The Greenest Generation: NVIDIA, Intel and Partners Supercharge AI Computing Efficiency

Accelerated NVIDIA Hopper systems with 4th Gen Intel Xeon Scalable processors — including NVIDIA DGX H100 and 60+ systems from NVIDIA partners — provide 25x more efficiency than traditional data center servers to save big on energy costs.

Author: Shar Narasimhan

AI is at the heart of humanity's most transformative innovations — from developing COVID vaccines at unprecedented speeds and diagnosing cancer to powering autonomous vehicles and understanding climate change.

Virtually every industry will benefit from adopting AI computing , but the technology has become more resource intensive as neural networks have increased in complexity. To avoid placing unsustainable demands on electricity generation to run this computing infrastructure, the underlying technology must be as efficient as possible.

Accelerated computing powered by NVIDIA GPUs and the NVIDIA AI platform offer the efficiency that enables data centers to sustainably drive the next generation of breakthroughs.

And now, timed with the launch of 4th Gen Intel Xeon Scalable processors, NVIDIA and its partners have kicked off a new generation of accelerated computing systems that are built for energy-efficient AI. When combined with NVIDIA H100 Tensor Core GPUs , these systems can deliver dramatically higher performance, greater scale and higher efficiency than the prior generation, providing more computation and problem-solving per watt.

The new Intel CPUs will be used in NVIDIA DGX H100 systems , as well as in more than 60 servers featuring H100 GPUs from NVIDIA partners around the world.

The coming NVIDIA and Intel-powered systems will help enterprises run workloads an average of 25x more efficiently than traditional CPU-only data center servers. This incredible performance per watt means less power is needed to get jobs done, which helps ensure the power available to data centers is used as efficiently as possible to supercharge the most important work.

Compared to prior-generation accelerated systems, this new generation of NVIDIA-accelerated servers speed training and inference to boost energy efficiency by 3.5x – which translates into real cost savings, with AI data centers delivering over 3x lower total cost of ownership.

Among the features of the new 4th Gen Intel Xeon CPU is support for PCIe Gen 5, which can double the data transfer rates from CPU to NVIDIA GPUs and networking. Increased PCIe lanes allow for a greater density of GPUs and high-speed networking within each server.

Faster memory bandwidth also improves the performance of data-intensive workloads such as AI, while networking speeds — up to 400 gigabits per second (Gbps) per connection — support faster data transfers between servers and storage.

NVIDIA DGX H100 systems and servers from NVIDIA partners with H100 PCIe GPUs come with a license for NVIDIA AI Enterprise , an end-to-end, secure, cloud-native suite of AI development and deployment software, providing a complete platform for excellence in efficient enterprise AI.

As the fourth generation of the world's premier purpose-built AI infrastructure, NVIDIA DGX H100 systems provide a fully optimized platform powered by the operating system of the accelerated data center, NVIDIA Base Command software.

Each DGX H100 system features eight NVIDIA H100 GPUs, 10 NVIDIA ConnectX-7 network adapters and dual 4th Gen Intel Xeon Scalable processors to deliver the performance required to build large generative AI models, large language models , recommender systems and more.

Combined with NVIDIA networking, this architecture supercharges efficient computing at scale by delivering up to 9x more performance than the previous generation and 20x to 40x more performance than unaccelerated X86 dual-socket servers for AI training and HPC workloads. If a language model previously required 40 days to train on a cluster of X86-only servers, the NVIDIA DGX H100 using Intel Xeon CPUs and ConnectX-7 powered networking could complete the same work in as little as 1-2 days.

NVIDIA DGX H100 systems are the building blocks of an enterprise-ready, turnkey NVIDIA DGX SuperPOD , which delivers up to one exaflop of AI performance, providing a leap in efficiency for large-scale enterprise AI deployment.

For AI data center workloads, NVIDIA H100 GPUs enable enterprises to build and deploy applications more efficiently.

Bringing a new generation of performance and energy efficiency to enterprises worldwide, a broad portfolio of systems with H100 GPUs and 4th Gen Intel Xeon Scalable CPUs are coming soon from NVIDIA partners, including ASUS, Atos, Cisco, Dell Technologies, Fujitsu, GIGABYTE, Hewlett Packard Enterprise, Lenovo, QCT and Supermicro.

As the bellwether of the efficiency gains to come, the Flatiron Institute's Lenovo ThinkSystem with NVIDIA H100 GPUs tops the latest Green500 list — and NVIDIA technologies power 23 of the top 30 systems on the list. The Flatiron system uses prior-generation Intel CPUs, so even more efficiency is expected from the systems now coming to market.

Additionally, connecting servers with NVIDIA ConnectX-7 networking and Intel 4th Gen Xeon Scalable processors will increase efficiency and reduce infrastructure and power consumption.

NVIDIA ConnectX-7 adapters support PCIe Gen 5 and 400 Gbps per connection using Ethernet or InfiniBand, doubling networking throughput between servers and to storage. The adapters support advanced networking, storage and security offloads. ConnectX-7 reduces the number of cables and switch ports needed, saving 17% or more on electricity needed for the networking of large GPU-accelerated HPC and AI clusters and contributing to the better energy efficiency of these new servers.

These next-generation systems also deliver a leap forward in operational efficiency as they're optimized for the NVIDIA AI Enterprise software suite .

Running on NVIDIA H100, NVIDIA AI Enterprise accelerates the data science pipeline and streamlines the development and deployment of predictive AI models to automate essential processes and gain rapid insights from data.

With an extensive library of full-stack software, including AI workflows of reference applications, frameworks, pretrained models and infrastructure optimization, the software provides an ideal foundation for scaling enterprise AI success.

To try out NVIDIA H100 running AI workflows and frameworks supported in NVIDIA AI Enterprise, sign up for NVIDIA LaunchPad free of charge.

Watch NVIDIA founder and CEO Jensen Huang speak at the 4th Gen Intel Xeon Scalable processor launch event.

Original URL: https://blogs.nvidia.com/blog/2023/01/10/intel-partners-ai-computing-efficiency/