

Huawei AI Server Series



Summary of Content

Huawei's AI server series includes the Atlas and TaiShan lines, powered by Ascend NPUs, designed for large-scale AI training and inference with SuperPoD architecture.

Overview of Huawei AI Servers

Huawei's AI servers are built around Ascend NPUs (Neural Processing Units), which are purpose-designed for AI workloads rather than adapted from graphics processing. The main server series include:

- Atlas 950 SuperPoD: Integrates up to 8,192 Ascend 950DT NPUs per SuperPoD, delivering up to 16 EFLOPS FP16 and 8 EFLOPS FP8 performance. It uses Huawei's UnifiedBus interconnect to operate thousands of NPUs as a single logical machine, enabling large-scale AI training and high-concurrency inference .
- Atlas 960 SuperPoD: A next-generation cluster designed to scale even further, supporting up to 500,000-1 million Ascend NPUs in a SuperCluster configuration, targeting the largest AI compute workloads globally .
- TaiShan 950 SuperPoD: Focused on general-purpose AI computing with low-latency memory pooling and high bandwidth, suitable for both training and inference in carrier and enterprise environments .
- TaiShan 500 and TaiShan 200 series: Cover medium and low computing power requirements, providing flexible deployment options for diverse AI workloads .
- Atlas 850E: Supports air-cooled data centers, scaling from 8 to 1,024 NPUs, ideal for smaller-scale inference clusters .

Ascend NPU Roadmap

Article Content

GPU Accelerated Cloud Server_ECS_HUAWEI CLOUD

GPU-accelerated cloud server (GACS) provides outstanding floating-point computing power that is great for real-time, highly

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 Launched Intelligent Servers at HUAWEI CONNECT 2018

[Shanghai, China, October 19, 2018] Huawei launched its Intelligent Server series at HUAWEI CONNECT 2018. The servers offer

Server Huawei: AI & Cloud Trends 2026

AI Training and Inference Servers (Ascend Series) Snapshot: High-performance computing nodes equipped with

Atlas 800 AI Server-Ascend Community

The Atlas 800 inference server (Model 3000) supports up to eight Atlas 300I/V Pro cards, delivering powerful real-time inference and

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 Ascend 950PR: The 1.56 PFLOP AI Chip vs Nvidia

Huawei has officially launched the Ascend 950PR, a 1.56-petaflop AI inference chip that delivers 2.8 times the FP4

AI-accelerated ECSs_x86 ECS Specifications_ECS Types and ...

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

Overview

It features high computing density, high energy efficiency ratio, high network bandwidth, easy expansion, and easy management. It is

Huawei wields growing influence in China's server industry with AI

Huawei Technologies, which was forced to divest its server subsidiary three years ago due to US tech sanctions, is

Huawei Releases FusionServer V5, its Latest Generation Server Series ...

Huawei has announced the global release of its latest-generation FusionServer series servers the FusionServer V5.

OceanStor AI Storage

In response to the demands of large AI model training and inference, Huawei redefines distributed file storage, focusing on six key

Huawei outlines roadmap for Ascend AI chips

SHANGHAI — Chinese vendor Huawei unveiled its roadmap for the next generation of Groundbreaking SuperPoD Interconnect: Leading a New Paradigm ...

This brings me to the topic of today's keynote: Groundbreaking SuperPoD Interconnect: Leading a New Paradigm for

All Products and Solutions

Directory of Huawei enterprise IT infrastructure products, solutions, and services.

Huawei Deutschland

Huawei ist ein führender globaler Anbieter von Informations- und Kommunikationstechnologie (ICT) -

Huawei's AI servers show tech advancing despite US curbs

TOKYO -- Huawei Technologies is steadily building up its own artificial intelligence infrastructure with homegrown chips and servers,

Ascend AI Cloud Service | Huawei Cloud

Ascend-native, cloud-based AI agents, such as search agents and big data agents, allow users to integrate the capabilities of AI

Huawei Enterprise Servers

Huawei RH Series Rack Servers Check out Huawei's next-generation FusionServer V3 and versatile Tecal V2 servers to meet

Overview

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

Huawei Cloud Announces Pangu Models 5.5 and All-new AI Cloud Service ...

Huawei Developer Conference 2025 (HDC 2025) kicked off today at Dongguan Basketball Center. Zhang Ping'an,

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

FusionServer Pro V6 Smart Server | Huawei Enterprise

Now, Huawei takes a leading role in the virtual event revolution, focusing on and promoting co-operation between

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

Huawei launches CloudMatrix 384 server as an alternative to Nvidia's AI infrastructure stack - SiliconANGLE

Contact Us

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

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

Email: sales@benniedesignstudio.co.za

Phone: +27 82 391 4765

Address: The Towers, 1 St Georges Mall, Cape Town, 8001, South Africa

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

