Winter 2010-11 Disk Storage Strategy

Oracle has pulled a Scrooge on us, voiding our November order for 320 TB of Thors - before they were discontinued. We didn't make it in time. Now we need to explore options and plans for our disk storage. Also to get us through this particular crunch - we have some 55 TB free at the time of writing, and owe 128 TB to KIPAC and BABAR!

Possible elements of the plan:

- clean up existing space
- explore replacing disks in existing fileservers with 2 TB ones
- find a new vendor(s)
- change storage model

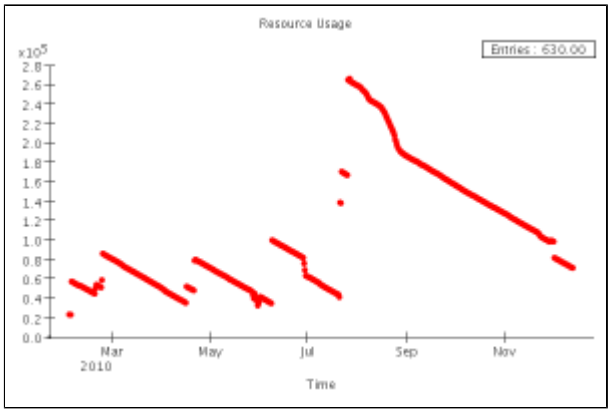
Log

- For a snapshot of current xroot usage, see http://www.slac.stanford.edu/~wilko/glastmon/xrddisk.html
- 16 Dec 2010 - Performed cleanup of /glast/Scratch and P110 reprocessing directories. Available xroot space increased from ~48 TB to ~120 TB.

Clean up

We have 1.05 PB of space in xrootd now. Some 640 TB of that is taken up by L1, and 90% of that space is occupied by recon, svac and cal tuples.

[added by Tom]
For the plot (12/14/10) - note that it does not include the "wasted" xrootd space per server, currently running about 21 TB.

Current xroot holdings: 

Data in tables current as of 12/8/2010

Total Space	1043 TB	35 servers (wain006 removed)
xrootd overhead	21 TB	(disk 'buffer' space)
Available Space	1022 TB	

Current usage:

Space used	966.778 TB	95 %
Space free	55.580 TB	5%

Consumption rate:

Level 1	783 GB /day	(averaged over period since 4 Aug 2008)

Note that L1 has a certain amount of retries. These are removed in a purge, which has not been performed for quite some time.

Therefore, with current holdings and usage, there is sufficient space for 71 days of Level 1 data (running out ~17 Feb 2011)

Usage distribution:

path	size [TB]	#files	Notes

/glast/Data	785	807914	
/glast/Data/Flight	767	792024	
/glast/Data/Flight/Level1/	713	535635	670 TB registered in dataCat
/glast/mc	82	11343707	
/glast/mc/ServiceChallenge/	56	6649775	
/glast/Scratch	51	358511	~50 TB recovery possible (removed 16 Dec 2010)
/glast/Data/Flight /Reprocess/	50	226757	~18 TB recovery possible
/glast/level0	13	2020329	
/glast/bt	4	760384	
/glast/test	2	2852	
/glast/scratch	1	51412	
/glast/mvmd5	0	738	
/glast/admin	0	108	
/glast/ASP	0	19072	

Detail of Reprocess directories:

path	size [GB]	#files	Notes	Removal Date
/glast/Data/Flight/Reprocess/P120	22734	65557	potential production	
/glast/Data/Flight/Reprocess/P110	14541	49165	removal candidate	16 Dec 2010
/glast/Data/Flight/Reprocess/P106-LEO	8690	1025		
/glast/Data/Flight/Reprocess/P90	3319	2004	removal candidate	
/glast/Data/Flight/Reprocess /CREDataReprocessing	631	22331		
/glast/Data/Flight/Reprocess/P115-LEO	533	1440		
/glast/Data/Flight/Reprocess/P110-LEO	524	1393	removal candidate	16 Dec 2010
/glast/Data/Flight/Reprocess/P120-LEO	503	627	potential production	
/glast/Data/Flight/Reprocess/P105	264	33465	**production**	
/glast/Data/Flight/Reprocess/P107	212	4523		
/glast/Data/Flight/Reprocess/P116	109	25122	**production**	
/glast/Data/Flight/Reprocess/Pass7-repro	22	16	removal candidate	
/glast/Data/Flight/Reprocess/P100	8	20089	removal candidate	

Detail of top-level Monte Carlo directory:

path	size [GB]	#files	Notes
/glast/mc /ServiceChallenge	57458	6649775	
/glast/mc/OpsSim	24691	4494890	
/glast/mc/DC2	1818	136847	
/glast/mc/OctoberTest	41	29895	
/glast/mc/XrootTest	33	18723	removal candidate
/glast/mc/Test	14	13403	removal candidate

Detail of Monte Carlo 'ServiceChallenge' directories is quite lengthy and appears in full here.

Data current as of 12/9/2010:
In /glast/mc/ServiceChallenge there are a total of 387 Monte Carlo tasks consuming 57276 GB.

- There are 99 "old" Monte Carlo tasks (GR v14 or older), consuming 26475 GB (46.2235491305 %)
- There are 232 "new" Monte Carlo tasks (GR v15 or newer), consuming 28363 GB(49.5198687059 %)
- There are 28 "GRB" Monte Carlo tasks, consuming 40 GB(0.0698372791396 %)
- There are 28 uncategorizable Monte Carlo tasks, consuming 2398 GB(4.18674488442 %)

It has been suggested that all MC datasets produced with Gleam versions v14 and older be removed (presumably along with any "HEAD" versions). If that policy is enacted, ~29 TB of space would be liberated.

Breakdown of space usage by Level 1 data products (from dataCatalog):

Filetype:	TotSize/#Runs =	Avg file size	(%)
RECON:	399257.600/13532 =	29.505 GiB	(58.2%)
CAL:	100147.200/13536 =	7.399 GiB	(14.6%)
SVAC:	71270.400/13535 =	5.266 GiB	(10.4%)
DIGI:	61542.400/13537 =	4.546 GiB	(9.0%)
FASTMONTUPLE:	28262.400/13537 =	2.088 GiB	(4.1%)
MERIT:	22528.000/13537 =	1.664 GiB	(3.3%)
GCR:	676.800/13537 =	51.196 MiB	(0.1%)
FASTMONTREND:	432.800/13537 =	32.739 MiB	(0.1%)
DIGITREND:	337.100/13537 =	25.500 MiB	(0.0%)
FILTEREDMERIT:	324.900/ 7306 =	45.538 MiB	(0.0%)
MAGIC7HP:	244.700/13315 =	18.819 MiB	(0.0%)
LS1:	183.700/13537 =	13.896 MiB	(0.0%)
CALHIST:	181.400/13536 =	13.723 MiB	(0.0%)
RECONTREND:	172.700/13537 =	13.064 MiB	(0.0%)
TKRANALYSIS:	172.400/13536 =	13.042 MiB	(0.0%)
LS3:	156.800/13537 =	11.861 MiB	(0.0%)
RECONHIST:	124.400/13536 =	9.411 MiB	(0.0%)
MAGIC7L1:	55.000/13315 =	4.230 MiB	(0.0%)
FASTMONHIST:	53.400/13535 =	4.040 MiB	(0.0%)
LS1BADGTI:	48.200/ 7306 =	6.756 MiB	(0.0%)
CALTREND:	45.500/13537 =	3.442 MiB	(0.0%)
FT1:	39.100/13537 =	2.958 MiB	(0.0%)
DIGIHIST:	32.600/13537 =	2.466 MiB	(0.0%)
etc...			

Cleanup summary

If the most aggressive policy discussed so far is enacted, one would liberate space as follows:

Scratch	50 TB	
Reprocessing	18 TB	
Monte Carlo	29 TB	
Total	**97 TB**	**(nearly 10% of total)**

Replacing Existing Disks

This option was received with a frown from CD. Of course, Oracle has thought of this and charges a lot for the replacement drives, which have special brackets. And then there is the manpower to replace and the shell game of moving the data around. We'd have to press a lot harder to get traction on this option.

From Lance Nakata:

> *It's possible to attach storage trays to existing X4540 Thors and create additional ZFS disk pools for NFS or xrootd use. The Oracle J4500 tray is the same size and density of a Thor, though the disks in the remanufactured configs below are only 1TB, not 2TB. The Thors would probably require one SAS PCIe card, one SAS cable, and the J4500 would need to be located within a few meters. Implementing this might require the movement of some X4540s.*
> *Another option is to attach the J4500 to a new server, but currently that would require additional server and Solaris support contract purchases. Please let us know if you'd be interested in this so Teri/Diana can investigate pricing and quantity available.*
> *Thank you.*
> *Lance*

New Vendors

Two obvious candidates are in use at the Lab now: DDN used by LCLS, and Dell el cheapo by ATLAS.

DDN

LCLS is not super thrilled by what they got, though we are told that other labs have them and are happy. CD is getting price quotes for thor-sized systems.

Dell

These are 5200 rpm & $160/TB, but low density. At the same density, 7200 rpm disks (and maybe better connectivity?) are $275/TB. This is less than half the thor costs.

Change Storage Model

The idea is to use HPSS as a storage layer to transparently retrieve xrootd files on demand. Wilko thinks the system can push 30-50 TB/day from tape. This is comparable to the rate needed for a merit reprocessing and so is thought not to be a major impediment. In this model, we would remove the big files from disk after making 2 tape copies. They would be retrieved back into a disk buffer when needed. So we would have a relatively fixed disk buffer, and a growing tape presence.

Here are some thoughts how to implement the new storage model.
The xrootd has the be configured to automatically stage a missing file from tape to disk. This is well known and in production for BaBar. However, instead of having a client somewhat randomly staging files it will be more efficient to sort the files by tape and then stage them in that order before the client needs the files.

If the xrootd cluster is retrieving files from HPSS a policy is needed to purge files from disks. In the beginning I assume that the purging will be run manually by providing the files that should be purged (either a list of files or
a regular expression filenames will be matched against). Purging will be done before and during a processing that requires a large amount of data to be staged.

Monitoring should be setup to have a record of how much data is being staged and of the available disk space in xrootd.
A warning should be sent if the disk space will be close to exhaustion. One has to look at the total xrootd disk space as some servers will be filled and some will have space.

Currently files are backed up to single tape copies. In order to create dual tape copies a new HPSS file family has to be created. Existing files have to be re-migrated to the new file family. One should also think about if one could have a file family for just the recon files. Currently they are (might be) mixed on the same tape with all the other L1 file types which might make the retrieving less efficient.

Below are the steps that were outlined above. I put some rough time estimates for the different steps.

1. Testing large scale staging. (2 weeks)
2. Deploy a new xrootd version (the current version supports only an old depreciated interface to HPSS) (1 week)
3. configure the xroot to stage files from HPSS to disk if they are missing (1 day)
4. Setup monitoring of the staging (partly comes with xrootd) (2 days)
5. Setup (I guess Nagios) to send out warnings if the xrootd disk space will be filled up ?
6. Develop code to do allow purging files using file lists or regEx (1 week)
7. Setup new dual copy file family (2 days)
8. Re-migrate files to have dual copies (1 day work, 2-3 weeks for migration)
9. Tools to sort files by tape and prestage them in order to increase the HPSS throughput. ?