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

Application Note

DA-11356-002_v04 | February 2024

Document History

DA-11356-002_v04

Version	Date	Authors	Description of Change
01	June 6, 2023	EC, MT, SM	Initial release
02	October 2, 2023	SB, SM	> Updated “Introduction” section > Updated “GPU STREAM” section > Added “CPU STREAM” section > Updated sustained GEMM > Added CuFFT and attachment
03	October 17, 2023	SB, SM	> Added DALI app > Removed Zero copy GEMM
04	February 15, 2024	SB, SM	Updated scripts attachment

Table of Contents

List of Tables

Introduction

This application note provides NVIDIA GH200 benchmark data in comparison to the NVIDIA® DGX™ H100 platform. This initial version of the application note provides benchmarks for low-level performance metrics for bandwidth and throughput. However, this application note will be updated over time to include more workloads and application performance data.

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, and memory coherent NVIDIA® NVLink® Chip-2-Chip (C2C) interconnect in a single superchip, and support for the new NVIDIA NVLink Switch System. The NVIDIA GH200 system is set with Ubuntu 22.04, NVIDIA® CUDA® 12.3, and NVIDIA Driver 545.14. All NVIDIA GH200 benchmark numbers provided in this application note are preliminary and subject to change.

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, CUDA 12, and NVIDIA Driver 525.85. All NVIDIA H100 SXM5 80 GB benchmark numbers are preliminary and are only presented for comparisons to GH200.

Partner benchmark results will vary based on a variety of factors such as ambient temperature, hardware, software, thermal design, and server configurations. These benchmark numbers are meant only as a reference data point.

Note: Run-to-run variation up to 3% in delivered performance on the same system is considered normal.

Important: All benchmark numbers are preliminary and represent performance at the launch of NVIDIA GH200 Grace Hopper Superchip and will be updated once the products become generally available. The CUDA and NVIDIA driver software stack along with DL frameworks and applications are continuously updated, and thus performance will vary over time. Refer to the NVIDIA performance page (https://developer.nvidia.com/deep-learning-performance-training-inference and https://developer.nvidia.com/hpc-application-performance) for the latest DL and HPC performance results.

Table 1. System Specifications

System Specification	NVIDIA DGX H100	NVIDIA GH200 Grace Hopper Superchip
GPU	8x NVIDIA H100 80 GB	1x NVIDIA H100 96 GB
CPU	Dual Intel Xeon Platinum 8480C, 2 GHz, 56 cores	Grace CPU, 3.1 GHz, 72 cores
System memory	2 TB DDR5	120 GB LPDDR5X 480 GB LPDDR5X

Libraries and Benchmarks

NVIDIA® CUDA-X™, built on top of 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 high performance computing (HPC).

There are also many CUDA code samples included as part of the CUDA toolkit. We have presented a few to highlight GH200.

GPU STREAM

NVIDIA provides an optimized CUDA implementation for the STREAM benchmark for measuring memory bandwidth on a single NVIDIA Hopper GPU. In addition to the four kernels included within STREAM, this implementation also includes basic load and store tests to measure read and write memory bandwidth.

Usage:

-n<elements>: number of double precision-floating point elements

-d<device>: determine which GPU to use

-r<random>: use random inputs or not

Command Line:

$ ./stream_vectorized_double_test -n1308622848

Interpreting Results:

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

Table 2. GPU STREAM Benchmark

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

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 a single CPU.

The following command downloads and compile STREAM with a total memory footprint of approximately 2.7GB, which is sufficient to exceed the L3 cache without excessive runtime. The general rule for running STREAM is that each array must be at least 4x the size of the sum of all the last-level caches used in the run, or 1 million elements, whichever is larger.

To run STREAM, set the number of OpenMP threads (OMP_NUM_THREADS) and the numactl flags according to the following example. Use OMP_PROC_BIND=spread to distribute the threads evenly over all available cores and maximize bandwidth.

Command Lines:

```
wget https://www.cs.virginia.edu/stream/FTP/Code/stream.c &&
gcc -Ofast -march=native -fopenmp \
        -DSTREAM_ARRAY_SIZE=120000000 -DNTIMES=200 \
        -o stream_openmp.exe stream.c

OMP_NUM_THREADS=72 OMP_PROC_BIND=spread numactl -m0 ./stream_openmp.exe
```

Interpreting Results:

NVIDIA GH200 Grace CPU has 120 GB of LPDDR5X with peak memory bandwidth of 512 GB/s or 480 GB of LPDDR5X with peak memory bandwidth of 384 GB/s.

Table 3. CPU STREAM Benchmark

STREAM	CPU Memory Bandwidth (GB/s)	
	GH200 120 GB	GH200 480 GB
Copy	459	328
Scale	450	345
Add	437	326
Triad	438	340

Basic Linear Algebra

The NVIDIA cuBLAS library is a fast GPU-accelerated implementation of the standard basic linear algebra subroutines (BLAS). NVIDIA provides the test binary, and cublasMatmulBench, that controls all test parameters for general matrix multiplication (GEMM) where the input data are all random numbers. This is provided for customers who would like to run GEMM tests for informational purposes only. However, we encourage customers to focus on delivered performance on real workloads. Even though Peak TFLOPs and GEMM performance are important to maximizing throughput, it is not a reliable predictor of application performance. There are many factors including software frameworks, memory bandwidth, kernel launch times, system CPU <-> GPU bandwidth, and architectural changes between GPU generations.

Usage:

-P=<bisb_imma,hsh,sss,ddd>: input precision, compute precision, output precision

-m=<int> -n=<int> -k=<int>: MNK parameters

-T=1000: number of times to run back-to-back

-ta=1: set GEMM layout to TN

-tb=1: set GEMM layout to NT

-B=0: set beta to zero

-za=1 -zb=1 -zc=1 -zd=1: allocate matrices in CPU memory and then zero-copy data to GPU memory

Command Lines:

FP8: ./cublasMatmulBench -P=qqssq -m=4224 -n=2048 -k=16384 -T=1000 -ta=1 -B=0

INT8: ./cublasMatmulBench -P=bisb_imma -m=8192 -n=4224 -k=16384 -T=1000 -ta=1 -B=0

FP16: ./cublasMatmulBench -P=hsh -m=12288 -n=9216 -k=32768 -T=1000 -tb=1 -B=0

BF16: ./cublasMatmulBench -P=tst -m=12288 -n=9216 -k=32768 -T=1000 -tb=1 -B=0

TF32: ./cublasMatmulBench -P=sss_fast_tf32 -m=8192 -n=4224 -k=16384 -T=1000 -ta=1 -B=0

FP64: ./cublasMatmulBench -P=ddd -m=4224 -n=2048 -k=16384 -T=1000 -tb=1 -B=0

FP32: ./cublasMatmulBench -P=sss -m=4224 -n=2048 -k=16384 -T=1000 -tb=1 -B=0

Interpreting Results:

GEMM results were run on a single NVIDIA Hopper GPU. Performance may vary if running problem sizes other than the ones provided, which were selected to provide the best performance on an NVIDIA Hopper GPU.

Table 4. Sustained GEMM TFLOPs

Datatype	GH200 Peak TFLOPs	GH200 GEMM TFLOPs	%SOL
FP8	1979	1514	76%
INT8	1979	1685	85%
FP16	989	848	86%
BF16	989	813	82%
TF32	495	480	97%
FP64	67	66	98%
FP32	67	54	80%

Fast Fourier Transforms

The CUDA Fast Fourier Transform library (cuFFT) provides GPU-accelerated FFT implementations. NVIDIA provides the test binary, cufftBench, that controls all test parameters and a test script that performs a specific FFT problem size.

Usage:

-R=<single|multi>: single or multi-GPU

-xsize=<int> -ysize=<int> -zsize=<int>: array of sizes

-type=<c2cf|c2ci|r2c|c2r>: complex and real-valued input and output

-precision=<half|single|double>

-rank=<1d|2d|3d>: 1D, 2D and 3D transforms

-ngpus=<int>: number of GPUs

–zerocopy=1: allocate data in CPU memory and then zero-copy data to GPU memory

Command Lines:

$ ./cufftBench -R=single -xsize=512 -ysize=512 -zsize=512 -type=c2cf -precision=double -rank=3d

Interpreting Results:

cuFFT results using a single NVIDIA Hopper GPU were run on NVIDIA DGX H100 and NVIDIA GH200.

Table 5. Sustained FFT GFLOPs

FFT Performance (GFLOPs)	
DGX H100	**GH200**
4065	4867

NVBandwidth

This application is a tool for bandwidth measurements on NVIDIA GPUs.

Commands to Run Test:

```
$ git clone https://github.com/NVIDIA/nvbandwidth
$ sudo ./debian_install.sh
$ cmake .
$ make
$ ./nvbandwidth
```

Interpreting Results:

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

Looking at “host_to_device_memcpy_sm” and “device_to_host_memcpy_sm,” each row represents measured single directional bandwidth between host and device for a single GPU.

Table 6. NVBandwidth

	NVLink-C2C Bandwidth (GB/s)
NVBandwidth	**GH200**
Host to device	419
Device to host	371

Attachments

The following files are attached to this application note:

- stream_test.nv7z
- cublasMatmulBench.nv7z
- cufftBench.nv7z
- scripts_for_apps_v3.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. Files with the .nv7z extension must be renamed to .7z and extracted using the 7-Zip file archive software.

Application Performance

DALI for ResNet50

Instructions to run this benchmark are within “scripts_ for_apps_v3.nv7z” attached file.

The NVIDIA Data Loading Library (DALI) is a portable, open-source library for decoding and augmenting images, videos, and speech to accelerate deep learning applications. DALI reduces latency and training time, mitigating bottlenecks by overlapping training and pre-processing. It provides a drop-in replacement for built-in data loaders and data iterators in popular deep learning frameworks for easy integration or retargeting to different frameworks.

Interpreting Results:

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

Table 7. DALI for ResNet50

DALI for RN50	Images/s	
	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

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 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, 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

© 2023, 2024 NVIDIA Corporation. All rights reserved.