

Server AI Chips



AI Overview

AI server chips are specialized processors, including GPUs, AI ASICs, and AI-optimized CPUs, designed to accelerate machine learning and deep learning workloads in data centers and cloud infrastructure.

Key Types of AI Server Chips

Graphics Processing Units (GPUs): GPUs remain the dominant hardware for AI workloads, particularly for training large language models and generative AI. Leading examples include NVIDIA's Hopper (H100/H200) and Blackwell (B200/B300) chips, which feature high-bandwidth memory exceeding 250GB and advanced packaging technologies like CoWoS-L and multi-die chiplets for exascale performance. AMD also competes with its MI300 series and upcoming MI400 series accelerators, optimized for inference-heavy AI workloads and large-scale data center deployments.

AI Application-Specific Integrated Circuits (ASICs): Custom AI ASICs are developed by hyperscalers and cloud providers to optimize performance for specific AI tasks. These chips are designed for efficiency, lower latency, and domain-specific computation, often outperforming general-purpose GPUs in targeted workloads.

AI-Capable CPUs: Modern server CPUs, such as AMD's Zen 5-based EPYC processors, are increasingly optimized for AI workloads. They can operate standalone for general-purpose tasks or in tandem with AI accelerators to handle large models and complex computations.

Field Programmable Gate Arrays (FPGAs): FPGAs are used for parallel processing of large datasets, offering flexibility for AI workloads that require custom data paths or low-latency inference.

Server Architecture and Connectivity AI servers integrate multiple accelerators and require high-speed interconnects like PCI Express (PCIe) 6.0 or the upcoming PCIe 7.0, which can deliver up to 512GB/s bandwidth over 16 lanes. These interconnects ensure efficient communication between CPUs, GPUs, memory, and storage, which is critical for large-scale AI training and inference.

Power management and server management chips, such as PMICs and Baseboard Manage...

Article Content

The State Of AI Infrastructure: Demand, Costs, And

As AI-driven compute demands grow, we expect custom silicon to grow as a share of compute spend, as the time and money required to design

AI Server Market Update: Vendors Shift from Silicon to

AI Server Market Update: Vendors Shift from Silicon to Services Dell, HPE, Lenovo, and Supermicro are riding record AI server demand, but winning

New Report Finds Semiconductors Account for 95% of an AI Data

The study also details how AI requires the full range of semiconductor technologies, including logic, memory, and analog and foundational chips. Additionally, the report projects annual revenue for

nvidia chips: Is Thailand-based OBON Corp really "Company-1", and ...

Is Thailand-based OBON Corp really "Company-1", and were Nvidia chips sent to Alibaba? A new report says the US suspects AI servers with Nvidia chips were routed through Southeast Asia

Malaysian Customs seizes AI chips worth \$16.7m at Kuala Lumpur

Malaysia customs seized AI chips worth RM52.9 million (S\$16.71 million) at Kuala Lumpur International Airport in June. The chips, falsely declared as "computer components," were to

What is an AI Server? AI Server Architecture Explained

Learn what AI servers are and how they power artificial intelligence. Complete guide to AI server components, architecture, and requirements for ML and AI.

Inside the NVIDIA Vera Rubin Platform: Six New Chips,

The Vera Rubin platform is built from six new chips, each engineered for a specific role in the AI factory and designed from the outset to operate as

Why liquid cooling will dominate AI data centres in 2026

As AI power demands surge into 2026, liquid cooling is becoming the essential technology keeping data centres efficient, stable, and future-ready.

New Report Finds Semiconductors Account for 95% of an AI Data Server ...

New Report Finds Semiconductors Account for 95% of an AI Data Server Rack's Value, Encompassing the Full Stack of Chip Technologies Monday, Jun 01, 2026, 5:00am by Semiconductor Industry

Samsung's profit triples, beating estimates as AI chip demand ...

Samsung posts record quarterly profit as AI-driven memory demand tightens supply. Results beat estimates and top Samsung's own guidance for the quarter. Samsung's memory

BIS Revises Export Review Policy for Advanced AI

Controls on Remote Access and IaaS If the chips are destined for cloud service providers, data centers, or other scenarios where AI compute

Amazon releases an impressive new AI chip and teases an

AWS has been building its own AI chips — and systems — for years now. It just released its third version, known as Trainium3, with some impressive specs.

Apple to Start Production of AI Server Chips in 2026,

Apple will produce AI server chips by 2026 and build new data centers by 2027, boosting domestic production and creating 20,000 jobs.

Apple plans to mass-produce its first AI server chips in

In a nutshell: Apple is preparing to shift more of its AI operations in-house, with plans to begin mass production of its first AI server chips in the

Advanced AI chips worth \$13 Million seized by Malaysia customs

Malaysia customs seized \$13 million worth of advanced AI chips at Kuala Lumpur airport, foiling an alleged smuggling attempt that violated export control regulations.

Apple AI Server: Why Apple's In-House AI Chips Mark a

Apple AI Server: Why Apple's In-House AI Chips Mark a New Silicon Era Apple's move toward proprietary AI server processors signals a major

Nvidia and D-Matrix Announce Joint AI Chip System

Nvidia and AI server chip startup d-Matrix are combining their respective hardware in a new system to power AI models, The Information reported. The move is part of Nvidia's strategy to

Arm Launches First Server CPU, Expanding Its Push

The move puts Arm deeper into the AI data center market and raises new questions about competition, customers and its relationships with existing

AMD's AI Chips Are Coming for Your Laptop, Desktop,

AMD unveiled Ryzen AI 400 and Embedded processors at CES 2026, touting up to 60 TOPS NPUs and new chips for laptops, desktops, and edge

Malaysian customs foil US\$13 million AI chip smuggling bid

Malaysia's customs department officers show a pallet of server units containing advanced AI chips during a press conference on the department's seizure of 72 server units containing

TSMC CoWoS packaging service

TSMC CoWoS packaging service is designed for high-bandwidth AI accelerators with multi-chip modules and advanced 2.5D integration.

AMD shares hit record high, spark global chips rally on

Advanced Micro Devices' shares jumped to an all-time high on Wednesday as its strong outlook boosted investor confidence about sustained

As AI Eats Up The World's Chips, Memory Prices Take

Memory prices are high because manufacturers have shifted factories toward lucrative AI and server chips, creating an artificial squeeze on everyday

AMD unveils AI server as OpenAI taps its newest chips

Advanced Micro Devices CEO Lisa Su on Thursday unveiled a new artificial intelligence server for 2026 that aims to challenge Nvidia's flagship

Huawei's Ascend AI chip ecosystem scales up as China

Huawei's AI server hardware, domestic packaging suppliers, and EDA partners now form the backbone of a China-first chip supply chain.

Alibaba Unveils New AI Chip, Flagship Model, and

Alibaba launched its most aggressive AI push yet, unveiling a new flagship large language model, a homegrown AI chip that triples the performance

Contact Us

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

Website: <https://www.sky-connect.co.za>

Email: info@sky-connect.co.za

Phone: +44 20 7946 0958

Address: 1 Cornhill, London EC3V 3ND, United Kingdom

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

