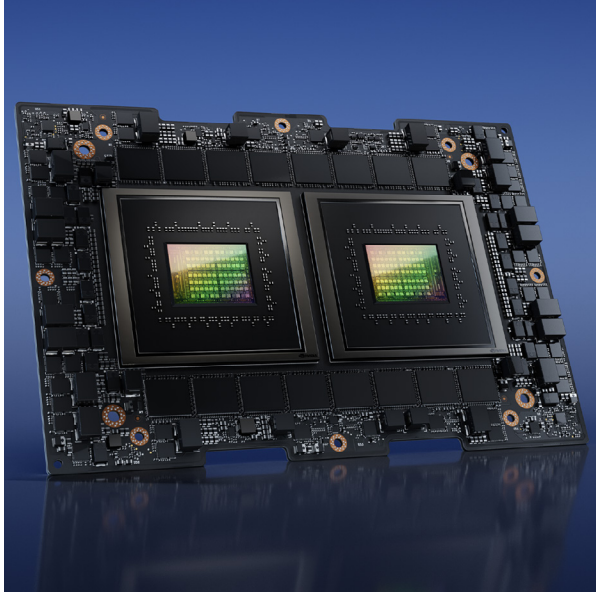
Datasheet

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 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 platforms, Grace is designed to offer best-in-class data center CPU 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.

Flexible Platform Design

The NVIDIA Grace CPU portfolio includes the NVIDIA Grace CPU Superchip that provides the heart of a dual-socket server with 144 Arm® Neoverse™ V2 cores and up to 960GB of LPDDR5X in a single compact module requiring only 500W of power for the CPU and memory—compared to almost 900W for conventional x86 servers.

To provide additional flexibility, the Grace CPU C1 delivers amazing performance in a single-socket configuration optimized for cloud, storage, edge, telco, and more to deliver up to 2X the performance per watt of conventional x86 servers.

NVIDIA Grace CPU simplifies the design of energy-efficient systems to deliver greater data center-scale performance. The Grace CPU combines 72 high-performance and power-efficient Arm Neoverse V2 cores, connected with the NVIDIA Scalable Coherency Fabric (SCF) that delivers 3.2TB/s of bisection bandwidth—double that of traditional CPUs. This architecture keeps data flowing between CPU cores, cache, memory, and system input and output (IO) to get the most out of the system's performance. This architecture provides good locality to data and memory, making the platform easy to optimize for developers.

The Grace CPU is the first data center CPU to utilize server-class high-speed LPDDR5X memory, balancing cost, power, bandwidth, and capacity. The wide memory subsystem in Grace delivers up to 500GB/s of bandwidth in only about 16W, about one-fifth that of conventional DDR5 memory. 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.

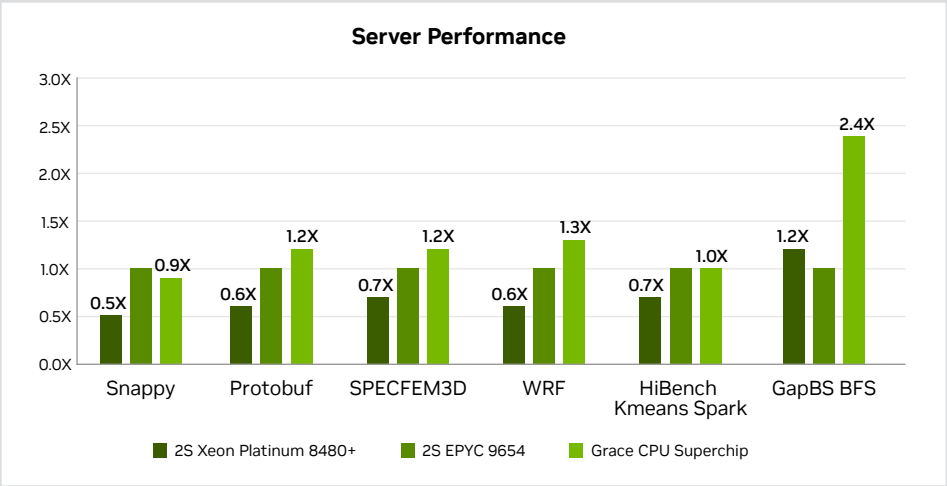
Key Features

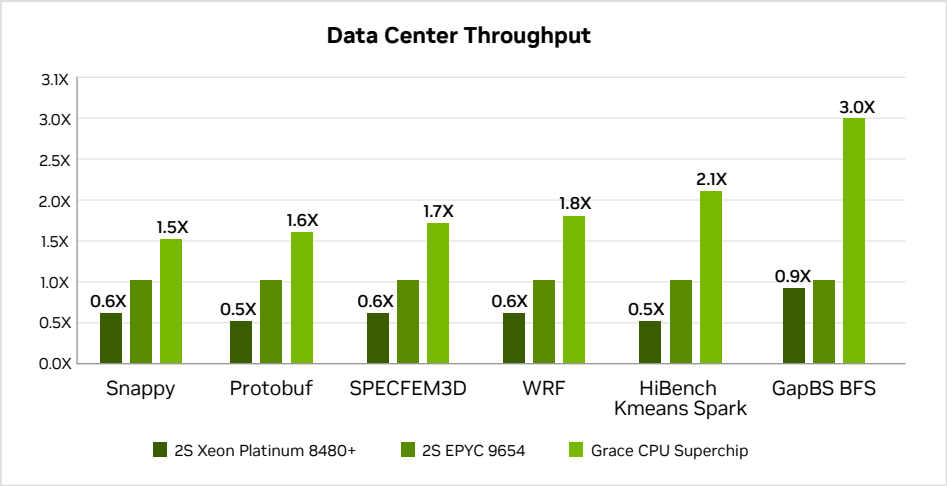
- 72 for NVIDIA Grace CPU C1 or 144 high-performance for NVIDIA Grace CPU Superchip Arm Neoverse V2 cores with 4x128b SVE2
- High-performance NVIDIA Scalable Coherency Fabric with 3.2 terabytes per second (TB/s) bisection bandwidth
- Up to 480 gigabytes (GB) of LPDDR5X memory with error-correction code (ECC) with up to 512GB/s of memory bandwidth per Grace CPU
- 900GB/s NVLink®-C2C
- Up to 250W for Grace CPU C1 or 500W for Grace CPU Superchip (CPU + memory)

Next-Generation Data Center CPU Performance Efficiency

The Grace CPU delivers outstanding per-core 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, data analytics, graph analytics, and simulation.

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. The NVIDIA Grace architecture is fully compatible with the Arm ecosystem, ensuring that any application designed for 64-bit Arm in data centers will also operate

1 NVIDIA Grace Superchip 480GB of LPDDR5X, AMD EPYC 9654 768GB of DDR5, and Intel Xeon Platinum 8480+ with 1TB DDR5. OS: Ubuntu 22.04 Compilers: GCC 12.3 unless noted below. Power for energy efficiency includes CPU + memory measured power.

Compression: Snappy (Commit af720f9a3b2c831f173b6074961737516f2d3a46 | N instances in parallel) Microservices: Google Protobufs (Commit 7cd0b6fbf1643943560d8a9fe553fd206190b27f | N instances in parallel) Seismic Data Proc: SPECFEM3D four_material_simple_model; HPC SDK 24.3 CFD: OpenFOAM Motorbike | Large v2212 MD: CP2K RPA 2023.1 Weather: WRF CONUS12km x86: ICC 2024.01; Data Analytics: HiBench+K-means Spark (HiBench 7.1.1, Hadoop 3.3.3, Spark 3.3.0; Grace: NVHPC 24.5, x86: Intel 2021.4) Graph Analytics: The Gap Benchmarks Suite BFS arXiv:1508.03619 [cs.DC], 2015.

seamlessly on Grace, and vice versa. 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™.

Product Specifications

	NVIDIA Grace CPU C1	NVIDIA Grace CPU Superchip
Core count	72 Arm Neoverse V2 cores with 4x128b SVE2	144 Arm Neoverse V2 cores with 4x128b SVE2
L1 cache	64KB i-cache + 64KB d-cache	64KB i-cache + 64KB d-cache
L2 cache	1MB per core	1MB per core
L3 cache	114MB	228MB
Base frequency | all-core SIMD frequency	3.1 GHz | 3.0 GHz	3.1GHz | 3.0GHz
LPDDR5X size	120GB, 240GB, and 480GB memory options	240GB, 480GB, and 960GB on-module memory options
Memory bandwidth	Up to 384 GB/s for 480GB Up to 512 GB/s for 120GB, 240GB	Up to 768 GB/s for 960GB Up to 1024 GB/s for 240GB, 480GB
NVIDIA NVLink-C2C bandwidth	n/a	900GB/s
PCIe links	Up to 4x PCIe Gen5 x16 option to bifurcate	Up to 8x PCIe Gen5 x16 option to bifurcate
CPU + memory power	Up to 250W TDP with memory	Up to 500W TDP with memory

Ready to Get Started?

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

© 2024 NVIDIA Corporation and affiliates. All rights reserved. NVIDIA, the NVIDIA logo, Grace, NVLink, and Omniverse are trademarks and/or registered trademarks of NVIDIA Corporation and affiliates in the U.S. and other countries. Other company and product names may be trademarks of the respective owners with which they are associated. All other trademarks are property of their respective owners. 3462416. OCT24

NVIDIA