Grid Deployment
Data challenge follow-up & lessons learned

Ian Bird
LCG Deployment Area Manager

LHCC Comprehensive Review

22nd November 2004

Outline

- Issues seen in the data challenges
  - Responses and requirements
- Lessons learned
  - Operations, certification, support
- Directions for future
  - Service challenges, etc

Problem summaries

- Outstanding middleware problems have been collected during the DCs
  - https://edms.cern.ch/file/495809//LCG2-Limitations_and_Requirements.pdf
- And in the GAG document
  - http://project-lcg-gag.web.cern.ch/project-lcg-gag/LCG_GAG_Docs/DCFeedBack.pdf
- 1st **systematic** confrontation of required functionalities with capabilities of the existing middleware
  - Some can be patched, worked around,
  - Most has to be direct input as essential requirements to future developments

Major issues seen in data challenges

- ## Job success rate
  - Due in large part to site mis-configurations, lack of fabric management (and expertise)
    - Many grid sites are new – never run a cluster before
    - Started to address this in the operations workshop
      - Create a fabric management handbook – recipes and best practices – leverage HEPiX, EGEE training, etc.
  - Now much improved – monitoring was significantly improved in the last 6 months, problems followed up direct to sites
    - Should improve more – process now better understood and agreed – being implemented in the EGEE CICs + Taipei GOC
      - Helps with more effort involved – rotating responsibility

Issues – 2

- ## Workload Management System

  - Simple Globus model does not match complexity of modern batch systems with fairshare schedulers

  - RB does not communicate full set of resource requirements to the batch systems

  - Current IS schema needs expanding to improve the description of batch capabilities, heterogeneity, etc

  - Many workarounds have been put in place – but these are hard to scale

  - WLM from gLite anticipated very soon → must make sure that these issues are addressed and resolved
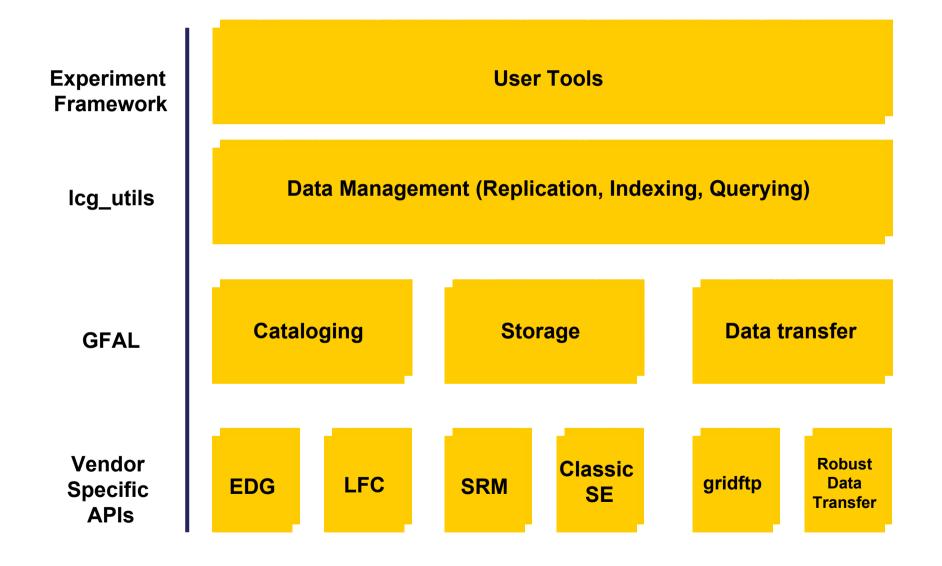
Issues – 3

- ## Information system
  - This was continually improved
  - Based on BDII – drop in replacement for MDS components
  - Can now scale to 100's of sites, with good robustness
  - Many issues that manifest as IS problems in fact due to underlying fabric problems
    - E.g. PBS hangs do not update IS

  - BDII allowed each experiment to have a "view" on the full system that could be tailored to its needs
    - Specify sites that are experiment-verified and LCG verified
    - Provided the experiment with control over which sites it used

Issues – 4

- ## Data management
  - ### Replica management tools
    - Edg tools replaced by lcg-utils – additional requested functionality and better performance
    - This was available early in the DCs but not used by all experiments
  - ### Replica catalogue performance and functionality
    - Re-engineering EDG RLS to address these issues
      - Same interfaces as now – lcg-utils
      - Propose also to provide gLite catalogue interface to provide simple implementation quickly with new interface
  - ### GFAL
    - originally a low-level IO interface to Grid Storage
    - Now provides:
      - File Catalog abstraction
      - Storage Element Abstraction (EDG SE, EDG 'Classic SE', SRM v1)
      - Connection to Information system

Layered Data Management APIs

Experiment Framework	User Tools
lcg_utils	Data Management (Replication, Indexing, Querying)
GFAL	Cataloging — Storage — Data transfer
Vendor Specific APIs	EDG — LFC — SRM — Classic SE — gridftp — Robust Data Transfer

Issues – 5

- ## Data transfer (gridftp) unreliability
  - Building reliable file transfer service – goal to demonstrate reliability and performance in series of service challenges
  - Currently in tuning phase with sites
    - Fermi achieving 3 Gbit/s sustained transfer rates
  - Service Challenges start in December
    - NIKHEF/SARA first, then Karlsruhe and Fermi

- ## Lack of managed storage elements
  - dCache seen as solution – but still not ready for simple widespread deployment, but significant effort has been invested to make this a solution for large sites
  - Lightweight alternative – provide basic space management and SRM interface, with data access via gridftp and rfio

- ## The disk-pool manager and file transfer service
  - Are not LCG-2 specific – needed for both LCG-2 and gLite based services

Lessons – Operations – 1

- (Lack of) hierarchy in:
  - Middleware distribution
  - Operation and user support
  - Communication (sites ↔ GDB; general),
    - mailing lists took over all discussions …
  - ⇨ EGEE-based structure with hierarchy of OMC, CIC, ROC
    - Still need fast reaction for (e.g.) data challenges – task forces?
- Installation and configuration tools
  - For middleware has to be as simple as possible & flexible
  - No single solution → small/large & established/novice sites
  - ⇨ Move towards generic procedures that can be used with any tools
- Certification and release preparation
  - Very effective, but slow and opaque to outside groups
  - Some problems only in real production and load
  - ⇨ Streamlined certification and release process
    - Separate core and "add-on" middleware
    - Advertise forthcoming changes – but expediency is not always democratic …

Lessons – Operations – 2

- **Monitoring and troubleshooting**
  - Many tools available for users, admins, operators
    - Big overlap/duplication of efforts, independent solutions
  - ⇨ Establishing a single monitoring infrastructure
    - Based on R-GMA, aggregate many sources of information; usable by many tools to produce customised views and displays
    - Already contains many of the existing tools
- **Fabric management**
  - Many (new) sites have no experience or expertise
  - ⇨ Produce a fabric management recipe/best practice document (part of EGEE "cookbook"), HEPiX, …
- **Testbed for early application access and testing**
  - ⇨ EIS testbed was provided
  - ⇨ Pre-production service gets sites and operators involved too
- **Middleware to be installed on WN → as close to 0 as possible**
  - Very important for co-existence with other grids etc.

Lessons – Certification – 1

- Certification process was crucial
  - Almost all installations worked 1st time
  - Need to tie in EIS testbed (and now pre-production system) as integral part of certification and release process
  - … and integrate deployment preparation
- Release cycles → monthly
  - But too short a cycle
- Public discussion of release contents
  - Did not happen fully – expediency during DC's – but need more input from site managers to improve the process
- Inclusion of external efforts
  - Current process does not easily allow inclusion of work done by external groups – needs to be modified
- Testing of experiment software
  - Was found to be extremely useful – investigate how to do this more effectively – pre-production service?
- Developers MUST be exposed to the middleware problems
  - Fixing must be their TOP priority – not the next piece of functionality

Lessons – support – 1

- Having the EIS group was essential
  - Support people who learned experiment environments and issues and followed the problems
  - Could certainly be expanded – needs backup people
- Experiments using LCG-2 for the 1st time only when the DC started!
  - A lot of time was lost finding problems that could have been seen earlier – vis LHCb experience
  - Only at this point we learned how experiments wanted to use the middleware
    - Meant changes and adaptations during the DC
    - Experiment interfaces were still being written after the DCs started …
  - Real functional/performance issues only surfaced then
  - LCG-2 was only ready in time for the DC's but LCG-1 had only been lightly used and not in the DC mode
- Suspect level of effort needed for interfacing experiments' software to middleware was underestimated

Lessons – support – 2

- EIS testbed (or pre-production service) needs to be an integral part of the release process with experiments' grid experts fully involved
  - Early feedback on the middleware

- Services and software (middleware and experiments) expect other services to be reliable
  - Not possible in a distributed system – must deal with failure and retries

- Must have a "one-stop-shop" for all problems
  - One place where all problems are reported and dispatched
  - Has not been clear who should report what and where
  - ⇨ outcome of Ops workshop: user support group

Future directions

- Deployment strategy
- Working groups
- Service challenges
- Planning and milestones

Deployment and services

- Current LCG-2 based service continues as production service for batch work
  - Experiments moving to continuous MC production mode
  - Together with work in-hand provides a well-understood baseline service
- Deploy in parallel a pre-production service
  - Deploy LCG-2 components, and
  - gLite components as they are delivered
  - Understand how to migrate from LCG-2 to gLite –
    - Which components can be replaced
    - Which can run in parallel
    - Do new components satisfy requirements – functional and management/deployment
  - Move proven components into the production system

LCG-2 vs gLite vs Alien

- ## LCG-2 is the current production service
  - Alien can use LCG-2 as a resource

- ## The "gLite prototype" is based on Alien
  - Has been deployed at a few sites – "by hand" and used by ARDA to start to address analysis on the grid

- ## gLite – development based on the ARDA prototype
  - Building the 2nd generation middleware, using ideas from Alien and EDG, as well as VDT etc – based on web services
  - Will be gradually deployed in pre-production as it is delivered

Interoperability issues

- LCG-2 and Grid3 – based on same tools and schema
  - See a way to reasonable interoperation
  - Not all issues are resolved (e.g. file catalogues)
- LCG/EGEE – gLite is the way forward – this is where the development effort is
  - What will OSG deploy – will it be compatible with gLite?
  - What will NorduGrid/ARC do – gLite?
  - How can we avoid divergence?
- LCG-2 is reviewed/monitored by all parties
  - OSG, NorduGrid deployment decisions are not reviewed by the LCG project
  - Risk is that interoperability is not possible in that situation

Operations management

- Following operations workshop
  - Have working groups on
    - Operations support
    - User support
    - Fabric management
    - Operational security group
  - With agreed strategies, currently being implemented
  - Planning longer term milestones, designed to improve service reliability and management
  - Check points with service challenges and data challenges
  - Many issues are common with Grid3/OSG – want to bring these together as far as possible

Service challenges

- Proposed to be used in addition to ongoing data challenges and production use:
    - Goal is to ensure baseline services can be demonstrated
    - Demonstrate the resolution of problems mentioned earlier
    - Demonstrate that operational and emergency procedures are in place

- 4 areas proposed:
    - Reliable data transfer
        - Demonstrate fundamental service for Tier 0 → Tier 1 by end 2004
    - Job flooding/exerciser
        - Understand the limitations and baseline performances of the system
            – May be achieved by the ongoing real productions
    - Incident response
        - Ensure the procedures are in place and work – before real life tests them
    - Interoperability
        - How can we bring together the different grid infrastructures?
- Now proposed expansion to Tier 1 readiness challenges
    - Set of milestones to build up to full data rates over next 2 years
    - Exactly what these should be is under discussion now

Planning next steps

- Is under way now
  - Goal to agree set of basic milestones by end 2004, covering
    - Service challenges
    - Operations, support, security, fabric management
  - These should be monitored by the GDA steering group
    - Tier 1 and large Tier 2 managers
  - … and the GDB

- Coordination and management
  - Grid Deployment Board – policy and high level agreements
  - GDA steering group – following deployment milestones and directions
  - Weekly operations meeting – coordinating operational issues
  - Operational security group
  - Working groups set up in the workshop

Summary

- Data challenges have run during 2004
  - Many issues have been raised, much has been learned
  - A lot has been addressed already
  - Experience is documented as input to new developments
- Strategy for improving operations/support is agreed
  - And implementation started
- Service and data challenges, ongoing productions
  - Will be used to show improvements and progress
- Preparing for gLite deployment
- Interoperability/co-existence of grids is a reality
  - Issues to be addressed at all levels – management/technical