Networking for AI

Unit 8 - Summary

- **AI Data Centers:** AI Data Centers designed for high performance, scalability and security for AI workloads. AI Data Center have four networks: Compute, Storage, In-band management, and Out-of-band management.
- **Networking Requirements:** Networking requirements for AI workloads include high throughput, low latency, and low processing overhead.
- **InfiniBand:** InfiniBand is a networking technology that features high throughput, low latency, and low processing overhead.
- **Ethernet:** Ethernet is the predominant LAN technology, with RoCE allowing RDMA over Ethernet networks.
- **NVIDIA Networking Portfolio:** NVIDIA Networking Portfolio designed to support high-performance, scalable, and secure data centers, particularly for AI-driven workloads. NVIDIA offers a networking portfolio including BlueField Platform, ConnectX SmartNICs, Spectrum Ethernet switches, Quantum InfiniBand networking switches, and LinkX cables.

What are the four networks in a typical AI data center?

A typical AI data center has four networks: Compute network, Storage network, In-band management network, and Out-of-band management network.

What are the key factors related to networking for AI workloads?

The key factors related to networking for AI workloads include network topology, bandwidth and latency, network protocols, data transferring techniques, and management methods.

What is InfiniBand?

InfiniBand is a networking technology that features high throughput, low latency, and low processing overhead. It is maintained by the InfiniBand Trade Association (IBTA).

What is Ethernet?

Ethernet is a networking technology that was introduced in 1979 and was first standardized in the 1980s as IEEE standards. It describes how network devices can format and transmit data to other devices on the same local area network. Ethernet is the predominant LAN technology.

What is the NVIDIA Networking Portfolio?

NVIDIA's networking portfolio is designed to support high-performance, scalable, and secure data centers, particularly for AI-driven workloads. The NVIDIA Networking Portfolio includes BlueField DPUs and SuperNICs, ConnectX SmartNICs, Spectrum Ethernet switches, Quantum InfiniBand switches, and LinkX cables.