Experiences with the NVIDIA Grace Hopper architecture in HPC systems

CURC Alpine: New User Seminar

Moderators: Andy Monaghan, Craig Earley

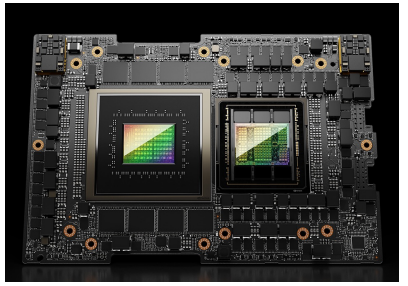
- Slides: https://github.com/ResearchComputing/rmacc_2024
  - In the directory “nvidia_GH200_experiences”

Overview

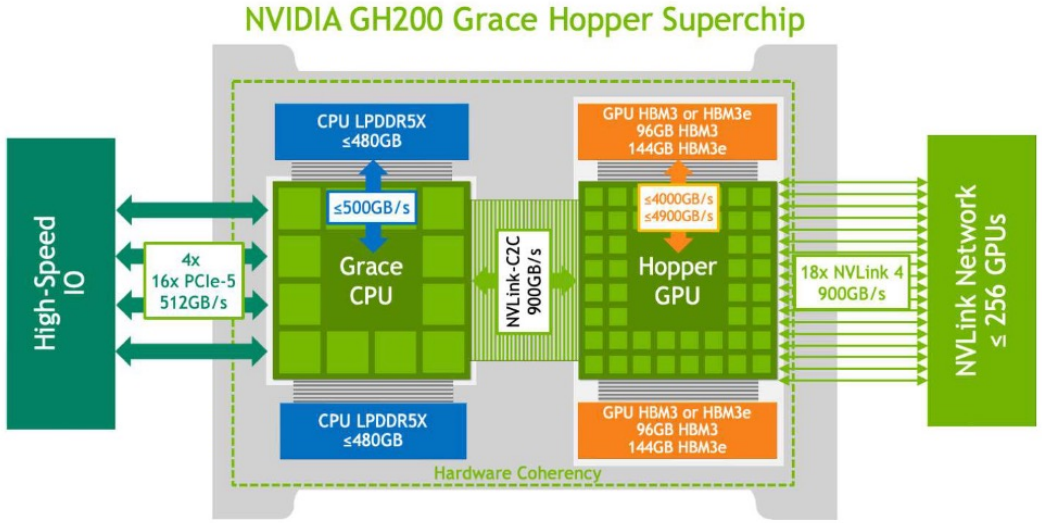
The NVIDIA GH200 Grace Hopper Superchip combines the NVIDIA Grace and Hopper architectures into a single CPU+GPU coherent memory model, with the ability to greatly accelerate data-intensive GPU workflows. This sixty-minute "Birds of a Feather" will enable participants to discuss their early experiences with the GH200, or to learn from others' experiences if they are considering the Grace Hopper architecture.

- First 30 minutes – user-facing experiences
- Second 30 minutes – system-facing experiences

The NVIDIA Grace Hopper 200



- Designed for “giant-scale” AI, HPC
- Combines Grace CPU w/ Hopper (H100) GPU via 900 GB/s NVLink
- 3x memory, bandwidth vs stand-alone H100
- ARM-based instructions
- Multi-instance GPU (MIG) capable



Source for image and specs: https://www.nvidia.com/en-us/data-center/grace-hopper-superchip/

Possible user-facing discussion topics

- If you or your users used the GH200 yet, what has the experience been?
- What workflows have worked best on the GH200s?
- What workflows have not worked well on the GH200s?
- What challenges and/or successes have you had with the ARM-based software on the GH200s?
- What has been your experience with NVIDIA support?

Possible system-facing discussion topics

- What were challenges of deploying the GH200 nodes?
  - Provisioning challenges?
  - Data center challenges?
  - Networking challenges?
- How have the nodes performed once deployed?
- What has been your experience with NVIDIA support?

Survey and feedback

http://tinyurl.com/curc-survey18