Vendor: Cisco

Exam Code: 300-640

Exam Name: Implementing Cisco Data Center AI Infrastructure

Version: DEMO

QUESTION 1
A medical company has existing third-party AI servers and NVMe storage, but they want to purchase Cisco network switches with a cloud managed feature for simplicity. Which solution best meets the requirements?

A. Bring Your Own AI (BYO AI) HyperFabric with Cisco 6000 Series Switches and Essentials license
B. Bring Your Own AI (BYO AI) HyperFabric with Cisco 6000 Series Switches and Premier license
C. Nexus Dashboard with Nexus 9000 Series Switches and Cloud license
D. HyperFabric AI with Cisco 6000 Series Switches and Premier license

Answer: B
Explanation:
Bring Your Own AI HyperFabric is the best fit when the customer already has third-party AI servers and NVMe storage but wants Cisco 6000 Series switching with simplified cloud-managed fabric operations. The Premier license aligns with the AI-focused HyperFabric use case, while Cisco describes Nexus HyperFabric as a cloud-managed AI infrastructure solution using Cisco 6000 Series switches for AI workloads.

QUESTION 2
A company is building a network fabric for an AI training cluster that will train large language models. The cluster includes 256 GPUs distributed across 32 servers. Security requirements mandate isolation between different training projects and allowing shared access to a central storage system. Which approach provides the required security isolation and maintains optimal performance for GPU-to-GPU communication?

A. Deploy separate physical networks for each project with dedicated switches and cross-connects to shared storage.
B. Use VXLAN with separate VNIs per project and Group Policy Option to control granular access to shared storage.
C. Implement port-based VLANs for each project with router ACLs controlling storage access at the distribution layer.
D. Configure QoS-based traffic steering with separate DSCP markings per project and policy-based routing to shared storage.

Answer: B
Explanation:
VXLAN provides scalable network segmentation by assigning separate VNIs to different training projects, preserving isolation without building separate physical fabrics. Group Policy Option adds policy-based control within the VXLAN overlay, allowing controlled shared access to central storage while maintaining high-performance GPU-to-GPU communication across the AI training fabric.

QUESTION 3
A bank is planning to deploy an AI inference solution at its branch locations. The solution must support workloads that require a balance of compute, memory, and potential GPU acceleration, and it must be suitable for installation by nontechnical onsite resources. The requirements are:

```
- high availability of the chassis management plane.
- support for up to 768 GB of memory for in-memory model storage and
processing
- remote server launch requiring no on-site IT staff
- use of Cisco Intersight for SaaS-based infrastructure lifecycle
management
```

Which solution meets the requirements?

A. Cisco UCS C845A M8 Server
B. Cisco UCS X-Series Direct
C. Cisco Unified Edge
D. Cisco UCS C240 Servers

Answer: C
Explanation:
Cisco Unified Edge is designed for distributed branch and edge AI deployments that need simplified remote operations through Cisco Intersight. It supports branch-friendly installation, SaaS-based lifecycle management, high-availability chassis management, up to 768 GB of memory, and balanced compute resources for inference workloads that may require GPU acceleration.

QUESTION 4
Which set of statements describes Quantized Congestion Notification?

A. Priority Flow Control reacts first to mitigate congestion, and Explicit Congestion Notification acts as a fail-safe to prevent traffic drops if Priority Flow Control is insufficient.
A congestion notification packet sent toward the source by a network switch in response to receiving Explicit Congestion Notification with congestion experienced bits set.
B. Explicit Congestion Notification reacts first to mitigate congestion, and Priority Flow Control acts as a fail-safe to prevent traffic drops if Explicit Congestion Notification is insufficient.
A congestion notification packet sent toward the source by hosts and the network switch in response to receiving Explicit Congestion Notification with congestion experienced bits set.
C. Explicit Congestion Notification reacts first to mitigate congestion, and Priority Flow Control acts as a fail-safe to prevent traffic drops if Explicit Congestion Notification is insufficient.
A congestion notification packet sent toward the source by a host in response to receiving Explicit Congestion Notification with congestion experienced bits set.
D. Priority Flow Control reacts first to mitigate congestion, and Explicit Congestion Notification acts as a fail-safe to prevent traffic drops if Priority Flow Control is insufficient.
A congestion notification packet sent toward the source by hosts and the network switch in response to receiving Explicit Congestion Notification with congestion experienced bits set.

Answer: C
Explanation:
Quantized Congestion Notification relies on Explicit Congestion Notification as the primary congestion signal so the sender can reduce its transmission rate before packet loss occurs. Priority Flow Control acts as a safety mechanism if congestion becomes severe, and the congestion notification packet is generated by the receiving host when it receives packets marked with congestion-experienced bits.

QUESTION 5
A financial company plans to deploy an AI training and inference infrastructure to support its customers. The team must use a solution that provides 96 GPUs, aligns with NVIDIA-compliant reference architectures, and enables cloud-based management of the entire on-premises network environment. Which Cisco solution meets the requirements?

A. Cisco HyperFabric AI Small AI Cluster with Cisco 6000 switches and AI servers on premises
B. Cisco HyperFabric AI Medium AI Cluster with Cisco 6000 switches and AI servers on premises

C. Cisco Nexus Dashboard deployed in AWS or Azure with Nexus 9000 switches and AI servers on premises
D. Cisco Intersight AI fabric with Nexus 9000 switches and AI servers on premises

Answer: A
Explanation:
Cisco HyperFabric AI Small AI Cluster matches the 96-GPU requirement because the validated HyperFabric reference architecture uses 12 Cisco UCS C885A M8 servers with NVIDIA HGX for a total of 96 GPUs. Cisco Nexus HyperFabric is also cloud-managed and compliant with NVIDIA Enterprise Reference Architecture, enabling centralized management of the on-premises AI network, GPU servers, and storage.

QUESTION 6
A server administrator must monitor the performance and utilization of GPUs in Cisco UCS servers running AI workloads. Which two metrics must be used in Cisco Intersight to accomplish these goals? (Choose two.)

A. GPU Utilization
B. Explorer
C. Monitor
D. GPU Performance
E. System CPU Utilization

Answer: AD
Explanation:
Cisco Intersight provides GPU-focused telemetry for UCS servers, including utilization and performance metrics. These measurements allow the administrator to monitor how heavily the GPUs are being used and whether the AI workload is receiving the expected GPU performance.

QUESTION 7
A customer deploys a UCS 885a server in a data center with two power grids. Three of the 3000W power supplies in the server connect to each grid. Due to an external event, one of the grids will be shut down for maintenance. How does the server respond to this event?

A. The server immediately shuts down and stops all operations.
B. The server operates with reduced performance and caps power to GPUs at 60%.
C. The server continues to operate at full performance using redundant power supplies, but the PSU status LED glows solid amber.
D. The server initiates a graceful shut down of the operating system.

Answer: B
Explanation:
With three 3000W power supplies lost from one power grid, the UCS C885A continues operating on the remaining grid but no longer has the full power budget available for maximum GPU operation. The server reduces performance by enforcing power limits, including capping GPU power, so workloads can continue running within the available power capacity.

QUESTION 8
A network operations team deploys HyperFabric AI and wants to monitor fabric health to proactively identify performance issues. The team must track metrics specifically relevant to AI training workloads. Which metric is most critical for identifying congestion issues that could

impact AI training performance in a HyperFabric AI deployment?

A. spanning tree topology change notifications
B. BGP EVPN route count per leaf switch
C. VXLAN tunnel packet encapsulation rate
D. PFC pause frame transmission rate per interface

Answer: D
Explanation:
PFC pause frame transmission rate is the key congestion indicator for lossless Ethernet fabrics used by AI training workloads. A high or increasing pause frame rate shows that interfaces are experiencing congestion and applying flow control, which can directly affect GPU-to-GPU communication performance in a HyperFabric AI environment.

QUESTION 9
A network engineer logs into Nexus Dashboard and sees an anomaly showing that a BGP peer connection has gone down. Next to the anomaly, a yellow bubble displays “Correlated.” What does Nexus Dashboard indicate by “Correlated” in this case?

A. It indicates that the Nexus Dashboard has correlated this anomaly to activity by specific user account.
B. It indicates that the Nexus Dashboard has seen this anomaly across more than one device in the fabric.
C. It indicates that this anomaly is a result of a higher-level root anomaly.
D. It indicates that the Nexus Dashboard has correlated the scope of the impact.

Answer: C
Explanation:
In Nexus Dashboard, a correlated anomaly means the displayed issue is associated with a higher-level root anomaly. The BGP peer-down event is being treated as a resulting or related symptom rather than the primary root cause, helping operators focus troubleshooting on the upstream issue first.

QUESTION 10
Server administrators receive reports of power utilization being near capacity across multiple UCS X-Series servers in the data center after a hardware upgrade. The chassis operates with six power supply units in redundant power mode. How can the administrator use additional power without impacting redundancy?

A. Implement a power capping policy.
B. Enable extended power capacity.
C. Add more power supplies.
D. Disable extended power capacity.

Answer: B
Explanation:
Extended power capacity allows a UCS X-Series chassis with six power supplies in redundant mode to make additional power available while still preserving redundancy. This is used when power demand increases after hardware upgrades and the administrator needs more usable power without changing the redundant power configuration.

QUESTION 11
A solutions architect must implement a monitoring strategy for a new AI/ML data center infrastructure built on Cisco UCS X-Series servers with NVIDIA H100 GPUs, managed by Cisco Intersight. The main objectives are to gain deep visibility into GPU performance, identify resource bottlenecks for AI training workloads, and ensure the overall health and efficiency of the compute environment. Which Cisco Intersight capability is most critical for effectively monitoring this AI-centric infrastructure?

A. Analyze feature automating the deployment and scaling of AI application containers and their underlying Kubernetes infrastructure, which reduces manual monitoring efforts
B. integration with Cisco Nexus Dashboard for advanced network flow analytics and anomaly detection across the data center fabric, which ensures high-speed data transfer between GPU nodes
C. basic server health reporting, which is limited to CPU and memory statistics, and requires specialized third-party tools for any GPU-specific performance monitoring or diagnostics
D. comprehensive health and performance monitoring, which provides real-time telemetry data for UCS servers, including granular metrics for NVIDIA GPUs, which enables direct insights into AI workload execution

Answer: D
Explanation:
Cisco Intersight provides health and performance monitoring for UCS servers, including real-time telemetry and GPU-specific metrics for NVIDIA GPUs. This visibility is essential for identifying bottlenecks, tracking AI training workload behavior, and maintaining the health and efficiency of GPU-accelerated UCS X-Series infrastructure.

QUESTION 12
A Cisco UCS C885A M8 server contains a GPU sled configured with six power supply units (PSUs) in an N+2 redundancy configuration. The workload on the GPU sled is currently drawing a load of 4500W. Using Power Save Mode, how many PSUs can be placed into standby mode to improve power efficiency without compromising redundancy?

A. 0
B. 1
C. 2
D. 3

Answer: C
Explanation:
With six PSUs in an N+2 configuration and a 4500W load, at least two PSUs must remain available to carry the active workload, and two additional PSUs must remain available to preserve redundancy. Therefore, four PSUs must stay active, allowing two PSUs to be placed into standby mode for power efficiency without compromising the N+2 redundancy requirement.

Thank You for Trying Our Product

Lead2pass Certification Exam Features:

★ More than **99,900** Satisfied Customers Worldwide.

★ Average **99.9%** Success Rate.

★ **Free Update** to match latest and real exam scenarios.

★ **Instant Download** Access! No Setup required.

★ Questions & Answers are downloadable in **PDF** format and **VCE** test engine format.

★ Multi-Platform capabilities - **Windows, Laptop, Mac, Android, iPhone, iPod, iPad**.

★ **100%** Guaranteed Success or **100%** Money Back Guarantee.

★ **Fast**, helpful support **24x7**.

View list of all certification exams: http://www.lead2pass.com/all-products.html

10% Discount Coupon Code: **ASTR14**