Storage Discussion

GACRC Advisory Committee Meeting

February 24, 2015

Current Inventory

ID	type	FS	Usable Capacity
sn0	Penguin “high-performance” IceBreaker	ZFS	32T
sn1	Penguin “high-performance” IceBreaker	ZFS	32T
sn2	Penguin “high-performance” IceBreaker	ZFS	32T
sn3	Penguin “high-capacity” IceBreaker	ZFS	280T
sn4	Penguin “high-capacity” IceBreaker	ZFS	280T
Lustre	Xyratex Lustre Appliance	Lustre	240T
Panasas	Panasas	PanFS	156T

- While the 10x ArchStors and 5x Thumpers are still in production, they are all slated for near-term decommissioning – weeks, not months.

Current Status (zcluster)

ID	type	Status	Used/Usable Capacity (TB)
sn4	“HC” IceBreaker	Installed & Operating. /SCRATCH	92/280
Panasas	Panasas	In Operation. /HOME	110/156
			202/436

Current Status (other)

ID	#groups	OFLOW	LAB STORAGE	COPIED to sn3?	Status	allocated/total
rccstor1	1		✓		Dying	20/22
rccstor2	5	✓		No. Lost	Dead	26/32
rccstor3	8	✓		Yes	Active	47/48
rccstor4	Drained of Data. Ready for Decommissioning					
rccstor5	4		✓	N/A	Dead	15/32
rccstor6	5	✓		In Progress	Active	19/32
rccstor7	6		✓		Dying	17/32
rccstor8	3		✓		Dying	13/32
rccstor9	4		✓	Yes	Active	8/32
rccstor10	Drained of Data. Ready for Decommissioning					

Current Status (other)

ID	#groups	NFS MOUNTS	COPIED to sn3?	Status	capacity
Thumper1	GACRC Galaxy MySQL DBs	✓	YES	Active	29/36
Thumper2	GACRC	✓	YES	Active	29/36
Thumper3	1	✓	N/A	Active	34/36
Thumper4	1	✓	N/A	Active	20/36
Thumper5	1	✓	NO	Active	25/36
					137/180

Current Status (Sapelo)

ID	type	Status	Used/Usable Capacity (TB)
sn0	“HP” IceBreaker	Installed & Operating. /HOME	1/32
sn1	“HP” IceBreaker	Installed & Operating. /HOME	4/32
sn2	“HP”IceBreaker	Installed & Operating. Not in use.	0/32
sn3	“HC” IceBreaker	Installed & Operating. /PROJECT	58/280
Lustre	Lustre	PO issued. Equipment not received	0/240

Some working definitions

- **Snapshot** - Copies of files that are stored on the same storage system as the original files. Snapshots are primarily used to recover files that have been accidentally deleted or corrupted within the recent past. Users are able to manage the file recovery tasks. Snapshots are not maintained beyond a defined rotation schedule, i.e., some number of hourly, daily, weekly, and monthly snapshots are kept on the storage system.
- **Backup** - Copies of files and/or snapshots kept on a storage system (disk/tape) other than the one that the original files reside on. Backups are primarily used to recover files following a catastrophic failure of the original file or storage system. Backups require administrators to perform file system recovery tasks. Like snapshots, backups have a defined rotation schedule.
- **Archive** - Copies of files that are not currently being accessed, on a resilient storage system dedicated to reliable long-term storage. Archives will be tape-based or disk-based, and typically part of a disaster recovery plan. The files may be copies of original data which is stored elsewhere (individual groups having their own copies), or the archive storage system may be fed by a dedicated "backup" storage system.

Notes for SCRATCH on zcluster & Sapelo

SCRATCH on Sapelo:

- Xyratex Lustre appliance (PO received, appliance being built)
- 240TB usable/320TB raw
- IB connected, will be mounted on Sapelo compute nodes
- running Robinhood Policy Engine
- files transfers in/out only through IB-connected Sapelo copy nodes.

SCRATCH on zcluster:

- “high-capacity” IceBreaker chain, currently in production
- 280TB usable/320TB raw
- specifically sn4 (escratch4)
- we would have to script the 90 day retention policy
- files transferred in/out only through zcluster copy nodes

Policy Statement for SCRATCH File System

The SCRATCH file system resides on a high-performance storage device and is to be used uniquely for temporary storage of files in use by actively running compute jobs. Files are to be removed from SCRATCH when a job completes, *e.g.* can be copied to the PROJECT file system. ***The SCRATCH file system is not backed up in any way and no snapshots are taken***.

Any file that is not accessed or modified by a compute job in a time period no longer than 90 days will be automatically deleted from the SCRATCH file system. Once deleted it will NOT be possible for the GACRC to recover the file. Measures circumventing this policy will be actively discouraged.

There is no storage size quota for SCRATCH usage. Space is only limited by the physical size of the scratch space being used. If usage across the entire file system is more than 80% of total capacity, the GACRC will take additional measures to reduce usage to a more suitable level. Amongst possible actions, request/force users to clean up their SCRATCH directories, reduce temporarily the 90 day limit to a lower limit, say 30 days.

Notes for HOME on zcluster & Sapelo

HOME on zcluster:

- Panasas – 103TB used/156TB capacity
- support of Panasas is till June 2016
- users are already there – just continue business as usual.

HOME on Sapelo:

- Currently sn0 is configured for HOME – (32TB usable / 48TB raw)
- in order to cover 1,000 users with 100GB allocations, we would need ~100TB
- this means we would require at least sn1 and sn2 in play, as we will be creating user accounts over time.

Backup of HOME:

In order to be able to backup both HOME on zcluster and Sapelo with regular frequency, we require at a minimum a device containing ((133*1.5) + (100*1.5))TB or 350TB (i.e. "high-capacity" IceBreaker chain with 3x fully populated (with 4TB drives) expansion cabinets rather than 2x) or equivalent capacity in tape.

Caveat: backing up on tape still creates the need to provision a new storage device in case of the loss of the source storage device. Backing-up to disk on an appropriate target device would allow the target to temporarily fill in while the broken source is fixed.

Policy Statement for HOME File System

The HOME file system resides on a high-performance storage device and is used for long-term storage of files, typically programs and scripts, needed for analysis on the GACRC computing clusters.

All users have 100GB allocated for their HOME usage. Groups may request a separate 100GB allocation for a directory under /usr/local/lab/, for shared use of common applications, libraries, and scripts.

HOME directories will have daily, weekly and up to 3 monthly snapshots kept on the same storage unit to protect against accidental file deletion. Users are strongly encouraged to make their own copies of critical files, while accepting any risks associated with its usage.

Notes for PROJECT on zcluster & Sapelo

- “high-capacity” IceBreaker chain – 280TB usable/ 320TB raw
- sn3 available initially
- files transfers in/out only through 10GigE-connected copy nodes
- PROJECT not mounted on compute nodes of Sapelo or zcluster
- additional capacity could be met through future acquisitions
- 280TB represents ~1TB per current group. Nothing more.

Note: To immediately increase PROJECT capacity, sn2 could be populated with 4TB drives (~$22k for a set of 80x 4TB drives, while the 80x 600GB drives could be repurposed, e.g. local scratch on compute nodes, JBOD, Hadoop cluster). This would make sn2 unavailable for Sapelo HOME, which we would have to replace eventually.

Backup of PROJECT:

In order to be able to backup PROJECT with regular frequency, we require at a minimum a device containing (280*1.5)TB or 420TB (i.e. high-capacity IceBreaker chain with three expansion cabinets rather than two) or equivalent capacity in tape, this per high-capacity IceBreaker chain.

Policy Statement for PROJECT File System.

The PROJECT file system resides on lower-performance/higher-capacity storage devices, accessible by all GACRC clusters' login and copy nodes. PROJECT will not be accessible on the clusters' compute nodes. This space is to be used by groups for storage of active projects using Sapelo and/or zcluster. PROJECT should not be seen as a long-term repository, as it is not designed as such. Once a project is completed, data should be moved from the PROJECT space to user-managed storage, freeing up capacity for the next active project.

Each group can request a PROJECT volume with an initial 1TB allocation, accessible by all users ascribed to the group, where the sharing of files will be enabled. Users are encouraged to consider their PROJECT space as the primary area to transfer compute job inputs/outputs. Additional space can be requested by a Faculty on behalf of his/her group, in increments of 1TB.

The GACRC reserves the right to establish a cost-recovery rate for PROJECT storage beyond the initial 1TB allocation. Appropriate communications will take place in such an event.

PROJECT directories will have daily, weekly and up to 3 monthly snapshots kept on the same storage unit to protect against accidental file deletion. Users are strongly encouraged to make their own copies of critical files, while accepting any risks associated with its usage.

Additional Text

Not currently stated in HOME policy statement:

A backup of the HOME directories will be made <TBD> onto a separate storage device, to protect against hardware failures.

Not currently stated in PROJECT policy statement:

A backup of the PROJECT directories will be made <TBD> onto a separate storage device, to protect against hardware failures.

Disclaimer to be placed on all storage policy statements:

Snapshot retention, data purge and quota allocation policies are subject to change based on available storage capacity, users' demand, equipment condition and availability, as well as other constraints.

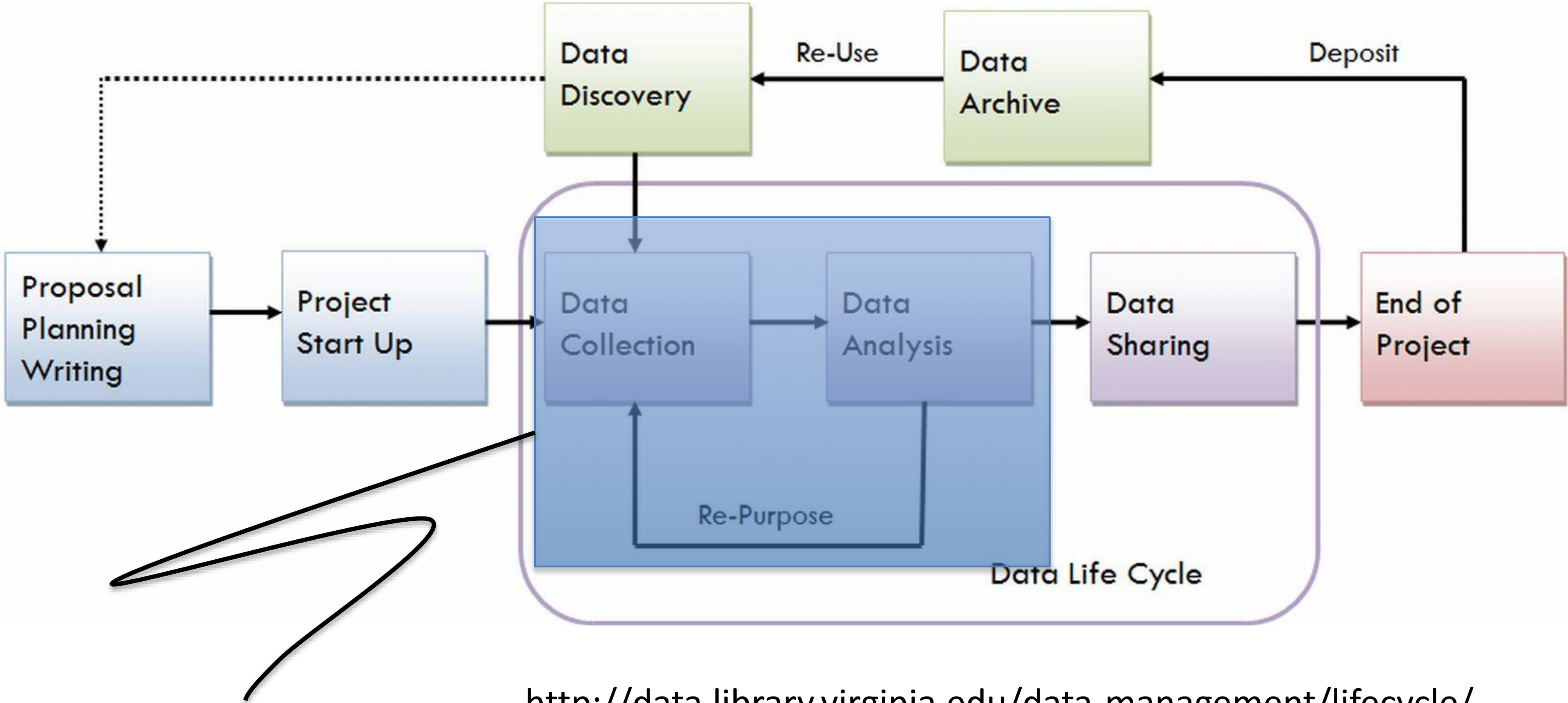
Please Note: Outside access to the PROJECT file system is not currently supported through NFS to a destination outside of the Boyd Data Center, Samba or CIFS. Transfer protocols available through the Sapelo and zcluster copy nodes are ftps, scp, rsync, amongst others.

More info is found at https://wiki.gacrc.uga.edu/wiki/Transferring_Files

Working Statements

- The GACRC is only able to provide storage services to those groups that are actively using the GACRC computational resources, zcluster and Sapelo.
- In this age of "Big Data", providing 1TB for PROJECT per GACRC research group is clearly insufficient. Some groups require 100TB+, which once properly backed-up will grow to near 300TB of capacity.
- By supporting only *active* projects, this brings a clear delimitation with respect to a research group's Data Management Plan. Since the GACRC is not involved in any discussions related to DMPs, it cannot <currently> take any responsibility towards DMP-related services.
- Capacity is but one aspect of storage environments. Other characteristics dictate what services can be supported on any given environment. There is no single storage device that can fulfill the many different services that are discussed over all DMPs currently in place at UGA.
- While the GACRC has two well qualified Linux HPC systems administrators, we do not have a staff position that can be described as an HPC storage engineer. Additionally, some DMPs will require administration of large-scale databases served through web services. Again, this is not an area of staff expertise at the GACRC.

Research Data Life Cycle



http://data.library.virginia.edu/data-management/lifecycle/

GACRC's CURRENT ROLE

Some Numbers of Importance

UGA investigators currently registered at the GACRC – across all current services	**214**
UGA investigators having actively used the zcluster in the past year, i.e. submitted at least 1 job on the zcluster “those using & generating data at the GACRC”	**140**
UGA investigators that received funding in 2013* “likely {using/generating/repurposing/making public} research data”	**903**
UGA Faculty (instruction, research, public service) “likely {using/generating/repurposing/making public} data, broadly defined”	**2,879**

*according to OVPR 2013 Annual Report

Storage Services – Present & Future

Initial Services (available on Sapelo & zcluster)

	Sapelo	zcluster
SCRATCH	Lustre (240TB)	ZFS (280TB)
HOME	ZFS (60 TB)	PanFS (156TB)
PROJECT	ZFS (280TB)	

Next-phase services

Service	$$$	Staff	Possible Responsibility
Backup environment for HOME & PROJECT	Yes	No	GACRC
New transfer and data management services for GACRC (e.g., Globus Online, GitLab)	No	No	GACRC
Centralized network-attached storage (NFS, CIFS mounts) available remotely across campus	Yes	Yes	GACRC (central) Colleges (users)

$$$: Investment required
Staff: Additional staff required, at the GACRC or elsewhere

Storage Services – Present & Future

Future Developments

Service	$$$	Staff	Possible Responsibility
Data management tools and improved support to DMPs	Yes	Yes	Libraries, OVPR, EITS
Basic data curation services	Yes	Yes	Libraries
Tools/technology to support data sharing (via cloud services, portals, gateways)	Yes	Yes	Libraries, EITS
Low-cost backup/archiving services for research data	Yes	Yes	GACRC
HIPAA/PII, FISMA protected data services	Perhaps	Yes	GACRC, InfoSec, Legal

This is a non-exhaustive list of possible services.

Backup of HOME and PROJECT volumes
Some Comments

The use of snapshot strategies (kept on same storage unit) for HOME & PROJECT allows for the restoring of accidentally deleted files (i.e. recovery from user error).

Further data protection would be brought by an appropriately chosen off-device backup strategy that would ensure recovery from a disaster (i.e. rebuilding a complete file system).

We propose a best practice approach taken by PACE at Georgia Tech, and described in the following document:

http://www.edtechmagazine.com/higher/article/2014/08/backup-metrics-hpc-environments

The following slide summarizes this approach.

Please note: backing-up the GACRC systems requires further investments.

Backups are file-based and use a disk-to-disk method

Initially implement an “expandable” system, not a “capacious” system

Deploy a sufficiently performing solution to be able to complete a backup cycle in a reasonable time-window

Backup solution is designed to function as the primary storage in case of complete failure, restoring temporarily full functionality without requiring large copies of data, while the failed storage device is repaired

Multiple backup servers, configured as high-capacity versions of the user-facing storage servers

Use internal checksums to help guarantee end-to-end data integrity

Use native file system compression

Use ‘rsync’ or ‘zfs send’ for file transfer

Backup frequency has an upper bound determined by the rate at which the target data can be copied from the primary storage to the backup device.

Regularly test the entire backup/restore process

For situations where there is increased risk, the disk-to-disk approach can be supplemented by enabling disk-to-disk-to-tape

Backup of Current Inventory

ID	type	Proposed Costs
sn0	Sapelo /HOME	$25k
sn1	Sapelo /HOME	
sn2	Sapelo & zcluster /PROJECT	$90k
sn3	Sapelo & zcluster /PROJECT	
sn4	zcluster /SCRATCH	N/A
Lustre	Sapelo /SCRATCH	N/A
Panasas	zcluster /HOME	$30k
Including software & installation:		**$150k**

These are preliminary costs, after only one quote. More exploration to follow with a few manufacturers. This does not include the ~$22k to transform sn2 into a high-capacity IceBreaker.

Current Lab Storage Groups

Group	Cluster Users?	Allocation
CPH	NO	20T
Steve Miller	NO	15T
Roberto Docampo	NO	11T
Jessica Kissinger	YES	7.5T
Justin Turney	NO	5T
Galaxy	YES	5T
Karl Lechtreck	NO	3T
Kojo Mensa-Wilmot	NO	1T
Kimberly Klonowski	NO	1T
Jacek Gaertig	NO	1T
Ping Shen	NO	1T
Paul Schliekelman	YES	1T
Michael Strand	YES	1T
Brian Haas	NO	1T
Amy Parks	NO	500G
Zachary Lewis	YES	500G
Kevin Ayres	NO	500G
	11xNO/5xYES	**75TB**

Current Oflow Groups

Group	Capacity Used	Group	Capacity Used
Thomas Mote	13.3T		
GGF	9.55T	Robert Schmitz	10T (6T lost)
Jim Leebens-Mack	8.36T	Travis Glenn	2T (lost)
Mary Ann Moran	7.74T	QBCG	4T (500G lost)
Ying Xu	6.53T	Shaying Zhao	2.4T (copied)
Katrien Devos	3.22T	Andrew Paterson	1T (copied)
David Hall	2.67T		
Shaying Zhao	1.45T		
David Landau	1.26T		
Jessica Kissinger	916G		
Jeffrey Dean	887G		
Zachary Lewis	393G		
Brendan Hunt	317G		
Total:	**58TB**		

Time for a vote

Policy Statement for SCRATCH File System

The SCRATCH file system resides on a high-performance storage device and is to be used uniquely for temporary storage of files in use by actively running compute jobs. Files are to be removed from SCRATCH when a job completes, *e.g.* can be copied to the PROJECT file system. ***The SCRATCH file system is not backed up in any way and no snapshots are taken***.

Any file that is not accessed or modified by a compute job in a time period no longer than 90 days will be automatically deleted from the SCRATCH file system. Once deleted it will NOT be possible for the GACRC to recover the file. Measures circumventing this policy will be actively discouraged.

There is no storage size quota for SCRATCH usage. Space is only limited by the physical size of the scratch space being used. If usage across the entire file system is more than 80% of total capacity, the GACRC will take additional measures to reduce usage to a more suitable level. Amongst possible actions, request/force users to clean up their SCRATCH directories, reduce temporarily the 90 day limit to a lower limit, say 30 days.

Policy Statement for HOME File System

The HOME file system resides on a high-performance storage device and is used for long-term storage of files, typically programs and scripts, needed for analysis on the GACRC computing clusters.

All users have 100GB allocated for their HOME usage. Groups may request a separate 100GB allocation for a directory under /usr/local/lab/, for shared use of common applications, libraries, and scripts.

HOME directories will have daily, weekly and up to 3 monthly snapshots kept on the same storage unit to protect against accidental file deletion. Users are strongly encouraged to make their own copies of critical files, while accepting any risks associated with its usage.

Policy Statement for PROJECT File System.

The PROJECT file system resides on lower-performance/higher-capacity storage devices, accessible by all GACRC clusters' login and copy nodes. PROJECT will not be accessible on the clusters' compute nodes. This space is to be used by groups for storage of active projects using Sapelo and/or zcluster. PROJECT should not be seen as a long-term repository, as it is not designed as such. Once a project is completed, data should be moved from the PROJECT space to user-managed storage, freeing up capacity for the next active project.

Each group can request a PROJECT volume with an initial 1TB allocation, accessible by all users ascribed to the group, where the sharing of files will be enabled. Users are encouraged to consider their PROJECT space as the primary area to transfer compute job inputs/outputs. Additional space can be requested by a Faculty on behalf of his/her group, in increments of 1TB.

The GACRC reserves the right to establish a cost-recovery rate for PROJECT storage beyond the initial 1TB allocation. Appropriate communications will take place in such an event.

PROJECT directories will have daily, weekly and up to 3 monthly snapshots kept on the same storage unit to protect against accidental file deletion. Users are strongly encouraged to make their own copies of critical files, while accepting any risks associated with its usage.