

PRACTICE QUESTIONS

Cisco

300-640 DCAI

Cisco Certified Specialist - Data Center AI Infrastructure

SHARED BY [u/Brave-Bison-9189](#) on aiexamcerts

EDITION	Free Set 1
QUESTIONS	10
SHARED	October 15, 2026
EXAM	Cisco 300-640 DCAI, Cisco Certified Specialist - Data Center AI Infrastructure

Shared by u/Brave-Bison-9189 on aiexamcerts

10 practice questions for the Cisco 300-640 DCAI exam, Cisco Certified Specialist - Data Center AI Infrastructure, each followed by its answer and a short explanation. More free sets and the full practice exam: www.aiexamcerts.com/cisco/300-640-dcai. Spotted a mistake? Email support@aiexamcerts.com with the exam code 300-640 DCAI and the question number.

1. Which access model best limits what an administrator can change on AI infrastructure?

- A Shared administrator accounts used by the whole team
- B Unrestricted root access on every switch
- C Role-based access with least-privilege roles
- D Anonymous access to the management interface

Answer: C

Role-based access with least privilege limits changes to what each role needs, and it stays auditable. Shared administrator accounts remove individual accountability, which makes them the most tempting shortcut. Unrestricted root and anonymous management access are plainly unsafe.

2. The collective operation in which every GPU ends up with the sum of all GPUs' gradients is called _____.

Answer: all-reduce (also accepted: allreduce, all reduce)

All-reduce combines values from all participants and returns the result to each one, which is the common step for synchronizing gradients. Broadcast sends data from one source to all others, which is the most tempting confusion. Collective libraries such as NCCL provide this operation.

3. Which interconnect links GPUs inside a single server with very high bandwidth, without passing through network switches?

- A InfiniBand
- B iSCSI
- C SNMP
- D NVLink

Answer: D

NVLink is a GPU-to-GPU interconnect within a server that provides much higher bandwidth than PCIe in many configurations. InfiniBand is a fabric between servers, which makes it the tempting confusion. iSCSI is a storage protocol, and SNMP is a management protocol.

4. A GPU with more memory always outperforms one with less memory on every workload, because data movement never limits performance.
- A True
 - B False

Answer: B (False)

Performance depends on whether the bottleneck is compute, memory bandwidth, data loading, or the network, so extra memory does not guarantee a speedup. Profiling identifies the real limit. The statement is False.

5. Which two actions improve GPU utilization in a training pipeline? (Choose two.)
- A Prefetch and preprocess data so the GPUs do not wait for input
 - B Lower the batch size to the smallest value on every run regardless of memory
 - C Overlap communication with computation where the framework supports it
 - D Disable all profiling so the GPUs run faster
 - E Use a single data loader thread on every node

Answer: A, C

Prefetching keeps GPUs fed, and overlapping communication with computation hides network delay. Shrinking the batch size on every run is the most tempting option, and it usually lowers utilization. Disabling profiling and using one loader thread do not improve utilization.

6. What is the role of Priority Flow Control (PFC) in a RoCE network?
- A It assigns IP addresses to GPUs automatically
 - B It pauses one priority class when its buffer fills, avoiding drops
 - C It encrypts traffic between the switches and the storage
 - D It limits each server to a single network interface for simplicity

Answer: B

PFC sends a pause for one priority class when a buffer threshold is reached, which prevents drops for RDMA traffic. Congestion notification is handled by ECN, a separate mechanism, which makes it the most tempting confusion. Address assignment, encryption, and interface limits are unrelated.

7. Which protocol carries NVMe commands over a network fabric for low-latency access to remote flash?
- A SMTP for mail delivery
 - B SSH for secure remote shell access
 - C NVMe over Fabrics (NVMe-oF)
 - D DHCP for address assignment

Answer: C

NVMe over Fabrics extends the NVMe command set across network transports such as RDMA or TCP. SMTP, SSH, and DHCP are not storage protocols. SSH is the most tempting confusion because it also connects to remote systems.

8. The two-tier topology in which every leaf switch connects to every spine switch is called a Clos or leaf-_____ fabric.

Answer: spine

Leaf-spine, also described as a Clos fabric, gives every server pair a predictable number of hops and scales by adding spines. Ring and bus layouts do not scale this way. The fabric depends on uplink capacity to avoid bottlenecks.

9. Cisco Intersight is best described as which kind of tool?
- A A compiler for GPU kernels
 - B A storage array firmware updater only
 - C A packet capture tool for wireless networks and their access points
 - D A cloud-based management platform for Cisco infrastructure

Answer: D

Intersight provides SaaS management for Cisco compute and other infrastructure, with an on-premises virtual appliance option for restricted sites. A GPU kernel compiler is a development tool, which makes it the most tempting confusion. Storage firmware and wireless capture are narrower tools.

10. When a training job slows down, what is the best first step to narrow the cause?
- A Correlate step times with GPU, network, and storage metrics
 - B Restart the training job immediately without collecting data
 - C Delete the job's logs to free disk space
 - D Ask the users to request more GPUs

Answer: A

Correlating step time with GPU, network, and storage metrics for the same window separates compute, communication, and I/O causes. Restarting first destroys evidence, which is the most tempting reaction. Deleting logs and requesting more GPUs do not diagnose anything.