Which is the world s first AI inference server

AI Overview

Lumai Iris Nova is recognized as the world's first optical computing AI inference server capable of running billion-parameter LLMs in real time.Lumai Iris Nova: Optical AI InferenceThe Lumai Iris Nova, launched in April 2026, is the first commercially recognized optical computing system designed for real-time inference of large language models (LLMs) with billions of parameters . Unlike traditional silicon-based servers, Iris Nova uses light (photons) instead of electrons to perform core mathematical operations, enabling massive spatial parallelism and millions of simultaneous computations . This optical approach allows the server to achieve up to 90% lower energy consumption compared to conventional GPU-based systems while maintaining high throughput and execution efficiency . The Iris Nova server employs a hybrid processor architecture, combining digital processing for system control with an optical tensor engine for computation. It has successfully run Llama 8B and 70B models in real time, demonstrating its ability to integrate seamlessly into existing data center workflows . The Iris family also includes Aura and Tetra, which are designed to extend performance and efficiency for hyperscale and enterprise deployments .Zoho Nathu La: Indigenous AI ServerIn parallel, Zoho launched Nathu La, India's first indigenously designed AI server, aimed at enterprise-grade LLM workloads . Developed over five years by Zoho's team in Nagpur, Nathu La features Intel Xeon 6 processors and custom-designed motherboards, chassis, BIOS, and network interface cards. It is optimized for virtualization, high-performance computing, storage, and AI inference, achieving 12–18% lower power consumption and 20–30% reduced total cost of ownership compared to conventional servers . Nathu La is currently deployed internally within Zoho's data centers and is not yet c...

Article Content

Local AI Inference Mini PC Now Runs 235B Models: AMD Ryzen AI

Local AI inference mini PC reached a milestone in June 2026: AMD Ryzen AI Max+ 395 runs 235B-parameter models on x86, letting developers cut $440-per-month cloud subscriptions.

Fortune Tech: OpenAI-Broadcom AI chips, Qualcomm-Modular deal,

The first chip is planned for deployment by the end of this year in custom servers made by Canada''s Celestica. Like its peers, this first Jalapeño chip marks what will be a multigenerational ...

Vista Equity Partners and Cambium Launch Vector

COMPUTEX — Vista Equity Partners and Cambium Capital today launched Vector Core Compute (VC2), the world''s first commercially-available

NVIDIA Triton Inference Server

Originally called TensorRT Inference Server, the project was renamed to Triton Inference Server in 2020 to better reflect its multi-framework support beyond TensorRT alone.

Nvidia Groq 3 LPU: Speeding AI Inference Tasks

At the 2026 Nvidia GTC conference, Jensen Huang announced an inference-specific chip, the Groq 3 LPU. The LPU will work in concert with the

Explained: Generative AI''s environmental impact

MIT News explores the environmental and sustainability implications of generative AI technologies and applications.

MAXER-5100: The World''s First 8L Dual-GPU AI Inference Server

AAEON''s MAXER-5100 is the world''s smallest industrial-grade AI inference server, uniquely equipped with 14th Gen Intel® Core™ processing, two integrated NVIDIA RTX™ 2000 Ada GPUs, and a

DeepSeek Is Developing Its Own AI Chip to Challenge Nvidia

Chinese artificial intelligence company DeepSeek is reportedly developing its own AI chip as it seeks to reduce its dependence on hardware supplied by Nvidia and Huawei.

Lumai Launches the World''s First Optical Computing System for Real

OXFORD, UK, April 28, 2026 – Lumai, the optical compute company addressing scalable AI, today announced its Lumai Iris inference server – the world''s first optical computing system to successfully

The Data Center Moves to Your Machine

Hybrid AI has been an industry ambition for a long time. Personal Computer with local inference, coming in July, is the first product that makes it

Explore AI Inference Platform | NVIDIA

NVIDIA Triton Inference Server is an open-source inference serving software that helps enterprises consolidate bespoke AI model serving infrastructure, shorten the time needed to deploy new AI

AI Agent Executes End-to-End Ransomware Attack

Sysdig''s Threat Research Team documented what it says is the first fully AI-agent-driven ransomware operation, an intruder it named **JADEPUFFER**, in a report published July 1, 2026.

Changes with Storage on Cloud and Storage Classification in VMM

First published on TECHNET on May 19, 2014 Storage Classification was introduced in System Center 2012 Virtual Machine Manager (VMM 2012) to provide the...

NVIDIA Kicks Off the Next Generation of AI With Rubin

NVIDIA today kickstarted the next generation of AI with the launch of the NVIDIA Rubin platform, comprising six new chips designed to deliver one

MAXER-5100 | AI Inference Server with Intel® Core™ CPU & Dual

AAEON''s MAXER-5100 is the world''s smallest industrial-grade AI inference server, uniquely equipped with 14th Gen Intel® Core™ processing, two integrated NVIDIA RTX™ 2000 Ada GPUs, and a

The $20 Billion Bet On Inference: What Every AI Infrastructure

Inference is now scaling rapidly and becoming the dominant cost for AI companies. Every ChatGPT query, every AI agent action, every generated video is based on inference.

Global Top Five Enterprise SSD Vendors Post Over

Meanwhile, enterprises accelerated upgrades to their general-purpose servers, while shortages in HDD supply pushed some demand toward

AAEON Unveils World''s First 8L Dual-GPU AI Inference

Leading provider of advanced AI solutions AAEON has released a new addition to its AI Inference Server product line, the MAXER-5100 - the

Better Artificial Intelligence (AI) Inference Stock: AMD vs. Intel

The growth of AI inference workloads in data centers is boosting demand for server CPUs, a market that''s dominated by AMD and Intel.

d-Matrix

d-Matrix is making Generative AI inference blazing fast, sustainable and commercially viable with the world''s first efficient memory-compute integration.

Qualcomm announces AI chips to compete with AMD

Qualcomm announced that it will release new AI accelerator chips. Nvidia has dominated the market for AI chips, with AMD seen as the second

NVIDIA BlueField-4 Powers New Class of AI-Native

NVIDIA today announced that the NVIDIA BlueField®-4 data processor, part of the full-stack NVIDIA BlueField platform, powers NVIDIA

AAEON Launches MAXER-5100: World''s First 8L Dual-GPU AI

AAEON unveils the MAXER-5100, the world''s first 8L AI inference server with dual NVIDIA RTX 2000 Ada GPUs, Intel i9 processor, and edge device security features.

Rename the current session (alias ''/title'') | | ''grok sessions list ...

'' | Rename the current session (alias ''/title'') | | ''grok sessions list'' | List recent sessions for this directory | | ''grok sessions search <query>'' | Search session titles and prompts | | ''grok sessions delete <id>''

We did the math on AI''s energy footprint. Here''s the

The emissions from individual AI text, image, and video queries seem small—until you add up what the industry isn''t tracking and consider where

SK hynix Announces 1Q26 Financial Results

SK hynix noted that despite the fact that first quarter is typically a seasonal downturn, strong demand persisted due to expanded investments in AI infrastructure. The company sustained

Contact Us

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

Website: https://shangases.co.za
Email: ekomedsolar@gmail.com
Phone: +86 13816583346
Address: No. 26 Heshun Middle Road, Economic Development Zone, Hai'an City, Jiangsu Province, China

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