ACCESS | Advancing Innovation

STEP-1: Data Movement
May 23, 2024

David Wheeler, NCSA / University of Illinois

Sean Stevens, NCSA / University of Illinois

OpenCI Award: OCI 21-38296

NSF

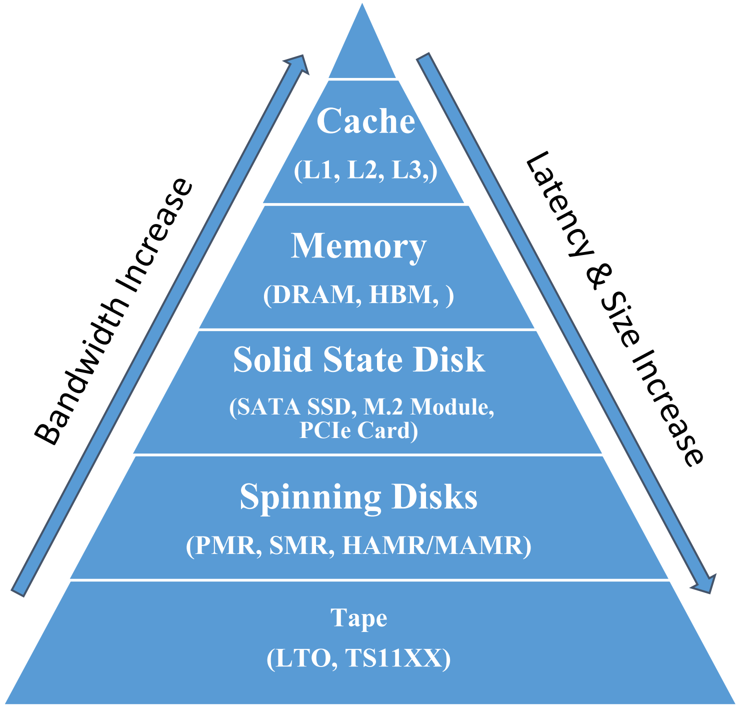
Outline

- Storage Overview
- Hardware and Software
- Data Transfer
- Data Movement Examples
- Globus Details

Storage Overview

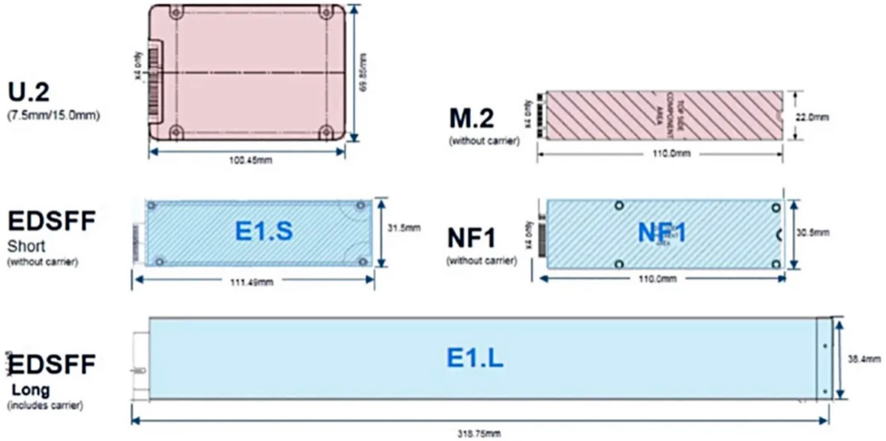
What is storage?

- Processor Cache
  - Fastest access; closest to the CPU; temporary (L1, L2, L3)
- System Memory (DRAM)
  - Very Fast access; close to CPU but not on it; temporary
- Solid State Storage
  - Fast access (esp. random)
  - Can be internal or part of an external storage system
  - Capable of high densities with high associated costs
- Spinning Disk
  - Slower; performance is tied to access behavior
  - Can be internal or part of an external storage system
  - Capable of extremely high densities
- Network / Cloud storage
  - Network - Can scale from slow to extremely fast, high density
- Tape
  - Extremely slow; typically used for cold storage



Common Storage Building Blocks

- Media
  - HDDs (SATA & SAS)
  - SSDs (SATA, SAS, NVMe)
  - Tape (LTO, TS11XX)
- Media Formats
  - HDD - 3.5” (2.5” dying out, some 10/15K SAS remain)
  - SSD - 2.5” (varying thicknesses), U.2/3, E1.S/L, E3.S/L, M.2
  - Tape - LTO (Open Standard), TS11XX (IBM format)
- Enclosures
  - JBOD & JBOF (Enclosures/Drawers)
  - Controller (Couplets)
  - Storage Servers



Common Connecting Fabrics

Network Fabrics

- **Ethernet**
  - Speeds: 1Gb - 400GbE (800GbE soon)
  - RJ-45, SFP+, QSFP+, QSFP-DD
  - TCP/RoCE
- **Infiniband**
  - EDR, HDR, NDR (100Gb, 200Gb, 400Gb) are modern versions in use today
  - RDMA support
- **Slingshot**
  - HPE specific (at present)
  - Ethernet based with "extra stuff"

Other Storage Fabrics & Carriers

- **Fiber Channel & SAS**
  - 24Gb SAS is now GA; 12Gb still common
  - 64GFC is now available (6.4GB/s per direction)
- **PCIe**
  - Gen 5 (32GT/s per lane) ~64GB/s per direction in a x16 slot (available in 2023)
  - Gen 4 (16GT/s per lane) ~32GB/s per direction in a x16 slot (available since 2019)
  - Carrier for CXL devices, NVME drives, NICs, etc.
  - Gen 6 (64GT/s per lane) ~128GB/s per direction in a x16 slot -- likely 2025/2026 GA

Storage Hardware and Software

Hardware - Data Transfer Node (DTN)

- CPU – the higher clock rate the better
- Memory - at least 64GB, more is better
- What connections (type, speed, count(cards/ports))
    - Storage:
        - local (direct connected to the DTN)
        - external (parallel FS, NAS, etc)
    - Fabric type(s)
        - Ethernet, Infiniband, Slingshot, SAS, FC, etc.
- PCI slots - correct type and number of slots for connection requirements
    - Form factor (number of lanes: x8, x16, etc)
    - PCIe-2/3/4 - most likely targeting gen 3 or 4 currently (2023)
- Software - use case dependent
    - Globus, rsync, scp/sftp, s3(different object tools), http

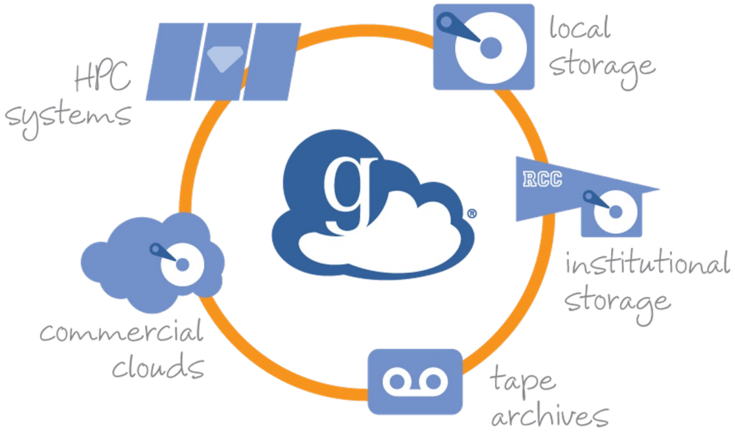
Software

- Specifics determine the best tool
  - Connectivity
  - Underlying storage (local FS, parallel FS, Flash/Spinning/Tape)
  - Access (CLI, DTN, etc)
  - Dataset
  - Transfer (one-time, repeated, etc)
- Common tools:
  - Globus
  - Rsync
  - S3 (different object tools)
  - SCP/SFTP
  - HTTP

- Specialized tools:
  - Tiered Filesystems
    - Typically FS specific policy tools
  - Backup/Archive/Tape systems
    - Typically system specific tools

Globus

- Used by researchers, universities, national labs, government, etc
- Simple interface for end user
- Move/Share/Discover Data where it lives: from a supercomputer, lab cluster, scientific instrument, tape archive, public cloud, or laptop.
- Share Data with external collaborators who might not have local account
- “Connectors” allow transfer to/from many storage systems/services
  - Local FS, parallel FS, AWS S3, Google Drive/Cloud, Microsoft Azure/OneDrive, Box, iRODS, HPSS, ceph, and more…
- Built-in parallel transfers (when configured) can transfer multiple files in parallel
- Scheduling options for repeated transfers (Globus Timer API)
- Automation for multi-step data movement processes (Globus Flows)



rsync

- Cli tool provides transfer/syncing
  - local FS (directory to directory)
  - Between different systems (over network typically using ssh as the transport)
- Common use cases
  - Sync dataset from one location to another within an FS or remote system
  - Migrate data from old to new system
    - For a very large migration, rsync might be used in conjunction with find or another indexing tool. Listings get split and then multiple rsync processes are run in parallel.
  - Scheduled sync(utilizing cron) to an offsite system for backup

SCP/SFTP - Secure Copy/File Transfer

- Included with SSH (Secure Shell)
- SCP
  - Simple CLI command `scp [[user@]host1:]file1 ... [[user@]host2:]file2`
  - Works like “cp” command, but source and/or destination can be a remote system
- SFTP
  - CLI command, connect to remote ftp server, then upload or download files
  - Secure File Transfer Protocol - secure version of FTP.
  - Some organizations still maintain ftp for upload/download.

S3 - multiple options

- AWS S3 tools
- rclone
- Globus (w/S3 connector)
- API/SDK access
- …

HTTP

- Backbone protocol of the internet
- Simple downloads to the world (everyone has a browser)
- **Not** optimized for large data transfer
- Alternatives to browser - Command Line
  - curl
  - wget

Tiered Storage Systems

- Some HPC/enterprise filesystems have the concept of tiered storage
- Storage Tier - group of like storage media providing specific level of service
  - Flash/SSD: Fast Tier
  - Spinning Disk: Slower Tier
  - Tape: Archival Tier
- Tiers can be
  - Internal to FS (pools)
  - External Storage System
    - Common examples HPSS
    - Spectrum Protect
- Data Movement may be manual or automated
- Hierarchical Storage Management(HSM) system
  - Policy driven/automated data movement between tiers
  - Built-in (Spectrum Scale Policies control HSM)
  - Bolt-on external solution (Lustre has hooks for external HSM tools)

Data Transfer

Data Movement - Who, What, When, Where… ?

Knowing the specifics will guide the choices of appropriate resources for managing and moving data…

Data Movement - Who, What, When, Where… ?

- Where is the data currently?
  - existing system or building new
  - to be generated by compute
- How much?
  - Size and number of files/objects
- Format?
  - POSIX, Object
- Who?
  - has/needs access
- Available connectivity?
  - External networks
  - Storage network/fabric
  - Existing systems for Data Movement?
- Restrictions?
  - CUI, CJIS, FERPA, FISMA, GLBA, HIPAA/PHI, ITAR,…

- Where is the data going to/from?
  - End user PC to/from storage system
  - Between tiers of a storage system
  - Between local storage systems: different clusters, high performance, archive
  - Between remote storage systems across the country/globe
  - Between local and cloud storage
  - Between cloud storage systems

Dataset(s) and packaging

- Dataset
  - collection of data to be transferred
- How a dataset is packaged can have a great impact on data movement
- Considerations:
  - Underlying storage
    - Source and Destination
  - Transfer software - parallel transfers?
  - Network Connection
  - Future use cases
  - Restrictions

Dataset(s) and packaging - considerations

- Underlying Storage:
    - Is there scratch space to utilize while packaging or can packaging be done on the fly?
    - Local direct connected vs Cluster Filesystems
    - Different Filesystems have different characteristics (even depending on how they're architected)
        - Multi-tier flash -> disk -> tape
        - All Flash
        - Disk only
        - etc…
    - Tape
        - Most efficient with a smaller number of large files (within reason)
            - files larger than a single tape = more planning
        - Usually considerations due to specific tape system as well

Dataset(s) and packaging - considerations

- Connectivity
  - Storage Fabric between storage and DTNs
  - Network connectivity between DTNs on source and destination
  - Find bottlenecks
    - Build to maximize where possible
    - Find the balance where things don't perfectly align
- How will the data be accessed
  - How will the data be utilized on the destination
- Restrictions
  - Privacy and other security restrictions can affect all of the above considerations

Data Movement Examples

Tiered Storage - Compute Job

- Move dataset to fast tier for use during compute job - possible triggers:
  - Compute job prologue
  - Hierarchical Storage Management (HSM) policy
- Results also initially land (written) on fast tier
- Once computation is complete dataset/results move back to slower tier(s) - possible triggers:
  - compute job epilogue code
  - HSM policy

Tiered Storage - Access Based

- HSM policy configuration
  - Move data to slower tiers as data ages (not accessed)
  - When accessed move data back to faster tiers
- Example
  - Researcher completes computations on a specific dataset
    - dataset ages (no access)
    - HSM moves it from Fast to Slow tier
  - 1 year later, the researcher is ready to publish utilizing the original dataset
    - Data is accessed
    - HSM pulls the data back to the Faster tier(s)

End user PC to/from storage system

- One-Time transfer
    - POSIX - CLI access allowed
        - scp/rsync
    - Object/other
        - rclone, AWS S3 tools, other
    - Globus DTN available (POSIX, S3, other)
        - Install GCP on PC and utilize Globus

- Repeated/Scheduled transfer
    - POSIX - CLI access allowed
        - cron + scp/rsync
    - Object/Other
        - cron + appropriate tool
    - Globus DTN available
        - Cron + Globus CLI/API
        - Globus Tasks

Between local storage systems

- Network/Site local storage systems
- Both FS available/mounted on same cluster/system
  - End user utilize common OS utilities for simple moves (cp, mv, rsync)
- Separate cluster FS's on local network (End User)
  - Account access assumed
  - Rsync or scp
  - Globus: if both systems have Globus DTNs
- Large Data Migration between filesystems
  - Likely to be executed by administrators
  - rsync or FS specific tooling
  - Both FS's mounted on DTN or separate DTN's with proper Network/Account Access

Between separate storage systems

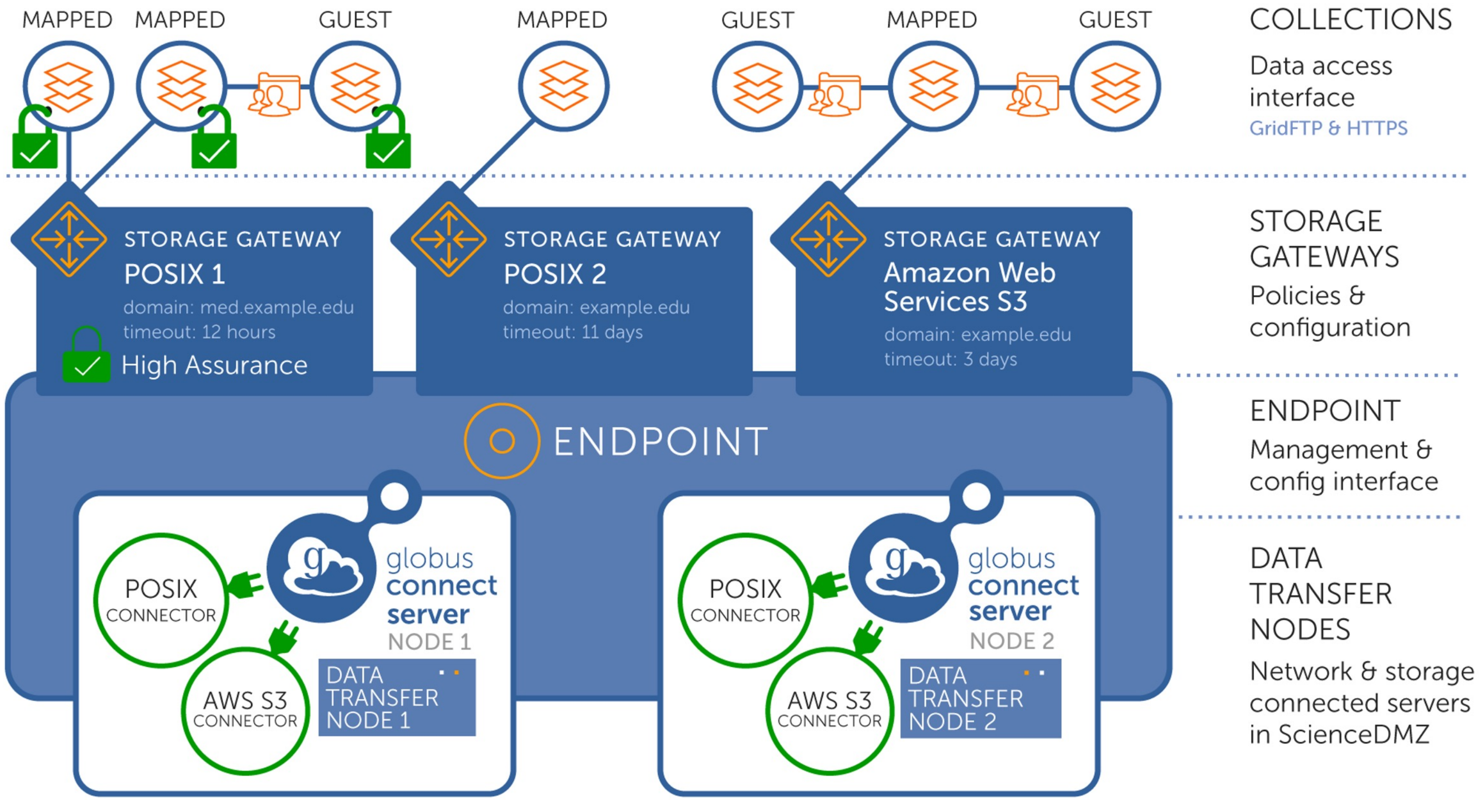
- Storage systems could be remote and/or local
- Ensure network path(s) between systems
- Need User Account Access on both sides:
  - For (scp/rsync)
    - account with shell access of some kind on each system
    - Somewhere to run from DTN node, local PC with access to both systems, etc.
  - For Globus/S3/other
    - Globus DTN on both sides configured appropriately for underlying storage
    - User account/access key to each system

Globus Details

Globus Detail - Terminology

- Globus Connect Personal
  - connect laptop/desktop to move/share data
- Globus Connect Server
  - runs on DTNs for multi-user endpoints
- Data Transfer Node (DTN)
- Endpoint
  - provides the interface for server management and configuration
- Storage-Gateway
  - provide the storage access policies for the endpoint's connected storage systems
- Collection
  - provide the data access interfaces
  - Mapped - user accessing must have local account
  - Guest
    - user can access without a local account
    - access based on permissions granted by an authorized user via Globus

Globus Detail - Terminology



COLLECTIONS

Data access interface

GridFTP & HTTPS

STORAGE GATEWAYS

Policies & configuration

ENDPOINT

Management & config interface

DATA TRANSFER NODES

Network & storage connected servers in ScienceDMZ

Globus Detail - Security

- Access Control
- Data remains at institutions, no storage/routing via Globus
- Integrity checks of transferred data
- Enforced encryption of Globus control communications
- Options for encryption of data in transit

Summary

- Storage solutions are made up of building blocks that are selected based on many design requirements and constraints
- Knowing the tradeoffs, features, and limitations of storage resources is essential to effective data movement
- There are many resources in the CI community that are available to support research data transfers

Resources

https://linuxclustersinstitute.org/ - LCI offers advanced technical training for those interested in deploying high-performance computing clusters through its workshops.

https://fasterdata.es.net - An Expert Guide for End-to-End Performance Tuning, Tools and Techniques

https://globus.org - Move, share, & discover data no matter where it lives.

Questions?