NCSA Lustre Site Update

J.D. Maloney | Lead Storage Engineer | NCSA

National Center for Supercomputing Applications
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Prepared for: LUG 2025

Outline

- Lustre Usage @ NCSA
- Experiences with Nodemap in Production
- Early Results of CSDC in production
- Experience with DNE3 with ”many” MDTs
- Versity S3 Gateway with Lustre

Lustre Usage @ NCSA

The Highlights

- 5 Lustre namespaces under management
- Aggregate Lustre capacity is 36PB usable
  - ~60% of NCSA's active storage capacity
  - ~5PB of this is NVME
- Aggregate IOR of ~1.2TB/s across all Lustre file systems
- Fastest FS at ~850GB/s; Largest FS is 25PB (not the same file system)
- All file systems running Lustre 2.15 (ExaScaler 6.3.X)
- All major fabrics represented (Ethernet, Infiniband, Slingshot)

Lustre Usage @ NCSA

Delta (”Delta” + “DeltaAI”)

- 12 x ES400NVX2
  - 24 x 15.36TB Gen4 NVME
  - 4 MDT + 8 OST per appliance
- 3 x ES7990X
  - 176 x 16TB HDD
  - 4 OST per appliance
- All appliances directly connected to HPE Slingshot 11 fabric
  - Leveraging the tcp driver for communication
  - Separate SS fabric for storage that peers at 1TB/s to each Delta/DeltaAI SS fabrics

Lustre Usage @ NCSA

Delta (”Delta” + “DeltaAI”)

- Single namespace (across all 48 MDTs) data placement controlled by PFL
  - /work/hdd/$alloc → 7990X OSTs
  - /work/nvme/$alloc → 400NVX2 OSTs

- OST pool quotas used to enforce each allocation’s usage within each tier

- Rolling out an altered PFL with compression on the NVME tier

Lustre Usage @ NCSA

Taiga

- 2 x ES400NVX
  - 12 x 15.36TB Gen3 NVME
  - 352 x 18TB HDD
  - 1 MDT + 16 OST (8 NVME + 8 HDD OSTs)
- 1 x ES 400NVX2
  - 12 x 15.36TB Gen 4 NVME
  - 440 x 18TB HDD
  - 1 MDT + 18 OST (8 NVME + 10 HDD OSTs)
- 1 x ES18KX
  - 24 x 7.68TB Gen 3 NVME
  - 900 x 12TB HDD
  - 1 MDT + 20 OST (8 NVME + 12 HDD OSTs)
- Backend HDR Infiniband Fabric

Lustre Usage @ NCSA

Taiga

- Makes extensive use of LNET routers
  - 17 routers connect Taiga to 6 distinct HSNs across various clusters (Eth, IB, SS)
- Leveraging the Lustre nodemap to enable I/O from clients leveraging Illinois Auth (instead of NCSA Auth)
- Made available via numerous mechanisms:
  - Native Lustre
  - Globus
  - NFS
  - SMB
  - S3
  - Kubernetes
  - Jupyter Notebooks
  - Web Gateway (soon)

Lustre Usage @ NCSA

Other

- Nightingale
  - Controlled data system (ePHI, CUI, etc.)
- ICCP
  - Illinois /scratch FS (very early ES400NVX3 system)
- Innovative Systems Lab
  - Providing /scratch for NCSA's hardware testbed

Experiences with Nodemap in Production

- NCSA operates systems in two different authentication domains
  - NCSA LDAP
  - Illinois Active Directory
- NCSA LDAP underpins our national and industry focused systems that serve users across the United States
- Illinois Active Directory is used on systems that are deployed specifically for Illinois research teams
  - Gives users a common identity that they already use on campus
  - Eases integration with other systems on the Urbana-Champaign campus, especially domain joined machines and workstations

Experiences with Nodemap in Production

- Historically file systems served either authentication domain but not both due to lack of uid/gid alignment between the systems
- In January 2025, all primary compute systems servicing Illinois researchers were moved into our NPCF facility
  - Allowed for a possible merge with Taiga for capacity storage (/projects)
- Illinois systems sit behind 4 LNET routers that translate between HDR200 & 200GbE
  - All clients behind these routers are subject to the nodemap policy
- Currently have ~9,000 UIDs being mapped and ~600 GIDs
  - These numbers growing daily

```
[root@tgio01 ~]# lctl nodemap_info IllinoisCC | grep "idtype: uid" | wc -l
8944
[root@tgio01 ~]# lctl nodemap_info IllinoisCC | grep "idtype: gid" | wc -l
591
```

Experiences with Nodemap in Production

- Automation for new users/groups to map is already in place and keeps up with changes every 30 minutes
  - Should be able to speed this up to every ~10 minutes soon
- Working through process for removing users from the nodemap when they leave the Illinois system
  - Handy to know who owned the data…but also need to keep the nodemap clean
  - Likely going to be a quarterly cleanup that we manually trigger and oversee
- Additional mapping of UIDS/GIDs is about to get underway as we deploy a system that supports the non-UIUC Illinois campuses (who each have own AD)
  - We look forward to the dynamic nodemap features slated to land in 2.17 ☺

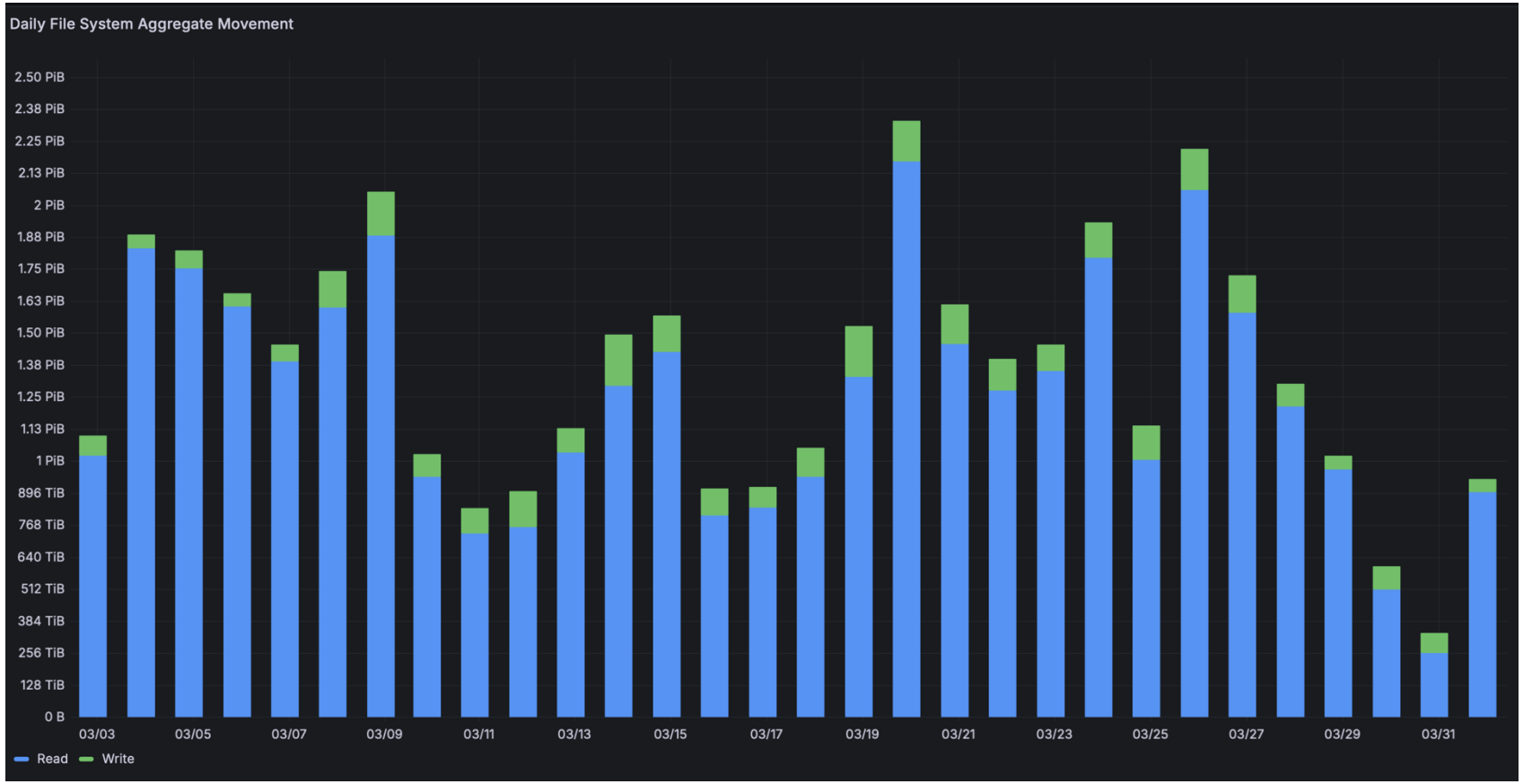
Early Results with Lustre Transparent Data Compression

- Broad appeal for this feature with the Delta Project Office, External Advisory Board, and users
  - Never enough NVME capacity for researchers
  - Required no user-facing code changes or Lustre “knowledge”…it came “free”
- Deployed on Delta file system as part of the PFL for /work/nvme
  - At present only a subset of allocations; doing a progressive roll out
- Usage started with early testing of the feature in early Fall 2024
  - Initially didn’t have support for ARM (Nvidia GraceHopper on DeltaAI) which delayed the ability to start full rollout
  - While waiting for ARM support did testing to choose optimal compression algorithm using Delta’s x86_64 (AMD Milan) nodes

Early Results with Lustre Transparent Data Compression

- Factors in choosing compression algorithm
  - Obviously want as much compression as possible without meaningful workload performance impact
  - Vast majority of compute nodes are quad or octo GPU equipped (A40, A100, H200, GH200), those are the compute engines used almost exclusively…CPUs on them are generally not fully taxed
  - Large read-write imbalance in favor or read for workflows on this system

Early Results with Lustre Transparent Data Compression



- Settled on lz4hc for our default algorithm
  - Ended up seeing better results than we'd anticipated
  - ~1.3x reduction on real user data on allocations we've scanned the results on
  - As expected, files generally either compress decently…or not at all

Early Results with Lustre Transparent Data Compression

- Performance testing with CSDC enabled
  - Focused on user-workload performance and not synthetic benchmark performance
  - NCSA's Application Support team ran 7 representative micro-benchmarks using the more popular applications that run on DeltaAI (pytorch, tensorflow, etc.)
  - Found no meaningful workload slow downs with compression enabled, all variances within error margins

- Coordinated rollout strategy
  - Pushing out to more active allocations first
  - Watching to ensure there are no major job run-time variances (so far none)
  - Hope to have fully rolled out by June 2025

Experience with DNE3 with ”many” MDTs

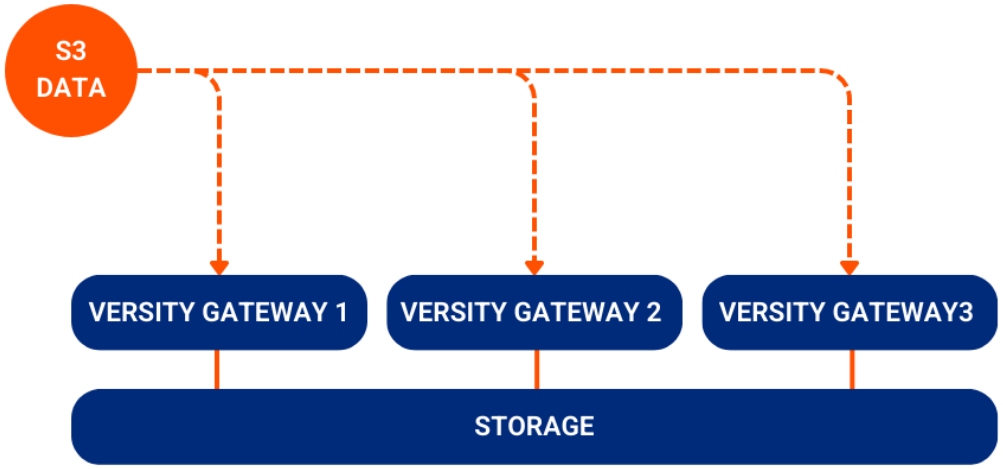
- Delta’s /work FS features 48 total MDTs
  - Equally balanced with 1 per stack on each ES400NVX2
  - ~257 million inodes per MDT (~12.3 billion total)

- Leveraging DNE3 with a max-inherit-rr of 10; working decently well with ~2.3 billion inodes on the FS

```
[root@cn057 ~]# lfs df -i /work | grep MDT | awk '{print $5}' | sort -n | uniq -c
      7 20%
     13 21%
      7 22%
      4 23%
      9 24%
      2 25%
      3 26%
      1 27%
      1 28%
      1 33%
```

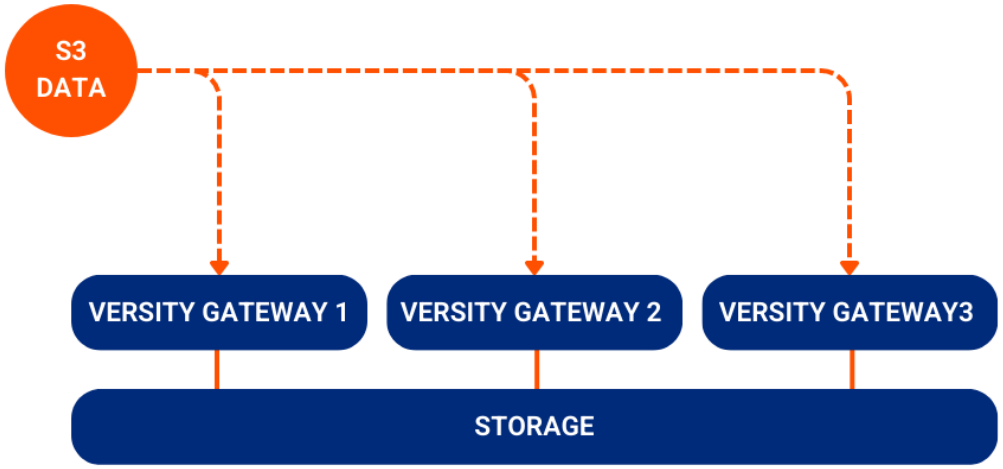
Versity S3 Gateway + Lustre

- The S3 protocol is almost a first-class citizen within our infrastructure
- Gaining popularity in a variety of use cases
  - Hosting and sharing of datasets for science gateways
  - Data transmission from remote data collection systems (eg. ag research)
  - Shared training datasets for AI pipelines
  - Integration with data backup and data movement tools



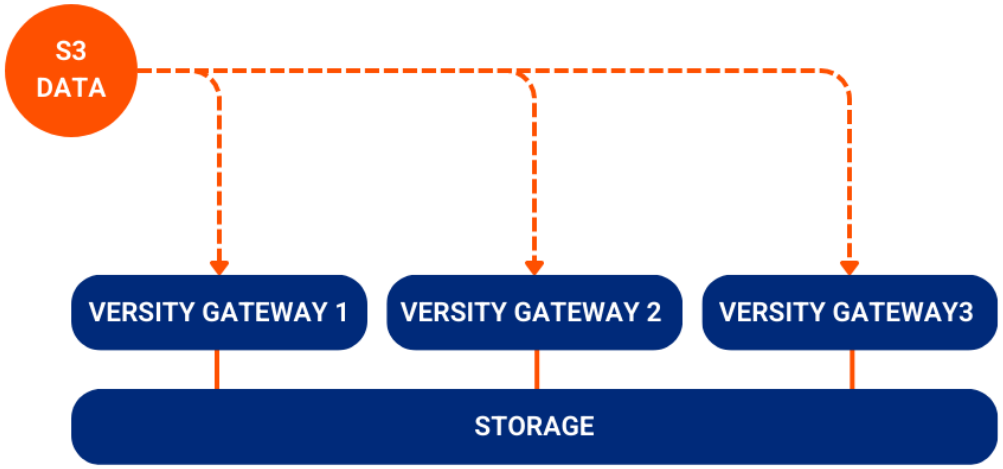
Versity S3 Gateway + Lustre

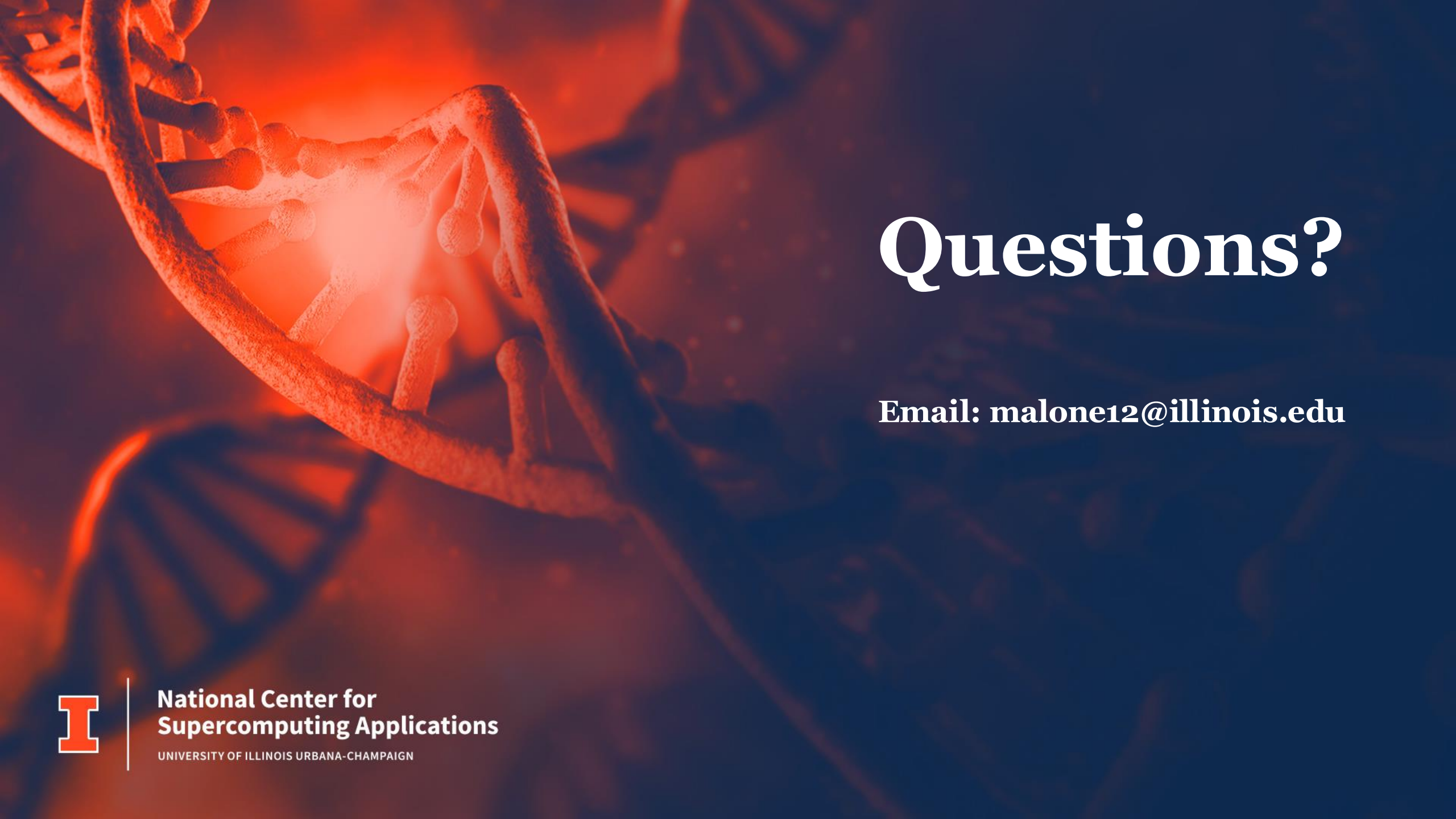
- Co-deployed with Globus on our DTN nodes for Taiga
- Allocations can file requests for S3 endpoints in front of parts/the whole of their allocation
  - Endpoints can be read/write or read-only
- Currently orchestrated via Ansible and systemd unit files/services



Versity S3 Gateway + Lustre

- A lot of active development around this service for us to scale it for broader access; multiple projects underway:
  - Orchestration via Kubernetes
  - User self-service key changes
  - More fine-grained access controls
  - Tie ins to allocation provisioning
- Plan to share as much of the management framework with the community as we can





Questions?

Email: malone12@illinois.edu

National Center for Supercomputing Applications
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN