Cisco.300-640.v2026-05-16.q29

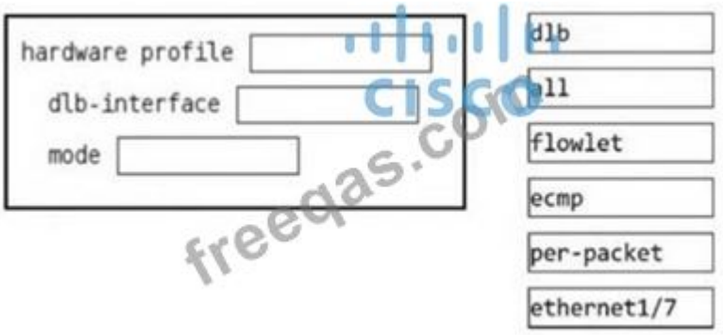
Exam Code:	300-640
Exam Name:	Implementing Cisco Data Center AI Infrastructure
Certification Provider:	Cisco
Free Question Number:	29
Version:	v2026-05-16
# of views:	114
# of Questions views:	290
https://www.freeqas.com/qa/Cisco/300-640/Cisco.300-640.v2026-05-16.q29.html	

NEW QUESTION: 1
Which result is provided through image recognition using AI?
A. anonymization of audit trails for compliance objectives
B. detection and labeling of visual elements from digital sources
C. scheduling of management tasks during maintenance periods
D. routing of multicast messages across network boundaries
Answer: (SHOW ANSWER)
Image recognition uses AI to analyze digital images or video frames and identify, detect, classify, or label visual elements such as objects, people, patterns, or scenes.

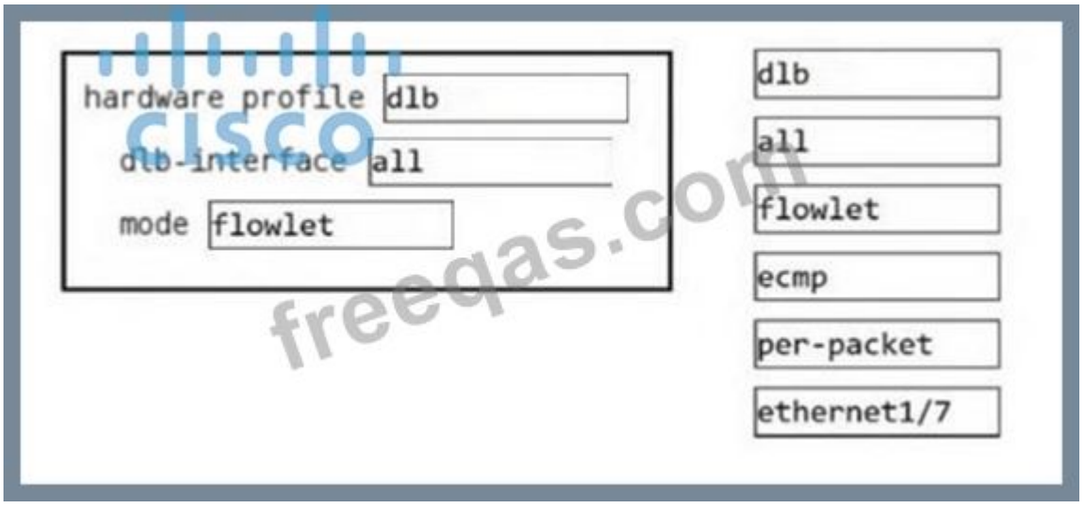
NEW QUESTION: 2
Drag and Drop Question

Port	Description	Intvl	Rx Mbps	Rx %%	Tx Mbps	Tx %%
Ethernet1/1	N/A	30/30	260305.8	32.5%	400967.0	50.1%
Ethernet1/2	N/A	30/30	379356.1	47.4%	360170.0	45.0%
Ethernet1/3	N/A	30/30	354389.4	44.3%	349509.0	43.7%
Ethernet1/4	N/A	30/30	362254.1	45.3%	400033.0	50.0%
Ethernet1/5	N/A	30/30	602293.9	75.3%	469556.0	58.7%
Ethernet1/6	N/A	30/30	372852.2	46.6%	469424.0	58.7%
Ethernet1/7	N/A	30/30	347895.3	43.5%	326102.0	40.8%
Ethernet1/8	N/A	30/30	471291.6	58.9%	328276.0	41.0%

Refer to the exhibit. When a new AI backend network that supports Cisco UCS-based GPU servers with NVIDIA Connect-X network adapters is benchmarked, the traffic distribution observed has unacceptable variability across the interfaces. Drag and drop the code snippets from the bottom onto the boxes in the code to configure the switch to resolve the issue.



Answer:



NEW QUESTION: 3

A global enterprise is deploying a new AI-driven analytics platform that requires high-performance GPU acceleration, large memory capacity, and robust virtualization support. The current compute environment must co-exist with the newly purposed GPU-enabled workload. This environment will continue to grow, so the customer wants to scale out the resources as needed. Which Cisco product meets the requirements?

A. UCS X-series

B. UCS M-series

C. Unified Edge

D. UCS C-series Standalone

Answer: (SHOW ANSWER)

Cisco UCS X-Series is designed for scalable, modular compute environments that support GPU- enabled workloads, high memory capacity, virtualization, and coexistence with existing infrastructure. Its modular architecture allows the enterprise to expand resources as AI workload demands grow.

NEW QUESTION: 4

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 (LEAVE A REPLY)

NEW QUESTION: 5
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: (SHOW ANSWER)
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.

NEW QUESTION: 6
A data center operator configures telemetry streaming on Cisco Nexus switches. What is the primary operational advantage over traditional SNMP polling?
A. Increased packet fragmentation
B. Real-time visibility with lower overhead
C. Elimination of routing tables
D. Mandatory hardware replacement
Answer: B (LEAVE A REPLY)
Streaming telemetry continuously exports operational data in near real time, improving visibility and reducing polling overhead compared to SNMP. AI infrastructures require rapid monitoring of congestion and latency conditions. Telemetry does not remove routing tables or require hardware replacement.

NEW QUESTION: 7
A data center is deploying an AI infrastructure for LLM training and model inference. This infrastructure requires high GPU density, low-latency interconnects, and the ability to scale compute resources. When

considering the Cisco UCS C885A M8 Rack Server for the deployment, which design aspect is the most critical for achieving optimal performance and scalability in this environment?
A. selecting the appropriate Nexus Fabric and its configuration to minimize latency
B. using the maximum number of CPU cores to support the virtualization overhead
C. implementing a redundant power supply configuration to ensure high availability
D. configuring the server with the minimum amount of RAM to reduce initial costs
Answer: A (LEAVE A REPLY)
For UCS C885A M8 servers running LLM training and inference workloads, the surrounding Nexus fabric design is critical because GPU-dense systems depend on low-latency, high- bandwidth, lossless network connectivity for scale-out performance. Selecting and configuring the correct Nexus fabric minimizes communication delays between compute nodes and enables the AI infrastructure to scale efficiently.

NEW QUESTION: 8
Which network design approach should be used to build a high-performance, nonblocking AI/ML data center fabric that supports 64 fully equipped UCS C885A nodes and provides the lowest possible latency and lossless Ethernet?
A. leaf-spine
B. flat Layer 2
C. rail optimized
D. traditional three-tier
Answer: C (LEAVE A REPLY)
A rail-optimized design is used for large AI/ML clusters because it aligns GPU-to-GPU communication paths across the fabric, minimizes latency, and supports high-throughput lossless Ethernet for distributed training workloads.

NEW QUESTION: 9
A corporate network administrator manages a HyperFabric AI environment and wants to receive email notifications about assertions in the fabric. The administrator sets up a recipient email endpoint under the notification endpoints in the controller, but notifications are not received.
Which two configurations cause the issue? (Choose two.)
A. The email server configuration is not configured using the DNS name.
B. The recipient email address is not verified by the HyperFabric controller.
C. The fabric of interest and email endpoint are not configured in the default notification policy.
D. The DNS server IP is not defined in the controller.
E. The HyperFabric controller cannot reach the email server of the Acme Corporation.
Answer: B,C (LEAVE A REPLY)
Email endpoints must be verified before they can receive HyperFabric notifications, and the endpoint must also be included with the relevant fabric in the notification policy. Creating the email endpoint alone does not activate assertion notifications for a fabric until both of these notification requirements are completed.

NEW QUESTION: 10

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 (LEAVE A REPLY)
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.

NEW QUESTION: 11
A network engineer must secure management access to Cisco Nexus AI switches. Which protocol should replace Telnet to provide encrypted administrative sessions?
A. FTP
B. SSH
C. SNMPv1
D. TFTP
Answer: B (LEAVE A REPLY)
SSH provides encrypted remote management sessions and protects credentials and configuration traffic from interception. Telnet transmits data in plaintext and is unsuitable for secure AI data center environments. FTP and TFTP are file transfer protocols, while SNMPv1 lacks modern security protections.

NEW QUESTION: 12
A customer is deploying a new AI fabric with all new NVIDIA GPUs and must verify the performance between the nodes in the network. Which tool must be used to validate the customer benchmarks?
A. NCCL
B. Spirent

C. iPerf
D. IXIA
Answer: A (LEAVE A REPLY)
NCCL is the standard tool used to validate GPU-to-GPU communication performance in NVIDIA- based AI fabrics. It measures collective communication behavior between GPU nodes, making it appropriate for benchmarking the AI training network performance.

NEW QUESTION: 13
A data center facility experienced a power outage, and the UCS X-Series servers did not restart automatically when power is restored. The configuration must be updated to prioritize these servers to boot as soon as possible to resume their work. Which set of power policy configurations must be used to meet these requirements?
A. Power Profiling: On
Power Priority: Default
Power Restore: Immediate
Processor Package Power Limit(PPL): Maximum
B. Power Profiling: On
Power Priority: High
Power Restore: Always On
Processor Package Power Limit(PPL): Default
C. Power Profiling: On
Power Priority: Default
Power Restore: Always On
Processor Package Power Limit(PPL): Default
D. Power Profiling: On
Power Priority: High
Power Restore: Immediate
Processor Package Power Limit(PPL): Maximum
Answer: B (LEAVE A REPLY)
The servers must be configured to power back on automatically after power is restored, so the restore behavior should be set to always power on. Setting the power priority to high ensures these UCS X-Series servers receive priority during power recovery, allowing them to resume workloads as quickly as possible.

NEW QUESTION: 14
Refer to the exhibit. Users report delayed response times and network latency when running certain AI workloads on a specific Cisco UCS Intersight Managed Mode Domain. Based on the fault message and reported symptoms, which troubleshooting action must be taken first to address the network latency issues?



A. Review the AI workload configuration to optimize resource allocation (CPU, memory, GPU) for the affected applications.
B. Examine the UCS fabric interconnect configuration for port-channel 11 on Fabric Interconnect B.
C. Check the physical cabling and SFP+ transceivers on port 11 of switch B for potential hardware failures or incorrect configurations.
D. Review the pause frame counters for signs of congestion on Fabric Interconnect B Port 11.
Answer: C (LEAVE A REPLY)
The fault shows a receive error statistics threshold crossing on switch B port 11. Because receive errors on a physical interface commonly indicate cabling, transceiver, or physical-layer issues, the first troubleshooting step is to inspect the physical connection and SFP+ components on the affected port before investigating workload or higher-layer configuration.

NEW QUESTION: 15
An engineer made changes to the corporate DNS, which caused a Cisco UCS X-Series domain to lose connectivity to Intersight. To restore connectivity, the management interface DNS and NTP settings must be updated using the device console UI. Which configuration change must be made in Intersight to allow the updated DNS configuration to persist?
A. Edit the IMC access policy associated to the fabric interconnects.
B. Create a device connector policy and apply it to the blade chassis.
C. Create a network connectivity policy and apply it to the blade chassis.
D. Edit the domain profile deployed to the fabric interconnect.
Answer: D (LEAVE A REPLY)
In Cisco Intersight Managed Mode, the fabric interconnect management DNS and NTP settings are controlled through the UCS domain profile. If the settings are changed locally from the device console UI, the deployed domain profile must also be updated in Intersight so the corrected configuration is not overwritten during the next policy deployment or compliance remediation.

NEW QUESTION: 16
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. Monitor
B. Insights
C. Active CPU Utilization
D. GPU Memory Utilization
E. Explorer
Answer: C,D (LEAVE A REPLY)
Cisco Intersight performance monitoring uses resource utilization metrics to identify bottlenecks in AI workloads. Active CPU utilization shows whether host processing is limiting workload execution, while GPU memory utilization shows how effectively GPU memory is being consumed during model processing and training or inference tasks.

Valid 300-640 Dumps shared by PrepPdf.com for Helping Passing 300-640 Exam! PrepPdf.com now offer the **newest 300-640 exam dumps**, the PrepPdf.com 300-640 exam **questions have been updated** and **answers have been corrected** get the **newest** PrepPdf.com 300-640 dumps with Test Engine here: https://www.preppdf.com/Cisco/300-640-prepaway-exam-dumps.html (**82** Q&As Dumps, **40%OFF** **Special Discount: Exam-Tests**)

NEW QUESTION: 17
What is a primary behavior that occurs during the model training stage of the AI lifecycle?
A. embedding the model into systems with APIs
B. establishing success criteria
C. adjusting internal weights and biases
D. detecting model drift
Answer: C (LEAVE A REPLY)
During model training, the model learns from training data by iteratively adjusting its internal weights and biases to reduce prediction error and improve performance against the defined objective function.

NEW QUESTION: 18
A network architect wants to provide high availability for AI compute nodes connected to dual Cisco Nexus switches. Which technology enables active-active uplink forwarding without STP blocking?
A. vPC
B. GLBP
C. VRRP
D. PVST+
Answer: A (LEAVE A REPLY)
Virtual Port Channel (vPC) allows devices to use active-active uplinks across two Nexus switches while avoiding spanning-tree blocking. This increases bandwidth utilization and resiliency for AI clusters. GLBP and VRRP provide gateway redundancy, while PVST+ still relies on blocking redundant paths.

NEW QUESTION: 19
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 (LEAVE A REPLY)

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.

NEW QUESTION: 20

Which type of Cisco Intersight profile, when deployed to fabric interconnects, includes the configuration settings for ports, VLANs, and VSANs?

A. Unified Edge profiles

B. UCS Chassis profiles

C. UCS Domain profiles

D. UCS Server profiles

Answer: C (LEAVE A REPLY)

A UCS Domain profile is deployed to fabric interconnects in Cisco Intersight Managed Mode and contains the domain-level configuration, including port policies, VLANs, VSANs, network connectivity, and other fabric interconnect settings.

NEW QUESTION: 21

An engineer notices excessive broadcast traffic inside a legacy AI server VLAN. Which technology should be implemented to reduce flooding and improve scalability?

A. VXLAN EVPN

B. RIP

C. PAT

D. GRE tunneling

Answer: A (LEAVE A REPLY)

VXLAN EVPN reduces unknown unicast and broadcast flooding by using control-plane learning with BGP EVPN. This improves scalability in modern AI data centers. RIP is outdated for large- scale fabrics, while PAT and GRE tunneling do not address broadcast domain limitations.

NEW QUESTION: 22

An engineer is deploying a Cisco AI POD environment to support large-scale AI workloads. The solution must provide a software-defined storage platform with a Disaggregated Shared- Everything architecture, allow compute nodes to run on Cisco UCS C225 M8 servers managed through Cisco Intersight, and deliver exabyte-scale, all-flash performance with integrated high- speed RDMA and GPU Direct Storage capabilities. Which solution meets the requirements?

A. NetApp All-Flash FAS (AFF)

B. Pure Storage FlashBlade

C. Local NVMe SSDs within each GPU server

D. VAST Data Universal Storage

Answer: D (LEAVE A REPLY)

VAST Data Universal Storage provides a software-defined, disaggregated shared-everything storage architecture designed for AI-scale workloads. It supports Cisco UCS-based deployments managed through Intersight and delivers high-performance all-flash storage with RDMA and GPU Direct Storage capabilities for large-scale AI training, fine-tuning, and inference environments.

NEW QUESTION: 23

What is a function of NVLink technology within a high-performance AI server or cluster?

A. to enable remote access and virtualization of GPU resources for distributed AI development teams

B. to manage power distribution and thermal regulation across all installed GPU units

C. to establish a secure, encrypted channel for data transfer between the AI server and external storage

D. to provide a high-speed, low-latency direct interconnect for GPU-to-GPU communication

Answer: D (LEAVE A REPLY)

NVLink provides a high-bandwidth, low-latency direct interconnect between GPUs. In AI servers and clusters, it accelerates GPU-to-GPU communication for training and inference workloads by allowing GPUs to exchange data much faster than through standard PCIe paths.

NEW QUESTION: 24

A national law firm is deploying a RAG solution for their legal research team to more efficiently use their 50 TB of confidential data and documents when preparing new contracts and briefs.

They decided on a UCS X-Series chassis equipped with four GPUs and six X210c servers. Which storage solution fits this deployment?

A. IP-based storage array

B. local storage in a JBOD configuration

C. tape storage

D. local storage in a RAID 5 configuration

Answer: (SHOW ANSWER)

A RAG deployment with 50 TB of confidential documents requires scalable shared storage that can be accessed by multiple UCS X210c servers. An IP-based storage array provides centralized capacity, shared access, and operational simplicity for storing and retrieving large datasets used by the RAG pipeline.

NEW QUESTION: 25

A network engineer is deploying an AI training fabric using Cisco Nexus switches with RoCEv2 enabled. During testing, packet drops occur when traffic bursts exceed buffer capacity. Which feature should be configured to minimize congestion and improve lossless Ethernet performance for GPU traffic?

A. BPDU Guard

B. Priority Flow Control

C. Storm Control

D. DHCP Snooping

Answer: (SHOW ANSWER)

Priority Flow Control (PFC) allows selective pausing of traffic classes during congestion, which is critical for RoCEv2 environments carrying GPU-to-GPU communication. AI workloads are highly sensitive to packet loss and retransmissions. BPDU Guard and DHCP Snooping are security features, while Storm Control only limits broadcast or multicast traffic and does not create lossless behavior for RDMA traffic.

NEW QUESTION: 26
What describes inference traffic patterns in AI deployments?
A. optimizes data storage solutions for data lakes
B. simplifies complex neural network architectures
C. provides real-time decision making
D. eliminates the need for data preparation
Answer: (SHOW ANSWER)
Inference traffic is associated with using a trained model to generate predictions or responses from new input data. In AI deployments, this commonly supports real-time or near-real-time decision making, such as classification, recommendations, detection, or automated responses.

NEW QUESTION: 27
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 (LEAVE A REPLY)
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.

NEW QUESTION: 28
An engineer deploys an AI fabric on Cisco Nexus 9000 Series Switches connected to Cisco UCS nodes with NVIDIA GPUs. The requirements call for optimal network performance for training workloads. Adaptive routing is enabled on the NICs, and per-packet load balancing is enabled on the switches. Which other configuration must be implemented for this integration to be fully operational?
A. Add the "adaptive-routing" keyword under the switch load balancing configuration.
B. Set the cisco_dlb=1 option in the NIC configuration so the NIC knows it can use adaptive routing.
C. Include switch configuration "hardware profile spectrum-x" in global configuration mode.
D. Include matching statements in your QoS policy so the switch knows which incoming packets to load balance.
Answer: B (LEAVE A REPLY)

The NIC must be explicitly configured with the Cisco dynamic load-balancing setting so it recognizes that the connected Cisco Nexus fabric supports adaptive routing. This completes the integration between NIC-side adaptive routing and switch-side per-packet load balancing for optimized AI training traffic.

NEW QUESTION: 29
An engineer configures Cisco Nexus Dashboard Fabric Controller for AI infrastructure automation. The company wants centralized policy management and simplified fabric provisioning across multiple pods. Which capability provides this benefit?
A. Distributed ACL processing
B. Intent-based fabric automation
C. Local VLAN switching
D. Manual CLI consistency checks
Answer: (SHOW ANSWER)
Intent-based fabric automation simplifies deployment and policy consistency across large AI infrastructures. Nexus Dashboard Fabric Controller automates provisioning, validation, and operational management. Distributed ACLs and local VLAN switching are operational features but do not centralize orchestration. Manual CLI verification is error-prone and unsuitable for rapidly scaling AI environments.

Valid 300-640 Dumps shared by PrepPdf.com for Helping Passing 300-640 Exam! PrepPdf.com now offer the **newest 300-640 exam dumps**, the PrepPdf.com 300-640 exam **questions have been updated** and **answers have been corrected** get the **newest** PrepPdf.com 300-640 dumps with Test Engine here: https://www.preppdf.com/Cisco/300-640-prepaway-exam-dumps.html (**82** Q&As Dumps, **40%OFF** **Special Discount: Exam-Tests**)