Huawei Intelligent AI Server



Overview

Huawei Cloud Announces Agentic AI Products, Shaping the Moving forward, Huawei Cloud will drive software-hardware synergistic innovation to build the found.

Key Takeaways

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. OverviewHuawei'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 ArchitectureAI 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.

© 2026 METRA FIBER MARKET - All rights reserved

Software and A

I IntegrationHuawei's AI servers are tightly integrated with Agentic AI platforms and ModelArts Next, providing: Reinforcement Learning as a Service (RLaaS) for enterprise AI tasks. Model routing and matrix capabilities to dynamically select optimal models for inference, improving efficiency and reducing costs by ~20%. Enterprise-grade agent platforms (AgentArts) for long-running tasks, security, and industry-specific AI deployment.

Elastic Memory Service to expand video RAM and reduce latency in multi-round AI interactions.

Mind

Spore and CANN frameworks for AI model development and deployment, offering a full-stack alternative to NVIDIA CUDA. Industry ApplicationsHuawei's Intelligent AI Servers are deployed across multiple sectors: Healthcare: Smart diagnostics and AI-assisted treatment planning. Manufacturing: Predictive maintenance and process optimization.

Scientific Computing: High-performance simulations and research workloads.

Enterprise I

T: AI-enhanced database operations, application servers, and content generation services.

Performance and Scalability

Supports clusters with over 100,000 AI cards, enabling 200 EFLOPS of total computing power. Unified scheduling for general and AI workloads ensures 90% resource utilization and rapid provisioning in under 30 minutes.

Cloud

Matrix 384 delivers 3–4x higher inference performance per card compared to previous Huawei AI servers. Strategic AdvantagesHuawei's AI server ecosystem emphasizes hardware-software synergy, domestic supply chain independence, and scalable AI infrastructure to overcome geopolitical constraints and provide a competitive alternative to Western AI hardware. The combination of Ascend NPUs, Kunpeng CPUs, and Agentic AI platforms positions Huawei as a leader in enterprise AI deployment, particularly in regions with restricted access to NVIDIA GPUs.

In summary, Huawei's Intelligent AI Servers, anchored by the CloudMatrix 384 and integrated with Agentic AI software, offer high-performance, scalable, and industry-ready AI computing solutions that support both general-purpose and specialized AI workloads across multiple sectors.

© 2026 METRA FIBER MARKET - All rights reserved

Article Content

Huawei Cloud Announces Agentic AI Products, Shaping the

Moving forward, Huawei Cloud will drive software-hardware synergistic innovation to build the foundation for enterprise-grade AI innovation, working alongside global customers, partners, and

Huawei Unveils Fully Upgraded AI WAN Solution with

At the summit, Huawei introduced its fully upgraded AI WAN Solution, built on an AI-Centric architecture. By redefining experience awareness,

Huawei Cloud Expanding AI Infrastructure for Industries

Huawei Cloud has outlined how it is building AI infrastructure and developing models for industry applications, with deployments spanning

Huawei''s CloudMatrix AI Supercomputer: A New Force in AI Compute ...

At the recent World AI Conference in Shanghai, Huawei unveiled the CloudMatrix 384, a massive AI cluster designed to serve China''s growing demand for large-scale model training—at a

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

China''s Huawei Technologies Co. Ltd. threw down the gauntlet to Nvidia Corp. Saturday when it revealed its most powerful artificial intelligence server system to date.

Huawei Cloud: Fostering the Fertile Ground for Compute,

In the AI era, Huawei Cloud is committed to pursuing innovation in AI computing power and technologies, helping every customer become an AI pioneer in their industry.

Advantages_Intelligent EdgeFabric-Huawei Cloud

It uses Huawei general-purpose servers and AI hardware and is deeply integrated with Huawei Ascend chips to provide high-performance, low-cost edge AI inference computing power. IEF

Huawei releases CloudMatrix 384 supernode: a milestone ...

CloudMatrix 384 supernode is Huawei''s latest masterpiece in the field of AI infrastructure. Its core innovation is to achieve the transformation of resource supply mode from

Huawei Connect 2025: Building AI Infrastructure In A Sanctioned World

© 2026 METRA FIBER MARKET - All rights reserved

It supports real-time inferencing, intelligent scheduling, and context-aware services, making it a strategic asset in Huawei''s push for technological sovereignty.

Huawei challenges Nvidia with rival ''CloudMatrix'' AI servers

Huawei''s CloudMatrix servers offer more memory and bandwidth than Nvidia''s GB200 NVL72 at the cost of considerably more power. Huawei is looking to take on Nvidia with a domestic

Huawei Unveils AI Talent Development Service

Huawei has introduced an integrated AI talent development service solution that encompasses "planning, training, and evaluation." This solution

China AI, kept current: policy, chips, models | AI in Asia

China AI, kept current: policy, chips, models, robots and the market The AIinASIA living hub for China: the Five-Year Plan and the amended Cybersecurity Law, the 41 per cent domestic chip

Huawei Launches Open-Access SuperPoD Architecture

Themed All Intelligence, HUAWEI CONNECT 2025 will delve into AI across three dimensions: strategy, technology, and ecosystems. You can expect

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

Redefining Computing-Network Limits: Intelligent Computing WAN Ensures Lossless Computing Efficiency To meet explosive enterprise demand for AI computing power, Huawei''s AI

Corporate Information

The intelligent world is approaching faster, and AI is poised to redefine the future of society more deeply and broadly than ever before. AI will continue to boost

Huawei Releases AI DC Products and Solutions for ...

During MWC Barcelona 2026, Huawei released a series of innovative products and solutions centered on AI Data Centers (AI DCs) at its AI DC Innovation Session.

Huawei Unveiled the Latest SuperPoD, Making an AI

With an innovative "cluster + SuperPoD" architecture, Huawei is delivering a resilient computing foundation, powered by an open and

Linewell and Huawei Launch Public Service AI Agent ...

© 2026 METRA FIBER MARKET - All rights reserved

At MWC Barcelona 2026, Linewell Software Group and Huawei have launched the Public Service AI Agent Solution. The launch marks an intelligent upgrade of public services from efficiency

Related Video Reference

This video was associated with the source search result. Verify technical details against current product documentation and project requirements.

Contact Us

Continue your research online:

Website: https://metramarket.es
Technical inquiry: https://metramarket.es/contact.html

This document is for preliminary research. Verify specifications against current standards, product documentation and project requirements.

© 2026 METRA FIBER MARKET - All rights reserved