast' problem

for the storage fabric. During large-scale training, every node frequently reads huge batches of data simultaneously. NVIDIA's reference architectures (BasePOD/SuperPOD) specify that for high-performance training, each node must be able to sustain a minimum throughput---often 8 GiB/s or more---to keep all 8 GPUs saturated. If the storage system can handle one node at high speed but chokes when all 32 nodes request data, the 'Scaling Efficiency' of the AI model will drop drastically as GPUs sit idle waiting for IO. Therefore, validating consistent per-node throughput under full cluster load is the most critical metric for an AI Factory. While IOPS (Option D) are important for small files, modern AI datasets are often sharded into large binary formats (like WebDataset or TFRecord) where sequential throughput becomes the primary bottleneck.

Question 12

Question Type: MultipleChoice

An administrator is configuring node categories in BCM for a DGX BasePOD cluster. They need to group all NVIDIA DGX H200 nodes under a dedicated category for GPU-accelerated workloads. Which approach aligns with NVIDIA's recommended BCM practices?

Options:

A- Assign nodes to the 'login' category to simplify Slurm integration.
B- Create a new 'dgx-h200' category, assign all DGX H200 nodes to it.
C- Use the existing 'dgxnodes' category without modification, as it is preconfigured for all DGX systems.
D- Avoid categories and configure each DGX node individually via CLI.

Answer:

B

Explanation:

NVIDIA Base Command Manager (BCM) uses 'Categories' as the primary organizational unit for applying configurations, software images, and security policies to groups of nodes. In a heterogeneous cluster---or even a large homogeneous one---creating specific categories for different hardware generations (like DGX H100 vs. H200) is a best practice. By creating a dedicated dgx-h200 category (Option B), the administrator can apply specific kernel parameters, driver versions, and specialized software packages (like specific versions of the NVIDIA Container Toolkit or DOCA) that are optimized for the H200's HBM3e memory and Hopper architecture updates. Using a generic dgxnodes category (Option C) makes it difficult to perform rolling upgrades or test new drivers on a subset of hardware without impacting the entire cluster. Furthermore, categorizing nodes allows for more granular integration with the

Slurm workload manager, enabling users to target specific hardware features via partition definitions that map directly to these BCM categories. This modular approach reduces 'configuration drift' and ensures that the AI factory remains manageable as it scales from a single POD to a multi-POD SuperPOD architecture.

Question 13

Question Type: MultipleChoice

A 24-hour HPL burn-in fails with "illegal value" errors during the first iteration. Which initial troubleshooting step resolves this without compromising burn-in validity?

Options:

A- Switch from FP64 to FP32 precision.
B- Disable GPU affinity.
C- Reduce test duration to 12 hours.
D- Verify the matrix size is divisible by block size.

Answer:

D

Explanation:

High-Performance Linpack (HPL) is the standard benchmark for stress-testing the computational stability and thermal endurance of an AI cluster. It solves a massive dense system of linear equations, and its mathematical configuration is highly sensitive. The HPL.dat configuration file defines the Problem Size (N) and the Block Size (NB). A fundamental requirement of the HPL algorithm is that the workload must be distributed evenly across the MPI processes and GPU threads. If the total matrix size N is not an exact multiple of the block size NB, or if the grid dimensions ($P \times Q$) do not align with the hardware topology, the solver may encounter an 'illegal value' error or a 'residual too large' failure at the very beginning of the run. This is a configuration error, not a hardware fault. Reducing the precision (Option A) would invalidate the test, as HPL must run in FP64 to be considered a standard 'burn-in.' Verifying that N is divisible by NB ensures the mathematical integrity of the test while allowing the hardware to be pushed to its theoretical performance limits.