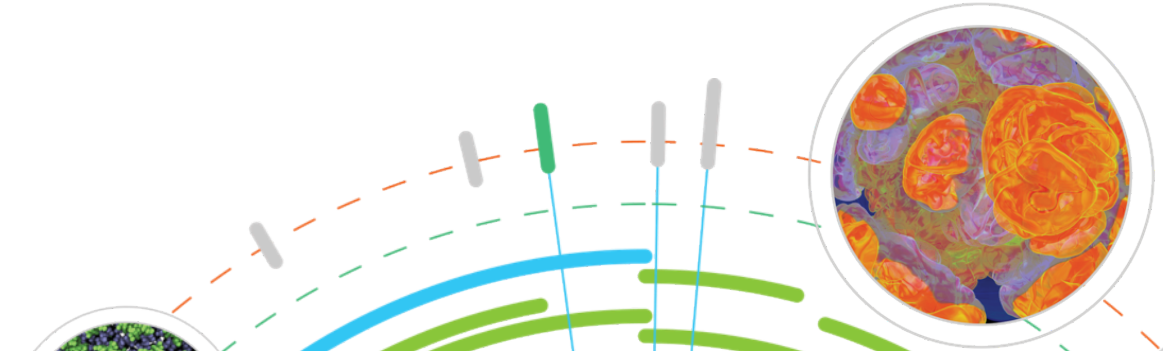
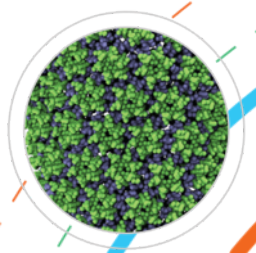
GPFS HPSS Interface (GHI)

Jack O'Connell
06/06/2016





Argonne Leadership Computing Facility

Argonne NATIONAL LABORATORY

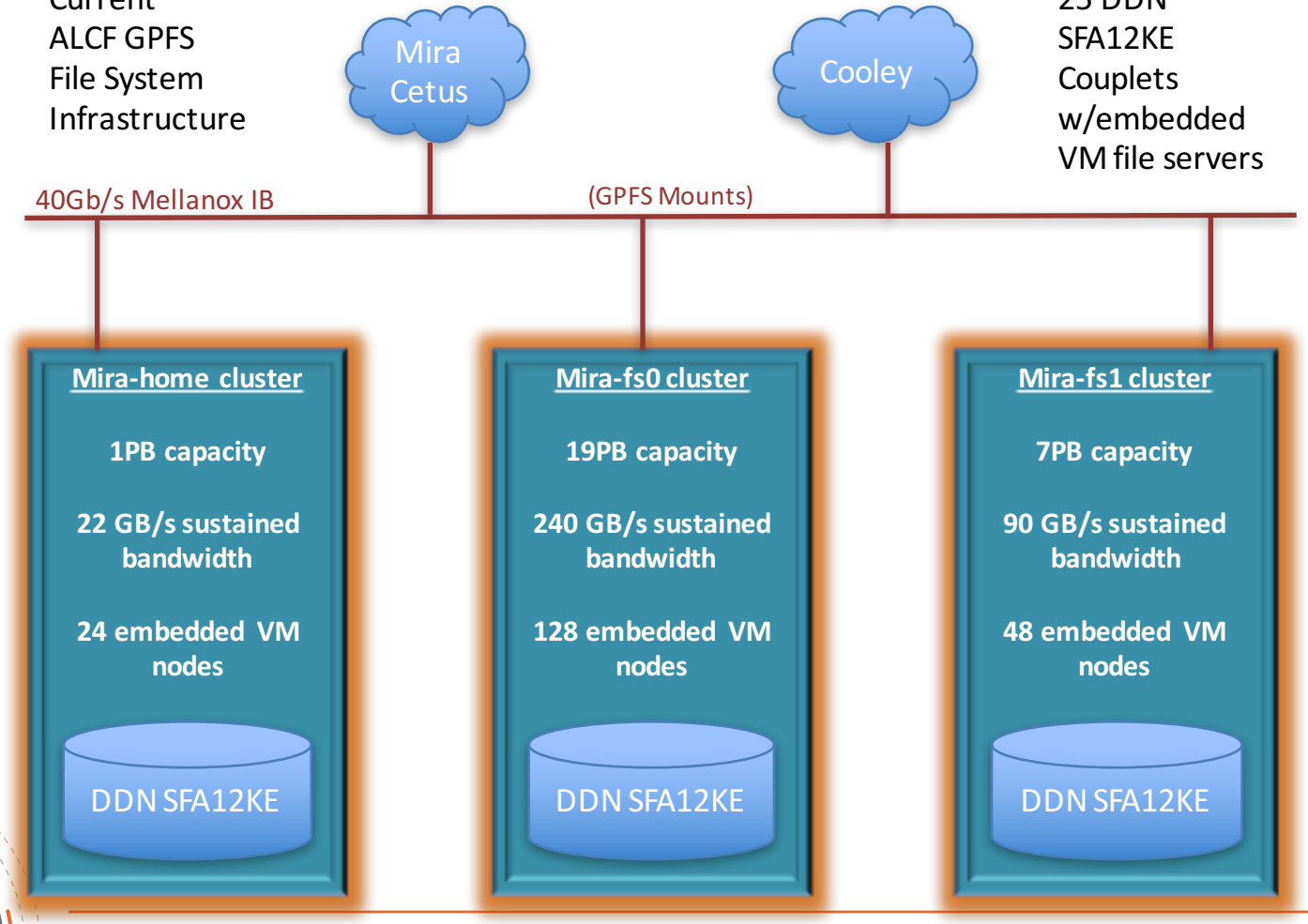
Overview

- GPFS file systems
- Current state of the ALCF Archive System
- A look towards the future
  - GHI

File System Overview

- Mira-home is intended to store executable files and configuration files.
  - User quota limits are enforced.
  - Data is fully protected thru GPFS metadata and data replication, GPFS snapshots and nightly backups.
- Mira-fs0 and mira-fs1 are intended as intermediate-term storage for Mira/Cetus job output such as checkpoint datasets.
  - Project associated fileset quota limits are enforced.
  - Only metadata replication is enabled.
  - The user is responsible for archiving their own data.
  - After project expiration, quota limits are reduced, data is archived and the fileset is removed.

Current
ALCF GPFS
File System
Infrastructure

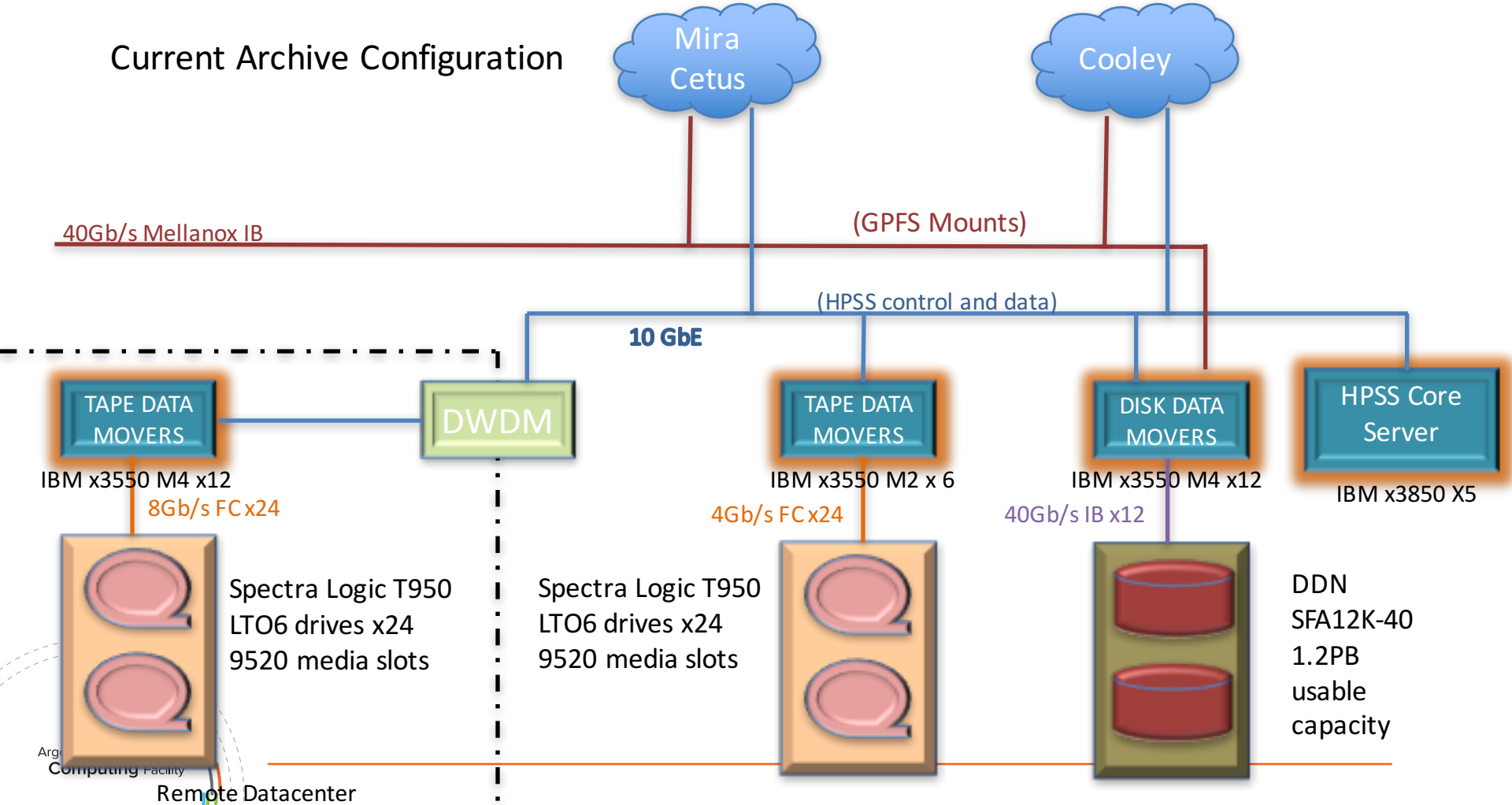


25 DDN
SFA12KE
Couplets
w/embedded
VM file servers

HPSS Archive Overview

- Redundant Core servers support manual failover
- NetApp resident DB2 metadata repository
- 1.2PB disk cache residing on DDN SFA12K-40 storage
- Twelve disk cache data movers
- Two 8 frame Spectra Logic T950 libraries
- Eighteen tape data movers

Current Archive Configuration



Supported Client Interfaces

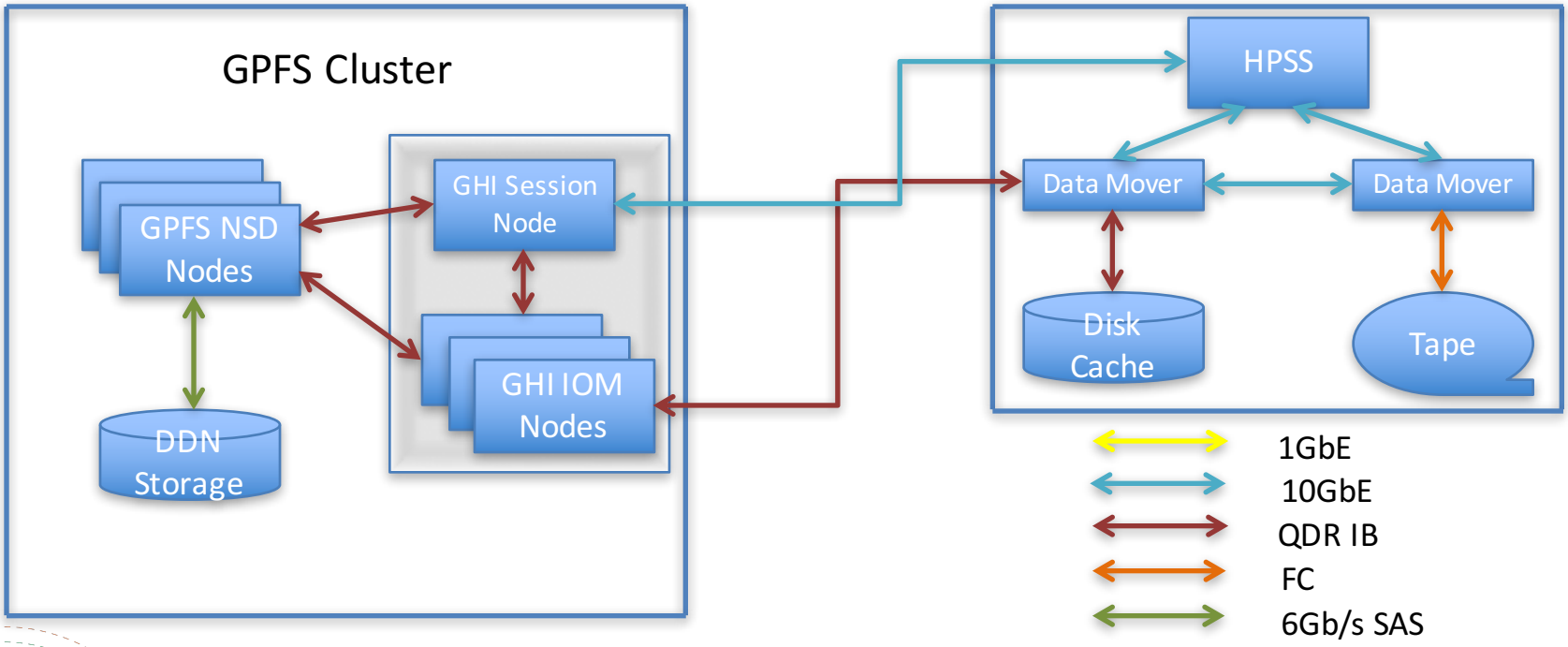
- HSI (Hierarchical Storage Interface)
  - Unix style command environment
  - Third party transfer agent reduces network traffic
  - Multiple concurrent transfers of smaller files

- Globus/GridFTP
  - Web based and command line interfaces
  - Globus provides hands-off management of transfers
  - GridFTP for script based workflows

GPFS HPSS Interface (GHI)

- Provides a hierarchical storage management and backup solution for GPFS
- GHI works quietly in the background providing two primary services
  - Space management
  - Disaster recovery
- GHI is fully automated and transparent to users

High Level GHI Integration View



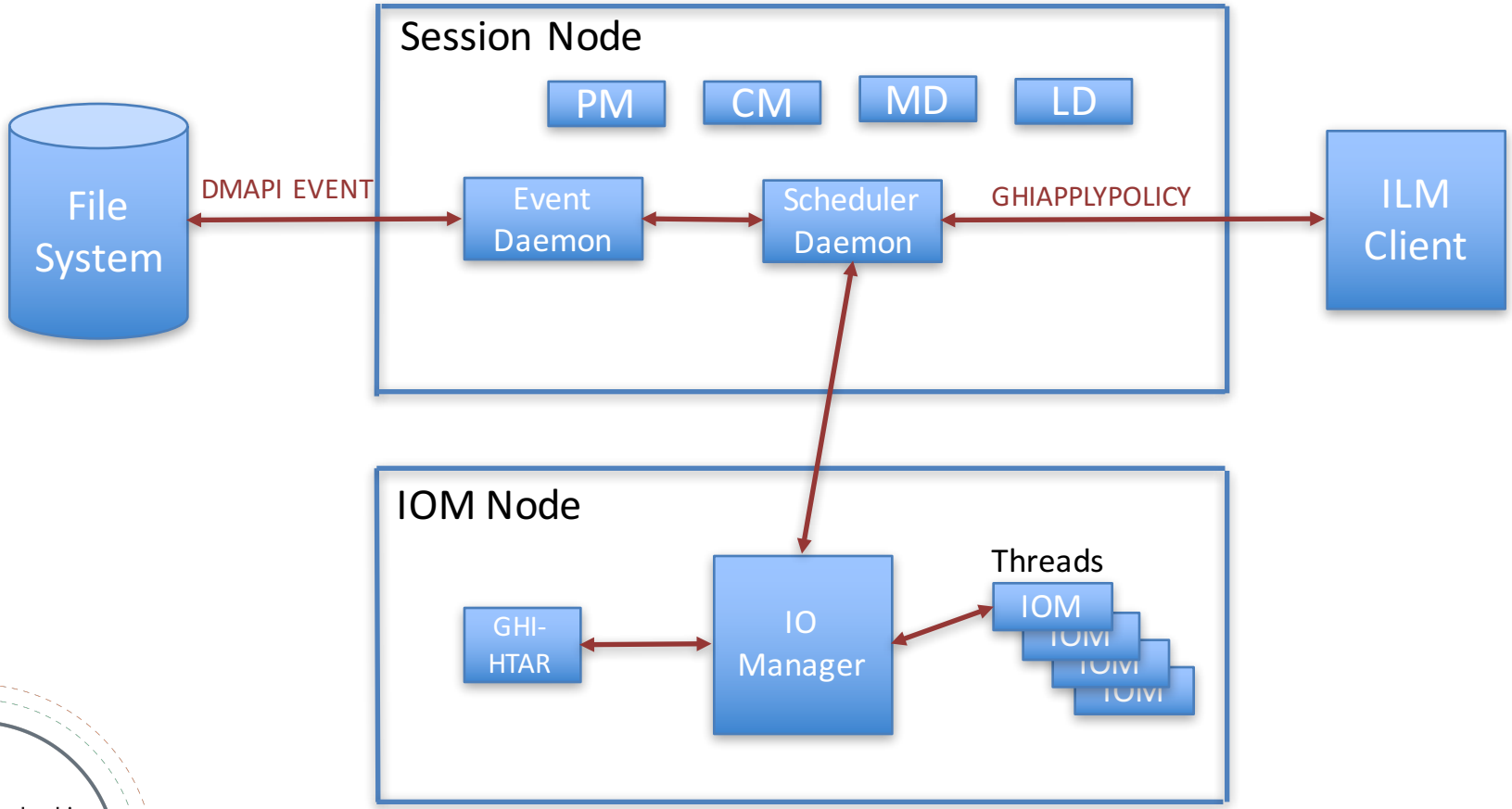
GHI Overview

- GHI is installed on the GPFS cluster quorum nodes and manages DMAPI enabled file systems
- The installation includes both the HPSS client and DB2 client
- GHI processes run on the following node types
  - Session node (Cluster manager node)
    - Configuration Manager (CM) – Processes GHI configuration changes
    - Process Manager (PM) – Starts and stops other GHI processes
    - Event Daemon (ED) – Registers for and processes DMAPI events
    - Log Daemon (LD) – Maintains rotational central logs
    - Mount Daemon (MD) – Captures mount and unmount events
    - Scheduler Daeamon (SD) – Schedules IO requests with IO Managers
  - IO Manager (IOM) node
    - IO Manager (IOM) – Performs non-aggregate transfers to/from HPSS
    - GHI-HTAR – Performs aggregate transfers to/from HPSS

General Workflow

- The Session node is where a DMAPI session exists and is registered to receive DMAPI events.
- GPFS generates DMAPI events when users perform operations such as read, write and truncate on managed files.
- The GHI Event Daemon captures the DMAPI events from the DMAPI session queue and submits requests to the GHI Scheduler Daemon.
- GPFS Information Life Cycle (ILM) policies are use to quickly identify files to migrate to HPSS or purge from GPFS.
- The Scheduler Daemon accepts requests from the Event Daemon and the ILM clients and communicates with the IO Managers to transfer data, or to process purge requests.

Workflow View



Administration

- Migration
  - Uses the GPFS ILM policy management interface to select files and write them to HPSS
  - Should be completed for all or most file system data before running a backup
- Backup
  - Uses the GPFS snapshot feature to capture the point in time state of the file system
  - Uses a GPFS ILM policy to migrate any un-migrated files to HPSS
  - Uses a second GPFS ILM policy to collect and write GPFS metadata to HPSS
- Restore (if necessary)
  - Restores the file system namespace and associated extended attributes
  - Only GPFS file metadata is restored
  - All files will have a residency of HPSS only
  - Files that will be immediately referenced can be recalled in bulk using a GPFS ILM policy

Admin. cont.

- Purge
  - Removes file data blocks from the file system to free space
  - Leaves the inode allowing the file to be visible to the user
  - Uses a GPFS ILM policy to select files to purge
  - Can be run at any time, but best practice is to automate.
- Recall
  - Uses a GPFS ILM policy to select files to be recalled from HPSS
  - Alternatively, the ghi_stage command may be used.
- Monitor
  - Scan /var/hpss/ghi/log files regularly
  - View /<FS mount point>/scratch/mon/mon_sd_out for real time scheduler daemon activity
  - View /<FS mount point>/scratch/mon/mon_iom_out for real time IO manager activity

End User Interaction

- GHI stores HPSS references as GPFS extended attributes
- Users are able to check the residency of their files with the GHI command ghi_ls
  - GPFS resident – The file has not been migrated to HPSS
  - Dual resident – The file exists in both GPFS and HPSS
  - HPSS resident – The file has been purged from GPFS but exists in HPSS
- Users can access HPSS only resident files by opening the file, which initiates a recall
  - At the administrative level, HPSS only resident files can be recalled using a GPFS ILM policy or the ghi_stage command

ALCF GHI Implementation

- Enable a threshold policy for the managed file system
  - Set high water mark at 90% and low water mark at 80%
  - Select larger files with the oldest atime to purge
- Migrate project filesets targeting expired projects first
  - Preserve the existing space management policies described earlier
    - Reduce quota limit, archive data and remove data
  - Once migrated, expired project fileset data blocks may be purged, freeing space
  - Continue adding policy rules to the migration policy until all filesets are migrated and continue at regular intervals
    - Adding too many at once may overload the HPSS infrastructure
- Once all project filesets have been migrated, backup the managed file system and continue at regular intervals

Tape Storage Estimate for HPSS/GHI

Current Archive size (PB)		28			
Enter current Mira FS usage (PB)		15.5			
Enter annual FS growth rate as % usage		30.0%			
Modify Compression Rates from "specs" tab					
		Estimated Number of LTO Cartridges			
		LTO6		LTO8	
Fiscal Year	Potential Archive size (PB)	Native	Compressed	Native	Compressed
2016	43.5	18245	14596		
2017	48.2	20196	16156		
2018	54.2	22731	18185	4440	3552
2019	62.1	26027	20822	5083	4067
2020	72.3	30312	24250	5920	4736

Questions?