



- Expert Verified, Online, **Free**.



CERTIFICATION TEST

- CertificationTest.net - Cheap & Quality Resources With Best Support

A company is implementing a new network architecture and needs to consider the requirements and considerations for training and inference. Which of the following statements is true about training and inference architecture?

- A. Training architecture and inference architecture have the same requirements and considerations.
- B. Training architecture is only concerned with hardware requirements, while inference architecture is only concerned with software requirements.
- C. Training architecture is focused on optimizing performance while inference architecture is focused on reducing latency.
- D. Training architecture and inference architecture cannot be the same.

Suggested Answer: C

Community vote distribution

C (100%)

🗨️ 👤 **cybe001** 5 months, 3 weeks ago

Selected Answer: C

C is correct

upvoted 1 times

For which workloads is NVIDIA Merlin typically used?

- A. Recommender systems
- B. Natural language processing
- C. Data analytics

Suggested Answer: A

Community vote distribution

A (100%)

 **Iledar951** 1 month, 2 weeks ago

Selected Answer: A

NVIDIA Merlin is a specialized application framework within the NVIDIA data center platform designed for a specific workload domain:

- Application Frameworks: NVIDIA offers application frameworks tailored for common domains.
- Merlin's Purpose: Examples of these frameworks include RIVA for conversational AI, Drive for Autonomous Vehicles, and MERLIN for Recommendation Systems, among many others.
- Broader Context: These frameworks are layered atop the software stack, including CUDA and DOCA, and numerous software libraries that transparently provide acceleration to developers

upvoted 1 times

Which NVIDIA parallel computing platform and programming model allows developers to program in popular languages and express parallelism through extensions?

- A. CUDA
- B. CUMML
- C. CUGRAPH

Suggested Answer: A

Community vote distribution

A (100%)

🗨️ 👤 **Iledar951** 1 month, 2 weeks ago

Selected Answer: A

CUDA is the NVIDIA parallel computing platform and programming model that serves as the foundation for accelerated computing on GPUs.

- The CUDA Toolkit is defined as an Nvidia groundbreaking parallel programming model that provides essential optimizations for deep learning, machine learning, and high-performance computing (HPC), leveraging NVIDIA GPUs.

- CUDA is part of the base layer of the software stack, encompassing the programming model for GPUs

upvoted 1 times

Which of the following aspects have led to an increase in the adoption of AI? (Choose two.)

- A. Moore's Law
- B. Rule-based machine learning
- C. High Powered GPUs
- D. Large amounts of data

Suggested Answer: *CD*

Currently there are no comments in this discussion, be the first to comment!

In training and inference architecture requirements, what is the main difference between training and inference?

- A. Training requires real-time processing, while inference requires large amounts of data.
- B. Training requires large amounts of data, while inference requires real-time processing.
- C. Training and inference both require large amounts of data.
- D. Training and inference both require real-time processing.

Suggested Answer: *B*

Currently there are no comments in this discussion, be the first to comment!

Which of the following statements is true about GPUs and CPUs?

- A. GPUs are optimized for parallel tasks, while CPUs are optimized for serial tasks.
- B. GPUs have very low bandwidth main memory while CPUs have very high bandwidth main memory.
- C. GPUs and CPUs have the same number of cores, but GPUs have higher clock speeds.
- D. GPUs and CPUs have identical architectures and can be used interchangeably.

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!

Which two components are included in GPU Operator? (Choose two.)

- A. Drivers
- B. PYTorch
- C. DCGM
- D. TensorFlow

Suggested Answer: AC

Currently there are no comments in this discussion, be the first to comment!

Which phase of deep learning benefits the greatest from a multi-node architecture?

- A. Data Augmentation
- B. Training
- C. Inference

Suggested Answer: *B*

Currently there are no comments in this discussion, be the first to comment!

Which architecture is the core concept behind large language models?

- A. BERT Large model
- B. State space model
- C. Transformer model
- D. Attention model

Suggested Answer: *C*

Currently there are no comments in this discussion, be the first to comment!

What is a key value of using NVIDIA NIMs?

- A. They provide fast and simple deployment of AI models.
- B. They have community support.
- C. They allow the deployment of NVIDIA SDKs.

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!

The foundation of the NVIDIA software stack is the DGX OS.
Which of the following Linux distributions is DGX OS built upon?

- A. Ubuntu
- B. Red Hat
- C. CentOS

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!

What is the name of NVIDIA's SDK that accelerates machine learning?

- A. Clara
- B. RAPIDS
- C. cuDNN

Suggested Answer: *C*



Community vote distribution

B (100%)

  **6e1e645** 1 week, 3 days ago

Selected Answer: B

I think C is incorrect. B - cuDNN accelerates deep learning primitives, but Rapids accelerates Machine Learning and data science.
upvoted 1 times

  **9320159** 3 months, 2 weeks ago

Selected Answer: B

cuDNN is a GPU-accelerated library specifically for deep neural networks, used primarily for deep learning, but RAPIDS focuses more broadly on accelerating machine learning and data science workflows.
upvoted 3 times

Which aspect of computing uses large amounts of data to train complex neural networks?

- A. Machine learning
- B. Deep learning
- C. Inferencing

Suggested Answer: *B*

Currently there are no comments in this discussion, be the first to comment!

Which of the following statements correctly differentiates between AI, Machine Learning, and Deep Learning?

- A. Machine Learning is a subset of AI, and AI is subset of Deep Learning
- B. AI and Deep Learning are the same, while Machine Learning is a separate concept.
- C. AI is a subset of Machine Learning, and Machine Learning is a subset of Deep Learning.
- D. Deep Learning is a subset of Machine Learning, and Machine Learning is a subset of AI.

Suggested Answer: *D*

Currently there are no comments in this discussion, be the first to comment!

How is the architecture different in a GPU versus a CPU?

- A. A GPU acts as a PCIe controller to maximize bandwidth.
- B. A GPU is architected to support massively parallel execution of simple instructions.
- C. A GPU is a single large and complex core to support massive compute operations.

Suggested Answer: *B*

Currently there are no comments in this discussion, be the first to comment!

What factors have led to significant breakthroughs in Deep Learning?

- A. Advances in hardware, availability of fast internet connections, and improvements in training algorithms
- B. Advances in sensors, availability of large datasets, and improvements to the "Bag of Words" algorithm.
- C. Advances in hardware, availability of large datasets, and improvements in training algorithms.
- D. Advances in smartphones, social media sites, and improvements in statistical techniques.

Suggested Answer: *C*

Currently there are no comments in this discussion, be the first to comment!

Which type of GPU core was specifically designed to realistically simulate the lighting of a scene?

- A. Tensor Cores
- B. CUDA Cores
- C. Ray Tracing Cores

Suggested Answer: *C*

Currently there are no comments in this discussion, be the first to comment!

Which GPUs should be used when training a neural network for self-driving cars?

- A. NVIDIA H100 GPUs
- B. NVIDIA L4 GPUs
- C. NVIDIA DRIVE Orin

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!

A customer is evaluating an AI cluster for training and is questioning why they should use a large number of nodes. Why would multi-node training be advantageous?

- A. The model is too large to fit into GPU memory
- B. The model is being used by a large number of users
- C. The model is being used for large scale inference workloads

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!

When should RoCE be considered to enhance network performance in a multi-node AI computing environment?

- A. A network that experiences a high packet loss rate (PLR).
- B. A network with large amounts of storage traffic.
- C. A network that cannot utilize the full available bandwidth due to high CPU utilization.

Suggested Answer: *C*

Currently there are no comments in this discussion, be the first to comment!

Which are three key features of InfiniBand networking technology?

- A. High reliability, high latency, and CPU offloads.
- B. High latency, high reliability, and high bandwidth.
- C. GPU offloads, low latency, high reliability.
- D. Low latency, high bandwidth, and CPU offloads.

Suggested Answer: *D*

Currently there are no comments in this discussion, be the first to comment!

An IT Professional is considering whether to implement an on-prem or cloud infrastructure. Which of the following is a key advantage of on-prem infrastructure?

- A. Lower upfront costs and capital expenditure
- B. Scalability and flexibility
- C. Ensure data security and sovereignty
- D. Easy remote management.

Suggested Answer: *C*

Currently there are no comments in this discussion, be the first to comment!

Which feature of RDMA reduces CPU utilization and lowers latency?

- A. Increased memory buffer size.
- B. Network adapters that include hardware offloading.
- C. NVIDIA Magnum I/O software.

Suggested Answer: *B*

Currently there are no comments in this discussion, be the first to comment!

What is one key advantage that Cloud GPU Infrastructure has over On-Prem GPU infrastructure?

- A. Lower cost barrier to entry.
- B. Reduced cost of I/O traffic.
- C. Greater flexibility for hardware orchestration.

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!

When training a neural network, what is the most common pattern of storage access?

- A. Random write
- B. Sequential read
- C. Sequential write

Suggested Answer: *B*

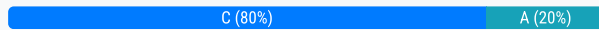
Currently there are no comments in this discussion, be the first to comment!

How many distinct network fabrics are in an AI cluster?

- A. 3
- B. 2
- C. 4
- D. 5

Suggested Answer: C

Community vote distribution



🗳️ 👤 **9320159** 3 months, 3 weeks ago

Selected Answer: A

Storage and in-band are considered to be one
upvoted 1 times

🗳️ 👤 **Titus169** 4 months, 2 weeks ago

Selected Answer: C

1- Compute Network
2- Storage Network
3- Inband Managment
4- Out of Band Managment Network
upvoted 1 times

🗳️ 👤 **cybe001** 5 months, 3 weeks ago

Selected Answer: C

The answer is C. 4.

In an AI cluster, you typically have four separate “fabrics”:

Scale-Up Fabric

Connects the GPUs within a single node (e.g., NVLink/NVSwitch).

Scale-Out Fabric

Links GPUs across nodes for distributed training (e.g., RoCE or InfiniBand).

Front-End Fabric

Carries storage traffic, application I/O, and in-band cluster management.

Management Network

Handles out-of-band tasks, such as provisioning, health checks, and firmware updates.

upvoted 3 times

How many 1 Gb Ethernet in-band network connections are in a DGX H100 system?

- A. 1
- B. 2
- C. 0

Suggested Answer: *C*

Currently there are no comments in this discussion, be the first to comment!

How many Mellanox ConnectX-6 Single Port VPI cards are in a DGX A 100 system?

- A. 8
- B. 16
- C. 4

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!

Which is the best PUE value for a data center?

- A. PUE of 1.2
- B. PUE of 3.5
- C. PUE of 5.0
- D. PUE of 2.0

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!

Which solution should be recommended to support real-time collaboration and rendering among a team?

- A. A cluster of servers with NVIDIA T4 GPUs in each server.
- B. A DGX SuperPOD.
- C. An NVIDIA Certified Server with RTX based GPUs.

Suggested Answer: *C*

Currently there are no comments in this discussion, be the first to comment!

What is a key benefit of using NVIDIA GPUDirect RDMA in an AI environment?

- A. It increases the power efficiency and thermal management of GPUs.
- B. It reduces the latency and bandwidth overhead of remote memory access between GPUs.
- C. It enables faster data transfers between GPUs and CPUs without involving the operating system.
- D. It allows multiple GPUs to share the same memory space without any synchronization.

Suggested Answer: B

Community vote distribution



B (100%)

  **junk** 1 month ago

Selected Answer: B

GPUDirect with RDMA is a technology developed by NVIDIA that enables direct memory access between NVIDIA GPUs and other devices, such as network adapters, storage systems, and other GPUs. It allows for efficient data transfers without involving the CPU, resulting in reduced latency and increased bandwidth.

upvoted 1 times

  **9320159** 3 months, 2 weeks ago

Selected Answer: B

NVIDIA GPUDirect RDMA (Remote Direct Memory Access) allows GPUs to communicate directly with each other across nodes without involving the CPU or operating system. This results in:

Lower latency for data transfers

Reduced bandwidth overhead

Improved scalability in multi-node AI training environments

It's especially beneficial in distributed deep learning, HPC, and real-time AI inference setups.

upvoted 4 times

What enables moving data between GPU memory and local or remote storage without using the CPU?

- A. NVLink
- B. GPUDirect P2P
- C. InfiniBand
- D. GPUDirect Storage

Suggested Answer: *D*

Currently there are no comments in this discussion, be the first to comment!

When using an InfiniBand network for an AI infrastructure, which software component is necessary for the fabric to function?

- A. Verbs
- B. MPI
- C. OpenSM

Suggested Answer: *C*

Currently there are no comments in this discussion, be the first to comment!

When deploying high density workloads in a data center, what are the three main resource constraints that need to be considered?

- A. Processing speed, storage capacity, and network connectivity.
- B. Power, cooling, and physical space.
- C. Bandwidth, security, and redundancy.

Suggested Answer: *B*

Currently there are no comments in this discussion, be the first to comment!

What is an advantage of InfiniBand over Ethernet?

- A. InfiniBand always provides higher bandwidth than Ethernet.
- B. InfiniBand supports RDMA while Ethernet does not.
- C. InfiniBand offers lower latency than Ethernet.

Suggested Answer: *C*

Currently there are no comments in this discussion, be the first to comment!

In a data center, what is the purpose and benefit of a DPU?

- A. A DPU is responsible for providing backup and disaster recovery solutions.
- B. A DPU is used for managing physical infrastructure, such as power and cooling.
- C. A DPU is responsible for managing network connections and security.
- D. A DPU is designed to offload, accelerate, and isolate infrastructure workloads.

Suggested Answer: *D*

Currently there are no comments in this discussion, be the first to comment!


Which of the following statements is true about Kubernetes orchestration?

- A. It is bare-metal based but it supports containers.
- B. It has advanced scheduling capabilities to assign jobs to available resources.
- C. It has no inferencing capabilities.
- D. It does load balancing to distribute traffic across containers.

Suggested Answer: *BD*

Community vote distribution

B (100%)

 **9320159** 3 months, 2 weeks ago

Selected Answer: B

Kubernetes features an advanced scheduler that assigns containers (pods) to appropriate nodes based on available resources, constraints, and workload needs, ensuring optimal resource utilization in a cluster.

upvoted 1 times

In an AI cluster, what is the purpose job scheduling?

- A. To gather and analyze cluster data on a regular schedule.
- B. To monitor and troubleshoot cluster performance.
- C. To assign workloads to available compute resources.
- D. To install, update, and configure cluster software.

Suggested Answer: *C*

Currently there are no comments in this discussion, be the first to comment!

In an AI cluster, what is the importance of using Slurm?

- A. Slurm is used for data storage and retrieval in an AI cluster.
- B. Slurm is responsible for AI model training and inference in an AI cluster.
- C. Slurm is used for interconnecting nodes in an AI cluster.
- D. Slurm helps with managing job scheduling and resource allocation in the cluster.

Suggested Answer: *D*

Currently there are no comments in this discussion, be the first to comment!

What NVIDA tool should a data center administrator use to monitor NVIDIA GPUs?

- A. NVIDIA System Monitor
- B. NetQ
- C. DCGM

Suggested Answer: *C*

Currently there are no comments in this discussion, be the first to comment!

What is the primary command for checking the GPU utilization on a single DGX H100 system?

- A. nvidia-smi
- B. ctop
- C. nvml

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!

What is a common tool for container orchestration in AI clusters?

- A. Kubernetes
- B. MLOps
- C. Slurm
- D. Apptainer

Suggested Answer: A

Currently there are no comments in this discussion, be the first to comment!



Which of the following NVIDIA tools is primarily used for monitoring and managing AI infrastructure in the enterprise?

- A. NVIDIA NeMo System Manager
- B. NVIDIA Data Center GPU Manager
- C. NVIDIA DGX Manager
- D. NVIDIA Base Command Manager

Suggested Answer: D

Community vote distribution

B (100%)

  **neta1o** 1 month, 3 weeks ago

Selected Answer: B

B. NVIDIA Data Center GPU Manager (DCGM)

DCGM is designed to simplify GPU administration in data centers by providing active health monitoring, diagnostics, system alerts, and governance policies including power and clock management. It helps administrators monitor GPU utilization, health, and system configuration across large clusters, improving reliability and operational efficiency. DCGM also integrates with Kubernetes and job schedulers for comprehensive AI infrastructure management.

- NVIDIA NeMo System Manager is not a standard NVIDIA tool for infrastructure monitoring.
- NVIDIA DGX Manager mainly manages DGX systems but is not as broad or specialized as DCGM for AI infrastructure.
- NVIDIA Base Command Manager is focused on AI workflow orchestration, not GPU hardware monitoring.

upvoted 1 times

Which NVIDIA software provides the capability to virtualize a GPU?

- A. Horizon
- B. vGPU
- C. virtGPU

Suggested Answer: *B*

Currently there are no comments in this discussion, be the first to comment!

When monitoring a GPU-based workload, what is GPU utilization?

- A. The maximum amount of time a GPU will be used for a workload.
- B. The GPU memory in use compared to available GPU memory.
- C. The percentage of time the GPU is actively processing data.
- D. The number of GPU cores available to the workload.

Suggested Answer: *C*

Currently there are no comments in this discussion, be the first to comment!