

目標を達成するNCP-AIO認定試験トレーニング:有難い問題NVIDIA AI Operations NCP-AIO日本語pdf問題



BONUS!!! PassTest NCP-AIOダンプの一部を無料でダウンロード: https://drive.google.com/open?id=1mgZN4zz1FJ2rbn0t-7I_wTFqUndqxwFG

証明書は私たちの日常生活で重要です。現在、NCP-AIO試験に合格するすべての受験者に、選択可能な3つの異なるバージョンを提供しています。NCP-AIO試験問題のAPPバージョンは、オフライン状態で機能します。クイズ準備を使用する場合、最新のNCP-AIO試験トレントをいつでもどこでも使用できます。オフライン状態のNCP-AIO実践ガイドを使用して、どのようにして学習を楽しむことができますか？オフライン状態で動作するバージョンをダウンロードするだけで、初めてNCP-AIOクイズトレントのバージョンをオンラインで使用する必要があります。

NVIDIA NCP-AIO 認定試験の出題範囲:

トピック	出題範囲
トピック 1	<ul style="list-style-type: none"> 管理: このセクションでは、システム管理者のスキルを評価し、データセンターにおけるAIワークロードの管理に不可欠なタスクを網羅します。受験者は、フリートコマンド、Slurmクラスタ管理、そしてAI環境に特有のデータセンターアーキテクチャ全体を理解していなければなりません。また、Base Command Manager (BCM)、クラスタプロビジョニング、Run.ai管理、そしてAIとハイパフォーマンスコンピューティングアプリケーションの両方に対応したマルチインスタンスGPU (MIG) の構成に関する知識も求められます。
トピック 2	<ul style="list-style-type: none"> トラブルシューティングと最適化: NVIこの試験セクションでは、AIインフラストラクチャエンジニアのスキルを評価し、高度なAIシステムで発生する技術的な問題の診断と解決に焦点を当てます。出題範囲には、Docker、NVIDIA NVlinkおよびNVSwitchシステムのFabric Managerサービス、Base Command Manager、Magnum IOコンポーネントのトラブルシューティングが含まれます。受験者は、ストレージパフォーマンスの問題を特定して解決し、AIワークロード全体で最適なパフォーマンスを確保する能力も実証する必要があります。
トピック 3	<ul style="list-style-type: none"> ワークロード管理: このセクションでは、AIインフラストラクチャエンジニアのスキルを評価し、AI環境におけるワークロードの効率的な管理に焦点を当てます。Kubernetesクラスターの管理、ワークロード効率の維持、システム管理ツールを用いた運用上の問題のトラブルシューティング能力を評価します。NVIDIAテクノロジーと連携し、異なる環境間でワークロードがスムーズに実行されることを重視します。

トピック 4	<ul style="list-style-type: none"> インストールと展開: このセクションでは、システム管理者のスキルを測定し、インフラストラクチャのインストールと展開に関するコプラクティスを扱います。受験者は、Base Command Manager のインストールと設定、NVIDIA ホストでの Kubernetes の初期化、NVIDIA NGC およびクラウド VMI コンテナからのコンテナの展開についてテストされます。また、AI データセンターにおけるストレージ要件の理解、DPU Arm プロセッサへの DOCA サービスの展開、AI ドリブン環境の堅牢な構築についても学習します。
--------	--

>> NCP-AIO 認定試験 トレーニング <<

NCP-AIO 日本語 pdf 問題 & NCP-AIO 専門知識訓練

PassTest の問題集は IT 専門家が NVIDIA の NCP-AIO 「NVIDIA AI Operations」 認定試験について自分の知識と経験を利用して研究したものでございます。PassTest の問題集は 真実試験の問題にとっても似ていて、弊社のチームは自分の商品が自信を持っています。PassTest が提供した商品をご利用してください。もし失敗したら、全額で返金を保証いたします。

NVIDIA AI Operations 認定 NCP-AIO 試験問題 (Q48-Q53):

質問 # 48

You are tasked with deploying a multi-tenant AI cluster using Base Command Manager (BCM). How would you best isolate tenant workloads to ensure security and resource utilization?

- A. Deploy individual VMS for each tenant's workloads, managed directly by BCM.
- **B. Utilize Kubernetes namespaces and resource quotas within a single cluster.**
- C. Create separate physical clusters for each tenant.
- D. Rely solely on user authentication and authorization for workload isolation.
- E. Use Docker containers without resource limits, relying on the OS to manage resources.

正解: B

解説:

Kubernetes namespaces provide a logical separation of resources within a single cluster. Resource quotas limit the amount of resources that a namespace can consume, providing isolation and preventing one tenant from monopolizing resources. Creating separate clusters is costly. User authentication/authorization isn't sufficient alone for resource isolation.

質問 # 49

You are using BCM to manage a Kubernetes cluster with multiple GPU nodes. You need to enable GPU monitoring using Prometheus and the NVIDIA DCGM exporter. Outline the steps required to accomplish this. Choose the correct sequence:

- A. 0 1. Deploy the NVIDIA DCGM exporter as a DaemonSet in your Kubernetes cluster. 2. Configure the NVIDIA DCGM exporter endpoints. 3. Install Prometheus in your Kubernetes cluster. 4. Verify GPU metrics are available in Prometheus.
- B. 0 1. Deploy the NVIDIA DCGM exporter as a DaemonSet in your Kubernetes cluster. 2. Configure Prometheus to scrape metrics from the DCGM exporter endpoints. 3. Install Prometheus in your Kubernetes cluster. 4. Verify GPU metrics are available in Prometheus.
- **C. 0 1. Install Prometheus in your Kubernetes cluster. 2. Deploy the NVIDIA DCGM exporter as a DaemonSet in your Kubernetes cluster. 3. Configure Prometheus to scrape metrics from the DCGM exporter endpoints. 4. Verify GPU metrics are available in Prometheus.**
- D. 0 1. Deploy the NVIDIA DCGM exporter as a Deployment in your Kubernetes cluster. 2. Configure Prometheus to scrape metrics from the DCGM exporter endpoints. 3. Install Prometheus in your Kubernetes cluster. 4. Verify GPU metrics are available in Prometheus.
- E. 0 1. Configure Prometheus to scrape metrics from the DCGM exporter endpoints. 2. Install Prometheus in your Kubernetes cluster. 3. Deploy the NVIDIA DCGM exporter as a DaemonSet in your Kubernetes cluster. 4. Verify GPU metrics are available in Prometheus.

正解: C

解説:

Prometheus must be installed first to enable metric collection. The DCGM exporter is then deployed as a DaemonSet (to ensure it runs on every node) and configured, enabling Prometheus to scrape the GPU metrics. Finally, the metrics availability is verified.

質問 # 50

You have a Docker container running a CUDA application. You notice that the container takes a long time to start, specifically when initializing the CUDA context. How can you troubleshoot and potentially improve the startup time?

- A. Use the NVIDIA CUDA Cache. Set the environment variable to a persistent volume to cache compiled CUDA kernels across container restarts.
- B. Use lazy loading techniques in your application to delay the initialization of CUDA-dependent modules until they are actually needed.
- C. Use a lighter base image. A smaller image will generally have a quicker startup time.
- D. Reduce the number of CUDA devices visible to the container using the environment variable to only expose necessary GPUs.
- E. Pre-initialize CUDA in the background. Launch a background process to initialize the CUDA context before the main application starts.

正解: A、B、D、E

解説:

CUDA context creation is time-consuming. CUDA cache (A) speeds up subsequent startups. Limiting visible devices (B) reduces the initialization overhead. Pre-initializing CUDA (D) amortizes the cost. Lazy loading (E) avoids unnecessary initializations. Using a lighter base image may help, but not as directly as the other options.

質問 # 51

You are deploying a VMI container on a cloud platform, and you need to set up automatic scaling based on the GPU utilization. Which of the following approaches is MOST appropriate for implementing this?

- A. GPU Utilization cannot be used for Autoscaling.
- B. Configure the container's application to automatically scale itself based on GPU utilization.
- C. Manually monitor GPU utilization and scale the number of containers using the cloud provider's CLI.
- D. Use Kubernetes Horizontal Pod Autoscaler (HPA) with a custom metric that monitors GPU utilization using the NVIDIA DCGM Exporter.
- E. Use Kubernetes Horizontal Pod Autoscaler (HPA) based on CPU utilization.

正解: D

解説:

Using Kubernetes HPA with a custom metric based on GPU utilization is the most robust and automated approach. The NVIDIA DCGM Exporter provides GPU metrics that can be used by the HPA to trigger scaling events based on actual GPU usage. Option A will not consider GPU Utilization.

質問 # 52

You are tuning the storage performance of a Kubernetes cluster that uses Longhorn as the storage backend. You've noticed inconsistent read latencies. What Longhorn specific configurations could you adjust to potentially improve the consistency and reduce latency for read operations?

- A. Increase the number of replicas for each Longhorn volume to improve data redundancy and read parallelism.
- B. Adjust the 'replica-auto-balance' setting to redistribute replicas more evenly across nodes.
- C. Disable Longhorn's built-in monitoring to reduce overhead.
- D. Reduce the size of the Longhorn volumes to decrease the amount of data that needs to be read.
- E. Enable 'data locality' to ensure that replicas are preferably placed on the same node as the pod accessing the volume, reducing network latency.

正解: A、B、E

解説:

