Arm Neoverse NVIDIA Grace CPU Superchips: Setting the Pace for the Future of AI

To handle data-intensive workloads like AI, we need specialized processors

By **Chris Bergey**, SVP/GM Infrastructure, Arm

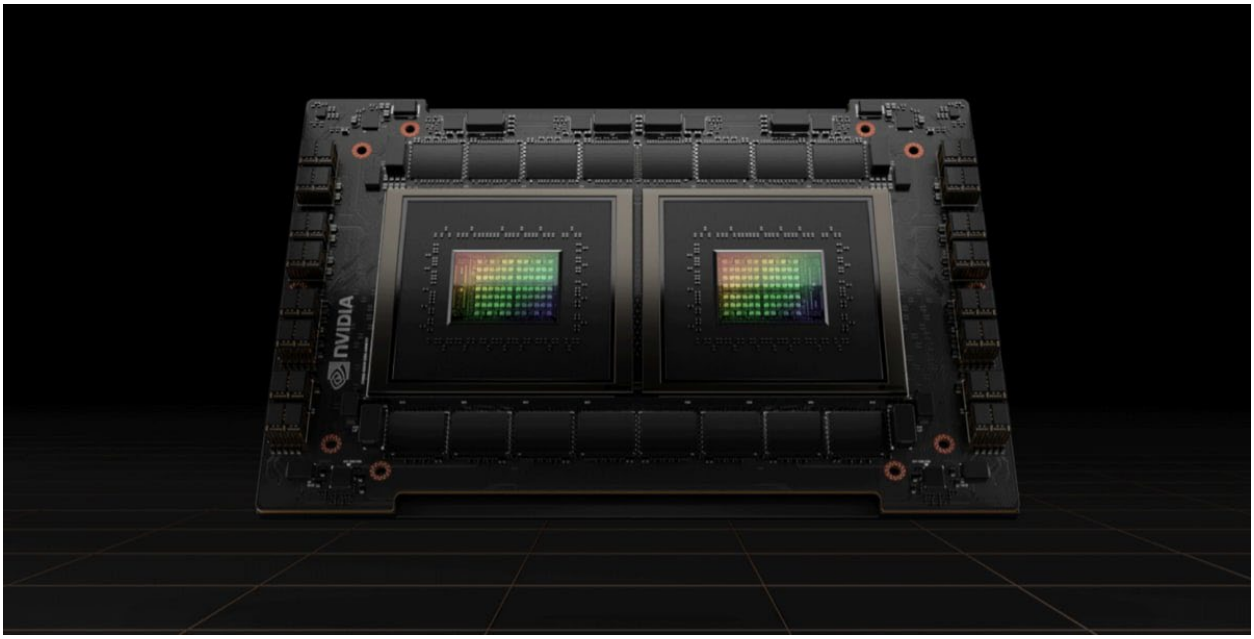
Artificial intelligence (AI) will change the world. But to realize its potential, we will have to change the way we design computing systems.

Tasks such as training neural networks in the cloud or performing pattern recognition at the edge in real-time will require specialized systems-on-chip (SoCs)—and ultimately servers and data centers—optimized for the unique power, performance, and data throughput demands of AI and machine learning (ML). Otherwise, the cost, power, and carbon required for AI will start to outweigh the gains.

NVIDIA, which pioneered the use of GPUs to enhance the performance, productivity, and efficiency of HPC and AI in data centers, recently revealed details behind its plans for the Arm Neoverse-based NVIDIA Grace CPU Superchip.

Highlights of the NVIDIA Grace CPU

- NVIDIA Grace CPU, based on our next-generation Armv9 architecture, will first come to market as part of the Grace CPU Superchip and the Grace Hopper Superchip. The Grace family of processors are designed to deliver up to a 10x performance leap in AI's most demanding tasks, such as training NLP models with more than a trillion parameters, while dramatically improving performance per watt.
- The NVIDIA Grace CPU Superchip will contain two Grace CPUs for a total of up to 144 Arm Neoverse cores, as well as up to 960GB of LPDDR5x memory inside the module for 1 TB/sec of memory bandwidth for the massive data throughput requirements for AI, HPC and hyperscale workloads.
- Benchmarks from NVIDIA show the Grace CPU Superchip achieving an estimated performance of 740 on the SPECrate®2017_int_base benchmark.
- The NVIDIA Grace Hopper Superchip combines a single Grace CPU with an NVIDIA Hopper-based GPU, the newest GPU architecture from NVIDIA. It will also feature 16 channels of PCIe Gen5 and up to 960GB of LPDDR5x memory.
- The CPUs and GPUs inside the Superchips will be linked by NVIDIA's new NVLink-C2C, offering 900 GB/sec of bandwidth between chips—7x more than PCIe Gen5 on NVIDIA. The result will be up to 30x higher aggregate system memory bandwidth to the GPU compared to today's leading servers.
- NVIDIA NVLink-C2C will support Arm's AMBA CHI protocol for direct chip-to-chip connectivity for faster communication between Arm-based CPUs within the same SoC, as well as communication between Arm-based SoCs like the Grace CPU Superchip with other Arm-based devices, such as NVIDIA Bluefield data processing units (DPUs), or with custom ASICs developed by NVIDIA's partners.



Arm Neoverse-based NVIDIA Grace CPU Superchip

The Grace Hopper Superchip will be the processor for Alps, a supercomputer being developed by the Swiss National Supercomputing Centre (CSCS) and Hewlett Packard Enterprise expected in 2023 designed for climate modeling, computational fluid dynamics, and other tasks. The Grace Hopper Superchip has also been selected for a supercomputer being developed by the U.S. Department of Energy's Los Alamos National Laboratory and HPE.

There are a number of firsts to unpack here. The Grace CPU Superchip is NVIDIA's first discrete data center CPU. The Grace CPU Superchips are also the first Arm Neoverse-based devices designed specifically for AI, HPC and hyperscale workloads, showing what is possible through the tight coupling of Neoverse CPUs as well as high-performance accelerators and memory systems. The Grace CPU also marks NVIDIA's first chip based on Armv9 and thus will be able to take advantage of many next-generation features. Finally, the new products substantially advance efforts to improve interconnect performance.

A Blueprint for AI Everywhere

The last decade was about what AI could accomplish. Going forward, the technology industry, our customers, and even the public at large will want to know more about *how* we get there, i.e. the changes needed in wireless networks, data center architecture, or device design to make autonomous vehicles, predictive health care, or other AI applications possible.

Fulfilling the promise of AI will draw on the talents across the Arm ecosystem, from system and chip designers to developers fine-tuning neural networks for speed, accuracy, and efficiency. We want to thank NVIDIA for helping set the course. It will be a long, and incredibly interesting voyage.