Going Green: New Generation of NVIDIA-Powered Systems Show Way Forward

Author: Chintan Patel

With the end of Moore's law, traditional approaches to meet the insatiable demand for increased computing performance will require disproportionate increases in costs and power.

At the same time, the need to slow the effects of climate change will require more efficient data centers, which already consume more than 200 terawatt-hours of energy each year, or around 2% of the world's energy usage .

Released today, the new Green500 list of the world's most-efficient supercomputers demonstrates the energy efficiency of accelerated computing, which is already used in all of the top 30 systems on the list. Its impact on energy efficiency is staggering.

We estimate the TOP500 systems require more than 5 terawatt-hours of energy per year, or $750 million worth of energy, to operate.

But that could be slashed by more than 80% to just $150 million — saving 4 terawatt-hours of energy — if these systems were as efficient as the 30 greenest systems on the TOP500 list.

Conversely, with the same power budget as today's TOP500 systems and the efficiency of the top 30 systems, these supercomputers could deliver 5x today's performance.

And the efficiency gains highlighted by the latest Green500 systems are just the start. NVIDIA is racing to deliver continuous energy improvements across its CPUs, GPUs, software and systems portfolio.

NVIDIA technologies already power 23 of the top 30 systems on the latest Green500 list.

Among the highlights: the Flatiron Institute in New York City topped the Green500 list of most efficient supercomputers with an air-cooled ThinkSystem built by Lenovo featuring NVIDIA Hopper H100 GPUs.

The supercomputer, dubbed Henri, produces 65 billion double-precision, floating-point operations per watt, according to the Green500, and will be used to tackle problems in computational astrophysics, biology, mathematics, neuroscience and quantum physics.

The NVIDIA H100 Tensor Core GPU, based on the NVIDIA Hopper GPU architecture , has up to 6x more AI performance and up to 3x more HPC performance compared to the prior-generation A100 GPU. It's designed to perform with incredible efficiency. Its second-generation Multi-Instance GPU technology can partition the GPU into smaller compute units, dramatically boosting the number of GPU clients available to data center users.

And the show floor at this year's SC22 conference is packed with new systems featuring NVIDIA's latest technologies from ASUS, Atos, Dell Technologies, GIGABYTE, Hewlett Packard Enterprise, Lenovo, QCT and Supermicro.

The fastest new computer on the TOP500 list, Leonardo, hosted and managed by the Cineca nonprofit consortium, and powered by nearly 14,000 NVIDIA A100 GPUs, took the No. 4 spot, while also being the 13th most energy-efficient system.

In total, NVIDIA technologies power 361 of the systems on the TOP500 list, including 90% of the new systems (see chart).

NVIDIA is also developing new computing architectures to deliver even greater energy efficiency and performance to the accelerated data center.

The Grace CPU and Grace Hopper Superchips, announced earlier this year, will provide the next big boost in the energy efficiency of the NVIDIA accelerated computing platform. The Grace CPU Superchip delivers up to twice the performance per watt of a traditional CPU, thanks to the incredible efficiency of the Grace CPU and low-power LPDDR5X memory.

Assuming a 1-megawatt HPC data center with 20% of the power allocated for CPU partition and 80% toward the accelerated portion using Grace and Grace Hopper, data centers can get 1.8x more work done for the same power budget compared to a similarly partitioned x86-based data center.

Along with Grace and Grace Hopper, NVIDIA networking technology is supercharging cloud-native supercomputing just as the increased usage of simulations is accelerating demand for supercomputing services.

Based on NVIDIA's BlueField-3 DPU, the NVIDIA Quantum-2 InfiniBand platform delivers the extreme performance, broad accessibility and strong security needed by cloud computing providers and supercomputing centers.

The effort, described in a recent whitepaper , demonstrated how DPUs can be used to offload and accelerate networking, security, storage or other infrastructure functions and control-plane applications, reducing server power consumption up to 30%.

The amount of power savings increases as server load increases and can easily save $5 million in electricity costs for a large data center with 10,000 servers over the three-year lifespan of the servers, plus additional savings in cooling, power delivery, rack space and server capital costs.

Accelerated computing with DPUs for networking, security and storage jobs is one of the next big steps for making data centers more power efficient.

Breakthroughs like these come as the scientific method is rapidly transforming into an approach driven by data analytics, AI and physics-based simulation, making more efficient computers key to the next generation of scientific breakthroughs.

By providing researchers with a multi-discipline, high-performance computing platform optimized for this new approach — and able to deliver both performance and efficiency — NVIDIA gives scientists an instrument to make critical discoveries that will benefit us all.

Take the Green Train: NVIDIA BlueField DPUs Drive Data Center Efficiency

NVIDIA Omniverse Opens Portals for Scientists to Explore Our Universe

NVIDIA H100 and Quantum-2 Systems Announced Worldwide

Going the Distance: NVIDIA Platform Solves HPC Problems at the Edge

Supercomputing Superpowers: NVIDIA Brings Digital Twin Simulation to HPC Data Center Operators

Speaking the Language of the Genome: Gordon Bell Finalist Applies Large Language Models to Predict New COVID Variants

Explainer: What Is Green Computing?

Original URL: https://blogs.nvidia.com/blog/2022/11/14/green/