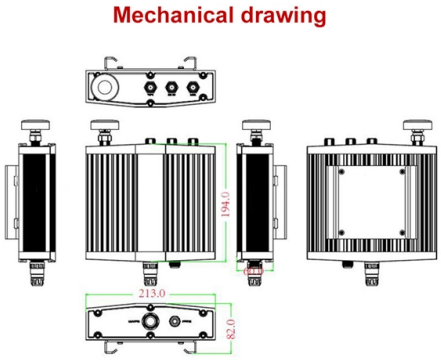
What are the Huawei AI server series



Overview

AI-accelerated ECSs, powered by Ascend processors and software stacks, are dedicated to accelerating AI applications.

Key Takeaways

Huawei's leading AI server models include the Atlas 950 and 960 SuperPoD servers, powered by Ascend 950/950PR AI chips, designed for ultra-large-scale AI workloads and LLM inference. Atlas 950 and 960 SuperPoD ServersHuawei's Atlas 950 and 960 SuperPoD servers are next-generation rack-scale AI clusters aimed at competing with NVIDIA's high-end AI systems. The Atlas 950 SuperPoD features 8,192 Ascend 950 AI chips, delivering 8 EFLOPS FP8 and 16 EFLOPS FP16 performance, with a total interconnect bandwidth of 16.

3 PB/s. These servers can be combined into the Atlas 950 SuperCluster, scaling up to 524,288 Ascend 950 chips, supporting between 500,000 and 1 million processors, making them among the largest AI compute clusters globally. The Atlas 960 is a further iteration designed for extreme scalability and high-performance AI workloads.

CloudMatrix384 ArchitectureThe CloudMatrix384 is Huawei's next-generation AI datacenter architecture, integrating 384 Ascend 910 NPUs and 192 Kunpeng CPUs with an ultra-high-bandwidth Unified Bus (UB) network. This architecture supports all-to-all communication, dynamic resource pooling, and optimized performance for large-scale mixture-of-experts (MoE) models. CloudMatrix384 also incorporates CloudMatrix-Infer, a solution for LLM serving with peer-to-peer scaling, expert parallelism, and INT8 quantization for efficient inference.

© 2026 CJC FIBRE CONSULTANTS - All rights reserved

Ascend A

I ChipsHuawei's AI servers are powered by the Ascend 950 and 950PR chips. The Ascend 950PR delivers 1. 56 PFLOPS FP4 performance, with 112 GB HiBL 1.

0 memory and 1. 4 TB/s bandwidth, optimized for large language model inference. These chips use a monolithic die design, achieving high compute density and efficiency, and are compatible with CUDA-like software stacks to facilitate migration from other platforms.

The Ascend 950 series provides the backbone for Huawei's SuperPoD and CloudMatrix servers, enabling multi-petaflop AI training and inference.

AI Ecosystem and Software Integration

Huawei A

I servers integrate with ModelEngine and Ascend AI Cloud services, offering zero-code model deployment, fine-grained resource scheduling, and high utilization of xPU resources. The ecosystem supports trillion-parameter model training, multi-modal data processing, and knowledge memory platforms for ultra-scale inference, ensuring high throughput and low latency for enterprise AI applications. SummaryHuawei's AI server lineup, including Atlas 950/960 SuperPoD, CloudMatrix384, and Ascend 950/950PR chips, is designed for large-scale AI training and inference, particularly for LLMs and multi-petaflop workloads.

These servers combine massive compute power, high-bandwidth interconnects, and advanced software integration, positioning Huawei as a major competitor in the global AI infrastructure market.

Article Content

Huawei AI Stack vs. NVIDIA DGX: MWC 2026 Deep Dive | Aixia

Looking further ahead, Huawei has publicly committed to a three-series roadmap: the Ascend 950 series (including the 950PR for inference and 950DT for training), the Ascend 960

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 ground, as their latest announcement includes unveiling

Huawei Unveiled the Latest SuperPoD, Making an AI Infrastructure

Huawei''s latest computing portfolio, including Atlas 950 SuperPoD, Atlas 850E, TaiShan 950 SuperPoD, and TaiShan 500, TaiShan 200 series, addresses diverse carrier compute demands,

© 2026 CJC FIBRE CONSULTANTS - All rights reserved

Huawei Ascend 950PR: The 1.56 PFLOP AI Chip vs Nvidia

Huawei will capture 35-40% of China''s AI chip market by revenue in 2027. The combination of the 950PR for inference and the 950DT for training, backed by CANN Next''s

Huawei outlines roadmap for Ascend AI chips

SHANGHAI — Chinese vendor Huawei unveiled its roadmap for the next generation of Ascend AI chips during the recent Huawei Connect event in Shanghai, China, outlining plans for

AI-accelerated ECSs_x86 ECS Specifications_ECS Types and ...

AI-accelerated ECSs, powered by Ascend processors and software stacks, are dedicated to accelerating AI applications. For details about the OSs supported by this type of ECSs, see OSs

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 almost 4x more, which shows that Huawei has developed the cluster to

Inside Huawei''s plan to make thousands of AI chips think like one

Imagine connecting thousands of powerful AI chips scattered in dozens of server cabinets and making them work together as if they were a single, massive computer.

Groundbreaking SuperPoD Interconnect: Leading a New Paradigm

With the Atlas 950 SuperPoD and Atlas 960 SuperPoD, we are confident in our ability to provide abundant computing power for rapid AI advancements, both today and tomorrow.

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 unit. That compares with just 72 B200 GPUs found in Nvidia''s GB200

DeepSeek launches V4 model adapted for Huawei AI

DeepSeek launched a V4 preview adapted for Huawei AI chips, with Reuters reporting Huawei chips were used in part of V4-Flash training.

Huawei Solar Inverter Complete Guide 2025: Models,

Comprehensive Huawei solar inverter guide covering SUN2000 series, performance data, pricing, installation tips, and expert reviews. Updated

© 2026 CJC FIBRE CONSULTANTS - All rights reserved

Alibaba''s AI Agent Chip Strategy Goes Beyond Nvidia

Alibaba''s Zhenwu M890 is built for AI agents, not just faster inference. With a chip roadmap to 2028 and Qwen 3.7-Max, this is a platform play.

Building AI-Optical Network to Drive New Growth in the AI Era

AI-ON Empowers Operators to Build AI-Centric All-Optical Target Networks, Driving Shared Growth in the AI Era To help operators accelerate transition into the era of Token

Huawei Plans Korea Launch for Ascend AI Chips

ETNews reported on July 2, 2026 that Huawei Korea is preparing a fourth-quarter launch for Ascend AI accelerator chips and the Atlas 950 SuperPod computing system in South Korea. The

Gartner Business Insights, Strategies & Trends For

Business and Technology Insights and Trends Unlock ROI from AI Build the bridge between AI hype, productivity gains and real business value.

Huawei''s Ascend and Kunpeng progress shows how China is

Huawei used its New Year message to highlight progress across its Ascend AI and Kunpeng CPU ecosystems, pointing to the rollout of Atlas 900 supernodes and rapid growth in

Welcome to Channel Dive | Channel Dive

Welcome to Channel Dive. We''re Informa TechTarget''s new publication, focused on delivering daily news and analysis for executives at North

© 2026 CJC FIBRE CONSULTANTS - All rights reserved

Contact Us

Continue your research online:

Website: https://cjcconsultants.co.za
Technical inquiry: https://cjcconsultants.co.za/contact.html

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

© 2026 CJC FIBRE CONSULTANTS - All rights reserved