Green Light: NVIDIA Grace CPU Paves Fast Lane to Energy-Efficient Computing for Every Data Center

Mainstream applications get 2x gains over x86 in energy-efficient performance on microservices, analytics, simulations and more.

Author: Ivan Goldwasser

The results are in, and they point to a new era in energy-efficient computing.

In tests of real workloads, the NVIDIA Grace CPU Superchip scored 2x performance gains over x86 processors at the same power envelope across major data center CPU applications. That opens up a whole new set of opportunities.

It means data centers can handle twice as much peak traffic. They can slash their power bills by as much as half. They can pack more punch into the confined spaces at the edge of their networks — or any combination of the above.

Data center managers need these options to thrive in today's era of energy efficiency .

Moore's law was a brilliant predictor of the last half-century of technical progress, but today we have reached the limits of the laws of physics. Gone are the days of semiconductor capabilities doubling every 18 months; we must do more with less. Engineers can no longer pack more transistors in the same space at the same power.

That's why new x86 CPUs typically offer gains over prior generations of less than 30%. It's also why a growing number of data centers are power capped.

With the added threat of global warming, data centers don't have the luxury of expanding their power, but they still need to respond to the growing demands for computing.

Compute demand is growing 10% a year in the U.S., and will double in the eight years from 2022-2030, according to a McKinsey study .

"Pressure to make data centers sustainable is therefore high, and some regulators and governments are imposing sustainability standards on newly built data centers," it said.

With the end of Moore's law, the data center's progress in computing efficiency has stalled, according to a survey that McKinsey cited (see chart below).

In today's environment, the 2x gains NVIDIA Grace offers are the eye-popping equivalent of a multi-generational leap. It meets the requirements of today's data center executives.

Zac Smith — the head of edge infrastructure at Equinix, a global service provider that manages more than 240 data centers — articulated these needs in an article about energy-efficient computing.

"The performance you get for the carbon impact you have is what we need to drive toward," he said.

"We have 10,000 customers counting on us for help with this journey. They demand more data and more intelligence, often with AI, and they want it in a sustainable way," he added.

The Grace CPU delivers that efficient performance thanks to three innovations.

It uses an ultra-fast fabric to connect 72 Arm Neoverse V2 cores in a single die that sports 3.2 terabytes per second in fabric bisection bandwidth, a standard measure of throughput. Then it connects two of those dies in a superchip package with the NVIDIA NVLink-C2C interconnect, delivering 900 GB/s of bandwidth.

Finally, it’s the first data center CPU to use server-class LPDDR5X memory. That provides up to 50% more memory bandwidth at similar cost but one-eighth the power of typical server memory. And its compact size enables 2x the density of typical card-based memory designs.

NVIDIA engineers are running real data center workloads on Grace today.

They found that compared to the leading x86 CPUs in data centers using the same power footprint, Grace is:

2.3x faster for microservices,

2x faster in memory intensive data processing

and 1.9 x faster in computational fluid dynamics, used in many technical computing apps.

Data centers usually have to wait two or more CPU generations to get these benefits, summarized in the chart below.

Even before these results on working CPUs, users responded to the innovations in Grace.

The Los Alamos National Laboratory announced in May it will use Grace in Venado, a 10 exaflop AI supercomputer that will advance the lab’s work in areas such as materials science and renewable energy. Meanwhile, data centers in Europe and Asia are evaluating Grace for their workloads.

NVIDIA Grace is sampling now with production in the second half of the year. ASUS, Atos, GIGABYTE, Hewlett Packard Enterprise, QCT, Supermicro, Wistron and ZT Systems are building servers that use it.

To dive into the details, read this whitepaper on the Grace architecture.

Learn more about sustainable computing from this session at NVIDIA GTC (March 20-23, free with registration): Three Strategies to Maximize Your Organization’s Sustainability and Success in an End-to-End AI World .

Read a whitepaper about the NVIDIA BlueField DPU to find out how to build energy-efficient networks.

And watch NVIDIA founder and CEO Jensen Huang’s GTC keynote to get the big picture.

Original URL: https://blogs.nvidia.com/blog/2023/03/21/grace-cpu-energy-efficiency/