

Recommended network cards for AI servers



AI Overview

For AI servers, high-performance NICs like AMD Pensando Pollara 400, NVIDIA Mellanox ConnectX series, and high-speed PCIe Ethernet cards are top choices, offering 200–400 Gbps bandwidth, low latency, and advanced offload features.

Key Considerations for AI NICs

- Bandwidth:** Dense GPU clusters often require 200–400 Gbps per server to prevent network bottlenecks during distributed training.
- Latency:** Low tail latency is critical for inference workloads to meet service-level agreements.
- Offload Features:** RDMA, GPU direct, and programmable packet processing reduce CPU overhead and accelerate GPU-to-GPU communication.
- PCIe Topology:** Ensure the NIC aligns with server PCIe lanes to maximize throughput.
- Port Configuration:** Dual-port NICs can separate AI traffic from storage or provide redundancy.

Top NIC Options for AI Workloads

- AMD Pensando™ Pollara 400 AI NIC**
 - Bandwidth:** Up to 400 Gbps Ethernet
 - Features:** Fully programmable P4 engine, RDMA support, OCP-3.0 form factor
 - Use Case:** Ideal for hyperscale AI clusters, large LLM training, and inference acceleration
- NVIDIA Mellanox ConnectX Series**
 - Bandwidth:** 200–400 Gbps
 - Features:** RDMA over Converged Ethernet (RoCE), GPU Direct RDMA, low-latency networking
 - Use Case:** Distributed AI training, HPC workloads, and GPU-intensive inference
- High-Speed PCIe NICs from Axiom and Other OEMs**
 - Bandwidth:** 100–400 Gbps
 - Features:** Cost-effective alternatives with reliability and compatibility for GPU nodes
 - Use Case:** Enterprise AI servers and GPU clusters where budget and flexibility are important
- AMD Instinct MI350P PCIe Cards (for AI acceleration)**
 - While primarily a GPU accelerator, these cards integrate high-bandwidth memory and PCIe connectivity, complementing NIC performance for inference and RAG pipelines.

Market Trends The AI NIC market is growing rapidly, dri...

Article Content

Efficient Log Management with Microsoft Fabric

Introduction In the era of digital transformation, managing and analyzing log files in real-time is essential for maintaining application health,...

How to choose a network card for a server

A practical 2026 guide to choosing a server NIC: speed, port count, PCIe, NUMA, RDMA, virtualization, latency, and compatibility. Learn how to avoid expensive design mistakes and hidden

QRZ Forums

Ham Made Software This is a curated forum - all posts must be approved by a moderator so please be patient for approval. ALL software written by QRZ Subscribers (including AI-writ

How to choose a network card for a server

A server NIC differs from a regular “network card” by far more than speed alone. In server scenarios, what matters is stable operation under long-term heavy load, mature drivers, support for

Compute Node Hardware — NVIDIA AI Enterprise: Software

The Software Reference Architecture is comprised of individually optimized NVIDIA-Certified System servers that follow a prescriptive design pattern to ensure optimal performance

Foxconn's Apple era fades as AI servers drive growth in Taiwan tech

Foxconn began producing reference designs for Nvidia's graphics cards around 2002 and started making general-purpose servers for cloud service providers' data centres as early as around 2009.

Network Interface Cards for AI Servers Market

These cards now support programmable pipelines that can offload AI acceleration frameworks, deep packet inspection, and container-to-container traffic orchestration, thereby reducing CPU overhead

Technological Advances in Network Interface Cards (NICs) for AI

AI-optimized NICs: Customized NICs designed specifically for AI workloads. Software-defined NICs (SDNICs): Software-controlled NICs that provide flexibility and programmability.

Unihost: Choosing the Right Server Specs for AI Workloads – CPU vs

A comprehensive guide to selecting the right server specifications (CPU, GPU, RAM) for AI workloads, covering deep learning, inference, and data processing."

Building the AI Server

AI accelerator servers are optimized to handle the processing required for different types of AI workloads, but what they have in common is the need to scale and connect multiple cards in a

Network Interface Cards (NICs) for AI Servers Projected to Grow at

This report offers a comprehensive analysis of the Network Interface Cards (NICs) for AI servers market, covering all aspects from market size and segmentation to key players and future

Quick Understanding GPU Server Network Card Configuration in AI Era

Delve into GPU server network configurations & optical communication solutions in the era of GenAI. Learn about effective computing power, PCIe bandwidth, and NADDOD's advanced

High-Performance Ethernet Networking for Artificial Intelligence Systems

Before digging into the details of how to maximize the network performance, it is critical to understand the server and network architecture basics. The diagram below shows a very high-level architecture

Azure OpenAI Service Multitenant Load Balancing and Token Per

This example shows how a multitenant service can distribute requests evenly among multiple Azure OpenAI Service instances and manage tokens per minute (TPM)...

Cisco Data Center Networking Blueprint for AI/ML Applications

This document provides a best practice blueprint for building a modern network environment that will allow AI/ML workloads to run at their best using shipped hardware and software

How to choose NICs for AI data centers: Bandwidth, latency and

Selecting network interface cards for AI data centers requires a balance of bandwidth, latency, off-load features and peripheral component interconnect express (PCIe) topology to optimize

Contact Us

For more information, pricing, or custom solutions, please contact us:

Website: <https://umele-pestovani.eu>

Email: info@umele-pestovani.eu

Phone: +49 152 287 6349

Address: Neue Mainzer Straße 66-68, 60311 Frankfurt am Main, Germany

This document is for informational purposes only. Specifications subject to change without notice.

