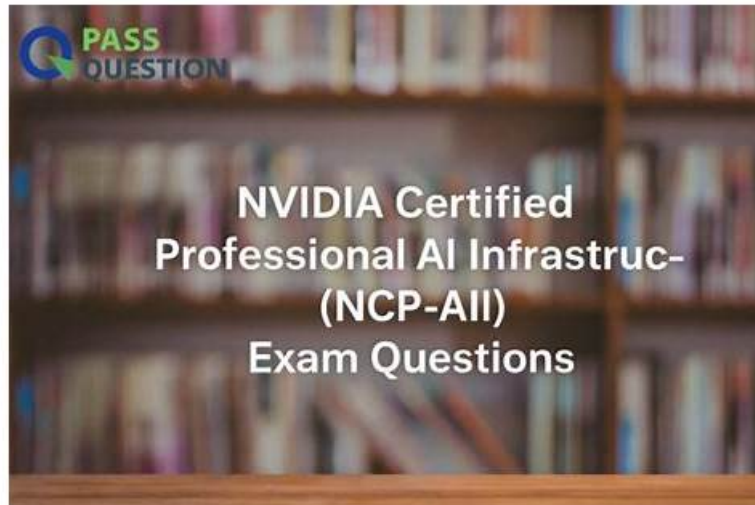


NVIDIA NCP-AII Reliable Exam Online - Valid NCP-AII Test Vce



P.S. Free 2026 NVIDIA NCP-AII dumps are available on Google Drive shared by PracticeTorrent:
<https://drive.google.com/open?id=16-0j7ZEg3ISPppzigiHvwpXu7lGYiFtk>

Our product's passing rate is 99% which means that you almost can pass the test with no doubts. The reasons why our NCP-AII study materials' passing rate is so high are varied. Firstly, our test bank includes two forms and they are the PDF test questions which are selected by the senior lecturer, published authors and professional experts and the practice test software which can test your mastery degree of our NCP-AII Study Materials at any time. The two forms cover the syllabus of the entire test. Our questions and answers include all the questions which may appear in the exam and all the approaches to answer the questions. So we provide the strong backing to help clients to help them pass the test.

NVIDIA NCP-AII Exam Syllabus Topics:

| Topic | Details |
|---------|--|
| Topic 1 | <ul style="list-style-type: none">• 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. |
| Topic 2 | <ul style="list-style-type: none">• 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 | <ul style="list-style-type: none">• Physical Layer Management: Covers configuring BlueField network platform devices and setting up Multi-Instance GPU (MIG) partitioning for AI and HPC workloads. |
| Topic 4 | <ul style="list-style-type: none">• 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 5 | <ul style="list-style-type: none">• 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. |

PracticeTorrent NVIDIA NCP-AII PDF Questions

Our NCP-AII Exam Torrent carries no viruses. We provide free update and online customer service which works on the line whole day. Our study materials provide varied versions for you to choose and the learning costs you little time and energy. You can use our NCP-AII exam prep immediately after you purchase them, we will send our product within 5-10 minutes to you. We treat your time as our own time, as precious as you see, so we never waste a minute or two in some useless process. Please rest assured that use, we believe that you will definitely pass the exam.

NVIDIA AI Infrastructure Sample Questions (Q27-Q32):

NEW QUESTION # 27

You have an NVIDIA A100 GPU and need to configure it for optimal performance across two distinct AI workloads: a large language model (LLM) training job and a computer vision inference service. The LLM benefits from maximum memory bandwidth, while the inference service requires low latency and high throughput. Which MIG configuration would best suit this scenario?

- A. Create two 7g.80gb MIG instances, one for each workload.
- B. Create one 14g.160gb MIG instance for the LLM and use CUDA MPS to multiplex the inference service.
- C. Create a single full-GPU instance and use Kubernetes resource quotas to isolate the workloads.
- **D. Create one 12g.120gb instance for the LLM and one 4g.40gb instance for inference.**
- E. Utilize Time-Slicing on a single full-GPU instance, allocating specific time slots to each workload using NVIDIA Vgpu technology

Answer: D

Explanation:

Creating a 12g.120gb instance for the memory-intensive LLM and a 4g.40gb instance for the inference service provides dedicated resources that cater to the specific needs of each workload, without the overhead or limitations of CUDA MPS or Kubernetes resource quotas. Option A is too conservative, potentially limiting the LLM performance. Option B sacrifices dedicated resources for inference, which may hurt latency. Option C does not leverage MIG and does not guarantee resource isolation and performance consistency. Option E introduces complexities associated with Time-Slicing and might not be suitable for real-time processing.

NEW QUESTION # 28

You're troubleshooting a DGX-I server exhibiting performance degradation during a large-scale distributed training job. 'nvidia-smi' shows all GPUs are detected, but one GPU consistently reports significantly lower utilization than the others. Attempts to reschedule workloads to that GPU frequently result in CUDA errors. Which of the following is the MOST likely cause and the BEST initial troubleshooting step?

- A. Power supply unit (PSU) overload, causing reduced power delivery to that GPU; monitor PSU load and check PSU specifications.
- **B. A hardware fault with the GPU, potentially thermal throttling or memory issues; run 'nvidia-smi -i -q' to check temperatures, power limits, and error counts.**
- C. A driver issue affecting only one GPU; reinstall NVIDIA drivers completely.
- D. A software bug in the training script utilizing that specific GPU's resources inefficiently; debug the training script.
- E. Insufficient cooling in the server rack; verify adequate airflow and cooling capacity for the rack.

Answer: B

Explanation:

While all options are possibilities, the consistently lower utilization and CUDA errors point strongly to a hardware fault. Running 'nvidia-smi -i -q' provides detailed telemetry data, including temperature, power limits, and ECC error counts, which are crucial for diagnosing GPU hardware issues.

NEW QUESTION # 29

You are evaluating different parallel file systems for an AI training cluster. You need a file system that supports POSIX compliance and offers high bandwidth and low latency. Which of the following options are viable candidates?

- A. GlusterFS
- B. Ceph
- C. Lustre
- D. NFS
- E. BeeGFS

Answer: C,E

Explanation:

BeeGFS and Lustre are designed for high-performance computing and AI workloads, offering high bandwidth, low latency, and POSIX compliance. GlusterFS and Ceph are more general-purpose distributed file systems. NFS is generally not suitable for demanding AI workloads due to its performance limitations.

NEW QUESTION # 30

You are troubleshooting an issue where a Docker container utilizing NVIDIA GPUs intermittently fails with a 'CUDA ERROR OUT OF MEMORY' error. The host system has sufficient memory and the individual GPU has enough memory as well. You suspect that the problem might be related to how memory is being allocated within the container environment. What steps can you take to investigate and potentially mitigate this issue?

- A. Lower the compute capability using '-compute' parameter on docker run.
- B. Adjust the environment variable inside the container to ensure consistent GPU ordering.
- C. Monitor GPU memory usage both inside and outside the container using 'nvidia-smi' to identify memory leaks or excessive allocation.
- D. Increase the shared memory size for the container using the '-shm-size' flag when running the container.
- E. Set the environment variable inside the container to limit the number of GPUs visible to the application.

Answer: C,D

Explanation:

A 'CUDA ERROR OUT OF MEMORY' error can occur due to insufficient shared memory within the container (A). Increasing the shared memory size allows the container to allocate more memory for inter-process communication and GPU data transfers. Monitoring GPU memory usage both inside and outside the container (D) is crucial to identify the source of the memory exhaustion. 'CUDA VISIBLE DEVICES' and (B & C) are primarily used for GPU selection and ordering, not memory management, although limiting GPU visibility could indirectly reduce overall memory consumption if the application is poorly designed and tries to allocate memory on all visible GPUs regardless of need. Lowering compute capability won't directly affect memory usage, although the application will need less memory to process, it might cause issues.

NEW QUESTION # 31

Which of the following statements accurately describe the benefits of using MIG (Multi-Instance GPU) in an AI/HPC environment? (Select all that apply)

- A. MIG eliminates the need for GPU virtualization software.
- B. MIG allows running multiple CUDA versions simultaneously on the same physical GPU.
- C. MIG increases the overall FLOPS of the GPU.
- D. MIG guarantees performance isolation between different workloads running on the same GPU.
- E. MIG allows a single GPU to be partitioned into multiple isolated instances, improving resource utilization.

Answer: D,E

Explanation:

MIG allows partitioning a single GPU into multiple isolated instances, thus optimizing resource utilization (A). It provides hardware-level isolation, ensuring performance consistency and preventing interference between workloads (B). While MIG simplifies GPU sharing, it doesn't eliminate the need for all virtualization software, especially in more complex environments. MIG doesn't increase FLOPS (D); it divides the existing compute power. While different containers running on different MIG instances could have different CUDA versions, MIG itself doesn't directly handle this (E); this is handled via containers or environments.

NEW QUESTION # 32

.....

