HIGH PERFORMANCE RESEARCH COMPUTING

ACES: Intro to the Grace Hopper Superchip

Dinesh S. Devarajan

April 21, 2026

High Performance
Research Computing
DIVISION OF RESEARCH

TAMU HPRC GH200 System Setup

System Specifications

- System is a S74-2U Grace-Hopper MGX, NVIDIA's modular GPU accelerated server design
- Purchased through the vendor QCT
- Operating system is RHEL 9.4 - Ubuntu 24.04 would not let us build Lustre & MOFED
- Kernel version is 5.14.0-427.31.1.el9_4.aarch64 (4k page - 64k page had issues with Lustre)
- File system is Lustre, version 2.15.5-1
- OFED version is 23.10.3.2.2.1

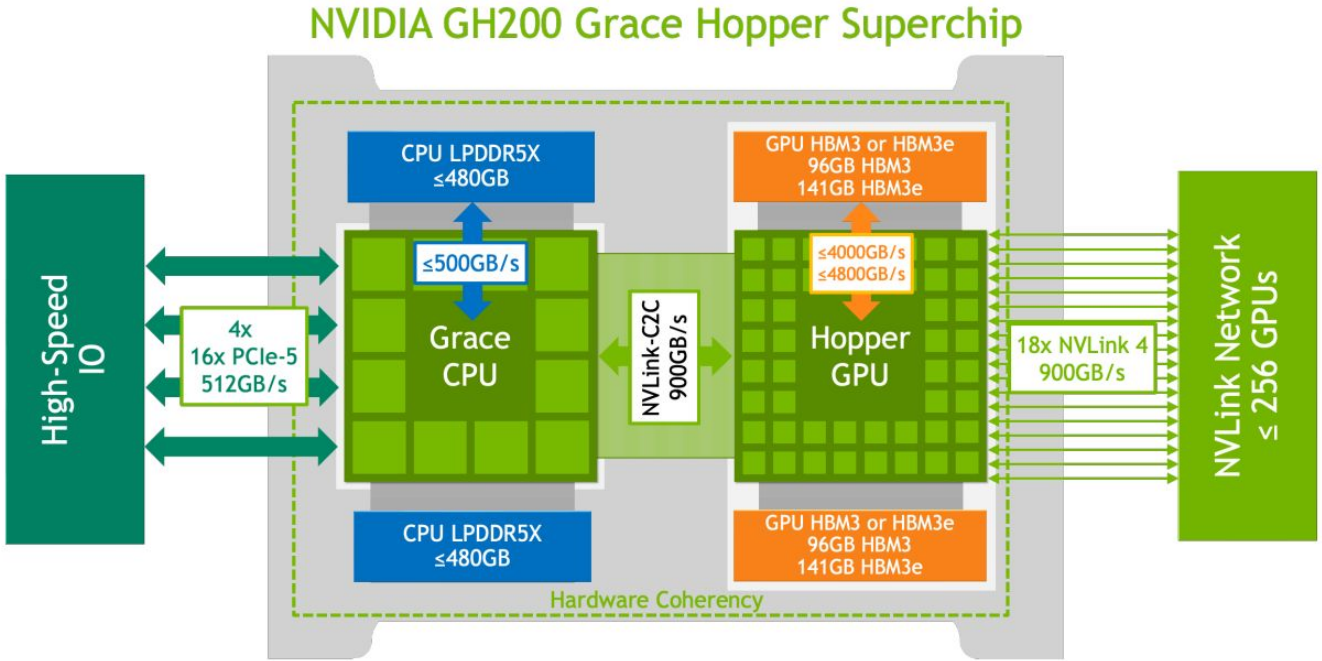
Overview of NVIDIA GH200 Specifications

Feature	Description
Grace CPU cores (number)	Up to 72 cores
CPU LPDDR5X bandwidth (GB/s)	Up to 500GB/s
GPU HBM bandwidth (GB/s)	4TB/s HBM3 4.9TB/s HBM3e
NVLink-C2C bandwidth (GB/s)	900GB/s total, 450GB/s per direction
CPU LPDDR5X capacity (GB)	Up to 480GB
GPU HBM capacity (GB)	96GB HBM3 144GB HBM3e
PCIe Gen 5 Lanes	64x

Table - GH200 Key Features [1]

https://www.amax.com/content/files/2023/12/NVIDIA_Grace_Hopper_Superchip_Architecture_Overview_hitepaper.pdf

Logical Overview of the NVIDIA GH200



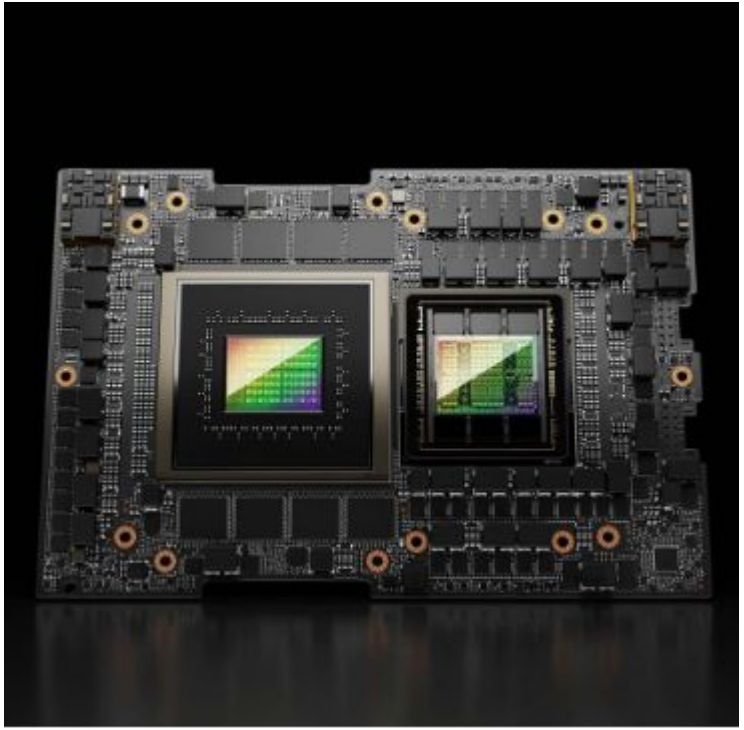
Logical diagram of GH200 Superchip

https://www.amax.com/content/files/2023/12/NVIDIA_Grace_Hopper_Superchip_Architecture_Overview_Whitepaper.pdf

NVLink-C2C

Fusion of Grace CPU & Hopper GPU

- NVLink-C2C Interconnect: The NVLink Chip-2-Chip (C2C) interconnect provides a high-bandwidth direct connection between a Grace CPU and a Hopper GPU
- Uses a 900GB/s chip-to-chip bandwidth for data transfer between the CPU and GPU
- NVLink-C2C provides 7x the bandwidth of x16 PCIe Gen links at lower latency
- NVLink-C2C uses 1.3 picojoules per bit transferred, which is greater than 5x more energy efficient than PCIe Gen 5

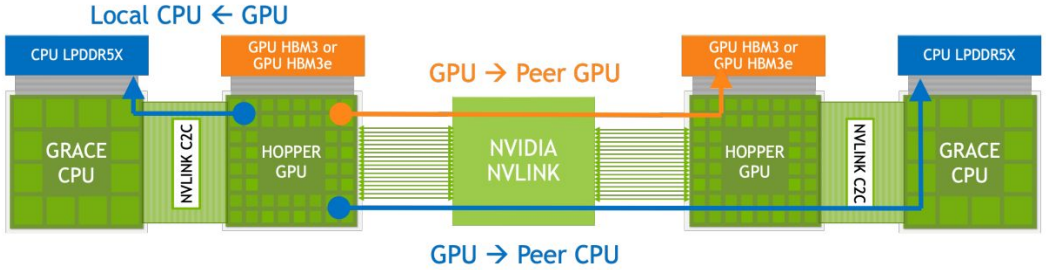


NVIDIA GH200 Supership

NVIDIA GH200 Architecture

Extended GPU Memory (EGM) & System Scalability

- Allows GPUs to access up to 144TB system memory across NVLink-C2C fabric
- Enables GPUs to efficiently access memory beyond the capabilities of single superchip's HBM3 or LPDDR5x memory
- EGM Access Speed: Minimum GPU-GPU NVLink or NVLink-C2C speed at 450GB/s for memory transfers

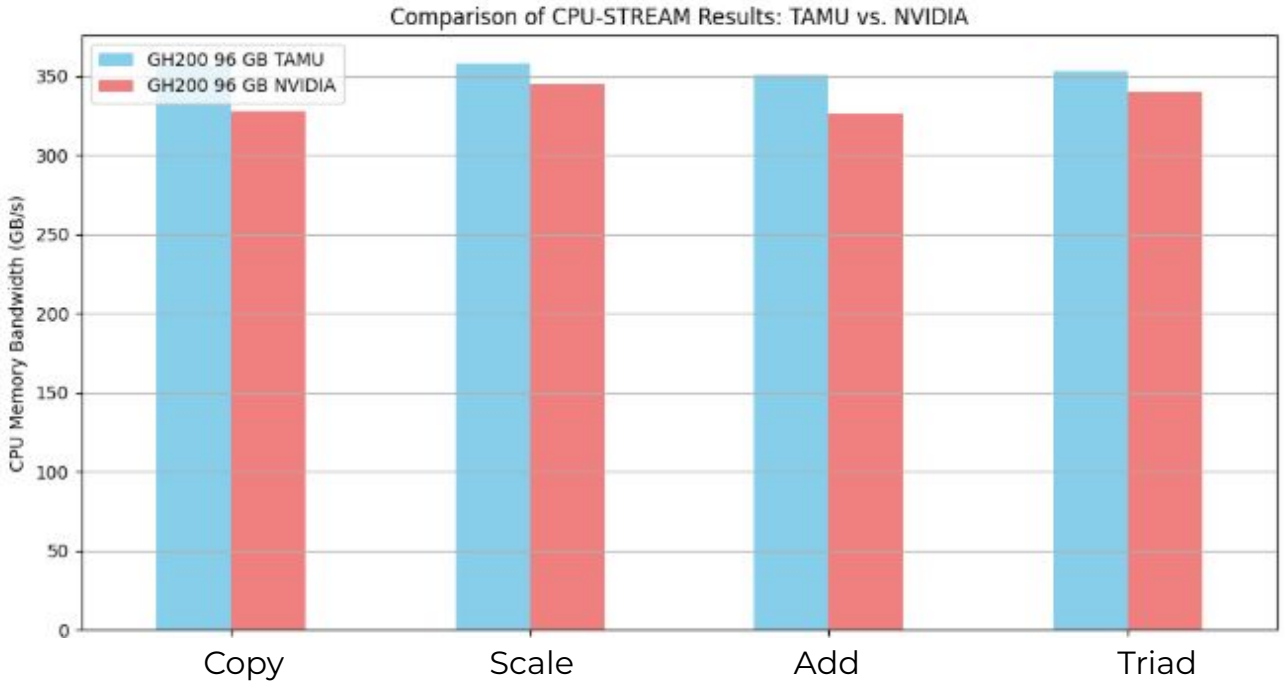


Memory across NVLink

NVIDIA GH200 Architecture

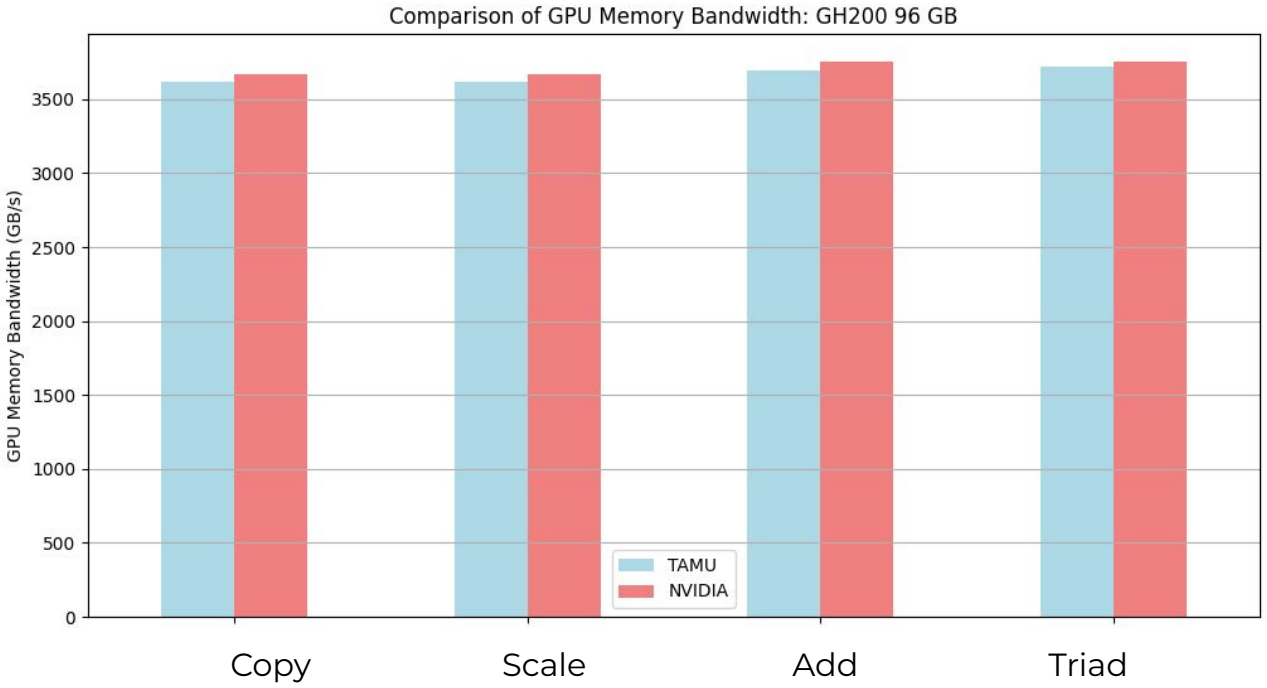
- **Balanced Power Between CPU & GPU**
  - The Grace CPU supports Memory Resource Partitioning and Monitoring (MPAM) feature that provides performance isolation between jobs. MPAM allows users to partition the available LPDDR5X bandwidth and CPU cache usage
  - The Hopper GPU supports Multi-Instance GPU (MIG), which can be used to partition the GPU into more instances (up to 7 “slices” - 12GB each)
  - Supports a 1:1 GPU-CPU ratio, ideal for heterogeneous workloads, where both CPU and GPU can operate efficiently together, offering power efficiency and improving overall performance.

CPU STREAM Benchmarking Results



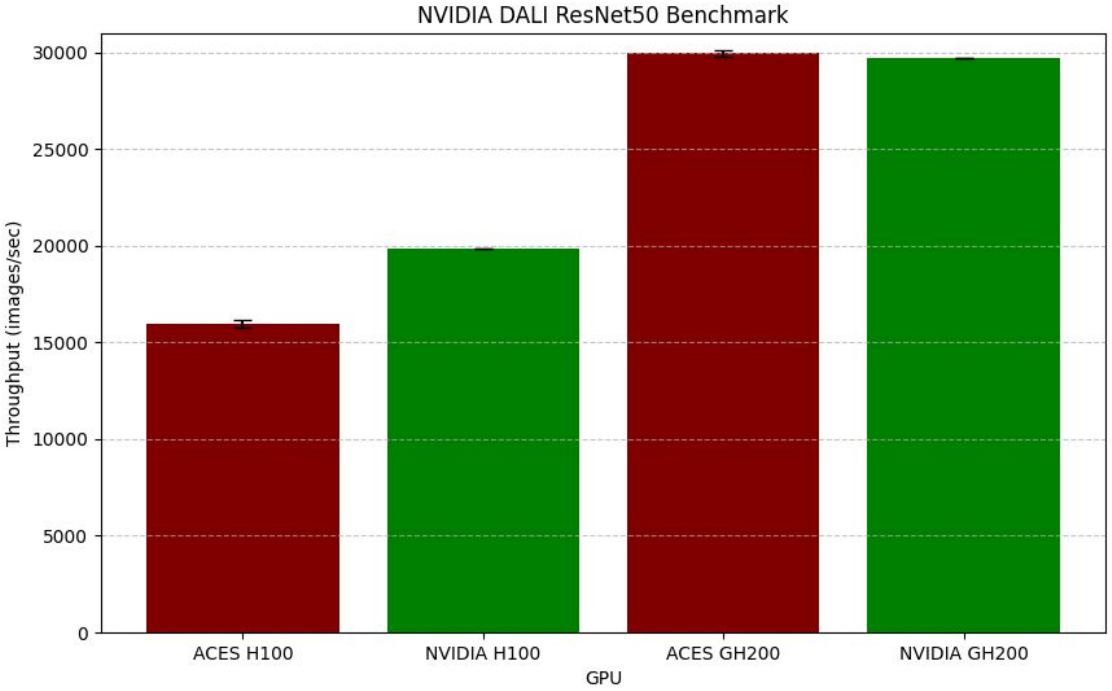
CPU-Stream comparisons GH200 96GB

GPU STREAM Benchmarking Results



GPU-Stream comparisons

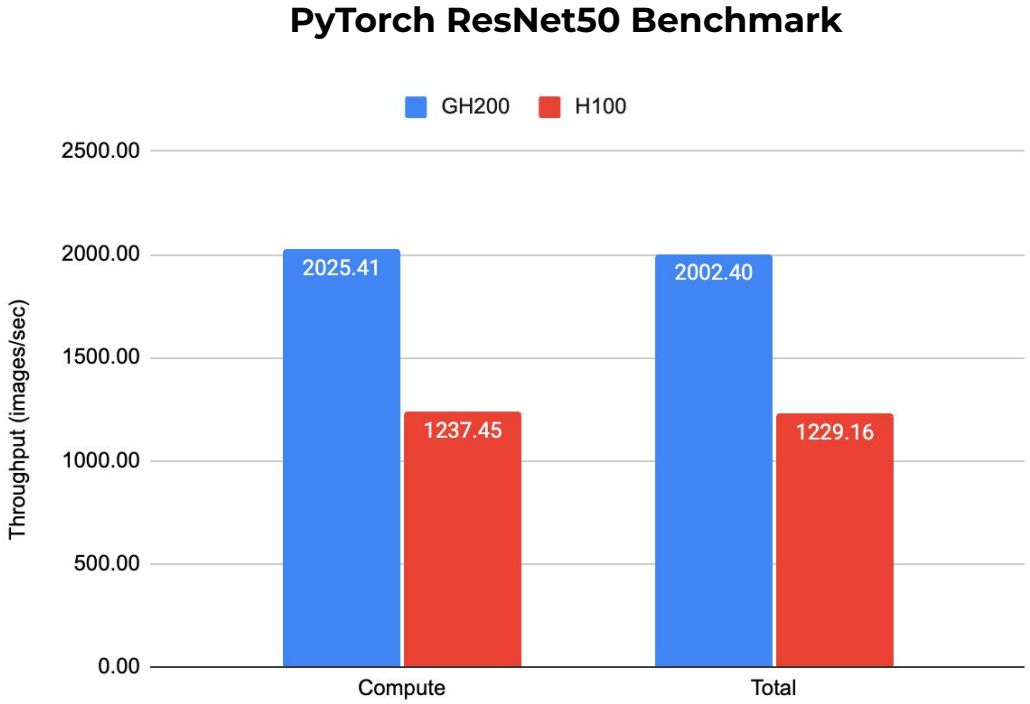
DALI ResNet50 Benchmarking Comparison



DALI ResNet50 Benchmark Comparison

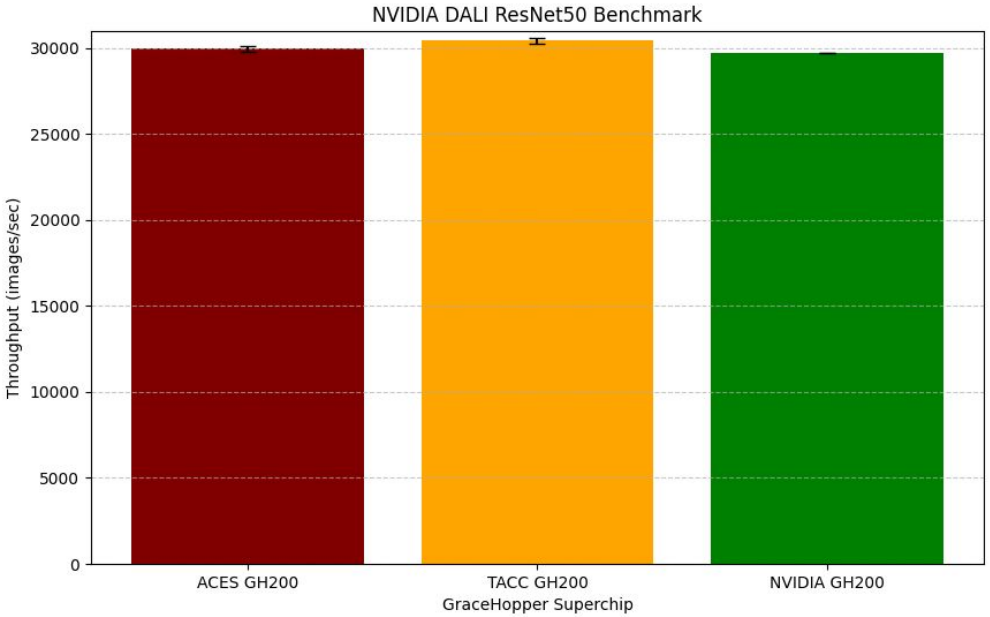
PyTorch ResNet50 Benchmarking Results

Containers used:
pytorch-24.08-arm64 (GH200);
pytorch-24.08-py3 (H100)



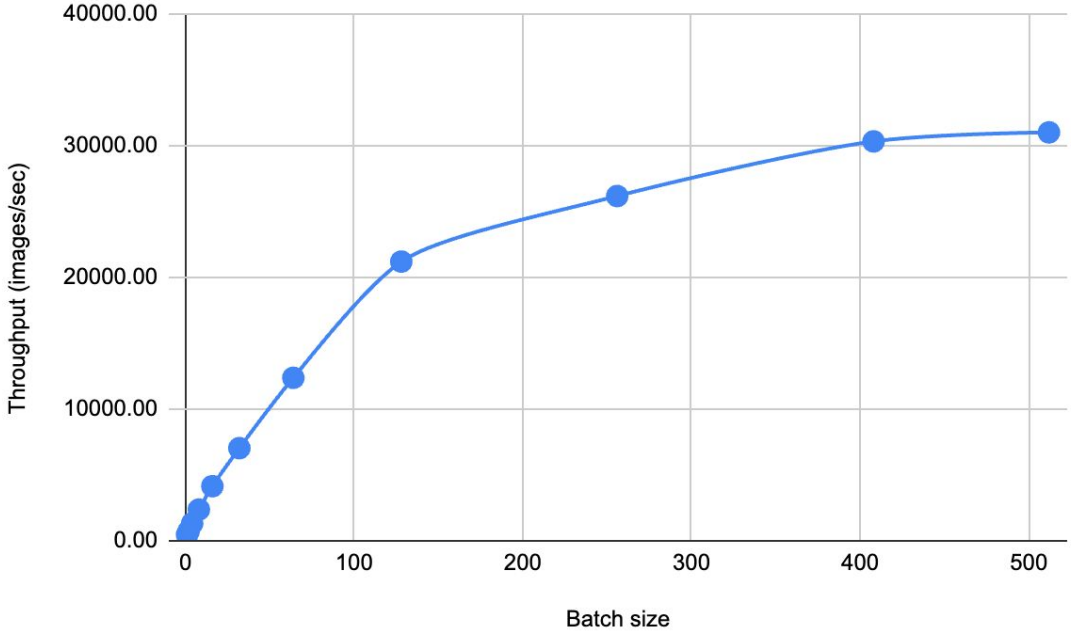
PyTorch ResNet50 Benchmark

Dali ResNet50 Benchmarking Comparison



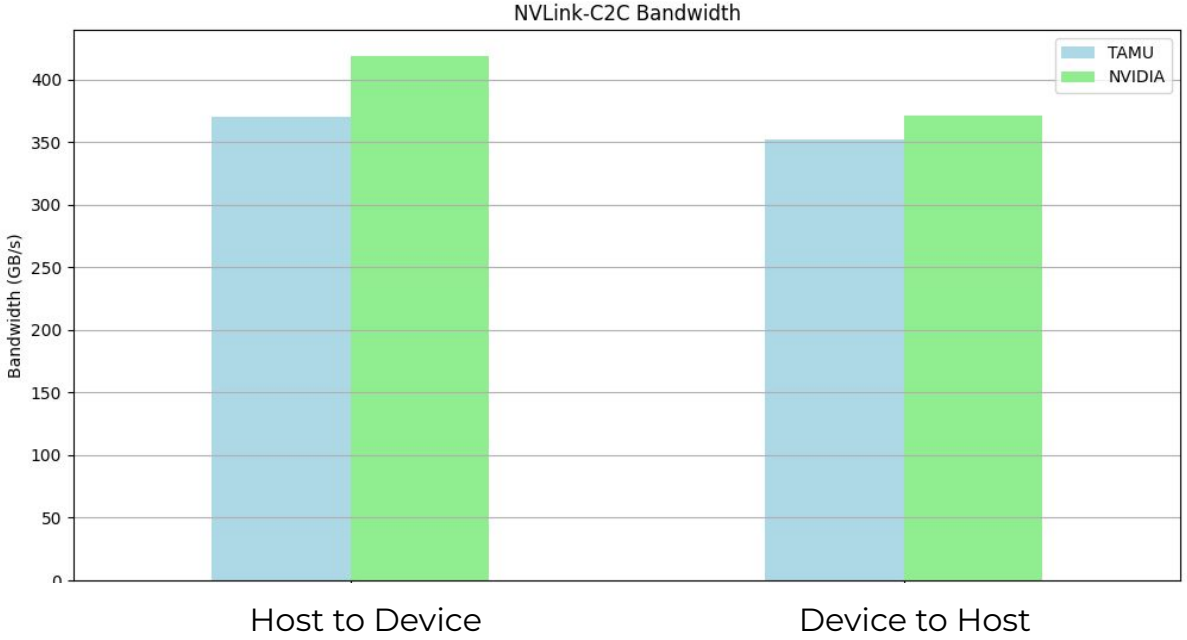
ResNet50 Benchmark Comparison

TAMU HPRC DALI ResNet50 Benchmarking Results



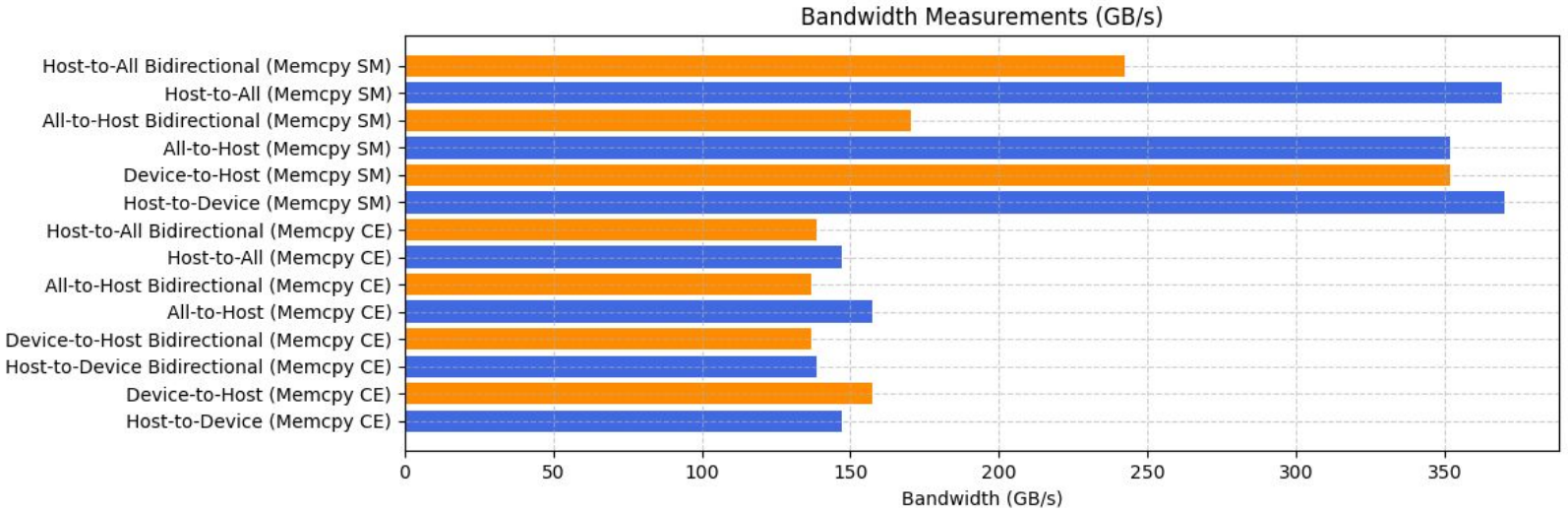
NVIDIA DALI ResNet50 Benchmark

C2C Bandwidth Benchmarking Comparison



NVbandwidth comparison on GH200

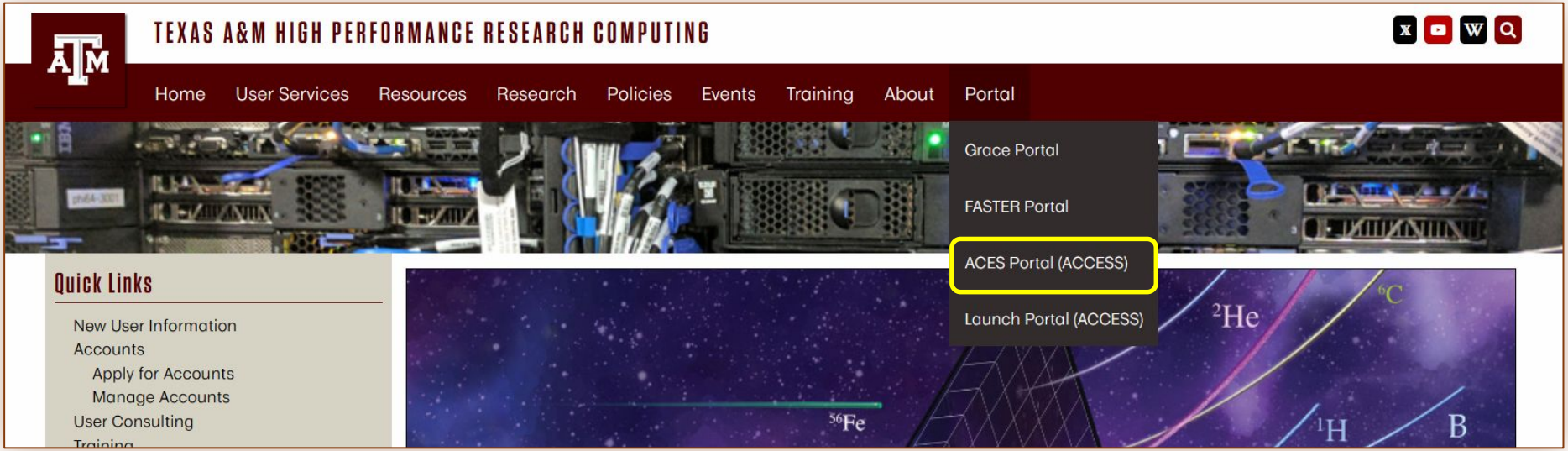
TAMU HPRC Bandwidth Benchmarking Results



TAMU bandwidth graphic

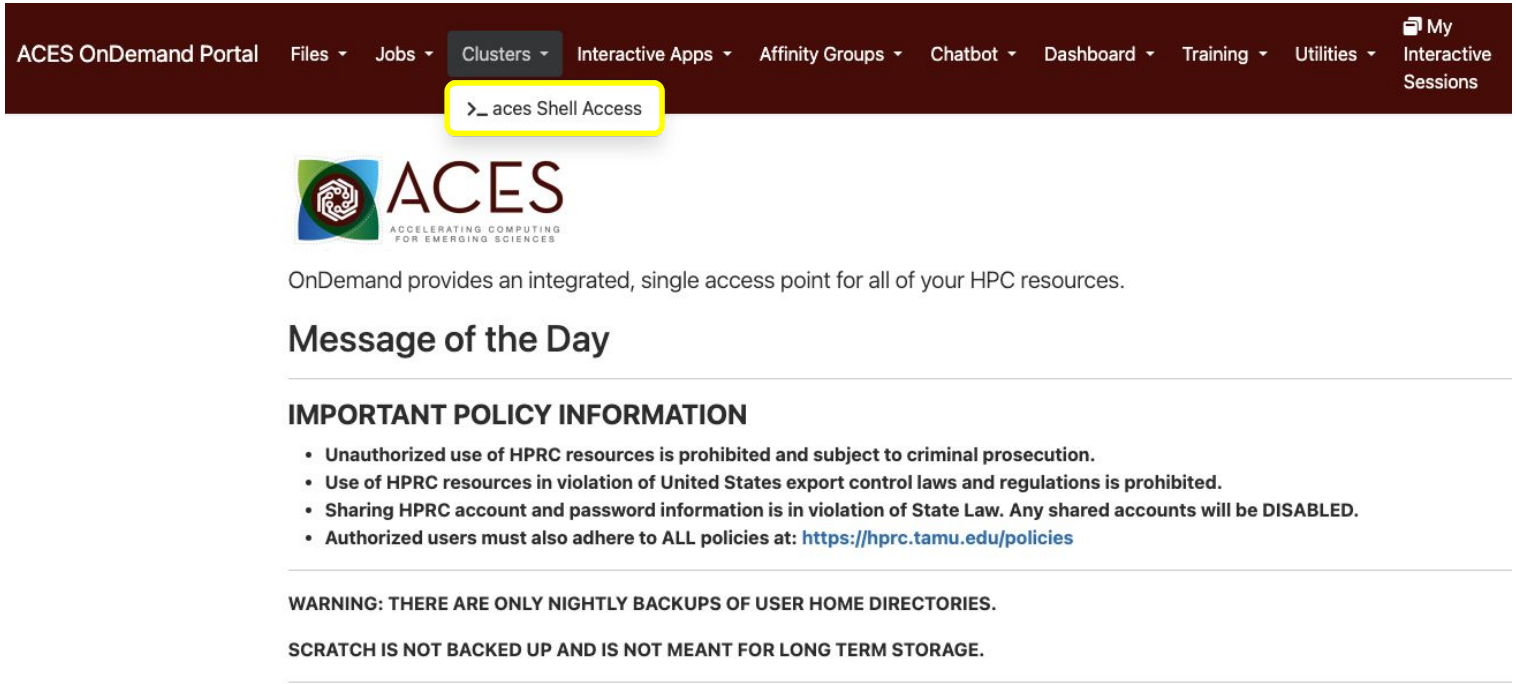
Accessing ACES via the Portal

- Go to https://portal-aces.hprc.tamu.edu/ and log in with your ACCESS credentials



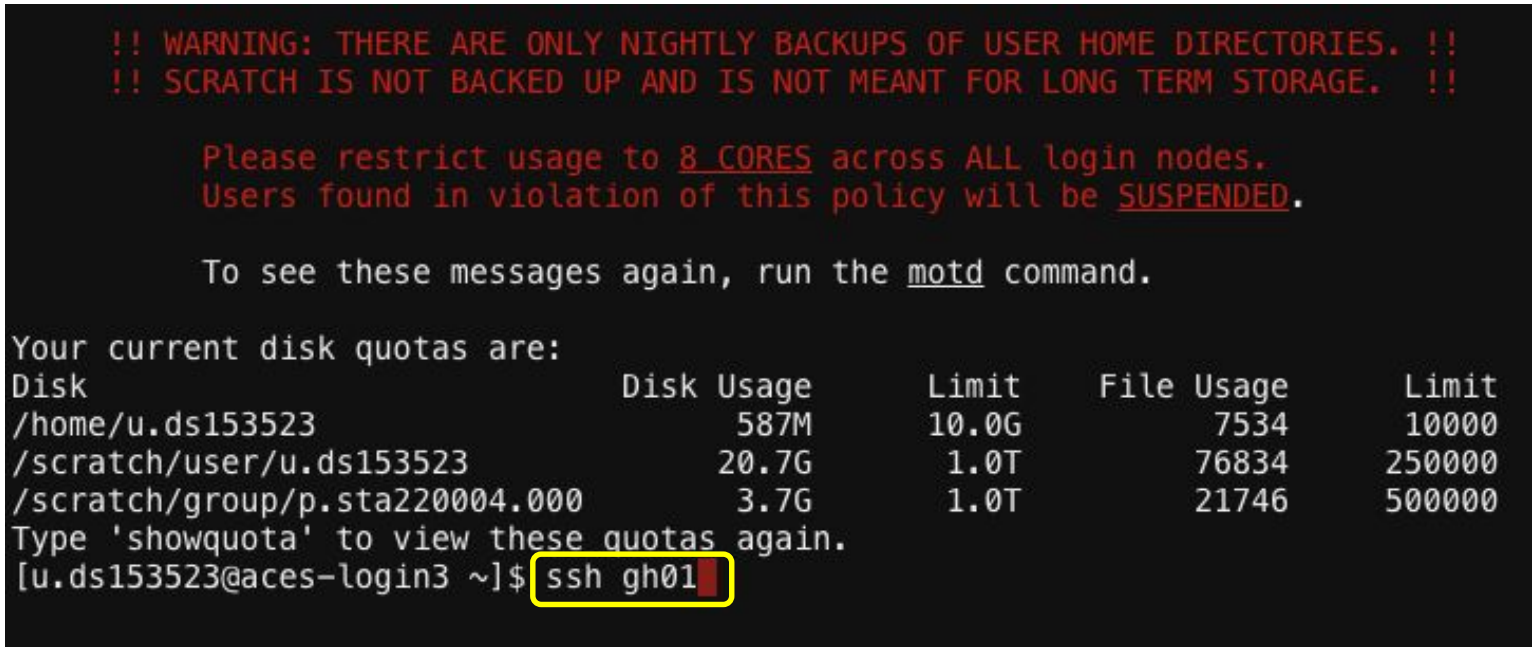
Shell Access

- Once logged into the portal, click on Clusters > aces shell access



ssh into GH200

- Once you are in an ACES terminal, issue the command `ssh gh01`



Exercise: DALI Resnet50 Interactive Benchmark

- cd /scratch/user/$USER
- ml CUDA/12.5.0
- python3 -m venv dali-benchmark-env
- source dali-benchmark-env/bin/activate
- pip install numpy
- pip install --extra-index-url https://developer.download.nvidia.com/compute/redist --upgrade nvidia-dali-cuda120
- wget https://raw.githubusercontent.com/NVIDIA/DALI/release_v1.31/tools/hw_decoder_bench.py
- git clone https://github.com/NVIDIA/DALI_extra.git
- python3 hw_decoder_bench.py --width_hint 6000 --height_hint 6000 -b 408 -d 0 -g gpu -w 10 -t 10000 -i DALI_extra/db/single/jpeg -p rn50 --hw_load 0.11

Request Access to TAMUs NVIDIA GH200

- If you wish to use TAMUs HPRC NVIDIA GH200 system, please write an email to help@hprc.tamu.edu and inform us to why you are inquiring to use our GH200
- Please let us know how long you intend to use the system, so that we can make the proper reservations
- In order to have access to the GH200, you will need **an active ACCESS** account on TAMUs HPRC ACES Cluster
  - https://hprc.tamu.edu/kb/User-Guides/ACES/Access/

Acknowledgments

- Staff and students at Texas A&M High-Performance Research Computing
- The National Science Foundation (NSF)

HIGH PERFORMANCE RESEARCH COMPUTING
TEXAS A&M UNIVERSITY

https://hprc.tamu.edu

HPRC Helpdesk:

help@hprc.tamu.edu
Phone: 979-845-0219

Take our short course survey!

HPRC Survey
https://u.tamu.edu/hprc_shortcourse_survey

Help us help you. Please include details in your request for support, such as, **Cluster** (ACES, FASTER, Grace, Launch), NetID (UserID), Job information (**JobID**(s), Location of your jobfile, input/output files, Application, Module(s) loaded, Error messages, etc), and Steps you have taken, so we can reproduce the problem.

References

[1] *Nvidia GH200 Grace Hopper Superchip Architecture*, www.amax.com/content/files/2023/12/NVIDIA_Grace_Hopper_Superchip_Architecture_Overview_Whitepaper.pdf. Accessed 2 Mar. 2025.

[2] *Nvidia GH200 Grace Hopper Superchip Benchmark Step- ...*, docs.nvidia.com/gh200-superchip-benchmark-guide.pdf. Accessed 3 Sept. 2024.

[3] *Nvidia Grace Performance Tuning Guide*, docs.nvidia.com/grace-performance-tuning-guide.pdf. Accessed 3 Sept. 2024.