True/False

FPGAs are more power-efficient and offer higher performance compared to CPUs for certain types of workloads.

True

Racks shelves are used to store servers and other IT equipments

True

In a data center for AI, servers includes both general purpose CPUs and accelerators (such as GPUs, TPUs and FPGAs)

True

NVlink is a low-bandwidth connection among Graphical Processing Units

False

Tower servers can be cooled easily

True

Racks shelves may contain switches

True

Rack servers need additional cooling systems due to their high overall component density

True

Rack, blade and towers are form factor of servers in data centers

True

In the context of rack servers, 1U is 44.45cm

False

Graphical Processing Units within a tray are connected using high-bandwidth interconnects, such as NVlink

True

Servers in data centers are like regular personal-computers with the same form factor

False

Tower servers have a cable management that is more complex than Rack ones

True

Racks servers have a cable management that is more complex than Tower ones

False

Tower servers consumes less space than rack and blade

False

Blade servers require expensive configurations

True

Graphical Processing Units support data-parallel computation

True

A deep neural-network cannot be trained on multiple Graphical Processing Units

False

In warehouse-scale computers the software running on these systems executes on clusters of hundreds to thousands of individual serve

True

TPUs are more power-efficient and offer higher performance compared to GPUs in deep learning tasks

True

GPUs require specialized software and programming frameworks to fully leverage their parallel processing capabilities

True

GPUs (Graphics Processing Units) are primarily used in datacenters for accelerating graphical applications and gamining

False

GPUs are generally less expensive to deploy and maintain compared to CPUs in datacenter environments

False

TPUs are widely available and can be easily procured for use in most datacenter environments

False

The motherboard in a server provides sockets and slots for installing CPUs, memory modules, storage devices, and network interface cards

True

Tower servers are the most common form factor used in warehouse-scale computers (WSCs)

False

Blade servers are known for their compact size and space-saving design

True

Blade servers are known for their compact size and high density, making them suitable for space-constrained environments

True

GPUs are primarily used for sequential computations and are not suitable for parallel processing

False

TPUs (Tensor Processing Units) are specialized hardware accelerators designed for machine learning tasks

True

TPUs (Tensor Processing Units) are custom-designed integrated circuits optimized for machine learning workloads

True

FPGAs (Field-Programmable Gate Arrays) offer high flexibility and can be reconfigured for specific computational tasks

True

Compared to CPUs, GPUs offer significantly higher performance for tasks like image and video processing

True

Compared to CPUs, GPUs are less efficient at handling simple models with small training sets

False

TPUs are less flexible than CPUs and GPUs but excel at dense vector and matrix computations

True

FPGAs are easier to program than CPUs and GPUs due to their high-level synthesis (HLS) capabilities

False

FPGAs (Field-Programmable Gate Arrays) are reconfigurable hardware accelerators that offer flexibility for various computational tasks

False

The majority of real-world machine learning systems consist of ML code.

False

Machine learning frameworks are exclusively designed for deep learning tasks

False

Deep learning frameworks focus on neural network topologies with many hidden layers

True

The main components of a server include the CPU, RAM, storage devices, and network interface cards

True

The motherboard in a server provides sockets and slots for connecting various components

True

Rack servers are designed to be mounted in racks, while tower servers resemble traditional desktop PCs

True

Tower servers offer better scalability and are easier to upgrade compared to rack and blade servers

False

Blade servers are generally more expensive than rack and tower servers

True

GPUs (Graphical Processing Units) are primarily used for graphics rendering and are not suitable for general-purpose computing

False

Machine learning and deep learning have significantly increased the demand for specialized hardware accelerators in data centers

True

Moore's Law accurately predicts the doubling of AI training compute requirements every 18-24 months

False

GPUs are not suitable for data-parallel computations where the same program is executed on many data elements in parallel

False

The performance of a synchronous GPU system for deep learning is limited by the slowest learner and slowest network messages

True

GPUs within a server rack are typically connected using low-bandwidth interconnects like PCIe

False

TPUs are designed to accelerate computations involving tensors, which are multidimensional arrays used in machine learning

True

TPUv1 is a training-focused accelerator, while TPUv2 is optimized for inference tasks

False

A TPU Pod can contain up to 512 TPU cores and 4 TB of total memory

True

TPUv3 is the first liquid-cooled accelerator in Google's data centers

True

FPGAs are less flexible than CPUs and GPUs but offer higher performance and lower power consumption for specific tasks

True

The majority of real-world machine learning systems consist of ML code, with configuration, data collection, and other tasks being less significant

False

FPGAs are typically programmed using high-level languages like Python and Java

False

Multiple choice

What is a Graphical Processing Unit (GPU)?

A A specialized hardware meant to provide storage in a data center

B A specialized hardware meant to support data-parallel computation

C An Internet-of-Things device meant to operate in the field

D A specialized hardware meant to provide networking in a data center