HealthManagement
Promoting Management and Leadership

NVIDIA Unveils Blackwell Platform, Towards Next-Gen Computing

NVIDIA has introduced during its latest keynote the Blackwell platform, which enables organisations to utilise real-time generative AI on large language models with significantly reduced cost and energy consumption compared to its predecessor. The Blackwell GPU architecture incorporates six key technologies for accelerated computing, promising breakthroughs in various fields including data processing, engineering simulation, electronic design automation, drug design, quantum computing, and generative AI, thereby tapping into emerging industry opportunities.

Paving the way towards new era of AI

“For three decades we’ve pursued accelerated computing, with the goal of enabling transformative breakthroughs like deep learning and AI,” said Jensen Huang, founder and CEO of NVIDIA. “Generative AI is the defining technology of our time. Blackwell is the engine to power this new industrial revolution. Working with the most dynamic companies in the world, we will realise the promise of AI for every industry.” Among the many organisations expected to adopt Blackwell are Amazon Web Services, Dell Technologies, Google, Meta, Microsoft, OpenAI, Oracle, Tesla and xAI.

Blackwell Innovations to Fuel Accelerated Computing and Generative AI

Blackwell comes with six game-changing technologies, which together enable AI training and real-time LLM inference for models scaling up to 10 trillion parameters:

- World’s Most Powerful Chip: Blackwell-architecture GPUs boast 208 billion transistors and utilize a custom-built 4NP TSMC process with two-reticle limit GPU dies connected by a 10 TB/second chip-to-chip link into a unified GPU.
- Second-Generation Transformer Engine: Equipped with micro-tensor scaling support and advanced dynamic range management algorithms, Blackwell doubles compute and model sizes and introduces 4-bit floating point AI inference capabilities.
- Fifth-Generation NVLink: With a groundbreaking bidirectional throughput of 1.8TB/s per GPU, the latest NVLink iteration accelerates performance for multitrillion-parameter and mixture-of-experts AI models, facilitating high-speed communication among up to 576 GPUs.
- RAS Engine: Blackwell-powered GPUs feature a dedicated engine for reliability, availability, and serviceability, incorporating AI-based preventative maintenance for diagnostics and reliability forecasting, ensuring uninterrupted operation for massive-scale AI deployments and reducing operating costs.
- Secure AI: Advanced confidential computing capabilities safeguard AI models and customer data while maintaining performance, with support for new native interface encryption protocols, crucial for privacy-sensitive industries like healthcare and financial services.
- Decompression Engine: A dedicated decompression engine supports the latest formats, accelerating database queries to deliver high performance in data analytics and data science, driving the GPU-accelerated transformation of data processing, a field with increasing investment.

A Massive Superchip

The NVIDIA GB200 Grace Blackwell Superchip connects two NVIDIA B200 Tensor Core GPUs to the NVIDIA Grace CPU via a 900GB/s ultra-low-power NVLink chip-to-chip interconnect. For optimal AI performance, GB200-powered systems can integrate with the NVIDIA Quantum-X800 InfiniBand and Spectrum-X800 Ethernet platforms, providing advanced networking at speeds up to 800Gb/s. The GB200 is a crucial component of the NVIDIA GB200 NVL72, a multi-node, liquid-cooled, rack-scale system designed for compute-intensive workloads. It combines 36 Grace Blackwell Superchips, each containing 72 Blackwell GPUs and 36 Grace CPUs interconnected by fifth-generation NVLink. The GB200 NVL72 also incorporates NVIDIA BlueField-3 data processing units for cloud network acceleration, composable storage, zero-trust security, and GPU compute elasticity in hyperscale AI clouds. Compared to an equivalent number of NVIDIA H100 Tensor Core GPUs, the GB200 NVL72 offers up to a 30x performance increase for LLM inference workloads while reducing cost and energy consumption by up to 25x. Acting as a single GPU, it delivers 1.4 exaflops of AI performance and 30TB of fast memory, serving as a foundational component for the latest DGX SuperPOD. Additionally, NVIDIA provides the HGX B200 server board, linking eight B200 GPUs via NVLink to support x86-based generative AI platforms,

with networking speeds up to 400Gb/s through the NVIDIA Quantum-2 InfiniBand and Spectrum-X Ethernet networking platforms.

Global Network of Blackwell Partners

Blackwell-based products will be available from partners later this year. Leading cloud service providers such as AWS, Google Cloud, Microsoft Azure, and Oracle Cloud Infrastructure will offer Blackwell-powered instances, along with NVIDIA Cloud Partner program companies like Applied Digital, CoreWeave, Crusoe, IBM Cloud, and Lambda. Sovereign AI clouds, including Indosat Ooredoo Hutchinson, Nebius, Nexgen Cloud, Oracle EU Sovereign Cloud, Scaleway, Singtel, and others, will also provide Blackwell-based cloud services. GB200 will be accessible on NVIDIA DGX Cloud, a platform developed with leading cloud providers for enterprise developers to deploy advanced generative AI models. Cisco, Dell, HPE, Lenovo, Supermicro, and other hardware manufacturers will deliver servers based on Blackwell products. Moreover, software makers like Ansys, Cadence, and Synopsys will utilize Blackwell-based processors to accelerate their engineering simulation software, enabling faster product development with lower costs and higher energy efficiency.

The Blackwell product lineup is complemented by NVIDIA AI Enterprise, an end-to-end operating system tailored for production-grade AI. This comprehensive solution encompasses NVIDIA NIM™ inference microservices, newly introduced, along with a suite of AI frameworks, libraries, and tools that enterprises can seamlessly deploy across NVIDIA-accelerated clouds, data centres, and workstations.

Source & Image Credit: NVIDIA

Published on : Tue, 23 Apr 2024