Leveraging Lustre as a Global File System for US and Illinois Researchers

J.D. Maloney | Lead Storage Engineer | NCSA

National Center for Supercomputing Applications
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Prepared for: LUG 2024

Outline

- Overview of NCSA & Compute Resources
- Taiga - NCSA’s Global File System
  - Design Goals
  - System Architecture
  - Handling Identities
  - Taiga’s Future
- Delta* File Systems

Overview of NCSA

- Operate the National Petascale Computing Facility (NPCF)
- Service over 9,500 active users
- Over 2,000+ compute nodes in production
- Over 100PB of storage capacity
- Store ~55PB of data across (7+ billion files)

NCSA Offices

NPCF

Campus

Major Compute Environments

Delta - 350 node HPE/Cray - AMD Milan + Nvidia A100 - NSF (ACCESS/NAIRR)

DeltaAI - ~140 node HPE/Cray - Nvidia Grace-Hopper - NSF (ACCESS/NAIRR)

Illinois Campus Cluster - ~850 node CPU+GPU Dell/HPE/Supermicro - Illinois

"Industry" - ~200 node Dell - Cascade Lake/Sapphire Rapids/H100

Radiant - 120 node Dell - Ivy Bridge (soon Bergamo)/Nvidia A100 - OpenStack

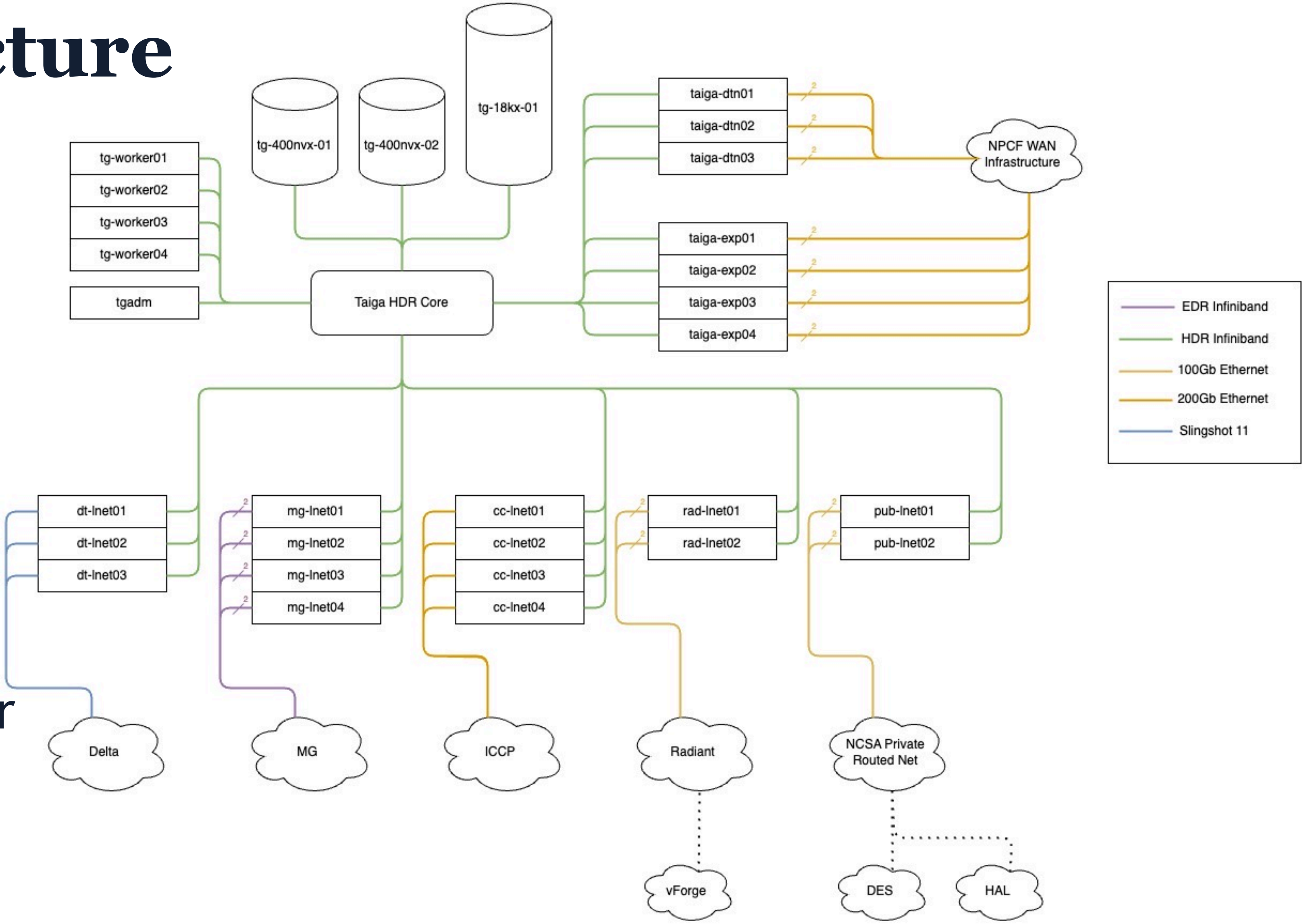
Nightingale/ACHE - 265 node Dell - Cascade Lake/Ice Lake/Milan/Nvidia

Design Goals for Taiga

- Easy to scale (both in capacity/performance) file system
  - Can add to the FS in reasonable chunks
- Hybrid NVME and HDD
  - More than just metadata on flash
- FS capable of peering with every compute environment natively via their HSNs
- User/Group/Project Quotas
- Support for multiple Authentication Systems
  - More on this soon
- Prevent users from having to copy files between compute systems
  - Users encouraged to compute actively against Taiga

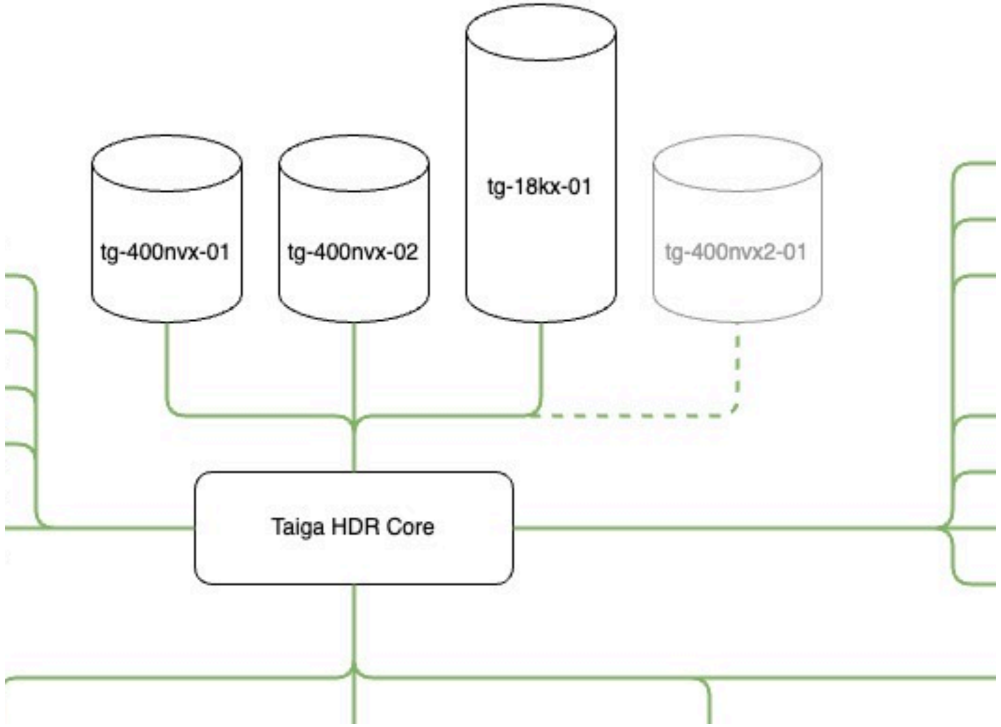
System Architecture

- Core network fabric is HDR Infiniband (200Gbps)
- File system is translated to every compute environment’s high-speed fabric
- A set of worker nodes perform data packaging/migration work
- Two sets of export nodes for Globus and NFS duties
- Connection to NPCF WAN at 2 x 100Gbps uplinks



System Architecture

- Originally built with 2 x ES400NVX units
  - Each with 4 x SS9012 (18TB Drives)
  - Each with 12 x 15.36TB Flash
- Expanded with 1 x ES18KX unit
  - 10 x SS9012 (12TB Drives)
  - 24 x 7.68TB Flash
- Soon to grow again with 1 x ES400NVX2 unit
  - 5 x SS9024 (18TB drives)
  - 12 x 15.36TB Flash
- File System Targets (current)
  - 3 x MDTs (4.2TB each)
  - 24 x Flash OSTs (12.4TB each)
  - 28 x HDD OSTs (16 x 553TB; 12 x 626TB)



System Architecture

- Running Exa 6.3 w/ hotfixes (near bleeding edge as of writing)
- Control data layout via a global FS PFL
  - First 64KB of each file on NVME (1 stripe)
  - 64KB to 256MB on HDD (1 stripe)
  - 256MB to 4GB on HDD (4 stripe)
  - 4GB to EOF on HDD (28 stripe)
- Mainly leverage project quotas to control allocations
  - Standard of 1.5 million inodes per 1TB of allocation
  - Can exception up to 4 million inodes per 1TB of allocation
- No DOM or Hot Pools implemented on this system
- Extensively monitored using NCSA’s in-house Lustre TIG implementation
- NFS exported via NCSA’s custom NFS HA stack (and soon SMB exported as well)

Handling Multiple Identity Systems

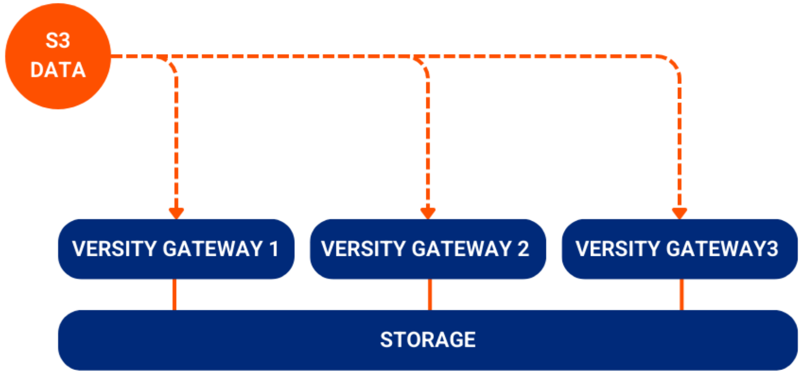
- NCSA maintains its own LDAP infrastructure to best accommodate thousands of US researchers that have no affiliation to University of Illinois
- Taiga leverages NCSA LDAP as its primary realm of identity information
- Illinois Campus Cluster + Other Illinois Computes services all authenticate and use the Illinois Active Directory system
  - By client count this will be ~50% of Taiga's clients
  - Users prefer the "single login"/no need for separate NCSA account with potentially different username
  - Works better when users SMB mount file system to domain-joined machines
  - Of course there is no UID/GID alignment between these systems
- NCSA is actively scale testing Lustre Nodemap

Handling Multiple Identity Systems

- Nodemap running via the “Illinois”/CC LNET routers, all other traffic not mapped
- Illinois users with existing NCSA accounts (on systems like Delta, etc.) will be mapped directly
  - Matching via email address
- Illinois users without NCSA accounts, an account will be silently provisioned
  - Will try to match username if possible
  - If unable to match username an alternative is chosen
- Illinois AD groups will get replicated into new NCSA groups
- Will check for updates every 30 minutes and update Nodemap accordingly
- Full config map is also maintained outside of Taiga so config can be easily re-provisioned if needed
- Users able to simultaneously access data between NCSA & Illinois systems

Access to Taiga via S3

- The S3 protocol is becoming more of a first-class citizen within our infrastructure
- A leading protocol in data sharing activities, especially active datasets
- We are deploying the Versity S3 Gateway in front of our *Granite* archive system now and *Taiga* will get it soon after
- Also becoming a valuable protocol for data ingestion from the edge for newer instruments and pipelines
- Teams will be able to request an S3 collection for a section/sections of their allocation



The Future of Taiga

- Is it Lustre? We're not yet sure
- System upgrade slated for 2026, seriously exploring a move to all NVME
- Would move Taiga from "projects" to "projects + scratch" status
  - A place where all "active" data is stored
- Need a much-improved policy engine and data manipulation tools and capabilities (both for admins & users)
- Want deeper quota support (within projects)
- Automation of dataset sharing/access (FAIR)
- Ideally tenant-based performance QoS capability
- We're testing options now; plan finalizing in mid/late 2025

Delta* Scratch/Work

- Can’t not touch on our fastest Lustre on site
- Delta Lustre had:
  - 3 x ES7990X units; 16TB drives; SAS SSD metadata
- Bringing in DeltaAI hardware:
  - 10 x ES400NVX2 units; 24 x 15.36TB NVME each
  - 40 total MDTs, sized for 30 billion inodes
  - Max-inherit depth of 5-7 levels
- Planning to merge all units into one file system
  - Separate “HDD” and “NVME” areas, controlled via PFL
  - Teams will be able to get quota on each pool separately

Questions?

Email: malone12@illinois.edu

National Center for
Supercomputing Applications
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN