

NVIDIA GH200 Grace Hopper Superchip Benchmark Step-by-Step Guide

Application Note

Document History

DA-11356-002_10

Version	Date	Description of Change
01	June 6, 2023	Initial release
02	October 2, 2023	> Updated the "Introduction" section.
		> Updated the "GPU STREAM" section.
		> Added the "CPU STREAM" section.
		> Updated the sustained GEMM.
		> Added CuFFT and the attachment.
03	October 17, 2023	> Added the DALI app.
		> Removed Zero copy GEMM.
04	February 15, 2024	Updated scripts attachment.
05	July 16, 2024	Removed the following attachments from the PDF file:
		> stream_test.nv7z
		> cublasMatmulBench.nv7z
06	September 5, 2024	> Added HPL Performance and the attachment.
		> Added HPL- MxP Performance and the attachment.
		> Added Llama-3 8B inference performance and the attachment.
		> Removed the cufftBench.nv7z attachment from the PDF file.
07	September 26, 2024	> Updated Table 8.
08	October 10, 2024	> Fixed the attachment files.
		> Removed the command lines for HPL and HPL-MxP and inserted into the HPL_HPL-MxP attachment.
		> Updated the GPU STREAM command.
09	December 9, 2024	> Updated the "CPU STREAM" section.
		> Updated the "NVBandwidth" section.
		> Updated the "NVIDIA DALI for ResNet50" section.
		> Updated the "Llama-3 8B Inference" section.
		> Updated the "Attachments" section.
10	March 21, 2025	> Updated the "Introduction" section.

Table of Contents

Introduction	. 1
Attachments	.2
Libraries and Benchmarks	.3
GPU STREAM	.3
CPU STREAM	.4
NVBandwidth	.4
High Performance LINPACK	. 5
High Performance LINPACK – Mixed Precision	. 5
Application Performance	.7
NVIDIA DALI for ResNet50	.7
Llama-3 8B Inference	.7

List of Tables

Table 1.	System Specifications	1
Table 2.	GPU STREAM Benchmark	3
Table 3.	CPU STREAM Benchmark	4
Table 4.	NVBandwidth	5
Table 5.	HPL Performance	5
Table 6.	HPL-MxP Performance	6
Table 7.	DALI for ResNet50	7
Table 8.	Llama-3 8B Inference at Latency Requirement	8

Introduction

This application note compares NVIDIA[®] GH200 benchmark data to the NVIDIA DGX[™] H100 platform.

DGX H100 benchmark numbers were measured on an Intel[®] Xeon[®] Platinum 8480C system. The clocks were set to the maximum at 1,980 MHz for GPU and 2,619 MHz for GPU memory with ECC enabled on Ubuntu 22.04, NVIDIA CUDA[®] 12, and NVIDIA Driver 525.85. All NVIDIA H100 SXM5 80 GB benchmark numbers are provided only as comparisons to GH200.

The NVIDIA GH200 Grace Hopper[™] Superchip architecture brings together the groundbreaking performance of the NVIDIA Hopper[™] GPU with the versatility of the NVIDIA Grace[™] CPU, connected with a high bandwidth, the memory coherent NVIDIA[®] NVLink[®] Chip-2-Chip (C2C) interconnect in a superchip, and support for the new NVIDIA NVLink Switch System. The NVIDIA GH200 system was set with Ubuntu 22.04, CUDA 12.3, and NVIDIA Driver 545.14. The maximum NVIDIA GH200 Grace Hopper Superchip power usage is 1000 W (for memory + CPU + GPU) with the GPU consuming up to 900 watts.

Application performance variation is expected for partner data center designs using NVIDIA products (GPUs, CPUs, NICs, and DPUs). Observed performance variations can be due to several factors, including ASIC production variation, general design, and component tolerances, partner thermal design implementations, environmental operating conditions, statistical run-to-run variation, and so on.

Important: All benchmark numbers are preliminary and represent performance at the launch of the GH200 Grace Hopper Superchip and will be updated after the products become generally available. The CUDA and NVIDIA driver software stack with Deep Learning (DL) frameworks and applications are continuously updated, so the performance will vary over time.

Refer to the following pages for the latest DL and high-performance computing (HPC) performance results:

> https://developer.nvidia.com/deep-learning-performance-training-inference

> <u>https://developer.nvidia.com/hpc-application-performance</u>).

System Specification	NVIDIA DGX H100	NVIDIA GH200 Grace Hopper Superchip
GPU	8x NVIDIA H100 80 GB	1x NVIDIA H100 96 GB
СРИ	Dual Intel Xeon Platinum 8480C, 2 GHz. 56 cores	1x Grace CPU, 3.1 GHz, 72 cores

Table 1. System Specifications

System Specification	NVIDIA DGX H100	NVIDIA GH200 Grace Hopper Superchip
System memory	2 TB DDR5	120 GB LPDDR5X
		480 GB LPDDR5X

Attachments

The following files are attached to this application note:

- > HPL_HPL-MxP.nv7z
- > DL_scripts_v4.nv7z

To access the attached files, click the **Attachment** icon on the left-hand toolbar on this PDF (using Adobe Acrobat Reader or Adobe Acrobat). Select the file and use the Tool Bar options (**Open, Save**) to retrieve the documents. Rename the files with the .nv7z extension with the .zip format and extract the files using an archive utility.

Libraries and Benchmarks

NVIDIA CUDA-X[™], which is built on CUDA, is a collection of libraries, tools, and technologies that deliver dramatically higher performance compared to CPU-only alternatives across multiple application domains from artificial intelligence (AI) to HPC.

There are also many CUDA code samples in the CUDA toolkit. To highlight GH200, the following sections provide information about a few options.

GPU STREAM

NVIDIA provides an optimized CUDA implementation for the STREAM benchmark to measure memory bandwidth on one Hopper GPU.

Usage

To run STREAM, use the hpc-benchmarks:24.06 container from https://catalog.ngc.nvidia.com/orgs/nvidia/containers/hpc-benchmarks. -n<elements>: number of double precision-floating point elements

Command Line

docker run --gpus all --shm-size=1g nvcr.io/nvidia/hpc-benchmarks:24.06 ./stream-gputest.sh --n 1308622848

Interpreting the Results

NVIDIA H100 SXM5 80GB in a DGX H100 system has 80GB of HBM3 with a peak memory bandwidth of 3,352 GB/s, and the GH200 Hopper GPU has 96GB of HBM3 with a peak memory bandwidth of 4,023 GB/s.

	GPU Memory B	andwidth (GB/s)
STREAM	DGX H100 80GB	GH200 96 GB
Сору	3067	3666
Scale	3060	3667
Add	3128	3754
Triad	3132	3755

Table 2. GPU STREAM Benchmark

CPU STREAM

The STREAM benchmark is a simple, synthetic benchmark program that measures sustainable main memory bandwidth in MB/s and the corresponding computation rate for simple vector kernels on one CPU.

Usage

To run STREAM, use the hpc-benchmarks:24.06 container from <u>https://catalog.ngc.nvidia.com/orgs/nvidia/containers/hpc-benchmarks</u>. -n<elements>: number of double precision-floating point elements

Command Line

```
docker run nvcr.io/nvidia/hpc-benchmarks:24.06 ./stream-cpu-test.sh --n 1308622848
```

Interpreting the Results

The GH200 Grace CPU has options for 120GB of LPDDR5X with a peak memory bandwidth of 512 GB/s or 480GB of LPDDR5X with a peak memory bandwidth of 384 GB/s.

	CPU Memory Ba	andwidth (GB/s)
STREAM	GH200 120 GB	GH200 480 GB
Сору	448	342
Scale	448	345
Add	442	336
Triad	444	340

Table 3. CPU STREAM Benchmark

NVBandwidth

For the source code for NVBandwidth, go to <u>https://github.com/NVIDIA/nvbandwidth</u>.

This tool measures bandwidth on NVIDIA GPUs.

Commands to Run Test

To run the test, run the following commands.

\$ git clone <u>https://github.com/NVIDIA/nvbandwidth</u>

```
$ cd nvbandwidth
```

```
$ sudo ./debian_install.sh
```

\$./nvbandwidth

Interpreting the Results

NVLink-C2C is an NVIDIA memory coherent, high-bandwidth, and low-latency superchip interconnect that delivers up to 900 GB/s of bidirectional bandwidth.

When looking at host_to_device_memcpy_sm and device_to_host_memcpy_sm, each row represents the measured single directional bandwidth between the host and the device for one GPU.

Table 4. NVBandwidth

	NVLink-C2C Ba	ndwidth (GB/s)
NVBandwidth	GH200 120GB	GH200 480GB
Host to device	419	335
Device to host	371	347

High Performance LINPACK

To run this benchmark, download the HPL_HPL-MxP.nv7z file that is attached to this PDF for instructions (refer to "Attachments" on page 2 for information about how to open the file).

NVIDIA has a GPU-accelerated implementation of High Performance LINPACK (HPL), which primarily stress tests the system's FP64 throughput.

Interpreting the Results

HPL was measured using the nvcr.io/nvidia/hpc-benchmarks:24.06 container. HPL solves the Ax=B liner system of equations, and HPL performance is bounded by DGEMM performance. The performance difference between HPL and DGEMM is due to different matrix sizes, input coefficients, type of initialization, and time of execution.

GH200 shows higher performance for the following reasons:

- > Larger GPU memory enabling larger problem sizes
- > Higher maximum power usage on GH200 and dynamic power sharing between CPU and GPU
- > Higher CPU core count per GPU

Number of	HPL Performa	ance (TFLOPs)
GPUs	DGX H100	GH200
1	47	52

High Performance LINPACK: Mixed Precision

To run this benchmark, download the HPL_HPL-MxP.nv7z file that is attached to this PDF for instructions (refer to "Attachments" on page 2 for information about how to open the file).

The HPL-MxP benchmark highlights the emerging convergence of HPC and artificial intelligence (AI) workloads. NVIDIA also has a GPU-accelerated implementation of High Performance LINPACK – Mixed Precision (HPL-MxP) that uses mixed-precision iterative and direct methods to utilize mixed-precision tensor cores.

Interpreting the Results

HPL-MxP was measured using the nvcr.io/nvidia/hpc-benchmarks:24.06 container. HPL-MxP solves the Ax=B liner system of equations with LU factorization in FP32, which uses FP16 or FP8 GEMM and the FP64 GMRES solver internally.



Like HPL, the performance difference between HPL-MxP and FP8 GEMM is due to different matrix sizes, input coefficients, type of initialization, and time of execution. GMRES is a bandwidth bound operation, which by default runs on CPUs and significantly affects performance.

GH200 shows higher performance for the following reasons:

- > Larger GPU memory enabling larger problem sizes.
- > Higher maximum power usage on GH200 and dynamic power sharing between CPU and GPU.
- > Higher CPU core count per GPU.
- > CPU to GPU communication is faster on GH200.

Table 6.HPL-MxP Performance

Number of	HPL-MxP Perfor	mance (TFLOPs)
GPUs	DGX H100	GH200
1	307	430

Application Performance

NVIDIA DALI for ResNet50

To run this benchmark, download the DL_scripts_v4.nv7z file that is attached to this PDF for instructions (refer to "Attachments" on page 2 for information about how to open the file).

NVIDIA Data Loading Library (DALI) is a portable, open-source library that is used to decode and augment images, videos, and speech to accelerate DL applications. DALI reduces latency and training time, which mitigates bottlenecks by overlapping training and preprocessing. It provides a drop-in replacement for built-in data loaders and data iterators in popular DL frameworks for easy integration or retargeting to different frameworks.

Interpreting the Results

DALI (v1.30) for ResNet50 using a single Hopper GPU was run on DGX H100 and GH200 using FP16. Faster data access to the CPU memory through NVLink-C2C and a higher CPU to GPU ratio with GH200 boost the data processing performance by 1.5x.

|--|

	Images/s		
DALI for ResNet50	DGX H100	GH200	
Typical ResNet50 data processing pipeline running on ImageNet like JPEG test data set (VGA, WXGA, HD). Image decoding > random resized crop > normalization and random flip to 224x224, NCHW format, FP16	19,885	29,757	

Llama-3 8B Inference

For the source code for NVIDIA TensorRT[™]-LLM, go to <u>https://github.com/NVIDIA/TensorRT-LLM</u>.

To run this benchmark, download the DL_scripts_v4.nv7z file that is attached to this PDF for instructions (refer to "Attachments" on page 2 for information about how to open the file).

Interpreting the Results

Llama-3 8B inference using one Hopper GPU was run on DGX H100 and GH200 using FP8 with various first token latency requirements.

	DGX H100		GH200	
Llama-3 8B Inference	First Token Latency (ms)	Throughput (Tokens/s)	First Token Latency (ms)	Throughput (Tokens/s)
BS=8, ISL/OSL=2048/128, TP=1, PP=1	220	1030	209	1114
(2s first token latency requirement)				
BS=64, ISL/OSL= 128/128, TP=1, PP=1	107	8732	101	9895
(1s first token latency requirement)				
BS=64, ISL/OSL= 128/2048, TP=1, PP=1	112	6424	126	6981
(1s first token latency requirement)				
BS=4, ISL/OSL= 2048/2048, TP=1, PP=1	114	681	108	752
(1s first token latency requirement)				

Table 8.Llama-3 8B Inference at Latency Requirement

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.

Customer 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.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

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 with out 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, CUDA, CUDA-X, DGX, Grace, Grace Hopper, NVIDIA Hopper, NVLink, and NVLink-C2C 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.

Copyright

© 2025 NVIDIA Corporation. All rights reserved.

