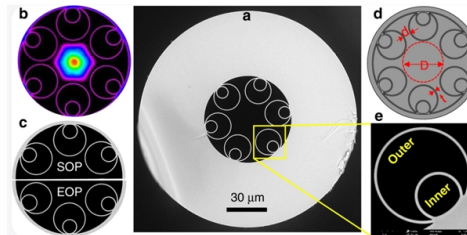


Huawei AI Server



Article Overview

Huawei's Intelligent AI Server, exemplified by the CloudMatrix 384, delivers high-performance AI computing with integrated hardware-software synergy for enterprise and industry applications.

Overview

Huawei's Intelligent AI Server ecosystem is designed to support large-scale AI workloads, including model training, inference, and industry-specific applications. The flagship CloudMatrix 384 server is powered by 384 Ascend 910C AI chips, forming a “supernode” architecture that enables up to 300 petaflops of compute performance, surpassing comparable NVIDIA systems in aggregate throughput. The server leverages high-speed interconnects and the UnifiedBus (UB) network to ensure low-latency communication between chips, achieving token generation latency under 10 milliseconds and throughput of 5 million tokens per second across 1,000 cards.

Hardware Architecture

- **AI Chips:** Ascend 910C NPUs provide high-efficiency AI computation, optimized for large-scale model training and inference.
- **Processors:** ARM-based Kunpeng CPUs power the servers, enabling tight integration with Huawei's AI software ecosystem and reducing reliance on x86 architectures.
- **Networking:** UB 2.0 interconnects and AI-centric switching support massive AI clusters, ensuring high bandwidth and low latency for distributed workloads.
- **Cooling & Power:** Fully liquid-cooled data centers support up to 80 kW heat dissipation per cabinet, with power usage effectiveness (PUE) of 1.1, accommodating AI workloads requiring up to 200 kW per cabinet.

Article Content

Huawei set to dominate China's AI server market, phasing out Nvidia

A new report states that Huawei and Cambricon, the leading chip suppliers in China's AI segment, will rule over the

Huawei Cloud

Huawei Cloud opens up Huawei's over 30 years of expertise in ICT and digitalization. We dive into digital and provide Everything as

Huawei Ascend 950PR: The 1.56 PFLOP AI Chip vs

At least two non-Chinese countries will sign major Huawei AI infrastructure deals by mid

Chinese tech firms from Huawei to Inspur push "all-in-one" DeepSeek AI ...

Dozens of Chinese hi-tech manufacturers – from Lenovo Group and Huawei Technologies to Inspur Group – are

Huawei's full stack AI data platform and agent foundry

Huawei announced a full stack of AI model training and inference data storage and pipeline

Huawei AI server out-performs Nvidia's most powerful AI server

At \$8 million, the price of the Huawei cluster is 3x the price of the Nvidia GB200 NVL72, while power consumption is

Atlas: Opening the door to AI with massive computing power

Huawei is accelerating its research in AI, and continues to embrace the data processing challenges of the AI era with a full-stack AI

Huawei Cloud announces agentic AI products

Moving forward, Huawei Cloud will drive software-hardware synergistic innovation to build the foundation for enterprise

Huawei Unveils World's Most Powerful SuperPoDs and SuperClusters

The TaiShan 950 SuperPoD, combined with Huawei's distributed GaussDB, can serve as a viable alternative to

Atlas 500 Pro AI Edge Server User Guide (Model 3000) 18

Purpose This document describes the Atlas 500 Pro AI edge server (model 3000), Atlas 500 Pro (model 3000) for short, in terms of

Huawei Unveils Fully Upgraded AI WAN Solution with AI-Centric ...

To address stagnant growth in the home broadband market and severe package homogenization, Huawei's AI WAN

Overview

The Atlas 800I A2 inference server is an AI inference device running on Kunpeng 920 and Ascend 910 AI Processor s. It features

Huawei Unveils Next-Gen Atlas 950 & 960 SuperPoD AI

Huawei claims to be taking the competition directly to NVIDIA's ground, as their

Atlas 500 Pro AI Edge Server User Guide (Model 3000) 18

The Atlas 500 Pro (model 3000) is a 2 U AI edge server powered by Huawei Kunpeng 920 processors, featuring superb computing

ModelArts AI Development Platform | Huawei Cloud

Huawei Cloud ModelArts is a one-stop AI development platform that enables developers and data

Atlas 800 Inference Server User Guide (Model 3000) 17

This document describes the appearance, structure, components, and specifications of the Atlas 800 inference server (model 3000),

Huawei Unveils Next-Gen Atlas 950 & 960 SuperPoD AI

Huawei claims to be taking the competition in the rack-scale segment directly to NVIDIA's

Overview

The Atlas 800 training server (model 9010) is an AI training server running on the Huawei Ascend 910 AI Processor and Intel

Ascend AI Cloud Service | Huawei Cloud

Ascend AI Cloud Service offers instant access to immense yet cost-effective AI computing power, a reliable

Huawei launches CloudMatrix 384 server as an alternative to Nvidia's

CloudMatrix 384 is powered by 384 of Huawei's most sophisticated AI chips, the Ascend 910C graphics processing

AI Server Boom: Huawei Atlas 950, Dell PowerEdge, HPE ProLiant,

The global AI computing infrastructure has entered a period of explosive growth in 2026. At the recent Mobile World

Huawei launches new AI server solution – Trivium China

What happened: On April 10, Huawei launched its new AI server system, CloudMatrix 384. Huawei claims it

Huawei and Cambricon lead China's AI server chip surge, piling

Huawei and Cambricon are set to dominate China's AI server market as geopolitics and self-reliance accelerate

Huawei Launches Its AI Data Platform to Power Faster AI Adoption for ...

At the Huawei AI DC Innovation Forum during MWC Barcelona 2026, Huawei unveiled its AI Data Platform, designed to address the

Huawei launches CloudMatrix 384 AI System to rival

Huawei says the system uses “supernode” architecture that allows the chips to interconnect

Elastic Cloud Server (ECS)

Elastic Cloud Server (ECS) provides secure, scalable, on-demand compute resources for you to flexibly deploy applications and

Huawei shows off AI computing system to rival Nvidia's

Huawei has become widely regarded as China's most promising domestic supplier of chips

Huawei Ascend 950PR: The 1.56 PFLOP AI Chip vs

Huawei launches the Ascend 950PR with 1.56 petaflops and 112GB HBM. How the 6,000 AI

Huawei AI Servers Test Effectiveness of US Curbs

Huawei AI Servers Test Effectiveness of US Curbs May 23, 2025 — 02:55 am EDT
Written by Hedder ->

Huawei Atlas AI Computing Platform: A Gift for Developers

The Huawei Atlas AI computing platform provides diverse product forms—such as modules, accelerator cards, edge

Contact Us

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

Website: <https://jrsekwele.co.za>

Email: info@jrsekwele.co.za

Phone: +31-20-794-2536

Address: Herengracht 47, 1015 BB Amsterdam, Netherlands

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

