Delta: Living on the Edge of Slingshot Support

NSF Award OAC-2005572

Brett Bode, David King, Greg Bauer, Galen Arnold, and Robert Brunner

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

Delta System Overview

Hardware Overview

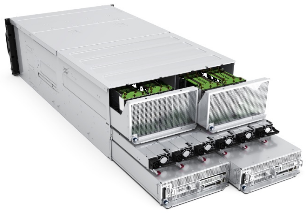
RESOURCE COUNTS

CPU NODES	124 x CPU Compute Nodes 8 x CPU Utility Nodes
GPU NODES	100 x 64-Bit x 4 GPU 100 x 32-Bit x 4 GPU 5 + 1 x 8 GPU & High Mem
STORAGE	7 PB HDD (Lustre) 3 PB SSD (non-POSIX)

SYSTEM TOTALS

CPUs	476 x AMD EPYC 7763 64 core “Milan”
GPUs	440 x NVIDIA A100 400 x NVIDIA A40 8 x AMD MI100
PERF	10 PF double-precision 100 PF single-precision 200 PF tensor

Hardware Overview



Mixture of Apollo 6500 and Apollo 2000 servers with DL385 utility nodes.

Four different GPU configurations in Apollo 6500

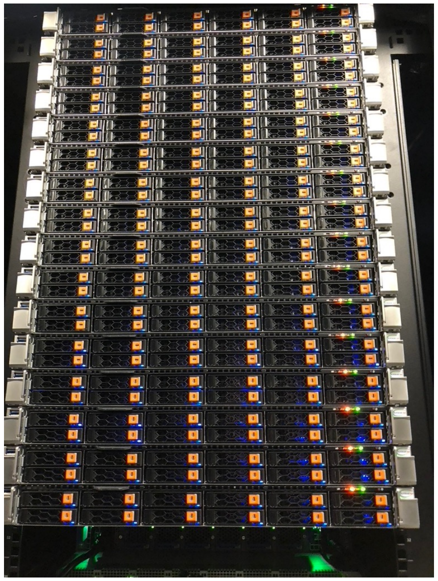
- Quad NVIDIA A100
- 8-way NVIDIA A100
- Quad NVIDIA A40
- 8-way AMD MI100

Apollo 2000 and A100 nodes use DLC cooling for CPUs and GPUs.

Single rail Slingshot throughout.

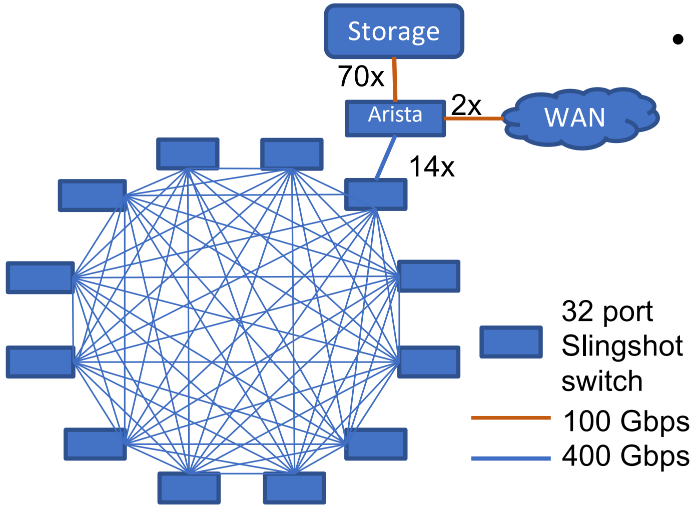
DDN Storage Environment

MODEL	DDN SFA 7990x
STORAGE	6 PB on Delta 1 PB on Center-wide FS
FILESYSTEM	Lustre
PERF	60 GB/sec for home + scratch.



STORAGE	3 PB Flash (raw)
FILESYSTEM	RED/IME
PERF	500GB/sec

Network Hardware



INTERCONNECT	HPE/Cray Slingshot
TOPOLOGY	switch-level all-to-all
FILESYSTEM	Lustre
LINK SPEED	100 Gbps per node now 200 Gbps per node after upgrade 400 Gbps switch to switch 2 x 100 Gbps WAN

- All Delta nodes are interconnected with an HPE/Cray Slingshot network arranged with the switches connected in an All-to-All configuration
  - Up to 32 nodes per switch
  - Slurm will attempt to keep a job on minimum number of switches
  - Ethernet plus QoS and congestion avoidance
  - Software is layered upon libfabrics
  - Arista 400Gbps switch provides connectivity to storage and to the WAN aggregation switch.
- Delta provides 200Gbps of external network connectivity with direct access from utility nodes and routed access for all compute nodes

Software

Operating System Software Stack

Operating System

- Red Hat ~~8.4~~ 8.6 Extended Update Support (EUS)

Kernel Drivers

- Storage
  - DDN Lustre/IME
- Networking
  - Nvidia Mellanox
  - HPE Cray Cassini
- Accelerators
  - Nvidia CUDA
  - AMD ROCm

Monitoring

- Telegraf

Scheduling

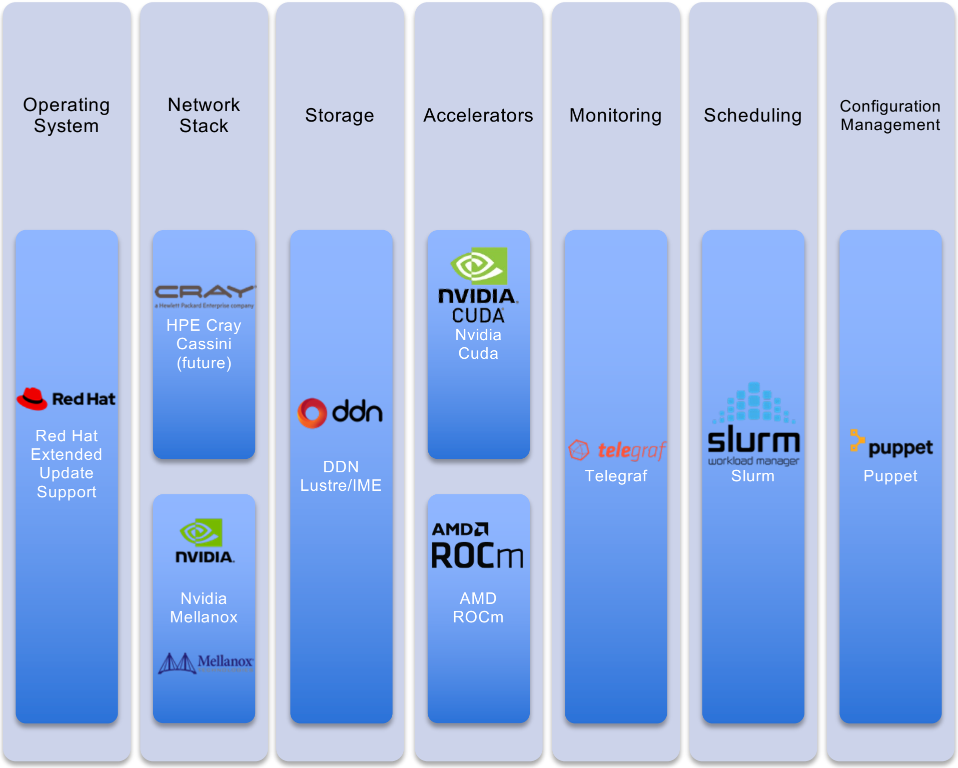
- Slurm

Provisioning

- xCAT

Configuration Management

- Puppet



Operating System
Red Hat Extended Update Support

Network Stack
HPE Cray Cassini (future)
Nvidia Mellanox

Storage
DDN Lustre/IME

Accelerators
CUDA
Nvidia Cuda
AMD ROCm

Monitoring
Telegraf

Scheduling
Slurm

Configuration Management
Puppet

Delta: Programming Environment

Spack for software installation

- Balance between what is provisioned in OS in memory, and what is provision by Spack on parallel file system
- Compilers: GCC, AOCC, NVIDIA HPC SDK, Intel and MPI: OpenMPI
- Support for OpenACC, OpenMP, CUDA, ROCm
- Vendor participation: AMD vested in Spack installation of their software

Modules

- lmod with two modtrees: CPU and GPU.

XSEDE and ACCESS integration

- CUE common user environment to assist with migration and porting.

Python

- Spack installs of "plain" python, Anaconda for CPU and GPU versions (outside of Spack)
- Users can locally install additional (newer) packages.
- **Container Support**
- Apptainer/Singularity provides container functionality. Operates within Delta security model
- NVIDIA NGC and AMD Infinity Hub containers provisioned as Singularity images on Delta, reduce user quota use

Slingshot 10 to Slingshot 11

Delta: Timeline

January 2022

- System installed and bring up started
- Initial install used Slingshot v1.5
- Significant stability issues and knowledge issues with the install team

February 2022

- Updated to Slingshot v.1.7.0
- Big improvement in stability

May 2022

- Updated to Slingshot v.1.7.2
- Slingshot environment has been very stable since then.

July-September 2022

- System acceptance
- Start of production operations

Issues with 1.7.2

- Slingshot switches do not have persistent default routes therefore we have to replicate services on the fabric managers (ntp and rsyslog)
- We have had nodes randomly stop communicating on switches. This gradually got worse and required a fabric reset.
- During a storage maintenance, we shut down LNET routers but didn't remove from from configs. This caused the the switch with the LNETs to lockup.

  - Don't shutdown the hosts, just disable the service.
- Before 2.0.1, ARPs needed to be statically set

Slingshot 11 Upgrade Plan

Hardware

- Cassini NICs delivered August 2022
- Installation on hold until software stack is complete

Software

- Decision made to wait for Slingshot software v2.0
- Started testing in November 2022

Preparing for the Migration

Test Environment

- Delta does not have a dedicated test system
- Built a test environment from a single loaner switch with virtualized fabric managers and four compute nodes
- On the same subnet as the production fabric to avoid downtime of production while allowing access to storage
- AMA macs must be different between fabrics due to conflicts
- Cannot communicate between fabrics due to changes between 1.7.2 and 2.0.x (packet duplication)

Software Stack

- Evaluated multiple builds of the user space stack with the goal of minimizing disruption to users.
- Prioritized OpenMPI
- Evaluate integrating the HPE/Cray PE into our environment.

Fabric Manager

Delta included two physical servers for fabric managers

- The hardware is significantly over-provisioned for Delta.
- Standalone installs require extra effort to maintain.

Upgrade Plan

- Migrate fabric managers to VMs utilizing Delta utility node image
- VMs are stateless with the FM database on local storage

Actual process was

- perform a fresh install of v2.0.1 in the VMs
- generate new configs based on old configs
- Upgrade the switch firmware
- HA is enabled between two VMs

Completed fabric manager/switch upgrade to v2.0.1 in late April 2023.

- Now capable of running Cassini and Mellanox on the same fabric

Issues

RHEL Kernel security updates

- HPE providing driver binary for only base kernel
- We MUST install security related kernel updates as quickly as possible.
- HPE providing driver source code such that we can build the driver on our own as needed – short term solution
- Long-term HPE will support DKMS

CrayPE is very difficult to integrate with our Spack built modules environment

CrayPE install via RPMs on bare metal is broken in multiple ways

- Cray-libsci install scripts are broken on CPE-22.12
- CrayPE dynamic modules don’t recognize the network therefore it never loads MPI

Low block size MPI ALL to ALL performance in SS11 is worse than SS10 with CrayPE and OpenMPI 5 beta

DDN currently has no support for kfilnd

- Will use TCP instead

HPE Support portal does not recognize our entitlements to download Slingshot software…

Current Blocking Issues

NCSA has a strong desire to continue with OpenMPI as the primary MPI to minimize user disruption. Our current issues are:

- Integration with Slurm is broken (srun does not work)
- GPU RDMA support
- Some performance degradations compared with Slingshot 10.
    - This is likely explained by Thomas Naughten’s OpenMPI talk yesterday.

Future

- We have had many very useful conversations here at CUG providing many suggestions on ways to move forward.
- Will have follow up discussions over the next several weeks to resolve our key issues.
- Once we have a satisfactory user space stack we will complete our migration to Slingshot 11.

Discussion

The Delta project is supported by the National Science Foundation (award OCI 2005572), and the State of Illinois. Delta is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications.