Unlocking the Future of AI with ASUS AI POD

Featuring NVIDIA GB200 NVL72

Introduction: Supercomputing

In the rapidly evolving world of artificial intelligence (AI) and high-performance computing (HPC), the ASUS AI POD stands out as a game-changing solution. Designed to meet the most demanding workloads, the ASUS AI POD integrates cutting-edge hardware, advanced networking, and a comprehensive software stack to deliver unparalleled performance, scalability, and efficiency. At its core lies the NVIDIA GB200 NVL72, a revolutionary AI server built to tackle large-scale AI applications with unmatched processing power and energy efficiency.

The ASUS AI POD is not just a hardware solution—it's a complete ecosystem. With its innovative architecture, the POD leverages NVIDIA's Blackwell GPUs and Grace CPUs, interconnected via high-speed NVLink to deliver extreme AI and HPC performance. This is further enhanced by a robust networking infrastructure, including NVIDIA Quantum-2 InfiniBand and Spectrum-4 Ethernet switches, ensuring low-latency, high-bandwidth communication across the cluster.

What truly sets the ASUS AI POD apart is its seamless integration of hardware and software. The ASUS AFS software solution simplifies deployment, management, and optimization of AI workloads, while the ASUS Infrastructure Deployment Center (AIDC) accelerates setup and ensures operational efficiency. Combined with advanced liquid cooling and scalable storage solutions like WEKA, the ASUS AI POD is designed to future-proof your AI infrastructure.

Whether you're training large language models, running hyperscale data centers, or pushing the boundaries of research and development, the ASUS AI POD with NVIDIA GB200 NVL72 offers a total solution that outpaces the competition. Its ability to deliver 30X faster real-time inference, coupled with energy-efficient cooling and end-to-end support, makes it the ideal choice for enterprises looking to harness the full potential of AI.

Continue reading to learn more about the ASUS AI POD system and deployment strategies. **The future of AI starts here.**

ASUS | BUSINESS

High-Performance AI POD: An Integrated Architecture for Demanding Workloads

ASUS AI POD is a fully integrated solution designed to accelerate demanding AI and high-performance computing (HPC) workloads. The system leverages cutting-edge hardware and a comprehensive software stack to deliver a scalable, robust, and highly efficient platform for research, development, and production environments.

Computing

- **GPUs:** The core computing units of the AI POD solution are **NVIDIA GB200 Blackwell GPUs**, NVIDIA's latest flagship GPUs delivering extreme AI and HPC performance. These GPUs are interconnected via high-speed NVLink.

- **AI Servers: ASUS AI POD AI Servers** form the foundation of the compute nodes. These servers are specially designed to house multiple GB200 GPUs and are optimized for AI and HPC applications, providing high-density computing power.

- **General-Purpose Servers:** In addition to AI servers, the architecture includes **ASUS RS Series Rack Servers**. These servers handle management and supporting workloads such as cluster management, monitoring and other tasks, enhancing the overall system's versatility.

Networking

High-Speed Interconnects: Three primary networking technologies are employed to optimize different communication needs.

- **NVLink Switch**: Designed for ultrafast communication between GPUs, ensuring data can be transferred rapidly, which is crucial for efficient AI model training.

- **InfiniBand Switch:** Powered by **ConnectX InfiniBand** network adapters, it provides high-bandwidth, low-latency networking for inter-node and storage communication, ensuring data can be quickly delivered to compute nodes.

- **Ethernet Switch:** Accelerated by the **BF3 DPU**, it handles management, monitoring, and external network communication, providing standard network connectivity.