



Introducing RC's New NVIDIA Grace Hopper Superchip



Research Computing
UNIVERSITY OF COLORADO BOULDER

Introducing RC's New NVIDIA Grace Hopper Superchip

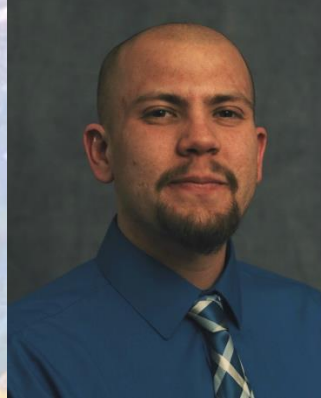
Instructor: Brandon Reyes

- Research Computing
- Website: www.rc.colorado.edu
- Documentation: <https://curc.readthedocs.io>
- Helpdesk: rc-help@colorado.edu
- Survey: <http://tinyurl.com/curc-survey18>

Meet the User Support Team



Layla
Freeborn



Brandon
Reyes



Andy
Monaghan



Michael
Schneider



John
Reiland



Dylan
Gottlieb



Mohal
Khandelwal



Ragan
Lee

Slides

https://github.com/ResearchComputing/introducing_rc_gh200_quick_byte



Session Overview

- GH200 architecture overview
 - Hardware specs
- What workflows will benefit from the GH200 architecture?
- GH200 software
- Beta testing phase
- How can you run on the GH200s?

GH200 architecture overview

The Grace Hopper Superchip (GH200) is a newer chip provided by NVIDIA. Its unique architecture allows the GPU and CPU to efficiently share and exchange memory.

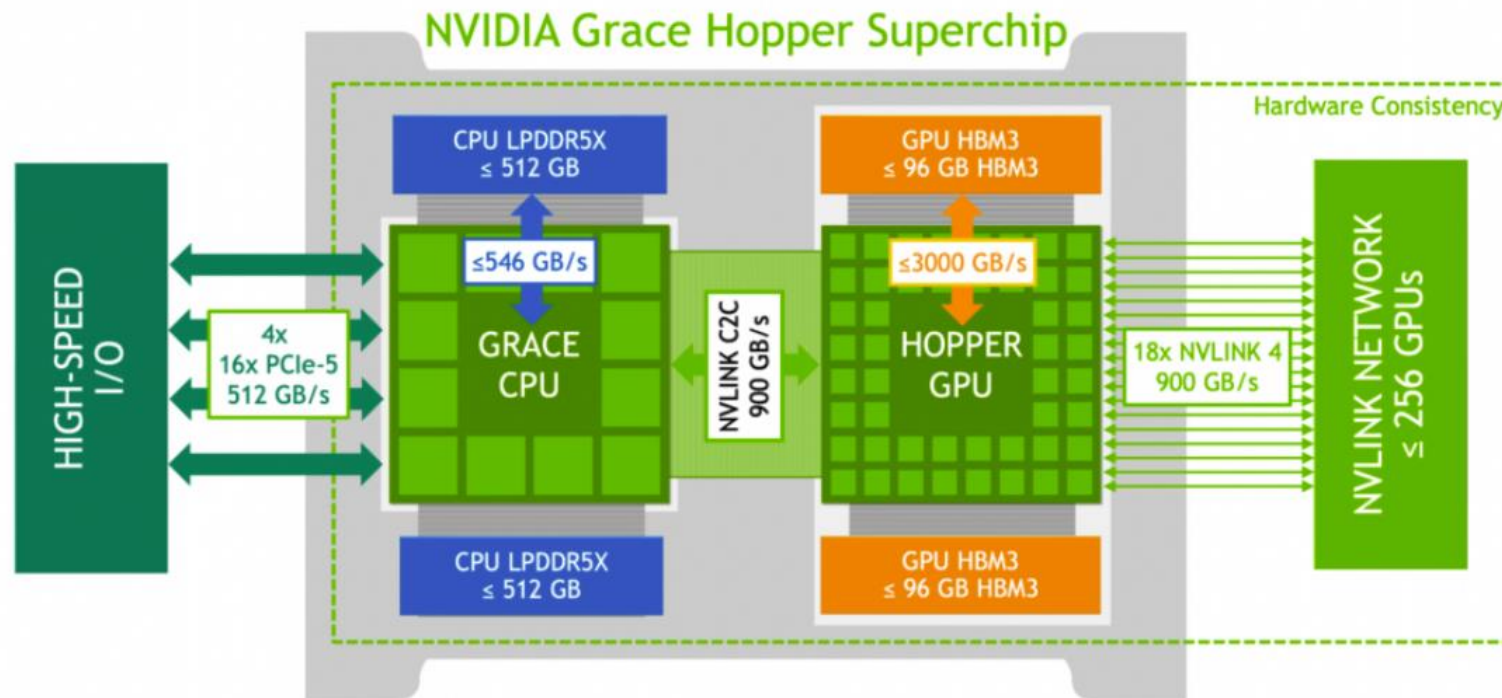


Image provided by <https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper>

Specifications for our 2 GH200 nodes

- Grace CPU has 72 cores and roughly 480 GB of RAM
 - Arm based (Neoverse V2)
- Hopper GPU has roughly 100 GB of VRAM
- CPU and GPU are connected via NVIDIA NVLink-C2C
 - This enables efficient and seamless memory transfer between the two components
- FAST I/O speed (512 GB/s)
 - Top I/O speeds only apply to the SSD
- Roughly 1.7 TB of usable SSD on the node

Note: CURC does not currently support multi-node GH200 workflows

Benefits of GH200 architecture

- NVLink-C2C provides seamless memory management between the CPU and GPU
 - Allows GPU threads to access RAM, extending the GPU memory
- 1:1 GPU-to-CPU ratio
 - GPU and CPU have roughly the same processing power dedicated to them
 - Ideal for workloads that utilize the CPU and GPU to a high degree
- High-speed I/O to SSD and large SSD storage space
 - Allows users to store a large amount of their data locally on the node (while it is in use) and quickly access it

What workflows will benefit from the GH200 architecture?

- Heterogenous software that uses both the CPU and GPU
- Inference for Large Language Models
 - We were able to run Llama 3.1 405b (requires ~300 GB of memory)
- Training large models and hyperparameter tuning
 - We have seen at least 2X speedup, some users reported 10X speedup
- NVIDIA also states that the following applications perform well on the GH200s:
 - Graph Neural Networks, ABINIT, OpenFOAM, GROMACS, FFTs, Multi-Grid Linear Solvers

GH200 software

Due to the different architecture, there is a curated software stack. We currently provide:

- CUDA compilers through NVIDIA HPC SDK e.g. nvcc, nvcc++
- CUDNN libraries
- Miniforge (mamba and conda)
 - If your specific library allows for aarch64 architecture with GPU capabilities, it will most likely work on the GH200s
- Apptainer
 - A large selection of compatible containers are available through NVIDIA's NGC catalog

We will work with you to install GH200 compatible software, in alignment with our software policies.

Beta testing phase

To identify issues and workflows that best run on the GH200s, we initiated a beta testing phase.

- This is directed towards users with established GPU workflows on either our A100 or MI100 GPUs
- An initial consultation is held to determine if the workflow is appropriate for the GH200s
- Once approved, we set up the software on the GH200s, onboard you to the nodes, and provide hands-on support for any issues encountered

How can you run on the GH200s?

- Submit a ticket to rc-help@colorado.edu
 - In subject include:
 - You are interested in running “X application” on the GH200s
 - In the email body:
 - Short description of what your workflow does
 - Why you believe the GH200s would be beneficial to your workflow
 - Provide us with Linux paths to the code you would like to run
 - If possible, any JobIDs of this workflow you have ran on a GPU node

Thank you!

Survey and feedback

<http://tinyurl.com/curc-survey18>



Slides

https://github.com/ResearchComputing/introducing_rc_gh200_quick_byte

