

PRACTICE QUESTIONS

Cisco

300-640 DCAI

Cisco Certified Specialist - Data Center AI Infrastructure

SHARED BY [u/throwaway_300640](#) on aiexamcerts

EDITION	Free Set 2
QUESTIONS	10
SHARED	October 15, 2026
EXAM	Cisco 300-640 DCAI, Cisco Certified Specialist - Data Center AI Infrastructure

Shared by u/throwaway_300640 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 accelerator property most often limits the size of a model that can be served from one GPU?

- A Available GPU memory versus the model's weights and activations
- B The number of CPU cores and host RAM in the server for data preparation
- C The clock speed of the host's management controller
- D The number of PCIe slots in the chassis

Answer: A

Model weights and activations must fit in accelerator memory, so available memory sets the practical size limit for one device. Host CPU cores matter for data preparation, which makes them a tempting but secondary factor. Management controller clocks and slot counts do not limit the model that fits on the GPU.

2. ECMP hashing sends each packet of a flow over a different link, so large flows can never overload a single link.

- A True
- B False

Answer: B (False)

Standard ECMP hashes each flow to one path, so a single large flow stays on one link and can overload it. Packet-level spraying is the approach that distributes the packets of one flow. The statement is False.

3. Which practice protects the management plane of AI infrastructure from unauthorized use?

- A Expose the management interface to the public internet for easy remote access
- B Management interfaces on a separate network with MFA and access control
- C Keep the default passwords on every device
- D Disable logging for management sessions

Answer: B

Isolating management networks and requiring strong authentication reduces exposure of administrative functions. Public exposure is the most tempting convenience, and it invites attack. Default passwords and disabled logs are common weaknesses that enable misuse.

4. A pipeline stage that loads the next batch of data while the GPU processes the current batch is called data _____.

Answer: prefetching

Prefetching overlaps input work with computation so the accelerator does not wait for data. Insufficient prefetching is a common cause of low GPU utilization. Caching the whole dataset in memory is a different tactic with its own memory cost.

5. Which two factors commonly limit the throughput of a multi-server training job? (Choose two.)

- A Network bandwidth and latency between servers
- B Collective communication patterns that wait on the slowest participant
- C The rack's power distribution unit model number
- D The number of GPUs that are powered off in the rack
- E The default login banner configured on each switch

Answer: A, B

Network bandwidth and latency limit how fast gradients move, and synchronization waits on the slowest worker in each collective. Powered-off GPUs and login banners do not affect step time. The power distribution unit model does not matter either when power is adequate.

6. What is a GPU node in an AI cluster?

- A A network switch port that links a GPU to the storage array
- B A software license that covers one GPU driver
- C A server with GPUs, host memory, and network adapters
- D A storage volume reserved for model checkpoints

Answer: C

A GPU node is a server equipped with accelerators, host memory, and network interfaces that participates in a training or inference job. A switch port is a network element, which makes it the tempting confusion. Licenses and storage volumes are different things.

7. Which component lets a Kubernetes cluster schedule GPU-backed pods correctly?

- A A larger DNS cache on the control plane for faster name lookups
- B A longer naming prefix for pods
- C Disabling the container runtime on GPU nodes
- D A GPU device plugin that advertises GPUs to the scheduler

Answer: D

A device plugin exposes GPUs as schedulable resources, so pods can request them and the scheduler can place them. A DNS cache does not advertise hardware, which makes it a plausible but wrong distractor. Pod naming and disabling the container runtime do not help scheduling.

8. Infrastructure-as-code tools make AI cluster configuration less repeatable, because every change is typed by hand on each device.
- A True
 - B False

Answer: B (False)

Declarative templates and playbooks record the desired state, which can be versioned, reviewed, and reapplied consistently. Hand changes on each device are the problem these tools are meant to remove. The statement is False.

9. Which mechanism marks packets to signal congestion so that senders can slow down before buffers overflow?
- A Explicit Congestion Notification (ECN)
 - B Spanning Tree Protocol
 - C Network Address Translation
 - D Dynamic Host Configuration Protocol for address leasing

Answer: A

ECN marks packets when a queue exceeds a threshold, and endpoints reduce their sending rate in response, as in DCQCN. Spanning Tree prevents loops, which is a different job and the most tempting confusion. NAT and DHCP handle addressing.

10. Why is object storage often used for large training datasets and model artifacts?
- A It is a transactional database with strict row-level locking for writes
 - B It scales capacity well and stores large objects with metadata
 - C It replaces GPU memory during training
 - D It runs the training code directly

Answer: B

Object storage scales to very large capacities and suits large, rarely changed objects such as datasets and artifacts. It is not a low-latency file system, which is the trade-off to plan for. Transactional databases and GPU memory have different roles.