GPU high-computing core switch

Overview

The NVLink Switch System, developed by NVIDIA, stands as a backbone of modern high-performance computing (HPC) and artificial intelligence (AI) infrastructure, revolutionizing how data is transferred across complex computational environments. Performance and efficiency at scale: How Vera Rubin converts architecture into real gains at scale, including one-fourth as many GPUs to train, 10x higher inference throughput, and 10x lower cost per token. Why Vera Rubin is the AI factory platform: How extreme co-design delivers predictable. The Rubin platform harnesses extreme codesign across hardware and software to deliver up to 10x reduction in inference token cost and 4x reduction in number of GPUs to train MoE models, compared with the NVIDIA Blackwell platform. NVIDIA Spectrum-X Ethernet Photonics switch systems deliver 5x. Through the collaborative work of multiple chips, Vera Rubin has built an AI supercomputing system that is comprehensively optimized in terms of performance, cost efficiency, and system-level security. Its core objective directly addresses the key challenge of the inference-dominated. The NVIDIA H100 Tensor Core GPU delivers unprecedented performance, scalability, and security for every workload. With NVIDIA® NVLink® Switch System, up to 256 H100 GPUs can be connected to accelerate exascale workloads, while the dedicated Transformer Engine supports trillionparameter language. VR200 NVL72 delivers 3. 3x inference performance over Blackwell Ultra GB300 NVL72, with HBM4 running past 3. For high-performance computing (HPC) applications, H100 triples the floating-point operations.

Article Content

Understanding Parallel Computing: GPUs vs CPUs Explained

We also use CPUs for general computing that would be almost too simple for GPUs. In certain scenarios, executing

NVIDIA NVLink and NVIDIA NVSwitch Supercharge Large Language

NVLink and NVSwitch provide high bandwidth communication between GPUs based on the NVIDIA Hopper

AMD patent explores "smart switch" solution to solve the ...

This switch can optimize memory access within nanosecond decision delays. After solving the data access problem,

NVIDIA Rubin GPU: 336B Transistors, T Orders

It is a complete compute architecture consisting of the Rubin GPU, the Vera CPU, a next

Infrastructure for Scalable AI Reasoning | NVIDIA Vera Rubin Platform

The NVIDIA HGX™ Rubin NVL8 integrates eight NVIDIA Rubin GPUs with sixth-generation high-speed NVLink interconnects to

Nvidia''s Vera Rubin platform in depth

Nvidia uses high-speed, low-latency NVLink fabric for linking CPUs (NVLink-C2C) and GPUs directly, while NVSwitch

Upgrading Multi-GPU Interconnectivity with the Third-Generation

This latest NVSwitch and the H100 Tensor Core GPU use the fourth-generation NVLink, the newest high-speed, point

GPU Architecture Explained: Key Components,

Discover the fundamentals of GPU architecture, from core components like CUDA cores,

Hopper GPU Architecture | NVIDIA

Learn about the next massive leap in accelerated computing with the NVIDIA Hopper™ architecture. Hopper securely scales diverse

What is a Tensor Core? The Nvidia GPU technology

What to read next As Nvidia transitioned from primarily a computing/hardware company to

NVIDIA Vera Rubin: AI Supercomputer with Seven New

The Rubin GPU is the core computing component of the Vera Rubin architecture; it is

What Do GPU Cores Mean and How Do They Work?

GPU cores handle thousands of tasks at once, but raw core count doesn''t tell the whole story. Here''s what they actually do and when

CUDA GPU Compute Capability | NVIDIA Developer

CUDA GPU Compute Capability Compute capability (CC) defines the hardware features and supported instructions for each NVIDIA

NVIDIA H100 Tensor Core GPU Architecture

The Hopper H100 Tensor Core GPU will power the NVIDIA Grace Hopper Superchip CPU+GPU architecture, purpose-built for

NVIDIA A100 Tensor Core GPU Architecture

NVIDIA meets these growing GPU computing challenges with the new NVIDIA A100 Tensor Core GPU based on the NVIDIA

Inside the NVIDIA Vera Rubin Platform: Six New Chips,

By combining Olympus CPU cores, second-generation SCF, high-bandwidth LPDDR5X

How to Switch Between Displays (dGPU, GPU, Intel and Nvidia)

When users face the problem of switching between displays (dGPU, GPU, Intel, and NVIDIA), they usually find that

Supermicro GPU Servers

PCIE GPU Access Choices Supermicro GPU servers are designed for applications requiring several GPUs within the server.

NVIDIA Kicks Off the Next Generation of AI With Rubin

NVIDIA will also offer the NVIDIA HGX Rubin NVL8 platform, a server board that links eight

NVIDIA Ampere Architecture

With MIG, each GPU can be partitioned into multiple GPU instances, fully isolated and secured at the

NVIDIA Turing Architecture In-Depth | NVIDIA Technical Blog

In addition to rendering highly realistic and immersive 3D games, NVIDIA GPUs also accelerate content creation

The state of GPUs is about to drastically change

The RTX 5080 will apparently have less than half of the cores of the RTX 5090, placing an

NVIDIA Grace Hopper Superchip Architecture In-Depth

The NVIDIA Grace Hopper Superchip architecture brings together the groundbreaking performance of the NVIDIA

NVIDIA NVLink Switch System: Powering the Future of

The NVLink Switch System extends this capability, connecting multiple NVLink-enabled

NVIDIA Vera Rubin Opens Agentic AI Frontier

NVIDIA today announced the NVIDIA Vera Rubin platform is opening the next frontier of

NVIDIA H100 Tensor Core GPU Datasheet

Built with 80 billion transistors using a cutting-edge TSMC 4N process custom tailored for NVIDIA''s accelerated compute needs,

Introducing NVIDIA HGX A100: The Most Powerful Accelerated Server ...

The HGX A100 8-GPU baseboard represents the key building block of the HGX A100 server platform. Figure 1 shows

NVIDIA | Datasheet

With second-generation Multi-Instance GPU (MIG), built-in NVIDIA confidential computing, and NVIDIA NVLink Switch System, H100

General-purpose computing on graphics processing units

General-purpose computing on graphics processing units (GPGPU, or less often GPGP) is the use of a graphics processing unit

NVIDIA Kicks Off the Next Generation of AI With Rubin

With built-in, in-network compute to speed collective operations, as well as new features for

Contact Us

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

Website: https://e-motional.eu
Email: info@e-motional.eu
Phone: +49-30-657-8924
Address: Friedrichstraße 101, 10117 Berlin, Germany

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