FIO Fabric Infrastructure and Operations

CERN IT Department

CASTOR Status

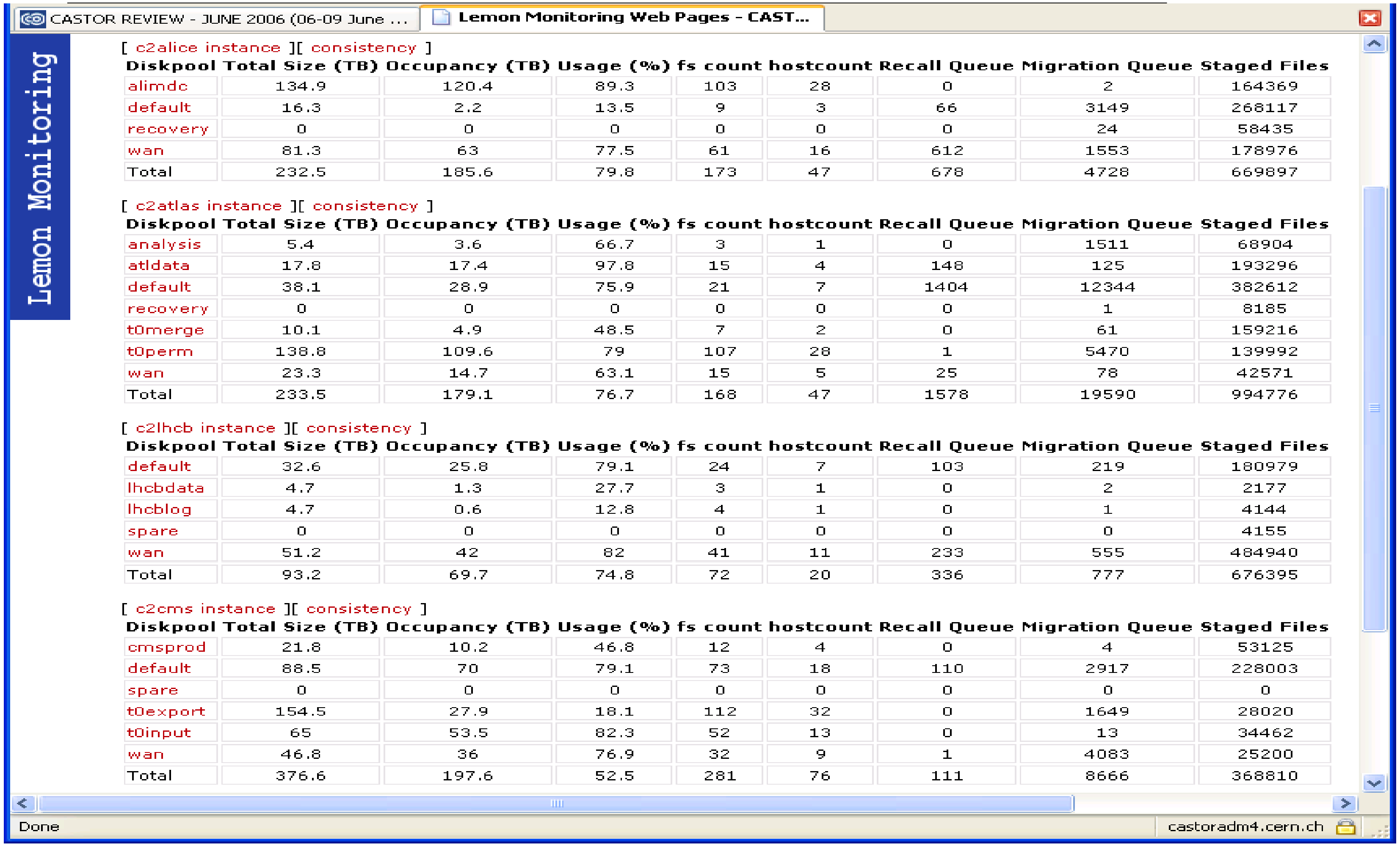
LCG GDB Meeting, 4th April 2007

Tony Cass

Leader, Fabric Infrastructure & Operations Group

IT Department

Desired CERN Configuration


CASTOR REVIEW - JUNE 2006 (06-09 June ...
Lemon Monitoring Web Pages - CAST...
Lemon Monitoring
Done
castoradm4.cern.ch

[c2alice instance][consistency]

Diskpool	Total Size (TB)	Occupancy (TB)	Usage (%)	fs count	hostcount	Recall Queue	Migration Queue	Staged Files
alimdc	134.9	120.4	89.3	103	28	0	2	164369
default	16.3	2.2	13.5	9	3	66	3149	268117
recovery	0	0	0	0	0	0	24	58435
wan	81.3	63	77.5	61	16	612	1553	178976
Total	232.5	185.6	79.8	173	47	678	4728	669897

[c2atlas instance][consistency]

Diskpool	Total Size (TB)	Occupancy (TB)	Usage (%)	fs count	hostcount	Recall Queue	Migration Queue	Staged Files
analysis	5.4	3.6	66.7	3	1	0	1511	68904
atldata	17.8	17.4	97.8	15	4	148	125	193296
default	38.1	28.9	75.9	21	7	1404	12344	382612
recovery	0	0	0	0	0	0	1	8185
t0merge	10.1	4.9	48.5	7	2	0	61	159216
t0perm	138.8	109.6	79	107	28	1	5470	139992
wan	23.3	14.7	63.1	15	5	25	78	42571
Total	233.5	179.1	76.7	168	47	1578	19590	994776

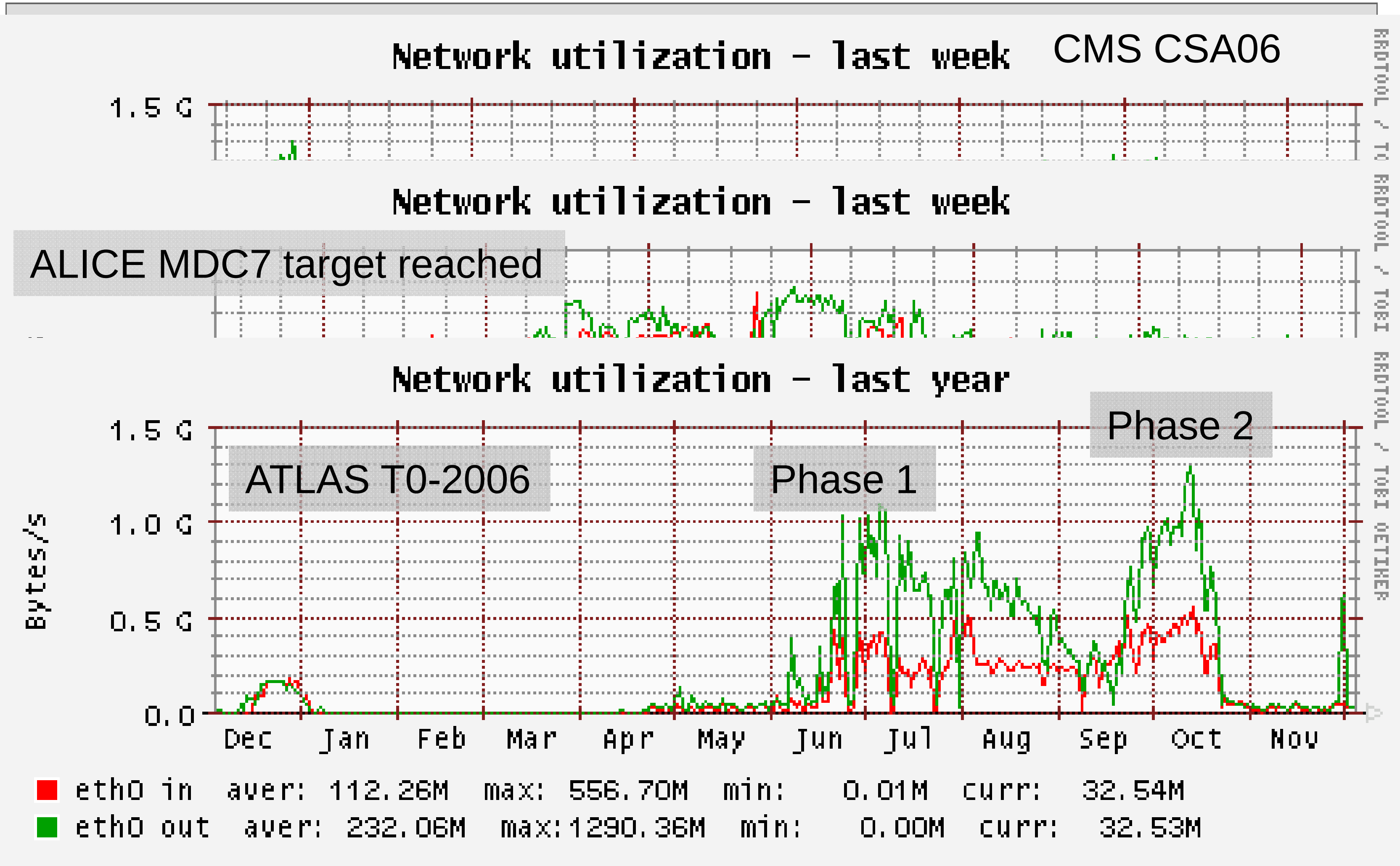
[c2lhcb instance][consistency]

Diskpool	Total Size (TB)	Occupancy (TB)	Usage (%)	fs count	hostcount	Recall Queue	Migration Queue	Staged Files
default	32.6	25.8	79.1	24	7	103	219	180979
lhcbdata	4.7	1.3	27.7	3	1	0	2	2177
lhcblog	4.7	0.6	12.8	4	1	0	1	4144
spare	0	0	0	0	0	0	0	4155
wan	51.2	42	82	41	11	233	555	484940
Total	93.2	69.7	74.8	72	20	336	777	676395

[c2cms instance][consistency]

Diskpool	Total Size (TB)	Occupancy (TB)	Usage (%)	fs count	hostcount	Recall Queue	Migration Queue	Staged Files
cmsprod	21.8	10.2	46.8	12	4	0	4	53125
default	88.5	70	79.1	73	18	110	2917	228003
spare	0	0	0	0	0	0	0	0
t0export	154.5	27.9	18.1	112	32	0	1649	28020
t0input	65	53.5	82.3	52	13	0	13	34462
wan	46.8	36	76.9	32	9	1	4083	25200
Total	376.6	197.6	52.5	281	76	111	8666	368810

Demonstrated Performance



CMS CSA06

ALICE MDC7 target reached

ATLAS T0-2006

Phase 1

Phase 2

Peak transfer of incoming data to tape at over 2GB/s

But…

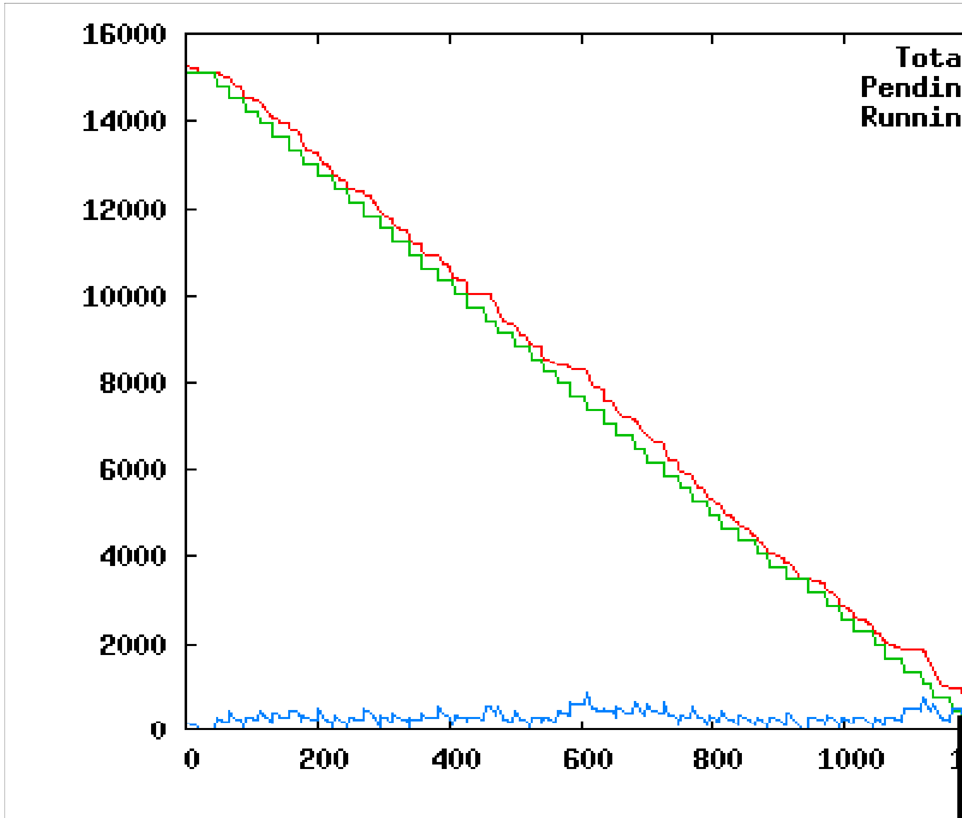
- Two significant software weaknesses
  - Scheduling of requests is greatly limited due to inefficiencies in the extraction of information from Oracle
    - maximum of 1,000 requests extracted and passed to LSF for scheduling; this leads to
      - interference between activities (service classes)
      - indeterminate scheduling time
  - Support for the “disk1” storage class is poor
    - CASTOR was designed to work with a tape archive and automatic garbage collection. It does not behave gracefully if disk pools fill up.
- Also
  - Inadequate stager hardware platform at CERN
    - Unreliable hardware leads to over-frequent Oracle problems
  - The software build & release process is complex (and there is no fully comprehensive test stage as yet)
    - limits turnround of versions at CERN
    - lack of support for a “stable release” means bug fixes bring in new features
      - a significant issue for external sites

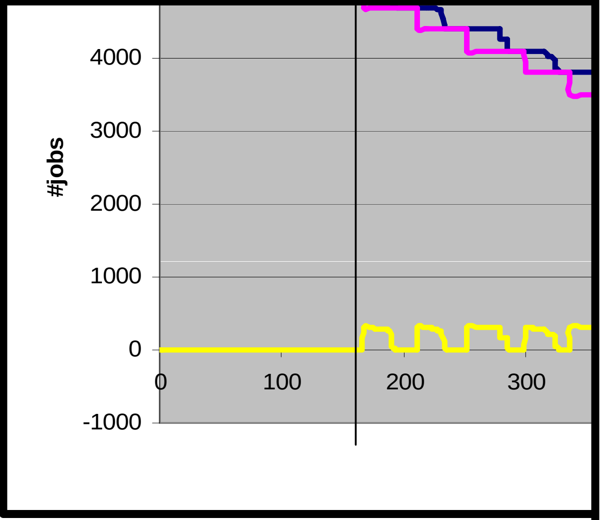
CERN IT Department

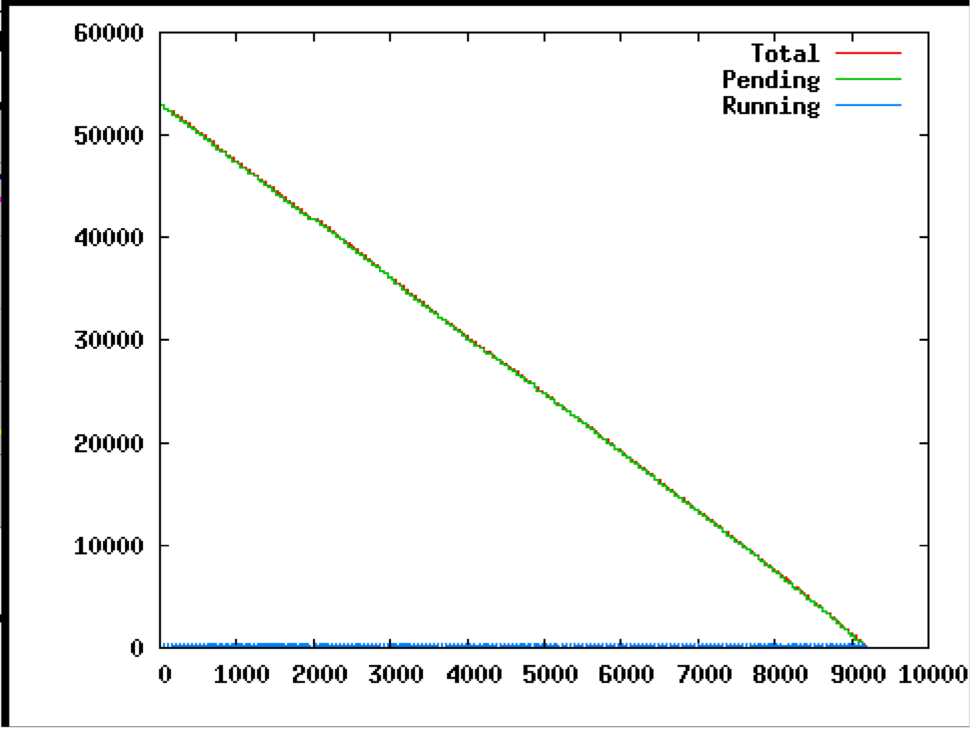
shared memory interface
esults in September
ady code took longer to

person,
monitoring), and
ory

0 slots
s







“Disk1” Support

- The changes needed to properly support disk1 storage classes are well understood:
  - Fail write requests (with clear error message) if a disk1 pool is full; at present these requests remain pending
    - … and pending requests cause problems today given the LSF plugin limitation
  - Disk pool access control
    - disallow requests that would result in decreased free space (either new files or replication/tape-recall of existing files) from non-production users
  - The effort required to fail write requests gracefully is relatively small (could be delivered in ~1 month), but more study is needed before providing an estimate for the access control work.
- … but this is not a priority at present; we need to work on the LSF interface and migrate to new database hardware as these are limiting our ability to meet the goals of the current ATLAS tests.
- … and anyway requires the SRM v2.2 interface to be in production
- Other disk1 issues
  - RAL noted an imbalance in the utlisation of filesystems within a disk1 pool. This is believed to be due to issues with the monitoring system which have since been fixed (but cannot be backported to the release used in production at RAL).

CERN stager h/w platform

CERN IT Department

- Standard disk servers are not an appropriate hardware choice for the critical stager database servers…
  - not a recent discovery: a migration to a SAN based Oracle RAC solution was planned in 2004!
  - Apart from reliability issues, the performance limitations of the hardware were highlighted recently: increasing system memory of the ATLAS stager from 2GB to 3GB lifted throughput to over 250 transactions/s compared to a previous limit of ~70.
    - As yet we have no clear idea of the throughput required for LHC production, though. (But the new LSF plugin leads to a reduced DB load for scheduling.)
- … but choice of the most appropriate platform took some time
  - during 2005 we learnt that the RAC solution did not provide the expected scalability given the CASTOR database access patterns
- Choice of (Oracle certified) NAS based system agreed in Autumn 2006
  - Still a RAC configuration, but also using Oracle DataGuard to ensure high availability
- New hardware being deployed now
  - CASTOR nameserver migrated on Monday
  - Stagers will follow; migration of the ATLAS stager is a high priority.

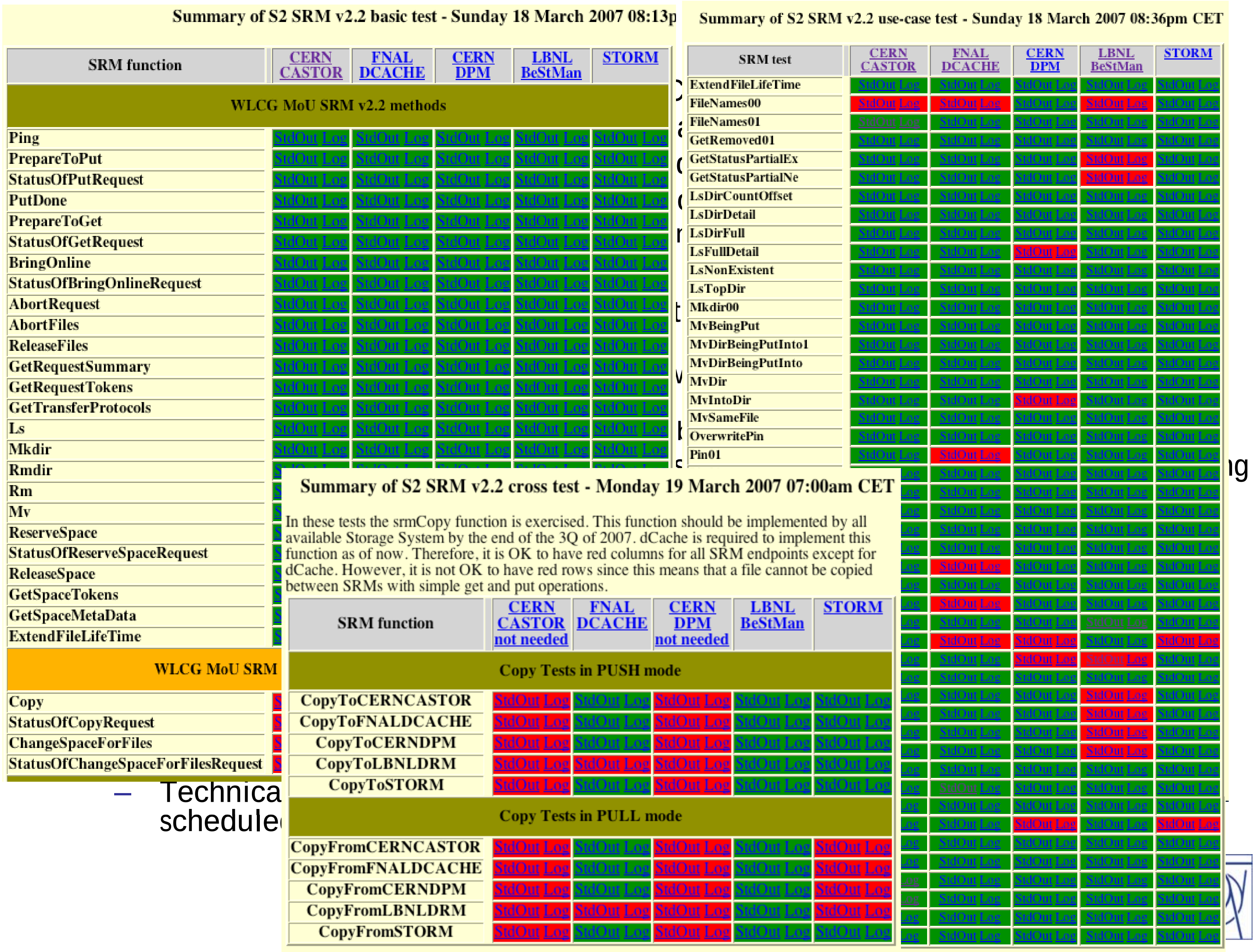
Software Release Process

- (Some) Problems with the current build process
  - Monolithic, so no easy way of only building selected packages
  - Imakefiles are difficult to maintain (accumulation of past settings)
  - there are many hardcoded values spread over the code
  - the Castor code base and build scripts need to be split up
- The need to support two (production quality) releases is recognised
  - an old, stable release (bug fixes only), and
  - a release integrating well tested new functionality as well as bug fixes.
- Addressing these issues will take time (fixing problems with production code always has priority), but planning this work has started
  - See slides 9-17 at http://indico.cern.ch/materialDisplay.py?contribId=7&amp;materialId=slides&amp;confId=7724
  - Testing has much improved over the past year although more automation is needed, as is a wider range of tests
    - We are currently collecting a list of tests performed by others and intend to integrate these into the pre-release testsuite
  - If progress elsewhere is satisfactory, CVS refactoring could start in late Q3 (i.e. after the summer)

Other Issues — I

- Strong Authentication
  - Required anyway, and a prerequisite for VOMS integration
  - Plan was
    - to produce a plan for the remaining work on strong authentication (for RFIO, CASTOR name server, and the CASTOR-2 stager) comparing impact of GSI and Kerberos 5 in Q2,
    - to reuse existing (DPM) ideas and developments, and
    - to be ready to deploy during Q3-Q4 (if compatible with run up to data taking)
  - This plan may have to be revised in the light of ongoing work with ATLAS
- VOMS integration
  - Plan was to build on the strong authentication work as from Q3/Q4.
  - Will follow DPM developments (virtual UIDs)
    - although there is an issue given the use of LSF for scheduling; this requires the UIDs to exist on the scheduling targets (i.e. disk servers). Workarounds are possible, though, and the issue has been discussed with Platform.
  - No production deployment before Q2 2008 (and probably later)

Other Issues — II



Summary of S2 SRM v2.2 basic test - Sunday 18 March 2007 08:13p

SRM function	CERN CASTOR	FNAL DCACHE	CERN DPM	LBNL BeStMan	STORM
WLCG MoU SRM v2.2 methods					
Ping	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
PrepareToPut	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
StatusOfPutRequest	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
PutDone	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
PrepareToGet	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
StatusOfGetRequest	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
BringOnline	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
StatusOfBringOnlineRequest	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
AbortRequest	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
AbortFiles	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
ReleaseFiles	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
GetRequestSummary	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
GetRequestTokens	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
GetTransferProtocols	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
Ls	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
Mkdir	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
Rmdir	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
Rm	🟩 S				
Mv	🟩 S				
ReserveSpace	🟩 S				
StatusOfReserveSpaceRequest	🟩 S				
ReleaseSpace	🟩 S				
GetSpaceTokens	🟩 S				
GetSpaceMetaData	🟩 S				
ExtendFileLifeTime	🟩 S				
WLCG MoU SRM					
Copy	🟥 S				
StatusOfCopyRequest	🟥 S				
ChangeSpaceForFiles	🟥 S				
StatusOfChangeSpaceForFilesRequest	🟥 S				

Summary of S2 SRM v2.2 use-case test - Sunday 18 March 2007 08:36pm CET

SRM test	CERN CASTOR	FNAL DCACHE	CERN DPM	LBNL BeStMan	STORM
ExtendFileLifeTime	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
FileNames00	🟥 StdOut Log	🟥 StdOut Log	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log
FileNames01	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
GetRemoved01	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
GetStatusPartialEx	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log
GetStatusPartialNe	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log
LsDirCountOffset	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
LsDirDetail	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
LsDirFull	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
LsFullDetail	🟩 StdOut Log	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
LsNonExistent	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
LsTopDir	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
Mkdir00	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
MvBeingPut	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
MvDirBeingPutInto1	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
MvDirBeingPutInto	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
MvDir	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
MvIntoDir	🟩 StdOut Log	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
MvSameFile	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
OverwritePin	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
Pin01	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟥 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟥 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟥 StdOut Log	🟥 StdOut Log	🟩 StdOut Log	🟥 StdOut Log
	🟩 og	🟩 StdOut Log	🟥 StdOut Log	🟥 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log	🟥 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
	🟩 og	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log	🟩 StdOut Log

Summary of S2 SRM v2.2 cross test - Monday 19 March 2007 07:00am CET

In these tests the srmCopy function is exercised. This function should be implemented by all available Storage System by the end of the 3Q of 2007. dCache is required to implement this function as of now. Therefore, it is OK to have red columns for all SRM endpoints except for dCache. However, it is not OK to have red rows since this means that a file cannot be copied between SRMs with simple get and put operations.

SRM function	CERN CASTOR not needed	FNAL DCACHE	CERN DPM not needed	LBNL BeStMan	STORM
Copy Tests in PUSH mode					
CopyToCERNCASTOR	🟥 StdOut Log	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
CopyToFNALDCACHE	🟥 StdOut Log	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
CopyToCERNDPM	🟥 StdOut Log	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
CopyToLBNLDRM	🟥 StdOut Log	🟥 StdOut Log	🟥 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
CopyToSTORM	🟥 StdOut Log	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log	🟩 StdOut Log
Copy Tests in PULL mode					
CopyFromCERNCASTOR	🟥 StdOut Log	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log	🟥 StdOut Log
CopyFromFNALDCACHE	🟥 StdOut Log	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log	🟥 StdOut Log
CopyFromCERNDPM	🟥 StdOut Log	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log	🟥 StdOut Log
CopyFromLBNLDRM	🟥 StdOut Log	🟥 StdOut Log	🟥 StdOut Log	🟩 StdOut Log	🟥 StdOut Log
CopyFromSTORM	🟥 StdOut Log	🟩 StdOut Log	🟥 StdOut Log	🟩 StdOut Log	🟩 StdOut Log

– Technica
schedule

CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Summary

- A single CASTOR2 instance today
  - **can** support the Tier0 requirements of each LHC experiment
  - **cannot** support mixed Tier0/analysis loads or guarantee non-interference if used to support multiple experiments.
- Demonstrating support for mixed loads is seen as a(n extremely) high priority
  - a task force has been setup to track this in the context of the ATLAS Tier0 and data export tests.
  - the key missing pieces, the LSF plugin and new database hardware, are now available
  - but time is short if these do not lead to a swift demonstration of adequate performance and reliability.
- The work needed to address other issues (notably disk1 support, but also strong authentication and VOMS integration) is understood, but will not start until adequate support for mixed loads has been demonstrated.