Version 2 - 28 February 2006
Ian Bird
Tony Cass
Les Robertson
Jamie Shiers

SC4 Workshop in Mumbai

Decisions, Conclusions, outstanding Issues

This note consists of three sections, prepared following the SC4 Workshop in Mumbai.

1. Conclusions on Data Management and Storage. This includes the conclusions of the storage management BOF which continued its work after the conference, concerning storage classes for SC4 and testing and deployment of SRM 2.1.
2. General points to be followed up.
3. Agreements, decisions and outstanding issues arising from the third day of the workshop – experiments' Tier-1 activity and plans for SC4.

Additional information is available from the agenda page for the workshop.

At the end of the workshop the middleware and software planning schedule for SC4 prepared by Flavia Donno was agreed.

A draft version of this note has been circulated to the people attending the workshop. It is now being sent to the GDB for feedback, following which it will go to the Management Board for approval and inclusion in the LCG plans for SC4.

An important further planning activity is the elaboration of a detailed schedule of the resources needed at each site by each experiment. A first version of this will be prepared for the Management Board during the next 4-6 weeks.

A specific point on which feedback from the GDB is requested is the request by ALICE for xrootd support at all sites. At present ALICE generally provides its own support for xrootd on the VO Box at the site. Note that in the case of CERN this will be considered after the discussion in March referred to in section 2 - Deployment of xrootd (page 5). Other GDB members should report on how they will respond to this request.

A new version of VOMS, with support for groups and roles, is scheduled for introduction before SC4 begins. This is a pre-requisite for other tools that will use groups and roles for data access control, job scheduling strategies, and other purposes. In SC4 only DPM is planned to provide storage access control support – dCache and CASTOR will provide this support in later releases. The new version of the WMS scheduled for release in SC4 supports a library which sites can use to implement group and role based scheduling – but there is no specific schedule at present for sites to implement such facilities. Feedback on this from the GDB members is requested.

The EGEE Workload Management System, including the Resource Broker and the new version of the CE supporting Condor C, is scheduled for availability for testing by experiments on March 15 and production at the beginning of SC4. The throughput requirements are as specified in the requirements of each experiment.

Document history

Version 1 (27feb06) - principle changes from the draft version (21feb06)

- An additional item - ***Distributed Database Deployment (3D)*** - has been added to section ***1. Data Management & Storage*** (page4)
- The target data rate at PIC for the April throughput tests has been reduced to match the available bandwidth (see table on page 10)
- A table giving the detailed ATLAS Tier-0/1 data rates for different datasets has been added (agreed at an ATLAS meeting in Mumbai) – see page 6.
- ALICE requires LFC at all sites (see page 7).
- Revised CMS planning has been prepared – see page 7, and a fuller presentation attached to the workshop agenda
- A section on gridFTP version 2 has been added – see page 4.

Version 2 (28feb06) – principal changes

- Statement on the EGEE Workload Management System (page 1)
- Addition of detailed LHCb planning (see page 9)

1. Data Management & Storage

1. FTS

- Agreed that multi-hop transfers are not required:
  - All transfers between any T0/T1/T2 can be point-point
  - Need to decide which server to use: propose T1 for T0-T2 etc
  - Some development is needed to manage and dynamically generate all possible channels
- Real requirement is transparency of where the service is
  - For example (fat) client that hides service details is OK - configures itself from information system.
- FTS does need to support both srm v1 and v2 in parallel

- FTS Development priorities
  a. Server selection by client
  b. Channel management
  c. SRM 1/2 support
  d. Pre/post transfer plug-ins – important for ATLAS

  Points (a) and (b) are scheduled for availability in SC4. Point (c) is required for the SRM version 2 testing and deployment programme (see page 3). Point (d) is expected to be available in SC4 for experimental use by ATLAS, installed on their VO BOXes.

2. LFC/DPM/GFAL/lcg-utils

a. Note: GFAL is the only SRM client library used by experiments

b. Priorities:
   a. SRM V2.1 compatibility tests between Castor-2/dCache/DPM
   b. Support for srm v2.1 in GFAL/lcg-utils
      - v1 & v2 MUST be supported in parallel
   c. Perl/python interfaces for GFAL/lcg-utils/DPM
   d. Draining of filesystems (Jean-Philippe finish interface; Graeme Stewart to provide tool)
   e. srmCopy and srmcp in dpm
      - No longer a high priority for DPM; Requested by large dCache sites to better optimize use of resources, but is only necessary on the dCache side as long as FTS understands the configuration.
   f. Global space reservation in DPM.

3. Castor-2

- The V1.1 Castor SRM endpoint is still needed, and will need to be provided in parallel with the V2.1 endpoint for some time.
  - Should have generically named endpoints – to avoid problems with specific names in SURLs in catalogs
- Castor-2 does support pools where files do not migrate to tape and garbage collection policy can be configured.
- Currently there is a distinction between local and WAN pools - but experiments want to make sure all data is visible everywhere.
  - This separation should go away over this year
  - For the moment files can be replicated between pools to make sure they are externally visible
  - Resources are currently segregated to ensure non-interference with T0-T1 transfers by local load

- This does not preclude sites using temporary buffer caches holding copies of files while they are being transferred to/from the network.

4. Storage Management

During the workshop it was agreed to establish a small group to complete the work on several issues concerned with storage management that had been discussed. This group met several times during the next few days and reached the following conclusions.

Storage Service Classes in SC4

For SC4 only SRM v1 will be available. It was agreed that only two storage classes will be supported:

- ***permanent***: single tape copy, system managed disk cache
  - this is the basic storage provided today by Castor and dCache + tape backend
  - the user knows that there is always a tape copy of the file and is charged for the tape storage used
  - the cache can usually be < < pool size
- ***durable***: no tape copy, user managed disk cache
  - storage system with no hierarchy (DPM. dCache or Castor without tape backend)
  - VO has full responsibility for managing the disk layer (jobs fail when disk pools full)
  - access performance is the same for all files
  - any tape backup decided by the site is not charged to the user

The implementation details must be agreed by all of the developers (CASTOR, dCache, DPM, GFAL, lcgutils) as a matter of urgency – responsibility of the Coordination Group mentioned below.

Storage Classes in SRM 2.1

Beyond SC4, SRM 2.1 will be used. Mass storage system providers, site representatives and experiments should agree a full classification of all the storage types required, assigning a name to each (see the Mumbai Workshop agenda for some discussion on this). Once this agreement is in place Storage Area Types can be defined for each of the classes, together with an efficient mechanism for users to select amongst the available classes.

Testing and Deployment of the SRM 2.1 Implementations

Maarten Litmaath is responsible for coordinating a group including the developers (of the SRM and Mass Storage Management implementations, GFAL, lcg-utils and FTS), and representatives of the experiments and major sites, with the following aims.

- agree on the full set of storage area types required and the way in which these will be supported (see above);
- agree on a common test set for SRM 2.1;
- plan and oversee the testing and deployment of the relevant implementations, and the migration from SRM 1.1 to SRM 2.1:
  - need to agree on concrete schedule for testing by experiments and other sites;
  - need to start reference testing of SRMs (Castor, DPM, dCache) with the available test suites (CERN, RAL, Berkeley)
    - can start end Feb/ after CHEP
  - Experiment testing needs several things to be done first: GFAL/lcg-utils and FTS (based on GFAL work)

- three testing activities can proceed in parallel:
  - testing with testing suite of different SRMs working together;
  - FTS – but development still needed;
  - GFAL/lcg-utils – once the libraries have been adapted to SRM V2.1
- only then can experiment testing start

This work must be done within the constraints of the October deadline for introducing an SRM 2.1 service in production.

gridFTP version 2

- gridFTP version 2 is backward compatible with the current version, and so it can be introduced progressively in different tools
- support is already in the development schedules of DPM and dCache
- CASTOR should also now include support for version 2 in their planning
- when versions of the storage management systems supporting gridFTP version 2 are available their deployment should be planned by the storage management coordination group mentioned above
- note that this will not be available at the beginning of SC4

User Files

A lower priority activity for this group is to study the requirements for and possible implementations of *User Files*.

During the work shop a requirement was stated by LHCb for User Files, created by end users during analysis. This is a general requirement, applicable to all experiments, including ATLAS. These files would have the following characteristics:

- generally small
- disk resident
- secure (rapidly backed up), and easily recoverable from the backup medium
- accessible by the owner from anywhere in the grid
- catalogued in the VO's grid catalogue
- shareable with other grid users
- limited by quota on a per-user basis

5. Distributed Database Deployment (3D)

- the initial set of sites (***ASGC, BNL, CERN, CNAF, GridKA, IN2P3, RAL***) should continue setting up for the partial production milestone at the end of March;
- the remaining sites (***NDGF, NIKHEF, PIC, TRIUMF***) must nominate representatives and start to attend the project meetings ***now***, to prepare for full deployment in October;
- experiments must now finalise their conditions data models and ramp up the scale of their distributed conditions/tag data deployment to enable proper sizing of the setup needed in October.

2 - Points from the Mumbai SC4 Workshop to be followed up (not agreed for deployment in SC4)

1. **User Files**
   See the note on Data Management and Storage issues. This will be followed up by the **Storage Management group coordinated by Maarten Litmaath**, when they have initiated the work on SRM 2.1. This is a long term issue.

2. **Deployment of *xrootd***
   a. this was requested to be made available at all sites for ALICE - not clear which sites have agreed to this for SC4 (GSI, Lyon?);
   b. several implementations of *xrootd* and the associated protocol have been discussed (dCache, Castor), which retain more or less of the functionality and performance of the SLAC implementation;
   c. the next step is to organise an in-depth discussion of the requirements and the different options when Andy Hanuchevsky visits CERN in March – should be attended by the storage system developers and interested sites and experiments – **to be organised by Peter Elmer**.

3. **rfio**
   a. there are two versions of rfio (Castor, DPM) that are incompatible but have the same names, requiring experiments to maintain separate executables;
   b. rfio timeouts are not caught by ROOT, causing the program to abort;
   c. there is a possibility that rfio can now be replaced by the rootd protocol with minimal work by developers and applications – **to be investigated by IT/GD group.**

3. Experiments & SC4

Decisions

- For support issues, it has been **agreed that we will use** helpdesk@ggus.org **(or www.ggus.org) from now on.** The existing support lists will be closed down latest May 2006.

Outstanding Issues

These refer to the experiment presentations during the third day of the workshop – available via the agenda page. A brief summary by experiment is given below (*Experiment Production Plans)*.

- The details of the T1<->T1 transfers still need to be finalised. A "dTeam" phase should be foreseen, to ensure that the basic infrastructure is setup. Similarly for T1->T2. A possible scenario follows:
  - All Tier1s need to setup an FTS service and configure channels to enable transfers to/from all other Tier1s.
  - *dTeam* transfers at 5MB/s (10MB/s?) need to be demonstrated between each T1 and all other T1s
  - These tests would take place during May, after the April throughput tests and before the SC4 service begins in June.

- **ATLAS**:
  - The details of the Tier0 exercise in March are under discussion – at present there is no agreement for any transfers to external sites - this is foreseen for June.
  - The data rates presented have been refined at an ATLAS Tier-1 meeting held in Mumbai, summarised in the following table. The total nominal aggregate data rate from the Tier-0 is rounded up to 780 MB/sec. RAW data goes to tape at Tier-1s and ESD+AOD data goes only to disk.

Tier-1	*Location*	*Fract.*	*RAW*	*ESD*	*AODm1*	*Total rate*
BNL	Brookhaven	24.0	76.8	100.0	20.0	196.8
SARA	Amsterdam	13.0	41.6	26.0	20.0	87.6
CCIN2P3	Lyon	13.5	43.2	27.0	20.0	90.2
FZK	Karlsruhe	10.5	33.6	21.0	20.0	74.6
RAL	Didcot	7.5	24.0	15.0	20.0	59.0
ASGC	Taipei	7.7	24.6	15.4	20.0	60.0
CNAF	Bologna	7.5	24.0	15.0	20.0	59.0
NDGF	distributed	5.5	17.6	11.0	20.0	48.6
PIC	Barcelona	5.5	17.6	11.0	20.0	48.6
TRIUMF	Vancouver	5.3	17.0	10.6	20.0	47.6
Total		100.0	320.0	252.0	200.0	772.0

  - Job submission rates per target grid need to be defined.
- **ALICE**:
  - An LFC service is required at all sites serving ALICE as a local file catalog (task force communication).
  - The "proof@caf" issue has not been discussed at the Tier0.
  - xrootd is requested at all sites. This is being negotiated on a site by site basis.
- **CMS**:
  - A new schedule has been produced taking into account the official dates for SC4 and gLite release schedule.

  - CMS stress the need to test the entire data management chain with files >2GB, to ensure that these are fully supported by all relevant components and services.

- **General**:
  - The detailed schedule and resource requirements need to be discussed and agreed once the above issues are resolved.

Experiment Production Plans

ALICE

The first point of this year's PDC'06/SC4 plan is the scheduled rerun of SC3 T0 disk – T1 disk transfers (max 150MB/s). These will be scheduled transfers through the FTD-FTS system and the target T1s are CNAF, IN2P3 Lyon, GridKa and RAL. Data generated during PDC'05 and available at CERN will be used. The amounts of data to be transferred to each centre will depend on the available storage capacity; however a possible scenario is to remove the transferred data on the target SE after it has been successfully transferred. The target duration of the exercise is 150 MB/s aggregate throughput during 7 days.
In parallel to the file transfers, we will continue to run jobs to test the stability of the complete system.

The requirement for LFC as a local catalog at all sites was clarified.

ATLAS

ATLAS' SC4 requests are summarised as follows:

- March-April (pre-SC4): 3-4 weeks in for internal Tier-0 tests (Phase 0)
- April-May (pre-SC4): tests of distributed operations on a "small" testbed (the pre-production system)
- Last 3 weeks of June: Tier-0 test (Phase 1) with data distribution to Tier-1s (720MB/s + full ESD to BNL)
- 3 weeks in July: distributed processing tests (Part 1)
- 2 weeks in July-August: distributed analysis tests (Part 1)
- 3-4 weeks in September-October: Tier-0 test (Phase 2) with data to Tier-2s
- 3 weeks in October: distributed processing tests (Part 2)
- 3-4 weeks in November: distributed analysis tests (Part 2)

LFC and VO Boxes are required at Tier-0 and Tier-1 sites, not at Tier-2s.

CMS

CMS emphasised the requirement to test the entire chain using files large than 2GB (to make sure that there are no hidden limitations still remaining...)

A revised plan was submitted after the workshop – summarised here - see revised plan for details.

Overall planning

- CMS is rolling out a new framework, deploying new data management system and new simulation production infrastructure
- CMS will have 10TB of new event data at the beginning of April
- CMS will be able to produce 10M with the new production infrastructure in May
- Beginning of June CMS would like a two week functionality rerun of goals of SC3

  - Demonstration and preparation for the 2006 Data Challenge CSA06
- July and August CMS will produce 25M events per month (roughly 1TB per day of data for CSA06
- September - October is CSA06

Data Transfer Goals

Tier-0 – Tier-1

- In 2008 CMS expects ~300MB/s being transferred from CERN to Tier-1 centers on average
  - Assume peaks to recover from downtime and problems with a factor of two (600MB/s)
  - Network Provisioning would suggest roughly twice that for raw capacity
- 2006 goal is to demonstrate aggregate transfer rate of 300MB/s sustained on experiment data by the end of year to tape at Tier-1
  - 150MB/s in the spring (April-June) -
    Tape rate goals for individual sites –
    - ASGC: 10MB/s; CNAF: 25MB/s; FNAL: 50MB/s
    - GridKa: 15MB/s; IN2P3: 15MB/s; PIC:30MB/s
    - RAL: 10MB/s
  - 300MB/s after October Service Challenge

Tier-1 - Tier-2

- Tier-1 → Tier-2 likely to be very bursty and driven by analysis demands
  - Network to Tier-2 are expected to be between 1Gb/s to 10Gb/s assuming 50% provisioning and a 25% scale this spring
  - Desire is to reach ~10MB/s for worst connected Tier-2s to 100MB/s to best connected Tier-2s in the spring of 2006 (1TB at all Tier-2 centers; up to 10TB per day for well connected centers)
- Tier-2 → Tier-1 transfers are almost entirely fairly continuous simulation transfers
  - The aggregate input rate into Tier-1 centers is comparable to the rate from the Tier-0
  - Goal should be to demonstrate 10MB/s from Tier-2s to Tier-1 centers → 1TB per day

Data Transfer Schedule

- March 15 FTS driven transfers from PhEDEx
- Starting in April CMS would like to drive continuous low level transfers between sites that support CMS - on average 20MB/s (2TB per day)
- In addition to low level transfers, CMS would like to demonstrate the bursty nature of Tier-1 to Tier-2 transfers → Demonstrate Tier-2 centers at 50% of their available networking for hour long bursts

Components needed

- Phedex integration with FTS -Expected middle of March
- CMS Data Management Prototype
  - Version 0 is released. Version 1 expected in the middle of Match
  - Currently has ability to define and track datasets and locations
    New version needed for New Event Data Model Data
- New Event Data Model Data
  - Expect first 10TB sample roughly on April 1
  - The ability to transfer data on files is a component, but the experiment needs to transfer defined groups of files, validate integrity and make them accessible.

Job Submission

- Roughly 200k job submissions per day in 2008
- Aim for 50k jobs per day during 2006.
- CMS will begin transitioning to gLite 3.0 in the pre-production system
    - For EGEE sites will use the gLite RB for submission (2/3rds of resources, 33k jobs per day)
    - For OSG CMS will test the gLite RB for interoperability, maintaining the direct Condor-G submission capability for scaling and speed if necessary
- By the beginning of March expect an analysis grid job submitter capable of submitting to new event data model jobs - version of the CMS Remote Analysis Builder (CRAB)
- Attempt job submission targets in the June exercise and again during the CSA06 preparation period in September
    - first two weeks of June → 25k jobs per day
    - during September → 50k jobs per day
- July and August - simulation at a high rate, but total number of jobs will not stress the simulation infrastructure

Local Mass Storage

- The CMS data model involves access to data that it stored in large disk pools and in disk caches in front of tape
- Goal in 2008 is 800MB/s at a Tier-1 and 200MB/s at a Tier-2
- 2006 Goal is 200MB/s at Tier-1 or 1MB/s per batch slot for analysis

Local Catalog

- The current baseline solution for SC4 for CMS is the trivial file catalogue.
    - In this mode the incoming application has logical file names and the existing mass storage name space is used to resolve the physical file names
- CMS has been evaluating LFC as an alternative.
- Need a catalog at CERN to serve as the basis for the Dataset Location Service

LHCb

SC4/DC06 preparation (April-May)

- Data simulation production at all sites
    - 100 Mevts (b-inclusive) + 100 Mevts (minimum bias)
    - ~4 MSI2k.months need for production
    - DIGI files (containing MCTruth) transferred to Tier-0
        - 125 TBytes only stored at CERN (Castor)
        - Using public transport gridFTP to CERN (no local storage)
        - Continuous data flow during the production period
        - Registration in the central LFC catalog
    - Use resources at all available sites
    - >~3000 jobs/day – duration ~36 hours

SC4/DC06 (June-July):

- Exercise the LHCb event model
    - Distribute simulated & digitised data DIGI
        - 125 TBytes in 60 days, i.e. 2 TB/day to all Tier-1's
        - Transfer using FTS, registration in LFC & processing DB
    - Data on *permanent* storage, pinned for ~1-2 days
    - Automatic reconstruction submission @ Tier-0 + 6 Tier-1's
        - ~1000 jobs /day – duration ~2 hours
    - Reconstruction produces rDST, stored on *permanent* storage locally
        - 60 TBytes at all Tier-1's (2 GB files)

        - Registration in LFC & processing DB
    - Automatic stripping submission @ sites where rDST are
        - ~100 jobs / day – duration ~3 hours
    - CPU requirements for reconstruction & stripping ~0.3 MSI2k.months
    - Stripping produces DST
        - Stored on *permanent and durable* storage locally
        - Shipped to *durable* storage at all Tier-1's + CAF
        - 2.2 TBytes on disk at each Tier-1 + CAF
        - Transfer using FTS, registration in LFC
    - Central LFC contents mirrored in the LFC instances at (some) Tier1 centres

Additional DC06 production (June onwards)

- Data simulation & reconstruction production
    - 200 Mevts (signal)
    - Use all sites except during June & July: exclude Tier 1 & CERN
        - ~2000 jobs/day – duration ~ 36 hours
        - Resources needed: ~7.5 MSI2K.months
    - rDST produced and transferred to associated Tier-1
        - ~100 TBytes at all Tier-1's on *permanent* storage
    - Automatic stripping submission @ Tier-1's
        - ~200 job/day – duration ~3 hours
    - Stripping produces DST
        - Stored on *permanent and durable* storage locally
        - Shipped to *durable* storage at selected Tier-1's + CAF
        - Total : 30 TB
        - ~90 TBytes on disk across Tier-1 + CAF (this will allow 3 copies of the data across the Grid. If disk available it is desirable to make all the 30 TB dataset available at each Tier-1- 210 TB)
        - Transfer using FTS, registration in LFC

Preparing for SC4 Disk-Disk and Disk-Tape Throughput Tests in April

These are the well-known rates that should be achieved in MB/s.

It is important to emphasise that these are daily averages sustained over extended periods - not one-time peaks.

Site	Disk-Disk	Disk-Tape
ASGC	100	75
TRIUMF	50	50
BNL	200	75 note 1
FNAL	200	75
NDGF	50	50
PIC	60	60 note 2
RAL	150	75
SARA	150	75
IN2P3	200	75
FZK	200	75
CNAF	200	75

Notes

1. The rate for BNL assumes that a full copy of the ESD is exported there.
2. The "nominal" rate for disk-disk at PIC is 100 MB/sec, but this will not be achievable before November 2006 due to limited WAN bandwidth.
3. As usual, we will first run the disk-disk throughput test and then disk-tape.
4. In July, the disk-tape rates go up to full-nominal, i.e. the disk-disk rates in the table above