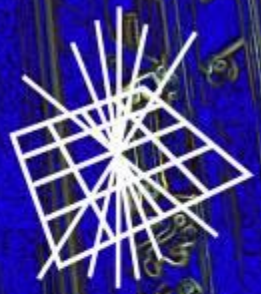


GridPP

UK Computing for Particle Physics

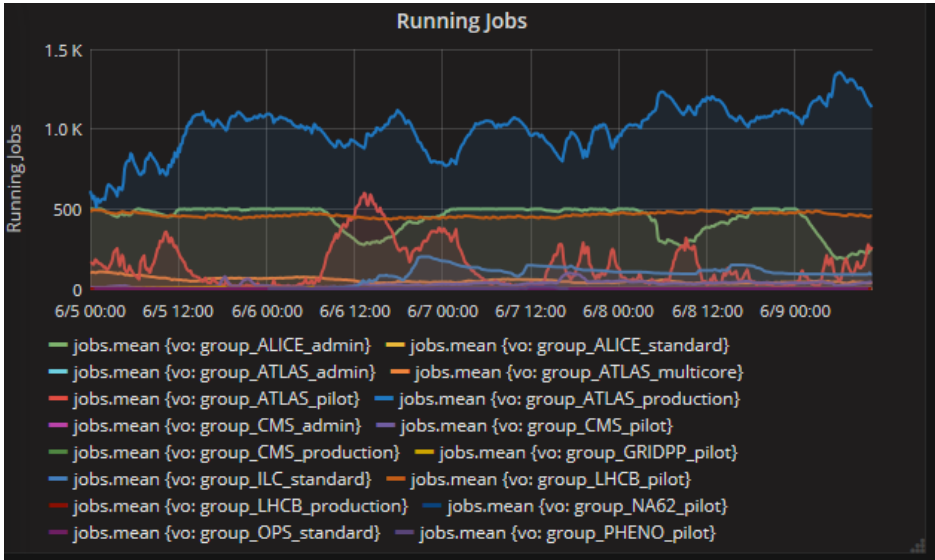
Oxford Site Report
HEPSYSMAN

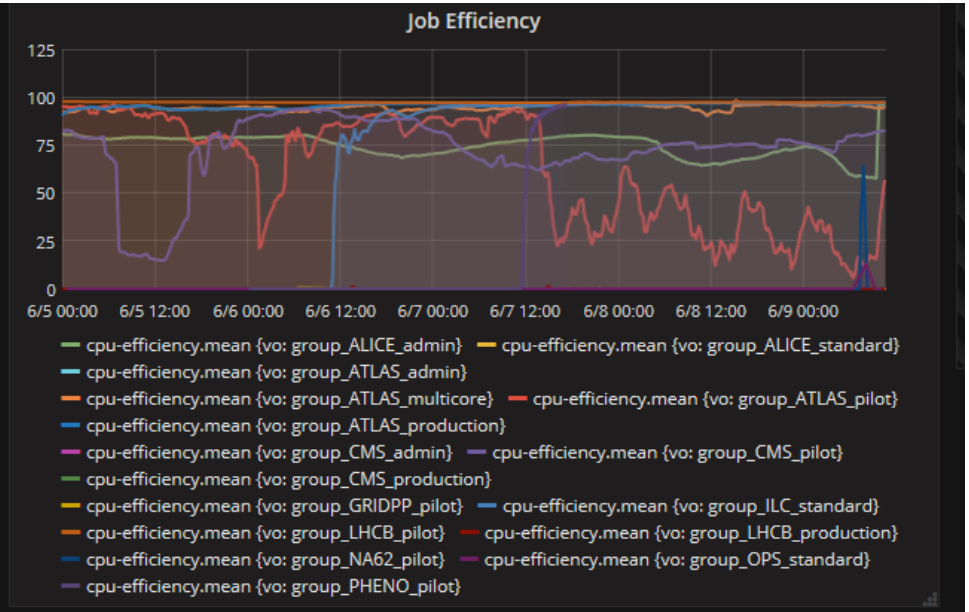
Kashif Mohammad, Pete Gronbech, Vipul Dawda

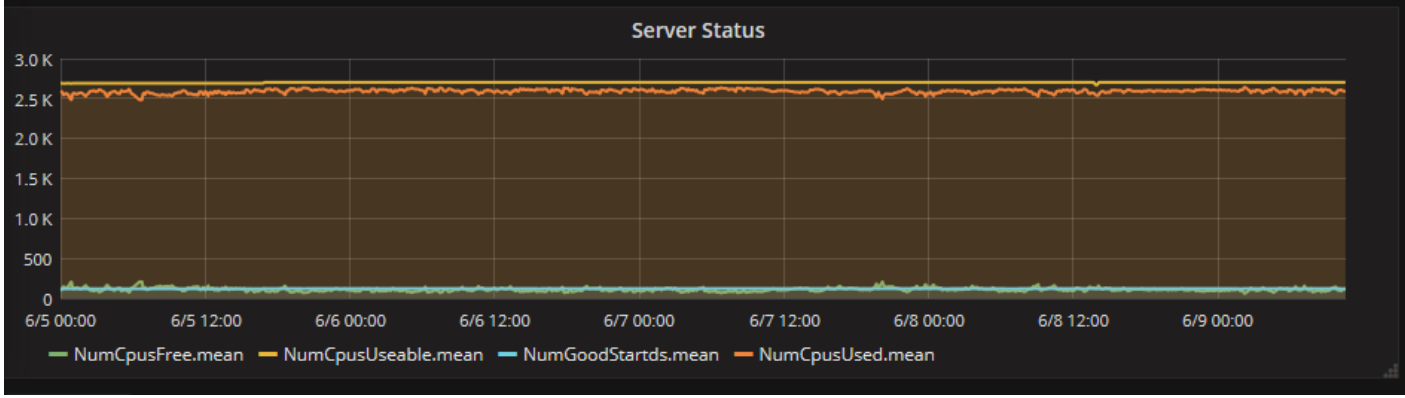
Grid Current Batch System

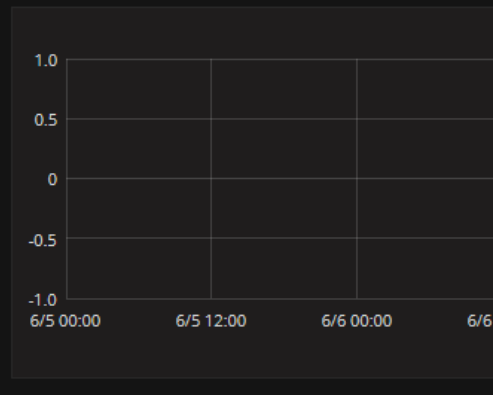
- HTCondor with ARC CE
  - t2arc01 : Production CE attached to 2.6K logical cpus
  - t2arc00 : Permanent test CE attached to two WNs
    - Planning to test CS7 on t2arc00 in next few weeks
    - Vip is going to do his hands dirty with this exercise
  - Completely configured with Puppet
  - condor_health_check with auto-nagios working very well
  - Automatic deletion of status files in ARC CE doesn't work all the time.
  - Keep removing old status files manually. /var/spool/arc is almost 100GB all the time.

Batch System Status









Storage

- Nothing exciting or new
- Running dpm 1.8.11
- Need to plan upgrade to 1.9 to fix hanging semaphore issue
  - Is it easy to upgrade ?

IPv6

- Frozen in Ewan's time
- Physics has been assigned /56 network and we have divided it into multiple /64 network
  - 2001:630:441:0900::/64 – 2001:630:0441:09ff/64
- Currently maintaining one ipv6 only UI and one dual stack UI.
- Oxford IT Services is not providing ipv6 DNS so we have our own ipv6 DNS
- Biggest blocking point is that university edge routers does not support ipv6 natively and support for ipv6 has been configured at software level.
- Routers were supposed to be upgraded last year but had been postponed three times since then

Local Batch System

- Torque and Maui on local SL6 batch system
  - Maui is almost broken
  - Torque is not well either
- Decided to move to HTCondor for CS7
  - Vip has setup a small cluster
  - Configured by Puppet
  - Waiting for some test users
-

Move from Lustre to Gluster

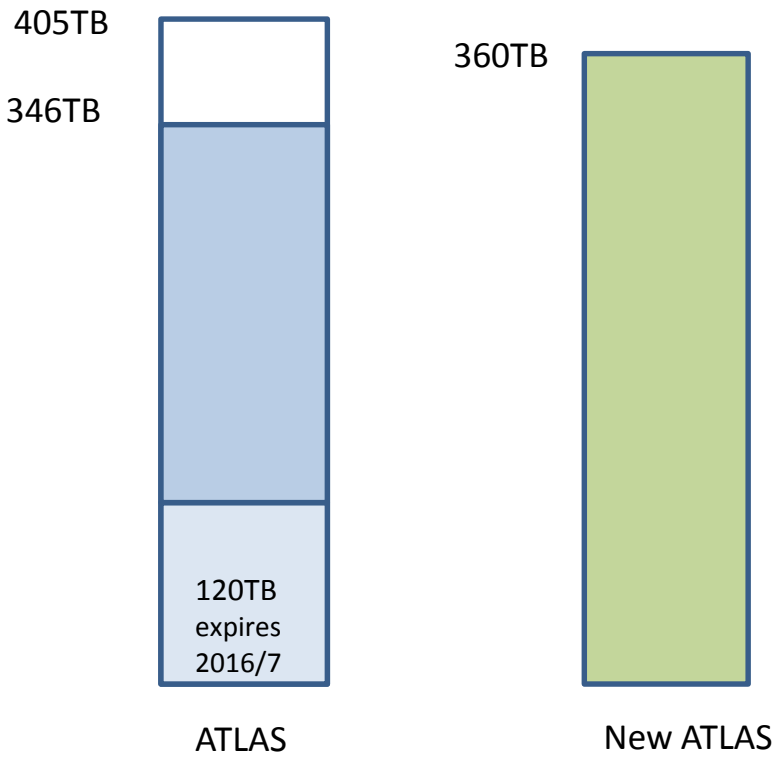
- Main issue was to upgrade Lustre 2.5 to 2.8/Higher
- No direct upgrade path available
- Not happy with the complexity and need to build a kernel module every time a new kernel becomes available.
- Open source future was not clear although this has changed recently
- So the option was to setup a parallel lustre kernel with new MDS and MGS or go for simpler gluster setup
- Budget constraints and loosing Sean and Ewan also played a role in the decision

Gluster

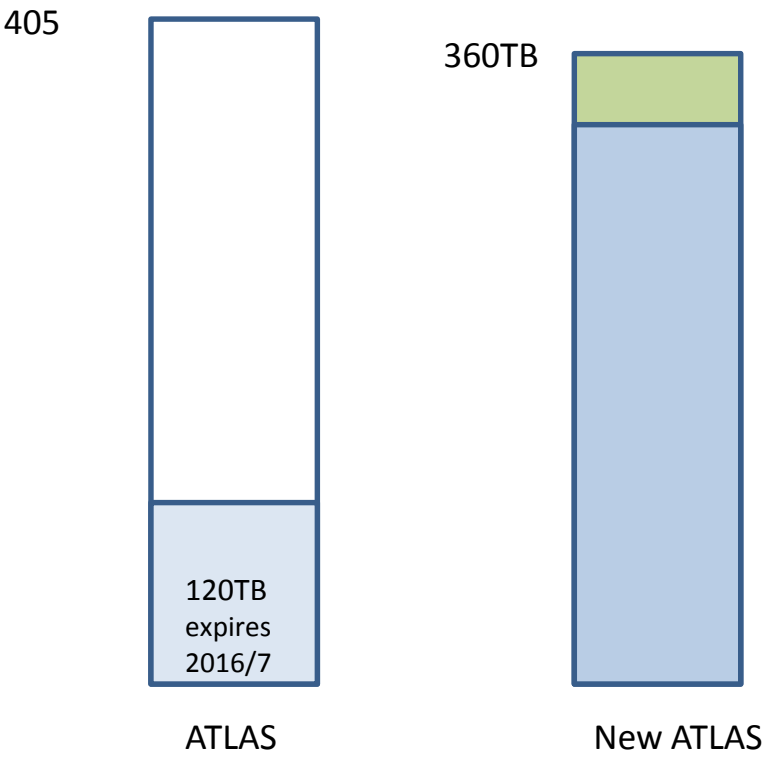
- Purchased six Dell R730xd with Twelve 8TB disks
- We got a very good deal from Dell
- Gluster has no meta data server and setup was very straight forward
  - Too easy, some times wondered whether I am doing something wrong ☺
- We have a 400TB Atlas file system and 200TB Lhcb file system, both with separate MDS.
- Needed to reuse some of the hardware from old file system
- Setup gluster file system with new hardware, rsync atlas data, rsync again and then a final rsync
  - That was the hardest part as some users kept writing tens of Tera bytes

Useful tool: Agedu

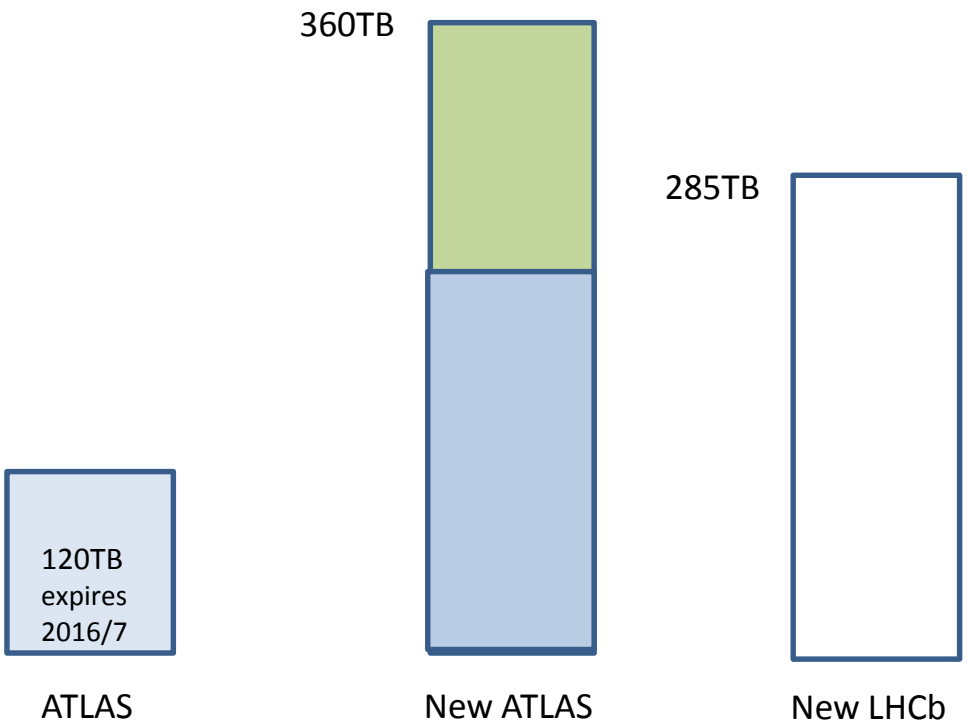
New servers setup running Gluster



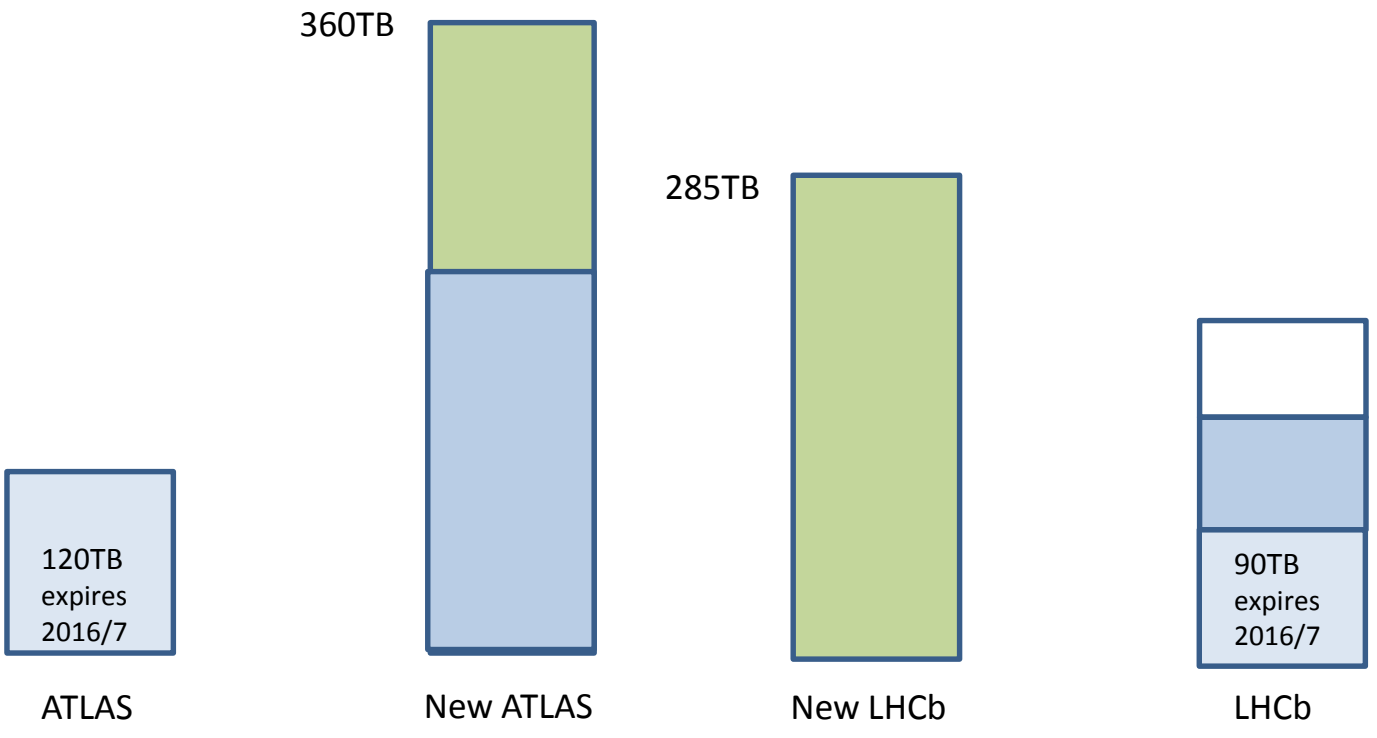
rsync / move Atlas data to new Gluster filesystem



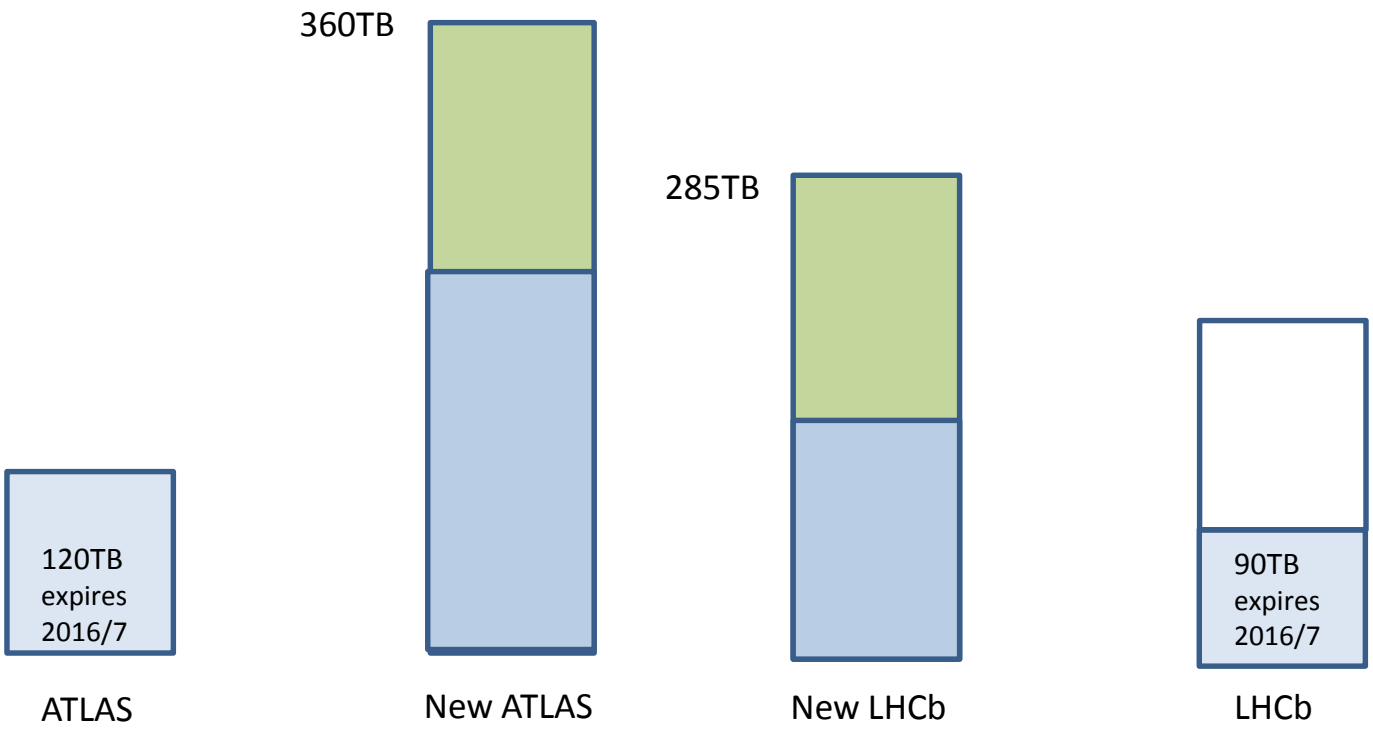
Re use ex Atlas (in warranty) servers to setup new Glustre files system for LHCb



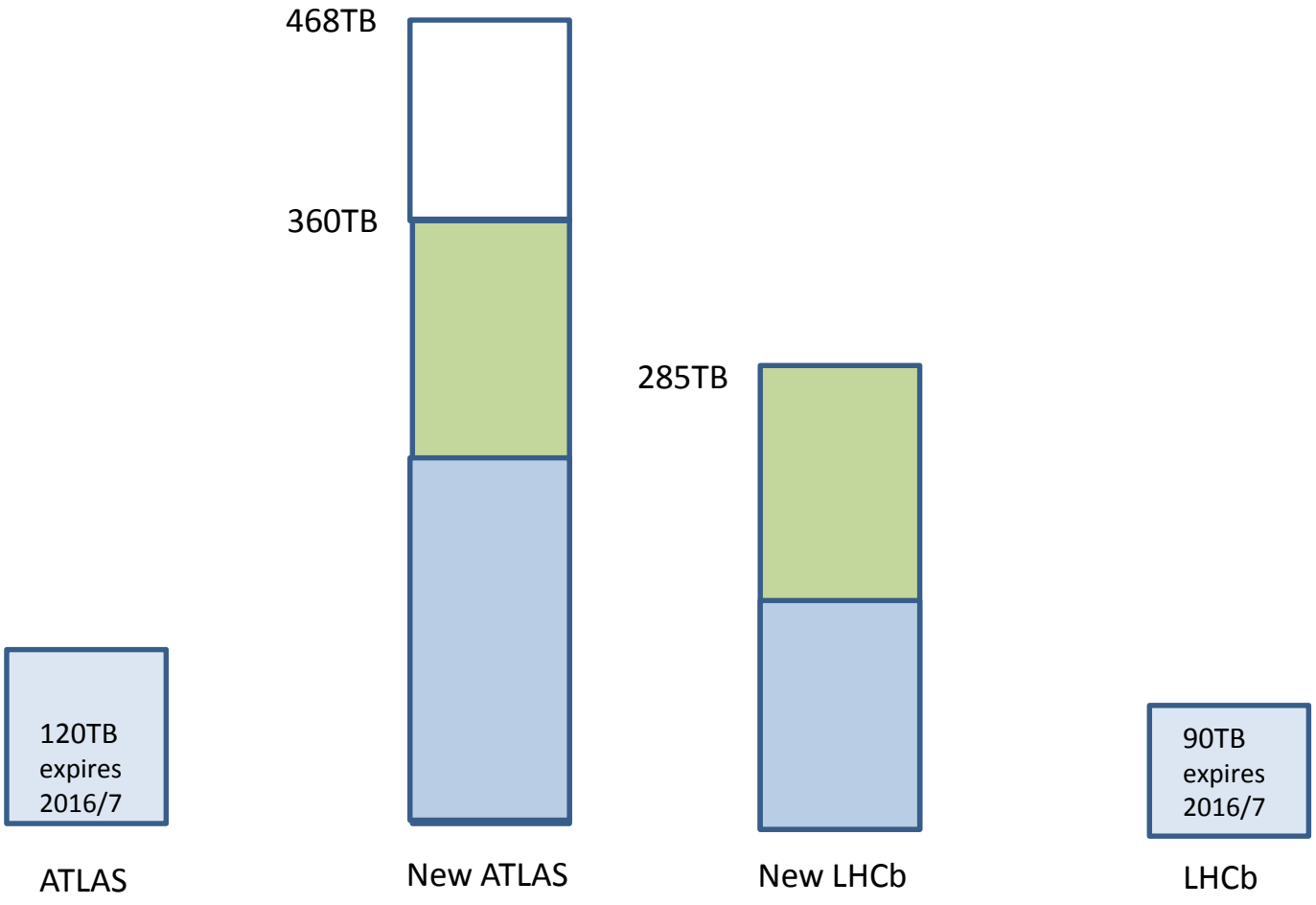
Migrate LHCb data to Gluster



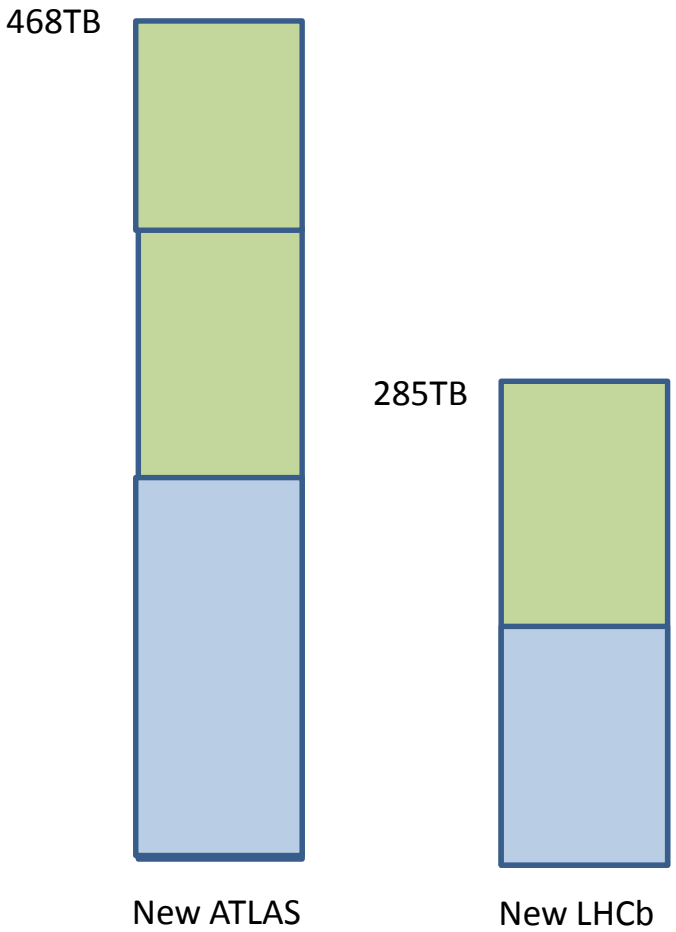
Migrate LHCb data to Gluster



Re use ex LHCb (in warranty) servers to expand Atlas Gluster filesystem



New Gluster setup for Atlas and LHCb for coming years.



Issues with Gluster

- Iozone results were more or less same for lustre and gluster
  - Tested gluster on a test setup from a single client
- Rsync is still going on for a user so I haven't run iozone after moving to production
- Copying and writing seems to be OK but 'du' and 'ls' are considerably slower
- Enable gluster client profiling as suggested on gluster mailing list
- Trying to understand output of gluster profiling output

Gluster profile output

%-latency	Avg-latency	Min-Latency	Max-Latency	No. of calls	Fop
0.00	0.00 us	0.00 us	0.00 us	2248	FORGET
0.00	0.00 us	0.00 us	0.00 us	6482	RELEASE
0.00	0.00 us	0.00 us	0.00 us	10138	RELEASEDIR
0.00	177.20 us	130.00 us	336.00 us	5	RMDIR
0.00	99.88 us	51.00 us	205.00 us	49	GETXATTR
0.00	89.01 us	38.00 us	270.00 us	80	SETXATTR
0.00	122.18 us	26.00 us	2613.00 us	67	FTRUNCATE
0.00	42.46 us	14.00 us	147.00 us	788	STATFS
0.01	49.94 us	15.00 us	1463.00 us	2098	INODELK
0.02	183.80 us	53.00 us	20566.00 us	915	LINK
0.03	133.23 us	58.00 us	10501.00 us	1421	RENAME
0.03	89.59 us	34.00 us	172.00 us	2337	SETATTR
0.03	3206.60 us	123.00 us	162298.00 us	70	FSYNC
0.04	70.83 us	24.00 us	13984.00 us	4485	OPEN
0.05	121.46 us	41.00 us	10571.00 us	2958	UNLINK
0.08	59.46 us	16.00 us	14160.00 us	10135	OPENDIR
0.11	10058.75 us	116.00 us	41438.00 us	80	MKDIR
0.17	56.97 us	8.00 us	48827.00 us	22615	STAT
0.32	13526.28 us	146.00 us	134667.00 us	177	SYMLINK
0.53	68.88 us	14.00 us	193130.00 us	57510	WRITE
0.77	35.01 us	6.00 us	29900.00 us	165414	FLUSH
0.86	99.25 us	11.00 us	240350.00 us	65222	FSTAT
0.89	46.96 us	17.00 us	22920.00 us	143087	READLINK
2.38	13498.51 us	113.00 us	550381.00 us	1323	MKNOD
3.50	11551.78 us	41.00 us	433438.00 us	2280	CREATE
11.87	3906.01 us	14.00 us	881696.00 us	22843	READDIRP
32.30	1263.61 us	26.00 us	1282841.00 us	192153	READ
46.00	107.85 us	22.00 us	1192496.00 us	3206311	LOOKUP

Duration: 549 seconds
Data Read: 25139914188 bytes
Data Written: 88562679 bytes