NVIDIA Introduces Grace CPU Superchip

144 High-Performance Cores and 1 Terabyte/Second Memory; Doubles Performance and Energy-Efficiency of Server Chips

GTC—NVIDIA today announced its first Arm® Neoverse™-based discrete data center CPU designed for AI infrastructure and high performance computing, providing the highest performance and twice the memory bandwidth and energy-efficiency compared to today's leading server chips.

The NVIDIA Grace™ CPU Superchip comprises two CPU chips connected, coherently, over NVLink®-C2C, a new high-speed, low-latency, chip-to-chip interconnect.

The Grace CPU Superchip complements NVIDIA's first CPU-GPU integrated module, the Grace Hopper Superchip, announced last year, which is designed to serve giant-scale HPC and AI applications in conjunction with an NVIDIA Hopper™ architecture-based GPU. Both superchips share the same underlying CPU architecture, as well as the NVLink-C2C interconnect.

"A new type of data center has emerged — AI factories that process and refine mountains of data to produce intelligence," said Jensen Huang, founder and CEO of NVIDIA. "The Grace CPU Superchip offers the highest performance, memory bandwidth and NVIDIA software platforms in one chip and will shine as the CPU of the world's AI infrastructure."

Introducing NVIDIA's CPU Platform
Created to provide the highest performance, Grace CPU Superchip packs 144 Arm cores in a single socket, offering industry-leading estimated performance of 740 on the SPECrate®2017_int_base benchmark.(1) This is more than 1.5x higher compared to the dual-CPU shipping with the DGX™ A100 today, as estimated in NVIDIA's labs with the same class of compilers.(2)

Grace CPU Superchip also provides industry-leading energy efficiency and memory bandwidth with its innovative memory subsystem consisting of LPDDR5x memory with Error Correction Code for the best balance of speed and power consumption. The LPDDR5x memory subsystem offers double the bandwidth of traditional DDR5 designs at 1 terabyte per second while consuming dramatically less power with the entire CPU including the memory consuming just 500 watts.

The Grace CPU Superchip is based on the latest data center architecture, Arm®v9. Combining the highest single-threaded core performance with support for Arm's new generation of vector extensions, the Grace CPU Superchip will bring immediate benefits to many applications.

The Grace CPU Superchip will run all of NVIDIA's computing software stacks, including NVIDIA RTX™, NVIDIA HPC, NVIDIA AI and Omniverse. The Grace CPU Superchip along with NVIDIA ConnectX®-7 NICs offer the flexibility to be configured into servers as standalone CPU-only systems or as GPU-accelerated servers with one, two, four or eight Hopper-based GPUs, allowing customers to optimize performance for their specific workloads while maintaining a single software stack.

Designed for AI, HPC, Cloud and Hyperscale Applications
The Grace CPU Superchip will excel at the most demanding HPC, AI, data analytics, scientific computing and hyperscale computing applications with its highest performance, memory bandwidth, energy efficiency and configurability.

The Grace CPU Superchip's 144 cores and 1TB/s of memory bandwidth will provide unprecedented performance for CPU-based high performance computing applications. HPC applications are compute-intensive, demanding the highest performing cores, highest memory bandwidth and the right memory capacity per core to speed outcomes.

NVIDIA is working with leading HPC, supercomputing, hyperscale and cloud customers for the Grace CPU Superchip. Both it and the Grace Hopper Superchip are expected to be available in the first half of 2023.

To learn more about the Grace CPU Superchip, watch Huang's GTC 2022 keynote. Register for GTC for free to attend sessions with NVIDIA and industry leaders.

About NVIDIA
NVIDIA's (NASDAQ: NVDA) invention of the GPU in 1999 sparked the growth of the PC gaming market and has redefined modern computer graphics, high performance computing and artificial intelligence. The company's pioneering work in accelerated computing and AI is reshaping trillion-dollar industries, such as transportation, healthcare and manufacturing, and fueling the growth of many others. More information at https://nvidianews.nvidia.com/.

(1) www.spec.org

(2) Pre-silicon Grace projection compared against SPEC CPU® 2017 compiled on GCC 10 with “-march=native -O3 -ffast-math -funroll-loops -flto” and run on production AMD EPYC™ 7742 systems, yielding a final estimated SpecRate2017_int_base score of 460 for a dual-socket design.”

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, specifications, performance and availability of the NVIDIA Grace CPU Superchip and the Grace Hopper Superchip; the Grace CPU Superchip offering the highest performance, memory bandwidth and NVIDIA software platforms in one chip and shining as the CPU of the world’s AI infrastructure; HPC applications being compute-intensive and their demands; and NVIDIA working with leading HPC, supercomputing, hyperscale and cloud customers for the Grace CPU Superchip are forward-looking statements that are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems; as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

Stephanie Matthew
Corporate Communications
NVIDIA
+1-408-646-3359
smatthew@nvidia.com