

First-hand NVIDIA NCP-AII Latest Exam - NVIDIA AI Infrastructure Free Practice



What's more, part of that Pass4cram NCP-AII dumps now are free: https://drive.google.com/open?id=1EUbV5jNHu9I0JdgNkEhkbTNoj0Sc_tmG

Our website aimed to helping you and fully supporting you to pass NCP-AII actual test with high passing score in your first try. So we prepared top NCP-AII pdf torrent including the valid questions and answers written by our certified professionals for you. Our NCP-AII Practice Exam available in three modes, pdf files, and PC test engine and online test engine, which apply to any level of candidates.

NVIDIA NCP-AII Exam Syllabus Topics:

Topic	Details
Topic 1	Physical Layer Management: Covers configuring BlueField network platform devices and setting up Multi-Instance GPU (MIG) partitioning for AI and HPC workloads.
Topic 2	- Troubleshoot and Optimize: Covers identifying and replacing faulty hardware components such as GPUs, network cards, and power supplies, along with performance optimization for AMD - Intel servers and storage.
Topic 3	Cluster Test and Verification: Covers full cluster validation through HPL and NCCL benchmarks, NVLink and fabric bandwidth tests, cable and firmware checks, and burn-in testing using HPL, NCCL, and NeMo.
Topic 4	- System and Server Bring-up: Covers end-to-end physical setup of GPU-based AI infrastructure, including BMC - OOB - TPM configuration, firmware upgrades, hardware installation, and power and cooling validation to ensure servers are workload-ready.
Topic 5	- Control Plane Installation and Configuration: Covers deploying the software stack including Base Command Manager, OS, Slurm - Enroot - Pyxis, NVIDIA GPU and DOCA drivers, container toolkit, and NGC CLI.

>> NCP-AII Latest Exam <<

NVIDIA NCP-AII Free Practice & Latest NCP-AII Test Dumps

Our NCP-AII guide questions have the most authoritative test counseling platform, and each topic in NCP-AII practice engine is carefully written by experts who are engaged in researching in the field of professional qualification exams all the year round. They have a very keen sense of change in the direction of the exam, so that they can accurately grasp the important points of the NCP-AII Exam. And you will pass the exam for the NCP-AII exam questions are all keypoints.

NVIDIA AI Infrastructure Sample Questions (Q13-Q18):

NEW QUESTION # 13

When updating the firmware on an NVLink switch transceiver, how can an engineer apply new firmware without interrupting the network?

- A. `mlxfwreset -d -lid 27 reset --yes` to reset the transceiver
- B. `flint -d -lid 27 --linkx --linkx_auto_update --activate`
- C. `nv action reboot` system to force immediate activation.
- D. Physically disconnect and reconnect the transceiver.

Answer: B

Explanation:

NVIDIA's LinkX optical transceivers and active copper cables often require firmware updates to ensure compatibility and performance optimizations. In a production DGX SuperPOD environment, interrupting the NVLink fabric can cause GPU-to-GPU communication failures and crash training jobs. To mitigate this, NVIDIA utilizes the `flint` utility (part of MFT) with specific flags for "Live" or "Seamless" updates. The `--linkx` flag targets the transceiver or cable specifically, rather than the switch ASIC itself. The `--linkx_auto_update` flag automates the sequence, while the `--activate` flag ensures the new firmware is applied to the module's active memory without requiring a full system reboot or a manual flap of the network link.

This "in-service" update capability is essential for large-scale AI clusters where uptime is measured in weeks or months of continuous training. By using the `-lid` (Logical Identifier) target, an administrator can address specific modules across the fabric from a central management node, ensuring that the high-bandwidth NVLink mesh remains stable while maintaining the latest hardware optimizations.

NEW QUESTION # 14

You are designing an AI infrastructure cluster for training large language models (LLMs). The dataset consists of 10TB of image data and 5TB of text data. You estimate that intermediate training data (checkpoints, temporary files) will require an additional 20TB of storage. You want to use a parallel file system for optimal performance. Considering a replication factor of 2 for data redundancy and a 20% overhead for file system metadata, what is the minimum raw storage capacity you should provision?

- A. 84 TB
- B. 100.8 TB
- C. 70 TB
- D. 42 TB
- E. 92.4 TB

Answer: E

Explanation:

Total data size: $10\text{TB} + 5\text{TB} + 20\text{TB} = 35\text{TB}$. With a replication factor of 2, the storage required is $35\text{TB} \times 2 = 70\text{TB}$. Adding 20% overhead for metadata, we get $70\text{TB} \times 1.2 = 84\text{TB}$. Therefore, the minimum raw storage capacity is $84 + 8.4 = 92.4\text{TB}$. Overhead needs to be calculated from after replication is implemented, so replication + 20% overhead.

NEW QUESTION # 15

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

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

Answer: A

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.

NEW QUESTION # 16

You are tasked with ensuring optimal power efficiency for a GPU server running machine learning workloads. You want to dynamically adjust the GPU's power consumption based on its utilization. Which of the following methods is the MOST suitable for achieving this, assuming the server's BIOS and the NVIDIA drivers support it?

- A. Enable Dynamic Boost in the NVIDIA Control Panel (if available), which will automatically allocate power between the CPU and GPU based on their current needs.
- **B. Use NVIDIA's Data Center GPU Manager (DCGM) to monitor GPU utilization and dynamically adjust the power limit based on a predefined policy.**
- C. Disable ECC (Error Correcting Code) on the GPU to reduce power consumption.
- D. Manually set the GPU's power limit using `nvidia-smi -pl` and create a script to monitor utilization and adjust the power limit periodically.
- E. Configure the server's BIOS/UEFI to use a power-saving profile, which will automatically reduce the GPU's power consumption when idle.

Answer: B

Explanation:

DCGM provides the most comprehensive and automated solution for dynamic power management. It can monitor GPU utilization in real-time and adjust the power limit based on predefined policies, ensuring optimal power efficiency without manual intervention. Manually adjusting the power limit is possible but requires scripting and continuous monitoring. Dynamic Boost is typically for laptops, and BIOS power profiles may not be fine-grained enough. Disabling ECC reduces power but compromises data integrity.

NEW QUESTION # 17

A large AI model is training using a dataset stored on a network-attached storage (NAS) device. The data transfer speeds are significantly lower than expected. After initial troubleshooting, you discover that the MTU (Maximum Transmission Unit) size on the network interfaces of the training server and the NAS device are mismatched. The server is configured with an MTU of 1500, while the NAS device is configured with an MTU of 9000 (Jumbo Frames). What is the MOST likely consequence of this MTU mismatch, and what action should you take?

- **A. Data packets will be fragmented, leading to increased overhead and reduced performance. Configure both the server and the NAS device to use the same MTU size (either 1500 or 9000).**
- B. The server will be unable to communicate with the NAS device. Reduce the MTU size on the server to match the MTU size of the NAS device.
- C. Data packets will be retransmitted, increasing the latency but still getting the full throughput. Configure the server to use Path MTU Discovery (PMTUD).
- D. The data transfer will be limited to the lowest common MTU size, but there will be no significant performance impact. No action is required.
- E. The connection between the server and the NAS device will be unreliable, resulting in data corruption. Increase the MTU size on both devices to the maximum supported value.

Answer: A

Explanation:

An MTU mismatch (option A) will cause fragmentation, where larger packets are broken down into smaller packets before being transmitted, adding overhead and reducing performance. The solution is to configure both devices to use the same MTU size. Choosing 1500 ensures compatibility, while 9000 requires the entire network path to support jumbo frames.

• • • • •

NCP-AII Free Practice: https://www.pass4cram.com/NCP-AII_free-download.html

- P.S. Free & New NCP-AII dumps are available on Google Drive shared by Pass4cram: https://drive.google.com/open?id=1EUbV5jNHu9t0JdgNkEhkbtNoj0Sc_tmG