2014 Fermi storage upgrade

Notes for planning the utilization of two new Dell 360 TB file servers for Fermi

- The new Dell servers each host 360 TB of space, intended for: growth (~1 TB/day), retirement of old servers, and to allow for possible reassignment of NFS servers.

Current situation

- As of 1 Aug 2014, Fermi has 184 TB of 'free space' in the xroot cluster, of which about 21 TB resides on READ-ONLY servers, leaving about 163 TB of accessible free space. At the current rate of consumption and ignoring future reprocessings, this storage could last as long as early January 2015.

- As of 1 Aug 2014 Fermi NFS cluster consists of four 32-TB wain-class servers (128 TB), of which 98 TB (77%) have been allocated. There are 30 TB of unallocated space and 34 TB of unused allocated space. Space, per se, is not an obvious issue with the NFS cluster, and groups of clients have been segregated to independent machines to prevent unwanted interactions between, say, the ISOC data acquisition and LAT user/group disk users. The four NFS user groups currently on their own servers are:
    - ISOC
    - MonteCarlo/RSP/ASP/Reprocessing
    - ReleaseManager
    - Users & Groups
- In the NFS domain, there is a considerable amount of "inactive" historical data which is seldom accessed. Some of this data is labeled as "Fill" on the spreadsheet describing the migration from sulky->wain machines two years ago (link).

- The four existing NFS servers are operating on equipment that is over 5 years old. Each of these servers hosts 32 TB of storage.
- wain025, the user/group disk, is routinely heavily loaded and not infrequently overloaded, sometimes leading to hangs and crashes.
- within the xroot cluster, the oldest four 32-TB servers are over 6 years old (wain017, 019, 020, 021)
- the 11 next oldest 32-TB xroot servers were purchased in 2009 (wain033, 034, 035, 036, 037, 038, 039, 053, 054, 055, and 056)
- six xroot servers have been made READ-ONLY by Wilko due to frequent problems (wain053, 054, 055, 056, 069, 071) This may be due to the extensive use of Seagate drives, which fail in a way that can hang the I/O bus. Note that Maintech has begun to use Hitachi drives for replacement.

Ordered Priorities:

1. Growth and reliability for Fermi on-orbit data.
2. Upgraded server for Fermi user and group disk partitions to increase performance – but retain segregation of the four NFS groups.
3. Retirement of old equipment
    a. oldest equipment
    b. unreliable equipment

Other factors:

- SCS is planning a strategic change to GPFS and is now experimenting with the two new Fermi servers. One possible outcome of this work is to retain GPFS and run xroot on top of it. If not, then the machines will be reinitialized, the disks reformatted to XFS and configured as earlier Dell super-servers.
- There is worry about mixing xroot and NFS on the same server so that practice has been avoided in the past.

Option 1:

- Dedicate the two new servers to xroot service.
- NFS upgrade
  - Migrate all xroot data from fermi-xrd001 (90 TB, installed 3/6/2012) to other servers
  - Reconfigure fermi-xrd001 for NFS use
  - Move contents of wain026, 031 and 032 to fermi-xrd001
  - Move user partitions on wain025 to wain031 (6-months newer than wain025)
  - Move group partitions on wain025 to wain032
  - Retire wain025 and wain026
- Select eight (8) additional wains in xroot service to vacate and retire.

Discussion:

- This idea might not be popular with ISOC as it bundles their currently independent file server, wain031, with other user groups on fermi-xrd001. They seem comfortable with the status quo so the remaining issue is the age of wain031 (installed 3/9/2009). (Question: *Could one use VMs, CHOS, cgroups or some other mechanism to effectively partition a machines resources and then run separate instances of NFS on each partition to achieve some guarantee of performance for each NFS group?*)
- While this option effectively doubles the performance for the user/group disks, as well as substantially increasing the amount of space available, it does not address the age of those servers (both installed March 2009).