NVIDIA Grace CPU Superchip Whitepaper

Performance and Efficiency for the Modern Data Center

V1.1

Table of Contents

A Superchip Built for Accelerated Computing 4
Grace CPU Superchip 4
High Performance Architecture 6
Grace CPU Core Architecture 9
NVIDIA Grace CPU Performance 11
Grace CPU Superchip Benchmarks 11
Superchip Application Case Studies 11
Computational Fluid Dynamics: OpenFOAM 12
Climate Modeling: NEMO 13
Numerical Weather Prediction: WRF Model 13
DNA Sequencing: Burrows-Wheeler Aligner 13
NVIDIA Grace CPU Software 15
Debugging and Profiling 17
Superchips for HPC, Cloud Computing, and Enterprise 19

List of Figures

Figure 1. Comparison of the Grace CPU Superchip with NVLink-C2C compared to traditional server architecture 6
Figure 2. NVIDIA Grace CPU and the NVIDIA Scalable Coherency Fabric, which join the Neoverse V2 cores, distributed cache and system IO in a high-bandwidth mesh interconnect 7
Figure 3. Grace CPU memory, SCF cache, PCIe, NVLink, and NVLink-C2C can be partitioned for cloud native workloads 8
Figure 4. NVIDIA Grace Arm Neoverse V2 Core is the highest performing Arm Neoverse core with support for SVE2 to accelerate key applications. 9
Figure 5. STREAM Copy and STREAM Triad Memory Bandwidth performance simulations for NVIDIA Grace CPU Superchip 11
Figure 6. Performance and energy savings for applications on the Grace CPU Superchip compared to dual socket Milan 7763 CPUs 12
Figure 7. The NVIDIA Grace CPU software ecosystem combines the full collection of NVIDIA software for CPU, GPU, and DPU with the complete Arm datacenter ecosystem 16

List of Tables

Table 1. NVIDIA Grace CPU Superchip Specifications 5

A Superchip Built for Accelerated Computing

NVIDIA® pioneered accelerated computing to tackle the world's greatest challenges. Accelerated computing requires innovation across the full stack, from hardware, to software, platforms, and applications across multiple domains. At the hardware level, the data center requires GPUs, DPUs, CPUs, and networking to scale up and scale out across cloud computing, high-performance computing (HPC), and enterprise computing. NVIDIA is known for incredibly powerful GPUs, DPUs, and networking. Now, the NVIDIA Grace™ CPU is the first CPU designed by NVIDIA for the data center.

The NVIDIA Grace CPU delivers twice the performance per watt of conventional x86-64 platforms and is the world's fastest Arm® data center CPU. The Grace CPU is found in two data center NVIDIA Superchip products. The first is the NVIDIA Grace™ Hopper Superchip that pairs a power efficient, high-bandwidth NVIDIA Grace CPU with an NVIDIA H100 Hopper GPU to maximize the capabilities for strong-scaling HPC and giant AI workloads.

The second is the NVIDIA Grace™ CPU Superchip, the first no-compromise Arm® platform for HPC, demanding cloud workloads and enterprise computing. Both these superchip products are made possible by NVIDIA NVLink® Chip-2-Chip (C2C), a coherent 900 GB/s bi-directional bandwidth interconnect.

This whitepaper highlights NVIDIA Grace CPU Superchip key features, provides an early look at the performance and capabilities of the platform, and the standards-based programming model for NVIDIA Grace. More information is available in the NVIDIA Grace Hopper Superchip Whitepaper.

Grace CPU Superchip

The NVIDIA Grace CPU Superchip represents a revolution in compute platform design by integrating the level of performance offered by a flagship x86-64 two-socket workstation or server platform into a single superchip. It enables 2X the compute density at lower power envelopes and improved TCO. The NVIDIA Grace CPU Superchip uses NVIDIA NVLink-C2C to pack 144 Arm Neoverse V2 cores and up to 1TB/s of memory bandwidth in a 500W power envelope.

Table 1. NVIDIA Grace CPU Superchip Specifications

	Grace CPU Superchip
Core Architecture	Neoverse V2 Cores: Armv9 with 4x128b SVE2
Core Count	144
Cache	L1: 64KB I-cache + 64KB D-cache per core L2: 1MB per core L3: 228 MB per superchip
Base Frequency | All-Core SIMD Frequency	3.1 GHz | 3.0 GHz
Memory Technology	LPDDR5X with ECC, Co-Packaged
Raw Memory BW	Up to 1 TB/s
Memory Size	Up to 960GB
FP64 Peak	7.1 TFLOPS
PCI Express	8x PCIe Gen 5 x16 interfaces; option to bifurcate Total 1 TB/s PCIe Bandwidth. Additional low-speed PCIe connectivity for management.
Power	500W TDP with Memory, 12V Supply

High Performance Architecture

The Grace CPU was designed to deliver high single-threaded performance, high memory bandwidth, and outstanding data movement capabilities with leadership performance per watt. These design goals required the development of several innovations to enable the Grace CPU Superchip.

Alleviate Bottlenecks with NVLink-C2C Interconnect

To create the NVIDIA Grace CPU Superchip with up to 144 Arm Neoverse V2 cores and avoid bottlenecks when moving data between the chips, the NVLink Chip-2-Chip (C2C) interconnect provides a 900 GB/s direct connection between chips.

A typical server architecture has two sockets, each composed of multiple dies and each die may represent multiple non-uniform memory (NUMA) domains. The Grace CPU Superchip uses a clean and simple memory topology. With only two NUMA nodes and the high-bandwidth NVLink-C2C, the Grace CPU Superchip helps alleviate NUMA bottlenecks for application developers and users.

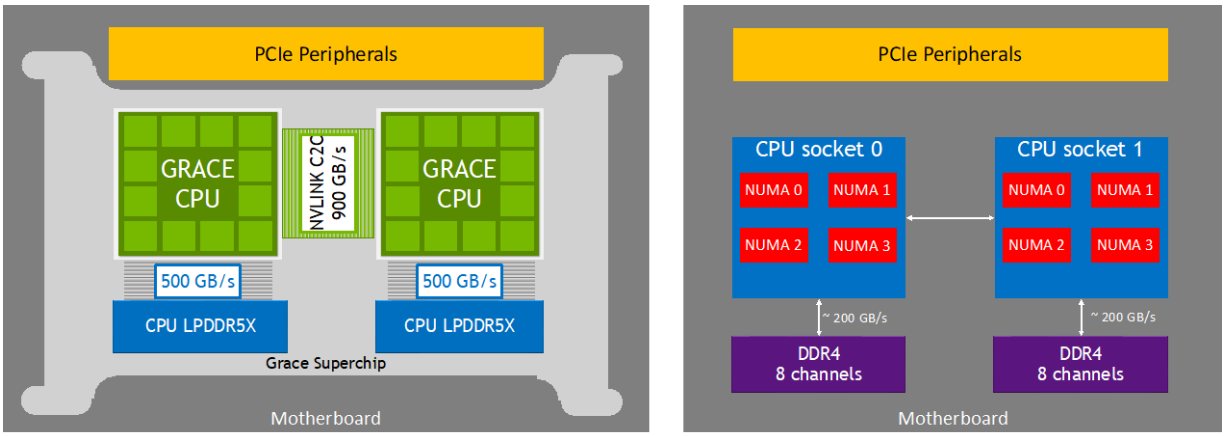

Figure 1. Comparison of the Grace CPU Superchip with NVLink-C2C compared to traditional server architecture

Scale Cores and Bandwidth with NVIDIA Scalable Coherency Fabric

NVIDIA Scalable Coherency Fabric (SCF), shown in Figure 2, is a mesh fabric and distributed cache architecture designed by NVIDIA to scale cores and bandwidth. It provides over 3.2 TB/s of total bi-section bandwidth to keep data flowing between the CPU cores, NVLink-C2C, memory, and system IO.

The CPU cores and SCF cache partitions are distributed throughout the mesh, while Cache Switch Nodes route data through the fabric and serve as interfaces between the CPU, cache memory, and system IOs. A Grace CPU Superchip has 228 MB of distributed L3 cache across the two chips.

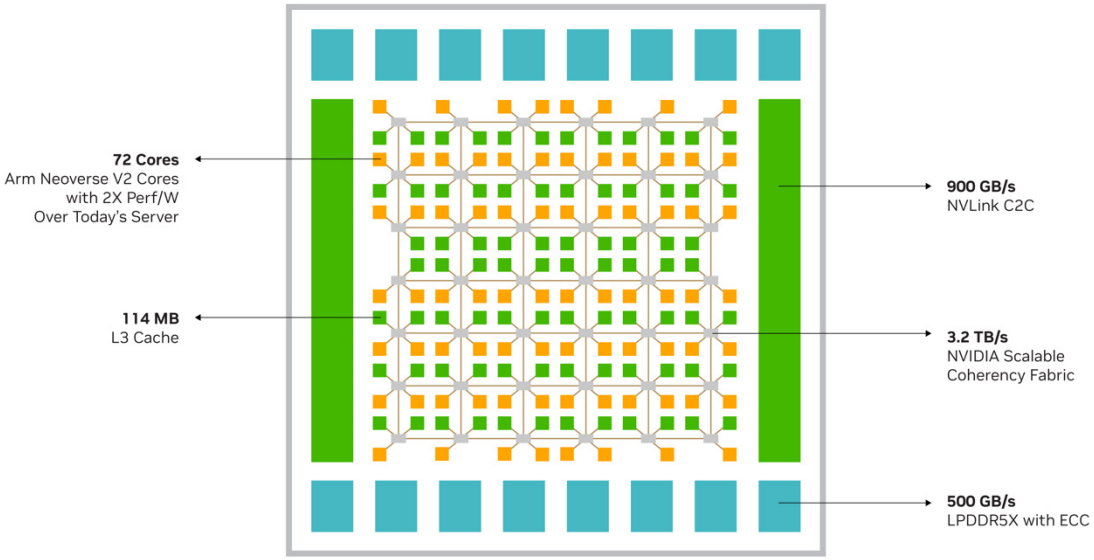

Figure 2. NVIDIA Grace CPU and the NVIDIA Scalable Coherency Fabric, which join the Neoverse V2 cores, distributed cache and system IO in a high-bandwidth mesh interconnect.

NVIDIA Grace CPU supports Arm's Memory Partitioning and Monitoring (MPAM), the Arm standard for partitioning system cache and memory resources to provide performance isolation between jobs. The NVIDIA-designed SCF Cache supports partitioning of cache capacity, I/O, and well as memory bandwidth using MPAM. It also supports the use of MPAM performance monitor groups (PMGs) for monitoring resources, such as cache storage usage and memory bandwidth utilization.

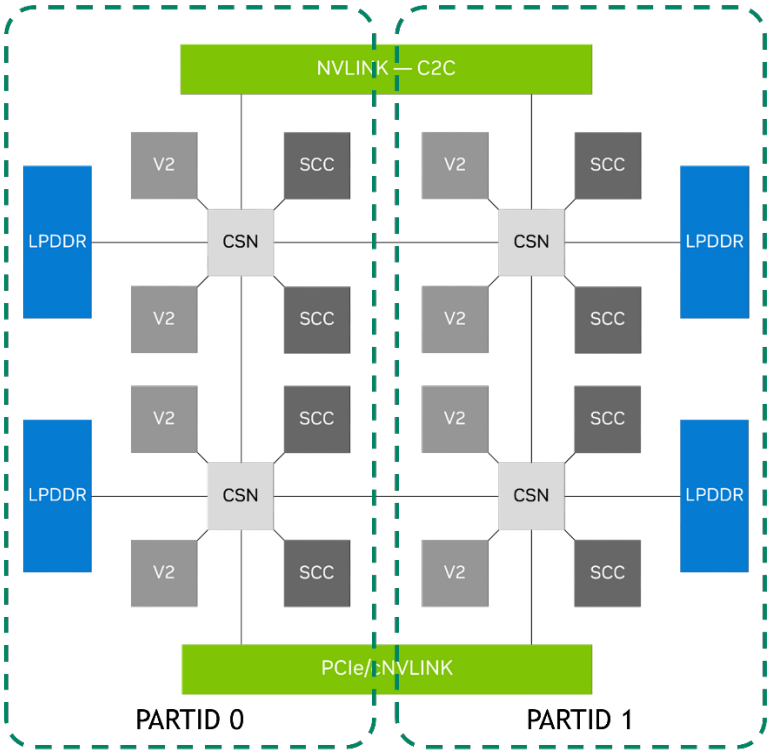

Figure 3. Grace CPU memory, SCF cache, PCIe, NVLink, and NVLink-C2C can be partitioned for cloud native workloads

LPDDR5X Memory Subsystem

The NVIDIA Grace CPU Superchip utilizes up to 960 GB of server-class LPDDR5X memory with Error Correction Code (ECC). This design strikes the optimal balance of bandwidth, energy efficiency, capacity, and cost for large-scale AI and HPC workloads.

Compared to an eight-channel DDR5 design, the Grace CPU LPDDR5X memory subsystem provides up to 53% more bandwidth at 1/8th the power per gigabyte per second while being close in cost. An HBM2e memory subsystem would have provided substantial memory bandwidth and good energy efficiency but at more than 3X the cost-per-gigabyte and only one-eighth the maximum capacity available with LPDDR5X.

The Grace CPU LPDDR5X architecture is the first data center class, resilient implementation of LPDDR technology. 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, allowing the Grace CPU to be deployed in scenarios where serviceability is difficult or costly.

The lower power consumption of LPDDR5X reduces the overall system power requirements and enables more resources to be put towards CPU cores. The compact form factor enables 2X the density of a typical DIMM-based design.

CPU I/O

The Grace CPU Superchip supports up to 128 lanes of PCIe Gen 5 for I/O connectivity. Each of the eight PCIe Gen 5 x16 link supports up to 128 GB/s of bi-directional bandwidth and can be bifurcated into 2x8's for additional connectivity. Additional PCIe interfaces are provided for system management purposes.

Server makers can use the standard expansion options for a variety of PCIe slot form factors with out-of-box support for NVIDIA GPUs and NVIDIA DPUs, NVIDIA ConnectX SmartNICs, E1.S and M.2 NVMe devices, modular BMC options, and more.

Grace CPU Core Architecture

NVIDIA Grace CPU uses the Arm Neoverse V2, Arm's highest performance data center core.

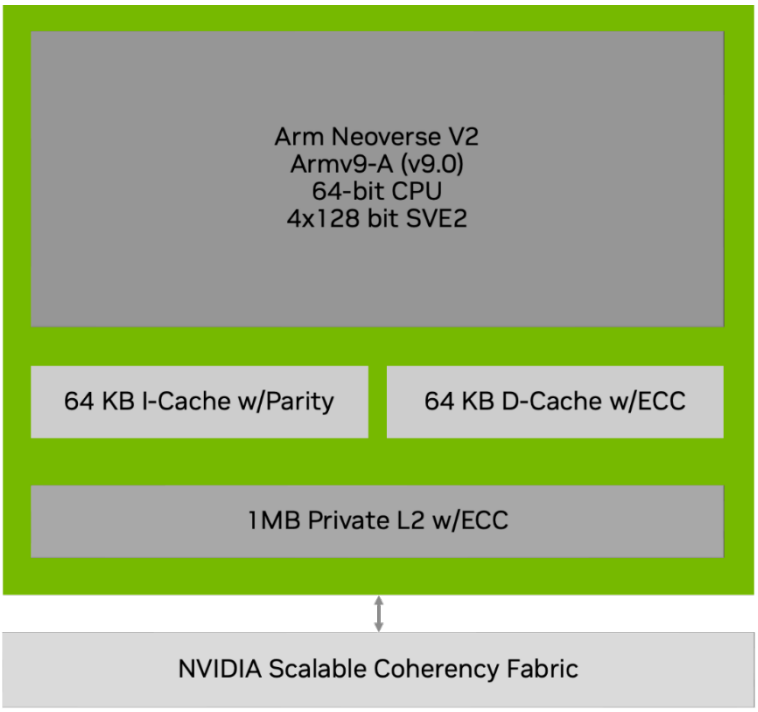

Figure 4. NVIDIA Grace Arm Neoverse V2 Core is the highest performing Arm Neoverse core with support for SVE2 to accelerate key applications.

Arm Architecture

The Grace CPU Neoverse V2 core implements the Armv9.0-A architecture, which extends the architecture defined in the Armv8-A architectures up to Armv8.5-A. Any application binaries built for an Armv8 architecture up to Armv8.5-A will execute on NVIDIA Grace. That includes binaries targeting CPUs like the Ampere Altra, the AWS Graviton2, and the AWS Graviton3.

SIMD Vectorization

The Neoverse V2 implements two single instruction multiple data (SIMD) vector instruction sets in a 4x128-bit configuration: the Scalable Vector Extension version 2 (SVE2), and Advanced SIMD (NEON). Each of the four 128-bit functional units can retire either SVE2 or NEON instructions. This design enables more codes to take full advantage of SIMD performance.

Many applications and libraries are already taking advantage of NEON. SVE is a length-agnostic SIMD ISA that improves NEON by supporting more datatypes such as FP16, enables more powerful instructions such as gather/scatter, and enhances support for long vector lengths. SVE is implemented in many flagship Arm implementations, ensuring compatibility of SVE optimizations for Grace CPU accrue toward portable binaries. SVE2 further extends the SVE ISA with advanced instructions that can accelerate key HPC applications like machine learning, genomics, and cryptography.

Atomic Operations

NVIDIA Grace CPU supports the Large System Extension (LSE) which was first introduced in Armv8.1. LSE provides low-cost atomic operations, which can improve system throughput for CPU-to-CPU communication, locks, and mutexes:

- Compare and Swap instructions, CAS, and CASP
- Atomic memory operation instructions, LD<OP> and ST<OP>, where <OP> is one of ADD, CLR, EOR, SET, SMAX, SMIN, UMAX, and UMIN
- Swap instruction, SWP

These instructions can operate on integer data. All compilers supporting NVIDIA Grace CPU will use these instructions automatically in synchronization functions like GCC's __atomic built-ins. The improvement can be up to an order of magnitude when using LSE atomics instead of load/store exclusives. For instance, a shared integer value can be incremented with a single atomic ADD rather than the sequence: load exclusive, add, attempt store exclusive, and repeat if the operation failed.

Armv9 Additional Features

The NVIDIA Grace CPU implements multiple key features of the Armv9 portfolio that provide utility in general-purpose data center CPUs, including but not limited to cryptographic acceleration, scalable profiling extension, virtualization extensions, full memory encryption, and secure boot.

NVIDIA Grace CPU Performance

The Grace CPU provides versatility for HPC and data center applications. The sections below highlight the performance of Grace Superchip for benchmarks and applications.

Grace CPU Superchip Benchmarks

Since it is designed to accelerate memory bound workloads, a single Grace CPU is expected to deliver up to 450 GB/s of realized STREAM copy and triad bandwidth. The Grace CPU Superchip can deliver a staggering 900 GB/s STREAM copy and triad memory bandwidth.

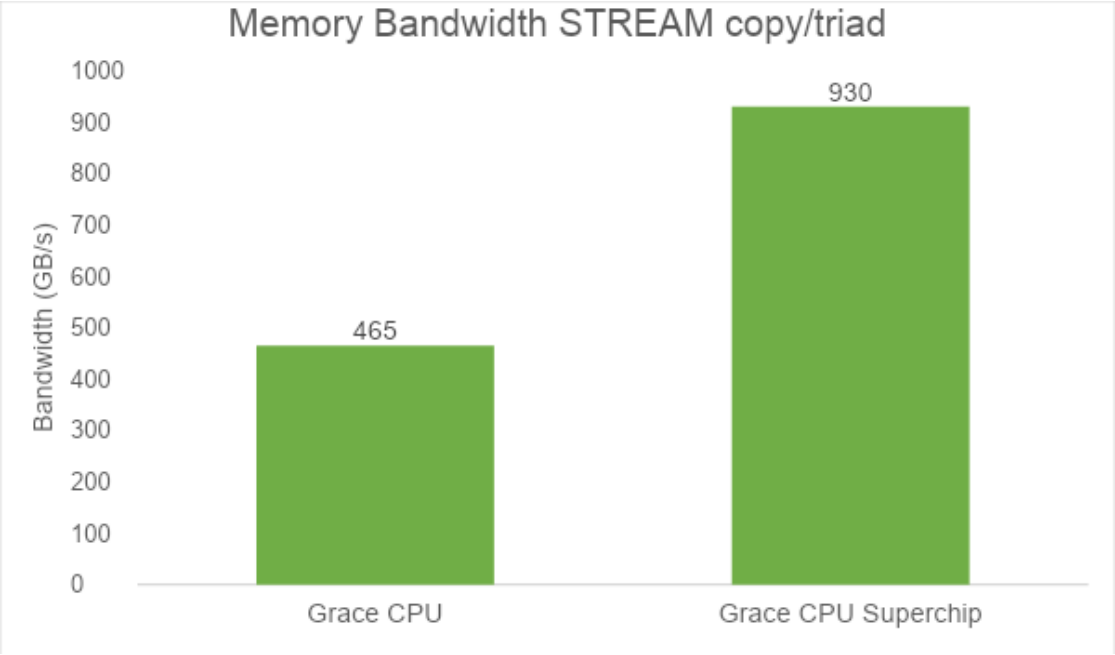

Figure 5. STREAM Copy and STREAM Triad Memory Bandwidth performance simulations for NVIDIA Grace CPU Superchip

Superchip Application Case Studies

The memory topology of a Grace CPU Superchip is clean and simple. With 144 high-performance Arm Neoverse V2 cores, up to 1 TB/s of memory bandwidth, only two NUMA nodes, and high-bandwidth NVLink-C2C, the NVIDIA Grace CPU delivers outstanding application performance and alleviates NUMA bottlenecks for application

developers and users. This straightforward design makes extracting more performance out of applications much easier with the Grace CPU Superchip.

Figure 6 shows performance projections of some example applications in HPC and data center that get accelerated by Grace CPU Superchip. In addition, we show the energy savings resulting from the reduction in power consumption and reduced runtimes.

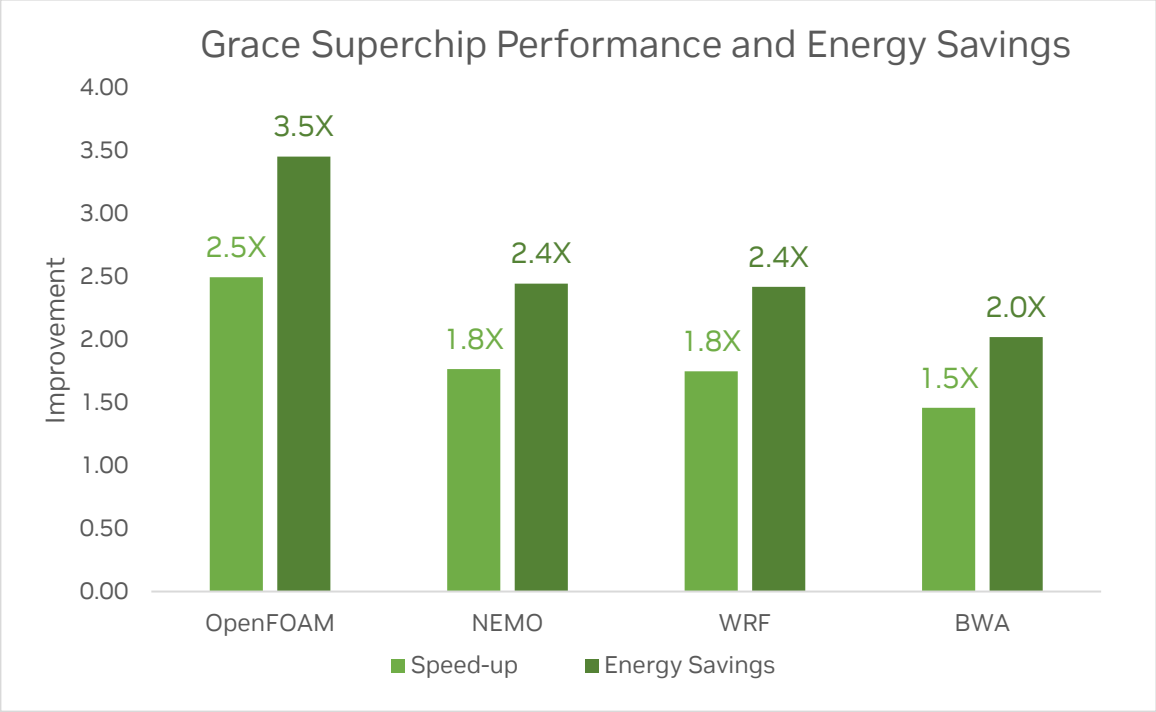

Figure 6. Performance and energy savings for applications on the Grace CPU Superchip compared to dual socket Milan 7763 CPUs

Computational Fluid Dynamics: OpenFOAM

OpenFOAM is a well-known open-source C++ library for implementing computational fluid dynamics and multiphysics solvers. We chose the OpenFOAM HPC motorbike L model 34 M (a scaled-up version of the OpenFOAM motorbike tutorial) to highlight the performance benefits of the Grace CPU Superchip for memory bound applications. The HPC motorbike L case is a representative of typical engineering workloads, with flatter profile and considerable time spent on the turbulence calculations. Our performance model projects a 2.5X speed up on the Grace CPU Superchip when comparing against 2x x86 CPUs, with 3.5X projected energy savings.

Climate Modeling: NEMO

Nucleus for European Modelling of the Ocean (NEMO) is a state-of-the-art modelling framework for research and forecasting activities in ocean and climate sciences. The NEMO code is developed by a European consortium, made up by the French National Centre for Scientific Research (CNRS) and Mercator-Ocean from France, Natural Environment Research Council (NERC), the Met Office from the United Kingdom, and the Euro-Mediterranean Center on Climate Change (CMCC) and the National Institute of Geophysics and Volcanology (INGV) from Italy. NEMO can solve ocean and sea-ice thermodynamics, oceanic tracer transport, and biogeochemical processes. NEMO can also be combined with atmospheric models like Icosahedral Nonhydrostatic Weather and Climate Model (ICON), and Weather Research and Forecasting (WRF) Model to build a digital twin earth.

We used the GYRE_PISCES test case, an idealized configuration representing a Northern hemisphere double gyre system. This specific configuration is used as a reference benchmark by the community due to its characteristics.

Our projections show that NEMO benefits from the estimated 800 GB/s STREAM memory bandwidth featured by the Grace CPU Superchip, which provides a total speed-up of 1.8X compared to 2x x86 CPUs, while projecting 2.4X energy savings.

Numerical Weather Prediction: WRF Model

The Weather Research and Forecasting (WRF) model is a state-of-the-art mesoscale Numerical Weather Prediction (NWP) system designed for both atmospheric research and operational forecasting applications. It is in widespread use across the globe with over 50,000 users and a top 10 application for many HPC centers.

As a community code, WRF has many contributors and has amassed numerous physics parameterization options over the last 20 years. The WRF community is diverse, and the users of WRF run it on a variety of platforms.

Like most NWP packages, WRF is mostly bandwidth-bound due to its low arithmetic intensity. Performance projections show 1.8X speed-up, and 2.0X energy savings on Grace Superchip when comparing against 2x x86 CPUs.

DNA Sequencing: Burrows-Wheeler Aligner

Burrows-Wheeler Aligner (BWA) is a short read mapping application widely adopted by the bioinformatics community as part of DNA sequence analysis pipelines. The application maps and aligns short sequences against a large reference genome to find their original locations.

The performance projections on this document will focus on the BWA-MEM algorithm and, more specifically, its BWA-MEM2 implementation. The internal pipeline of BWA-MEM2 can be split into five stages that feature different performance limiters. The first three stages in the computation are limited by a mix of memory bandwidth, cache

latency due to query operations, and a large data index. It is in these kinds of scenarios where the high memory bandwidth of the Grace CPU Superchip brings further performance. In the last two regions, performance is led by data computation. These two stages based on alignment methods can be highly vectorized and benefit from the SVE2 units present in the Grace CPU Superchip, increasing performance when comparing to current state-of-the-art x86 solutions.

We use the HG002 dataset from Illumina paired-end sequencers, which consists of the complete human genome at 30x coverage. It is treated as reference data in the bioinformatics community. Considering the sensitivities from different regions in the code and the Grace CPU Superchip characteristics, our performance projection predicts a speed-up of 1.5X in performance, and 2.3x energy savings.

NVIDIA Grace CPU Software

NVIDIA Grace CPU builds on Arm standards, such as Arm Server Base System Architecture (SBSA) and the Base Boot Requirements (BBR) of the Arm SystemReady Certification Program. Grace CPU also uses the popular Neoverse microarchitecture, which means that this CPU is supported by an enormous software ecosystem. All major Linux distributions, and the vast collections of software packages they provide, work perfectly - and without modification - on NVIDIA Grace.

Compilers, libraries, tools, profilers, system administration utilities, frameworks for containerization, and virtualization are available today and can be trivially installed and used on NVIDIA Grace exactly as on any other datacenter CPU.

In addition, the whole NVIDIA software stack is available for NVIDIA Grace. The NVIDIA HPC SDK and every CUDA component have Arm-native installers and containers. The NVIDIA GPU Cloud™ (NGC) also provides deep learning, machine learning, and HPC containers optimized for Arm.

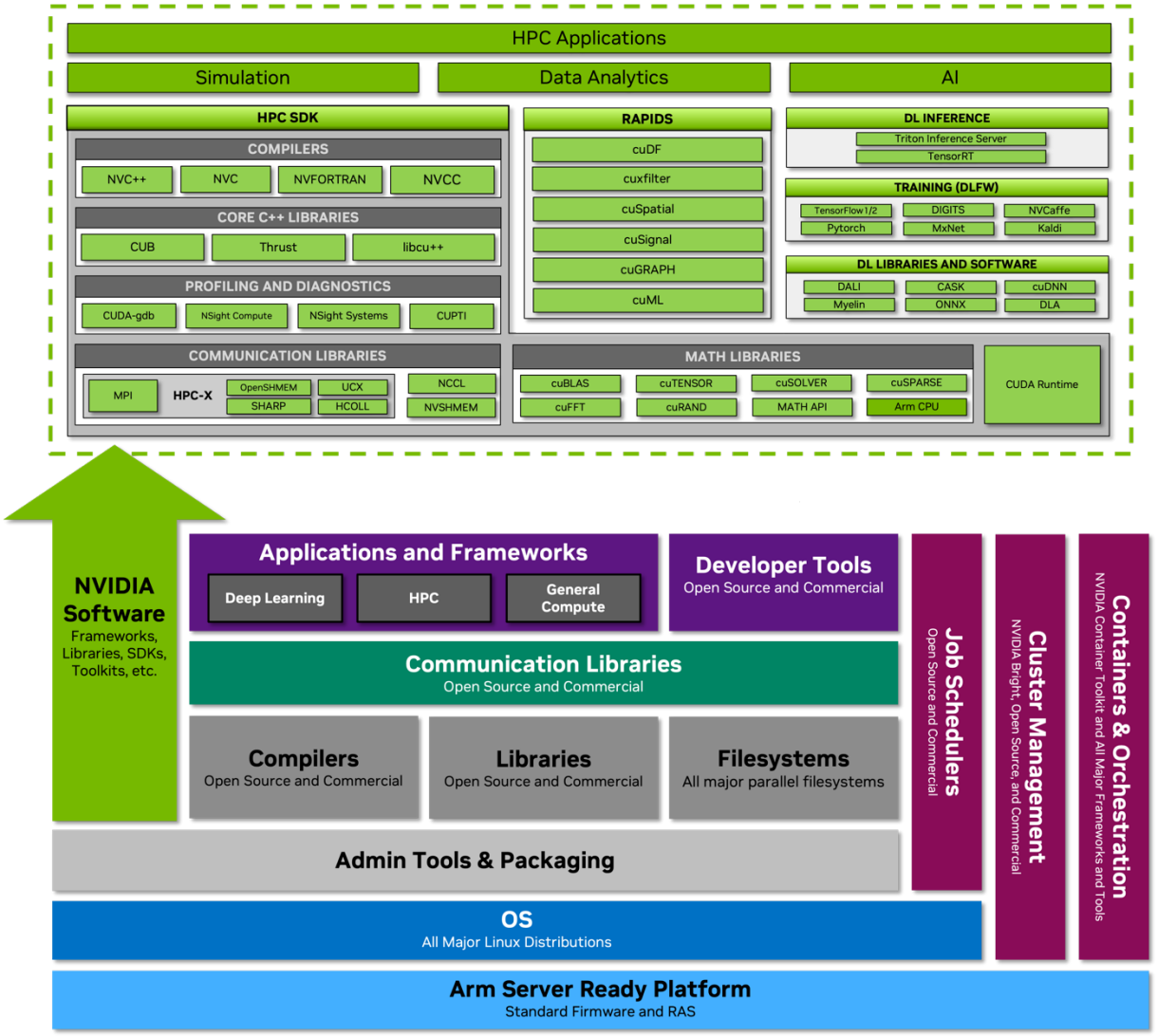

Figure 7. The NVIDIA Grace CPU software ecosystem combines the full collection of NVIDIA software for CPU, GPU, and DPU with the complete Arm datacenter ecosystem

NVIDIA Grace CPU follows mainstream CPU design principles and is programmed just like any other server CPU. Broadly speaking, applications built using interpreted or JIT'd languages such as Python, Java, PHP, and Node.js, run as-is on Grace CPU.

Applications using compiled languages including C/C++, Fortran, Go, and Rust can use existing binaries targeting Armv8 or later. These applications can be recompiled to improve performance. For maximum performance on NVIDIA Grace CPU, compiled

applications should use compiler flags that target the Armv9 ISA and tune for the Neoverse V2 microarchitecture using any open source or vendor compiler. This includes GCC, LLVM, NVHPC, Arm Compiler for Linux, and HPE Cray Compilers.

The specific advantages of Grace CPU for programmability include:

- **Strong support for open-source and commercial toolchains.** The GCC and LLVM compilers, and the NVIDIA HPC compilers (part of the NVIDIA HPC SDK) are tuned for Grace CPU. These compilers can auto-vectorize code with SVE2, leverage LSE atomic instructions in compiler synchronization built-in functions, and they provide command line options to tune instruction selection and ordering for the Grace CPU microarchitecture. The NVIDA HPC SDK further accelerates applications targeting Grace CPU by providing accelerated math libraries. These libraries provide tuned kernels that leverage Grace SVE2, NVLink-C2C, and high bandwidth memory.
- **Excellent SIMD instruction-level parallelism.** Grace CPU achieves a total SIMD vector bandwidth of 512-bits without the need to restructure code for wide vectors, and without paying a high energy cost when not fully utilizing them. All existing Arm SIMD code written using NEON or newer code targeting SVE2 will run well.
- **Best-in-class memory bandwidth.** The NVIDIA Grace Superchip has one type of memory, and all cores can access it with high bandwidth. Legacy applications can realize the full benefit of the advanced LPDDR5X memory technology.

Applications built on MPI, OpenMP, OpenACC, OpenSHMEM, POSIX threads, or any parallel programming model that currently executes on a data center CPU, are supported on the Grace CPU Superchip. Additionally, modern Python frameworks are increasingly exploiting parallelism using high-level constructs, which are fully supported on Grace CPU. Lastly, applications developed using modern ISO language parallelism in C++ and Fortran will run well on both the Grace Superchip and Hopper GPU platforms with a single, standards-based source code. Adopting modern, scalable programming models enables application performance to improve over time as new hardware advances are realized, while also improving developer productivity and reducing code complexity.

Debugging and Profiling

As a standards compliant Armv9 CPU, NVIDIA Grace is fully supported by many commercial and open-source developer tools. Grace implements the Arm CoreSight Performance Monitoring Unit (PMU), which exposes a standard set of hardware performance events. This includes counters for SVE2 and NEON instructions and L1/L2/L3 cache operations. Profilers can measure these events through standard APIs such as Linux's perf_events, to characterize CPU core performance. Grace CPU also implements Arm Statistical Profiling Extensions (SPE), which provides hardware-based statistical sampling for low-overhead, high resolution performance measurement.

Community maintained and supported tools like HPCToolkit, LIKWID, TAU Performance System®, Score-P, and mpiP, work well on Grace CPU. These tools are built on libraries and technologies like Linux Perf, PAPI: Performance Application Programming Interface, libunwind, and GNU Binary Utilities, that have been part of the Arm software ecosystem for years and are actively used by developers and Arm community members worldwide.

Commercial tools typically found today at major HPC centers worldwide support NVIDIA Grace CPU, including Arm Forge, TotalView by Perforce, and HPE CrayPat. These high-quality commercial tools support parallel applications written in all major programming languages and have been demonstrated on thousands of CPU cores.

The NVIDIA Nsight family of performance analysis tools helps users identify coarse- and fine-grained optimization opportunities in their applications. NVIDIA Nsight Systems is a system-wide performance analysis tool designed to visualize an application's algorithms across many GPUs, CPUs, DPUs, Memory, Network I/O, and File I/O. NVIDIA Nsight Systems is fully integrated with the NVIDIA software ecosystem; supporting tracing, sampling, and visualizing system, library, and framework API calls such as CUDA, CUDA-X, RAPIDS, Magnum-IO, GPU Direct, MPI, UCX, OpenSHMEM, OpenMP, OpenACC, OS events, and even call-stack sampling. NVIDIA tools provide the same workflow experience on the NVIDIA Grace CPU as on other CPUs.

Superchips for HPC, Cloud Computing, and Enterprise

The NVIDIA Grace CPU was designed to serve a wide array of use cases, including accelerated and tightly-coupled use cases featuring the NVIDIA Grace Hopper Superchip, as well as the broader high-performance and power efficiency needs of the general-purpose CPU market.

The NVIDIA Grace CPU Superchip builds on the existing Arm ecosystem to create the first no-compromise Arm CPU for high performance computing (HPC), demanding cloud workloads, and last-mile compute usages that require high-performance and power efficient dense infrastructure.

The NVIDIA Grace CPU Superchip combines two NVIDIA Grace CPUs connected over 900 GB/s bi-directional bandwidth NVLink-C2C to deliver 144 high-performance Arm Neoverse V2 cores with up to 1TB/s bandwidth of data center class LPDDR5X memory with ECC. Looking at a range of HPC applications including OpenFOAM, NEMO, WRF, and BWA, NVIDIA Grace CPU delivers up to 2.5X in speedup and over 3X improvement in energy efficiency compared to existing high-performance data center CPUs.

Grace CPU Superchip systems will be available from an array of OEMs and vendors with form factors and feature sets similar to popular one and two-socket x86-64 server designs. These superchips also feature novel configurations that allow users in the cloud computing, HPC, and enterprise spaces to capitalize on Grace CPU Superchip's category leading performance-per-watt and packaging integration. Users with more specific needs can start from and customize several NVIDIA reference designs, including conventional 1U and 2U enterprise servers, 2U 4-node dense HPC systems, and more through their original design manufacturer of choice.

Finally, for the most flexibility, platform teams can easily and rapidly build the Grace CPU Superchip directly into their own designs. Grace integrates the heart of a conventional server - including some of the most difficult aspects of system design like voltage regulation, cross-socket connectivity, and DRAM routing - in an easy-to-consume prepackaged module.

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation (“NVIDIA”) makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customers should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer (“Terms of Sale”). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, NVIDIA Grace GPU, CUDA, NVLink, NVIDIA GPU Cloud, and NSight are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

VESA DisplayPort

DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

Arm

Arm, AMBA, and Arm Powered are registered trademarks of Arm Limited. Cortex, MPCore, and Mali are trademarks of Arm Limited. All other brands or product names are the property of their respective holders. "Arm" is used to represent Arm Holdings plc; its operating company Arm Limited; and the regional subsidiaries Arm Inc.; Arm KK; Arm Korea Limited.; Arm Taiwan Limited; Arm France SAS; Arm Consulting (Shanghai) Co. Ltd.; Arm Germany GmbH; Arm Embedded Technologies Pvt. Ltd.; Arm Norway, AS, and Arm Sweden AB.

OpenCL

OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc.

Copyright

© 2024 NVIDIA Corporation. All rights reserved.

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com