Huawei AI Computing Cluster Server



Overview

In the future, CloudMatrix super nodes can build large clusters of more than 10,000 slices to provide computing power, which will greatly lower the threshold for large-scale commercial use of AI.

Key Takeaways

Huawei's CloudMatrix 384 is a high-density AI computing cluster featuring 384 Ascend 910C NPUs, delivering 300 PFLOPs of BF16 compute performance, designed to rival Nvidia's top AI servers.

Overview and Architecture

The Cloud

Matrix 384 is Huawei's flagship AI server system, unveiled at the World Artificial Intelligence Conference in Shanghai. It integrates 384 Ascend 910C AI chips and 192 Kunpeng CPUs, interconnected via an ultra-high-bandwidth Unified Bus (UB) network, enabling direct all-to-all communication and dynamic resource pooling. The system employs a supernode architecture with optical links for low-latency, high-bandwidth communication, allowing large-scale AI workloads to run efficiently.

Huawei emphasizes a super-organized setup that reduces failure risks during large-scale training.

© 2026 KENHARDT OPTICAL NETWORKS - All rights reserved

Performance

The Cloud

Matrix 384 delivers 300 PFLOPs of dense BF16 computing, roughly double the compute performance of Nvidia's GB200 NVL72 system. Despite using more energy—559 kW per hour, nearly four times Nvidia's comparable system—the cluster achieves superior throughput and memory bandwidth. For large language model (LLM) workloads, Huawei's CloudMatrix-Infer solution demonstrates 6,688 tokens/s prefill throughput per NPU and 1,943 tokens/s decode throughput per NPU, maintaining low latency (

Article Content

Huawei Unveiled the Latest SuperPoD, Making an AI

For AI computing, the Atlas 950 SuperPoD, powered by UnifiedBus, integrates 64 NPUs per cabinet and can scale up to 8,192 NPUs, delivering

Huawei Cloud announces agentic AI products

At this event, Huawei Cloud officially introduced Agentic Infra, a new paradigm featuring an efficient token factory, continuous learning, unified general & AI compute scheduling, and secure

Ascend AI Cloud Service | Huawei Cloud

Ascend AI Cloud Service offers instant access to immense yet cost-effective AI computing power, a reliable platform for training and running models and

Huawei launches CloudMatrix 384 AI System to rival

The Huawei CloudMatrix 384 is a high-density AI computing system featuring 384 Huawei Ascend 910C chips, designed to rival Nvidia''s GB200

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

Nvidia acknowledges Huawei challenge after rival

BEIJING — Chinese telecommunications giant Huawei announced Thursday new computing systems for powering artificial intelligence with its in

Huawei Unveils Next-Gen Atlas 950 & 960 SuperPoD AI

© 2026 KENHARDT OPTICAL NETWORKS - All rights reserved

Huawei claims that the Atlas 950 SuperPoD will be the world''s largest AI cluster, in terms of the chips onboard, since systems with 500K to 1

Huawei releases CloudMatrix 384 supernode: a milestone ...

In the future, CloudMatrix super nodes can build large clusters of more than 10,000 slices to provide computing power, which will greatly lower the threshold for large-scale commercial use of AI.

Huawei unveils Atlas 950 SuperCluster — promises 1 ZettaFLOPS

Huawei has unveiled its next-generation data-center scale AI solution that can offer 1 FP4 ZettaFLOPS performance for AI inference and 524 FP8 ExaFLOPS for AI training at its Huawei...

Huawei unveils new high-performance AI supercomputing clusters

Huawei Technologies Co on Thursday announced the launch of what it calls the world''s most powerful AI computing clusters—SuperNode and SuperCluster—marking its latest

Huawei announces computing strategy and releases

At HUAWEI CONNECT 2019, Huawei announced its strategy for the computing market and released Atlas 900, the world''s fastest AI training cluster.

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 chip compares to Nvidia and reshapes China''s 0B market.

Huawei Launches Enhanced AI-Centric Network

The company also launched enhanced AI-Centric Network solutions that will help carriers prepare for the agentic era by enabling intelligent services,

Huawei''s new computing cluster, world''s first AI agent phone to debut ...

Huawei initially unveiled the computing cluster, powered by 8,192 Ascend neural processing unit (NPU) cards, last year. The product is optimised for AI neural networks and machine learning tasks ...

China''s Huawei aims to outpace global leaders with

Huawei Technologies has announced plans to launch what it says will be the world''s most powerful AI computing clusters over the next two years.

Key products in Huawei''s AI chips and computing power roadmap

© 2026 KENHARDT OPTICAL NETWORKS - All rights reserved

China''s Huawei (HWT.UL) has ended years of secrecy to detail its product roadmap for chips and computing power systems, discussing for the first time how it intends to challenge global

Huawei Unveils AI Agent Phone, Computing Cluster at WAIC

Huawei Technologies is set to unveil its next-generation computing cluster and what has been described as the world''s first artificial intelligence agent phone. This will occur at the upcoming World Artificial

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

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://kenhardtprimary.co.za
Technical inquiry: https://kenhardtprimary.co.za/project-desk.html
African project coordination: Century City, Cape Town, Western Cape, South Africa

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

© 2026 KENHARDT OPTICAL NETWORKS - All rights reserved