EIT Opto-Routing

Large-capacity video memory AI server

-Powered by EIT Opto-Routing

Overview

We strongly recommend a server grade platform like Intel Xeon® or AMD EPYC™ for hosting LLMs and applications using them. Those platforms have key features like lots of PCI-Express lanes for GPUs and storage, high memory bandwidth/capacity, and ECC memory support. Running large language models (LLMs), high-resolution Stable Diffusion or FLUX generations, or complex voice and video AI workflows efficiently requires a significant amount of GPU Video RAM (VRAM). This is one of the most important hardware specifications when choosing a graphics card for any kind. A server for local AI inference should not be chosen by the most expensive graphics card, but by whether the model, working cache and parallel requests fit into video memory, and whether the system has enough CPU resources, PCIe lanes, power and cooling. By the end of this article, readers will be equipped with the knowledge to make informed decisions about their AI.

Large-capacity video memory AI server

Server with GPU: for your AI and machine learning

Whether a server is suitable for AI training depends on the size (number of parameters) of the AI model used. The GEX131 has 96GB of VRAM and

NVIDIA GPU Servers for AI, Deep Learning , ASA

ThecombinationofmassiveGPU-acceleratedcompute,state-of-the-artserverhardware, and software optimizations enable organizations to scale to hundreds

Building Your Own AI Rig: More Memory, More

Power

In this guide, we explore the importance of memory capacity in AI workloads and provide recommendations for building your own AI rig with more

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."

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

With Unihost's dedicated servers, you get access to cutting-edge hardware combinations optimized for AI workloads, including high-performance GPUs with substantial VRAM, powerful multi

GPU Memory Essentials for AI Performance , NVIDIA

To run AI models locally, the GPU memory size is crucial as it directly impacts the size and complexity of the models, with larger models requiring more

Best Unified Memory Computers for Local LLMs (2025):

Unified memory has become one of the most important features for anyone running local LLMs in 2025. Instead of splitting memory between CPU

AI Storage and Servers: Meeting the Demands of

Discover how AI storage solutions integrated into powerful AI servers optimize artificial intelligence workflows, from training to archiving.

Hardware Recommendations for Large Language Model

Our Large Language Model Servers are tested and optimized to give you the best performance and reliability. View our hardware recommendations.

Choosing the Right Storage for Enterprise AI Workloads

Effective enterprise AI requires the right storage for specific workloads. Storage decisions based on performance and affordability are key.

How to Pick the Right Server for AI? Part Two: Memory

How to Pick the Right Memory for Your AI Server? Also known as RAM, memory is used in a server to store programs and data for the processors'

Local AI Inference Server 2026: How to Choose GPU, CPU and VRAM

Learn how to size VRAM, CPU, PCIe lanes, memory, power and cooling for a reliable local AI inference server. A practical guide for avoiding GPU overkill and planning around real workloads

Global Memory Shortage Crisis: Market Analysis and

AI servers and enterprise environments require far more memory per system than consumer devices, so the AI build-out is pulling a disproportionate

GPU Servers for AI: A Comprehensive Guide

Explore the essentials of GPU servers in AI development. Learn about their architecture, benefits, and how to choose the right server for your AI

How to Choose the Best GPU Server for AI Workloads

Learn how to select the ideal GPU server for your AI workloads, considering use cases, hardware specs, scalability, and operational costs.

ITPro Today, Network Computing, IoT World Today combine

ITPro Today, Network Computing and IoT World Today have combined with TechTarget. The page you are looking for may no longer exist.

GPU Server for AI: Practical Component Choices

A clear guide to hardware choices, explaining when a GPU server for AI fits, how to size VRAM, RAM, and NVMe, and how to avoid wasted capacity in

TechInsights Platform

Trusted by 125,000 semiconductor professionals. You're one step away from the most authoritative semiconductor intelligence. Take the final step--log in now to unlock:

12 Best High VRAM GPU Options In 2026 (Consumer

Running large language models (LLMs), high-resolution Stable Diffusion or FLUX generations, or complex voice and video AI workflows

Microsoft 365 Roadmap , Microsoft 365

The Microsoft 365 Roadmap lists updates that are currently planned for applicable subscribers. Check here for more information on the status of new features and

How to Pick the Right Server for AI? Part Two: Memory

In this section, we look at how memory, storage, power supply units (PSUs), thermal management, expansion slots, and I/O ports may affect the

AI infrastructure stocks Lumentum, Celestica, Seagate beat

While Nvidia has been the biggest infrastructure winner during the AI boom, other data center stocks have performed better this year.

Contact Us

For datasheets, pricing, or custom optical networking solutions, please visit: https://www.entrenamientointeligente.es