



Cisco 300-640

Implementing Cisco Data Center AI Infrastructure

Thank You for Downloading 300-640 Updated Exam Questions

<https://certs4sale.com/cisco/300-640-pdf-exam-dumps>

IMPLEMENTATION GUIDE & PRACTICE EXAM

Implementing Cisco Data Center AI Infrastructure

Cisco Certified Network Professional & Cisco Certified Network Professional Data Center Track

VENDOR

Cisco

EXAM CODE

300-640

CERTIFICATION TRACK

CCNP Data Center

TOTAL QUESTIONS

5 Questions & Answers

Question 1

SINGLE CHOICE

Your organization is planning to deploy a large-scale generative AI infrastructure to support multiple concurrent fine-tuning workloads using transformer models. The data science team has informed you that the current bottleneck is inter-GPU communication during backpropagation across multiple nodes. Which combination of technologies should you recommend to address this performance constraint?

- A. Implement NVLink for intra-node GPU communication and RoCEv2 with PFC for inter-node GPU synchronization
- B. Deploy multiple isolated clusters with Ethernet switching and implement data parallelism across separate networks
- C. Use standard TCP/IP networking with QoS policies to prioritize training traffic over other data center workloads
- D. Configure high-bandwidth storage with Fibre Channel SAN to cache model gradients and reduce network overhead

Correct Answer: A**EXPLANATION**

The correct answer addresses the actual bottleneck: inter-GPU communication during distributed training. NVLink provides high-bandwidth, low-latency GPU-to-GPU communication within a single node, while RoCEv2 (RDMA over Converged Ethernet v2) extends this capability across nodes with microsecond-level latency. PFC (Priority Flow Control) prevents packet loss on congested links, which is critical for RDMA reliability. This combination directly solves the backpropagation synchronization problem. The other options are suboptimal: isolated clusters would exacerbate the communication problem; standard TCP/IP lacks the latency characteristics needed for synchronized gradient updates; and caching gradients in storage doesn't address intra-GPU synchronization. This scenario reflects real architectural decisions in modern AI PODs like Cisco AI Infrastructure for transformers.

Question 2

SINGLE CHOICE

Your organization is planning to deploy a distributed AI training workload across multiple GPU-accelerated nodes in an on-premises data center. The training job requires high-bandwidth, low-latency communication between nodes to synchronize gradients during backpropagation. Which Cisco infrastructure components and technologies should you prioritize to meet these requirements?

- A. RoCEv2 for RDMA connectivity, PFC and ECN for congestion control, and NVLink for GPU-to-GPU communication
- B. Standard Ethernet with TCP/IP, QoS marking at Layer 2, and PCIe Direct for GPU interconnect
- C. Fibre Channel SAN for data movement, PFC for flow control, and multi-hop Ethernet switching
- D. Cloud-based GPU instances with high-speed internet connectivity and standard data synchronization protocols

Correct Answer: A**EXPLANATION**

For distributed AI training with multiple GPU-accelerated nodes, RoCEv2 (RDMA over Converged Ethernet v2) is the correct choice because it provides the low-latency, high-throughput kernel-bypass networking needed for synchronized gradient updates across nodes. PFC (Priority Flow Control) and ECN (Explicit Congestion Notification) are essential congestion control mechanisms to prevent packet loss in lossless networks required by RDMA. NVLink provides the fastest GPU-to-GPU communication for multi-GPU systems. Option 2 is incorrect because standard TCP/IP adds unacceptable latency for synchronization. Option 3 misses the point—Fibre Channel is for storage, not GPU interconnect. Option 4 describes a cloud deployment, which is not what was requested in the scenario (on-premises).

Question 3

SINGLE CHOICE

A financial services company wants to build a hybrid AI infrastructure that will run inference workloads both on-premises and in public cloud, with monthly retraining on the largest models in the cloud. The models and datasets must remain synchronized between environments, and network traffic must be encrypted end-to-end. Which approach addresses both the secure connectivity and data synchronization requirements of this hybrid deployment?

- A. VPN tunnels for encrypted connectivity, scheduled batch file synchronization via object storage, and cloud-managed model registry
- B. Public internet connectivity with TLS encryption, real-time filesystem replication, and manual model version management
- C. Direct fiber peering with no encryption, continuous data streaming between data centers, and model versioning in Git
- D. Site-to-site IPsec tunnel with asymmetric encryption, Fibre Channel SAN replication, and local model storage only

Correct Answer: A**EXPLANATION**

Hybrid AI deployments require secure, encrypted connectivity (VPNs meet this requirement) and a practical data synchronization strategy. VPN tunnels provide encrypted end-to-end communication, scheduled batch synchronization via object storage is a standard pattern for hybrid cloud that avoids continuous overhead, and a cloud-managed model registry (like ML model stores) provides version control and centralized management across environments. Option 2 fails because real-time filesystem replication is impractical across cloud boundaries and Git is not designed for large model artifacts. Option 3 is insecure (no encryption) and uses storage protocols (Fibre Channel) inappropriate for wide-area hybrid connections. Option 4 combines conflicting technologies and misses the synchronization requirement by specifying local-only storage.

Question 4

SINGLE CHOICE

You are using Cisco Intersight to monitor a data center with 200 GPU-accelerated UCS servers running inference workloads. You need to implement proactive alerting when GPU utilization exceeds 85% or when network latency to a backend storage cluster exceeds 50ms. Which operational telemetry and monitoring capabilities should you configure to meet these requirements?

- A. Custom alert rules on GPU utilization and network latency metrics, log correlation to correlate alerts with system events, and operational telemetry collection from UCS domains and network endpoints
- B. Manual polling of GPU metrics every 5 minutes, setup of threshold-based logs in syslog, and periodic manual review of Intersight dashboards
- C. Intersight advisory notifications only, SNMP traps from network devices, and historical trend analysis via Nexus Dashboard
- D. Real-time packet capture filtering on GPU processes, alert suppression on high-frequency events, and direct SSH access to servers for metric validation

Correct Answer: A**EXPLANATION**

Cisco Intersight supports custom alert rules based on performance thresholds (GPU utilization, latency), operational telemetry collection from compute and network infrastructure to gather the necessary metrics, and log correlation to correlate alerts with root-cause system events—this is the modern, proactive monitoring approach expected in the 300-640 exam. Option 2 is outdated and reactive (manual polling, unstructured logs). Option 3 mixes unrelated tools and misses the requirement for proactive alerting thresholds. Option 4 describes low-level debugging techniques (packet capture, SSH), not infrastructure monitoring, and suppression is counterproductive when trying to detect problems early.

Question 5

SINGLE CHOICE

A Generative AI model serving company is evaluating their infrastructure for token-per-second performance to support real-time inference requests. They have two options: purchasing on-premises NVIDIA H100 GPU clusters connected via high-speed Ethernet, or leveraging a hybrid approach with baseline load on-premises and burst traffic offloaded to cloud. Which factors should influence this evaluation of network, compute, and cost efficiency requirements?

- A. Consistent request volume and SLA response times, GPU compute capacity and memory bandwidth, network latency between user and inference endpoint, and power/cooling costs of on-premises versus cloud variable costs
- B. Availability of commercial AI frameworks only, number of total GPUs regardless of model type, raw network throughput in Gbps, and up-front capital expenditure alone
- C. Preference for cloud-only deployments, inference batch size optimization, storage capacity in TB, and cloud provider naming conventions
- D. GPU memory clock speeds, total employees in IT operations, internet bandwidth capacity in Mbps, and licensing costs of monitoring tools

Correct Answer: A**EXPLANATION**

Evaluating deployment architecture for Generative AI inference requires understanding workload characteristics (consistent vs. burst traffic), compute requirements (GPU capacity, memory bandwidth for token generation), network latency (critical for real-time inference), and total cost of ownership (power/cooling for on-premises versus cloud variable costs). These are the key evaluation criteria in section 2.4-2.5 of the exam. Option 2 incorrectly emphasizes capital expenditure alone and ignores operational/variable costs. Option 3 ignores the core compute and network requirements needed for inference. Option 4 conflates unrelated factors (clock speeds, employee counts, tool licensing) that don't influence architecture evaluation for inference workloads.

Exam Practice Document

Implementing Cisco Data Center AI Infrastructure (300-640)

THANK YOU FOR DOWNLOADING 300-640 UPDATED EXAM QUESTIONS

Note: Thanks For Trying The Demo Of Our 300-640 Exam Product

Visit Our Site to Purchase the Full Set of Actual 300-640 Exam Questions With Answers.

Money Back Guarantee



Click The Link Below

<https://certs4sale.com/cisco/300-640-pdf-exam-dumps>