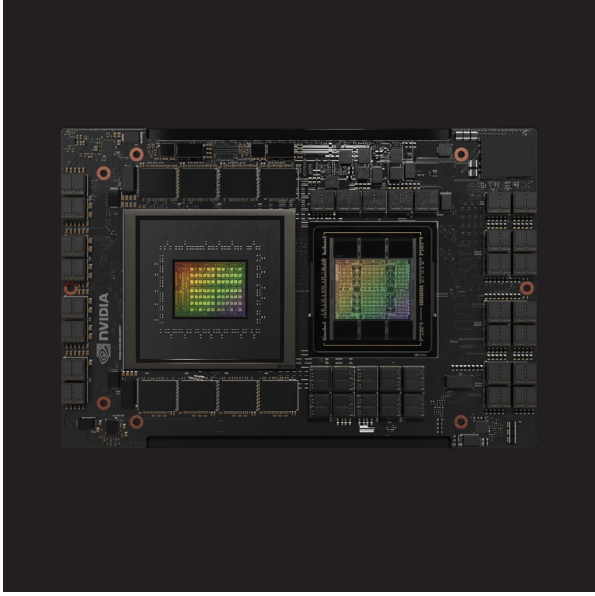
Datasheet

NVIDIA Grace CPU Superchip

The world's first no-compromise data center CPU.



The World's Most Efficient Data Center CPU

Built for demanding applications that run on traditional CPUs, the NVIDIA **Grace™ CPU Superchip** is the first no-compromise CPU for cloud, enterprise, and high-performance computing (HPC). Featuring twice the performance per watt of the latest conventional x86-64 platforms, Grace is designed to offer best-in-class data center compute throughput. Grace also delivers TCO for demanding compute-optimized cloud instances, HPC and supercomputing deployments, enterprise compute infrastructure, data analytics, intelligent edge platforms, and more.

A New Approach to Platform Design

The NVIDIA Grace CPU Superchip simplifies the design of energy-efficient systems to deliver greater data center-scale performance and functionality than today's multi-socket servers. The NVIDIA Grace CPU Superchip includes high-performance, power-efficient Arm Neoverse V2 cores connected with the high-bandwidth NVIDIA Scalable Coherency Fabric to keep data flowing between CPU cores, NVIDIA® NVLink®-Chip-to-Chip (C2C), memory, and system input and output (IO). This architecture provides good locality to data and memory, making the platform easy to optimize for developers.

The NVIDIA Grace CPU Superchip integrates the supporting components traditionally found elsewhere in servers, improving both power efficiency and density, while dramatically simplifying system design.

The Grace CPU Superchip is the first data center CPU to utilize server-class high-speed LPDDR5X memory, balancing cost, power, bandwidth, and capacity. The co-packaged memory employs a novel provisioning and error-detection technique, which eliminates the need to service or replace failed memory in the field. This offers improved reliability and allows Grace to be deployed in scenarios where serviceability is difficult or costly.

Finally, with the high-speed NVLink-C2C path between non-uniform memory access (NUMA) domains, system designers can provide flexible utilization of peripherals and help alleviate NUMA bottlenecks for application developers and users.

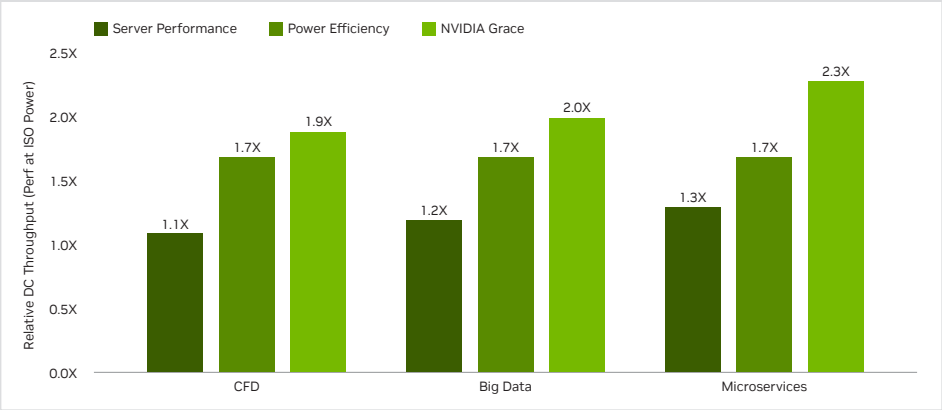
Key Features

- Up to 144 high-performance Arm Neoverse V2 Cores with 4x128b SVE2
- High-performance NVIDIA Scalable Coherency Fabric with 3.2 terabytes per second (TB/s) bisection bandwidth
- Up to 960 gigabytes (GB) of LPDDR5X memory with error-correction code (ECC) with up to 1TB/s of memory bandwidth
- 900GB/s NVLink-C2C
- 500W module (CPU + memory)

Next-Generation Data Center CPU Performance Efficiency

The Grace CPU Superchip delivers outstanding performance, memory bandwidth, and data-movement capabilities with leadership performance per watt to deliver generational gains in energy-efficient CPU computing for the data center. The NVIDIA Grace CPU Superchip provides versatility and performance for foundational data center workloads such as microservices, memory-intensive data processing for big data, and non-accelerated memory intensive HPC applications such as computational fluid dynamics (CFD).

2X More Data Center Throughput in Cloud and HPC Apps[1]



The New Standard for Software Infrastructure

The NVIDIA Grace CPU follows mainstream CPU design principles, is programmed just like any other server CPU, and is backed by the full NVIDIA ecosystem. All major Linux distributions and the vast collections of software packages they provide work perfectly and without modification on the NVIDIA Grace CPU.

To enable developers to jump-start their work, the NVIDIA Grace CPU Superchip is supported by the full NVIDIA software stack, including NVIDIA HPC, NVIDIA AI, and NVIDIA Omniverse™.

[1] Data center-level projection of NVIDIA Grace Superchip vs. x86 flagship dual-socket data center systems (112 and 192 core systems). CFD: OpenFOAM (Motorbike | Small) Big Data: HiBench+K-means Spark (HiBench 7.1.1, Hadoop 3.3.3, Spark 3.3.0) and Microservices: Google Protobufs (Commit 7cd0b6fbf1643943560d8a9fe553fd206190b27f | N instances in parallel).

NVIDIA Grace Superchip performance based on engineering measurements. Results subject to change.

Preliminary Product Specifications

	Feature
Core count	144 Arm Neoverse V2 Cores with 4x128b SVE2
L1 cache	64KB i-cache + 64KB d-cache
L2 cache	1MB per core
L3 cache	234MB
LPDDR5X size	240GB, 480GB and 960GB on-module memory options
Memory bandwidth	Up to 1TB/s
NVIDIA NVLink-C2C bandwidth	900GB/s
PCIe links	Up to 8x PCIe Gen5 x16 option to bifurcate
Module thermal design power (TDP)	500W TDP with memory
Form factor	Superchip module
Thermal solution	Air cooled or liquid cooled

Ready to Get Started?

To learn more about NVIDIA Grace CPU Superchip, visit **nvidia.com/grace-cpu**

NVIDIA