LCG-2 Deployment Issues

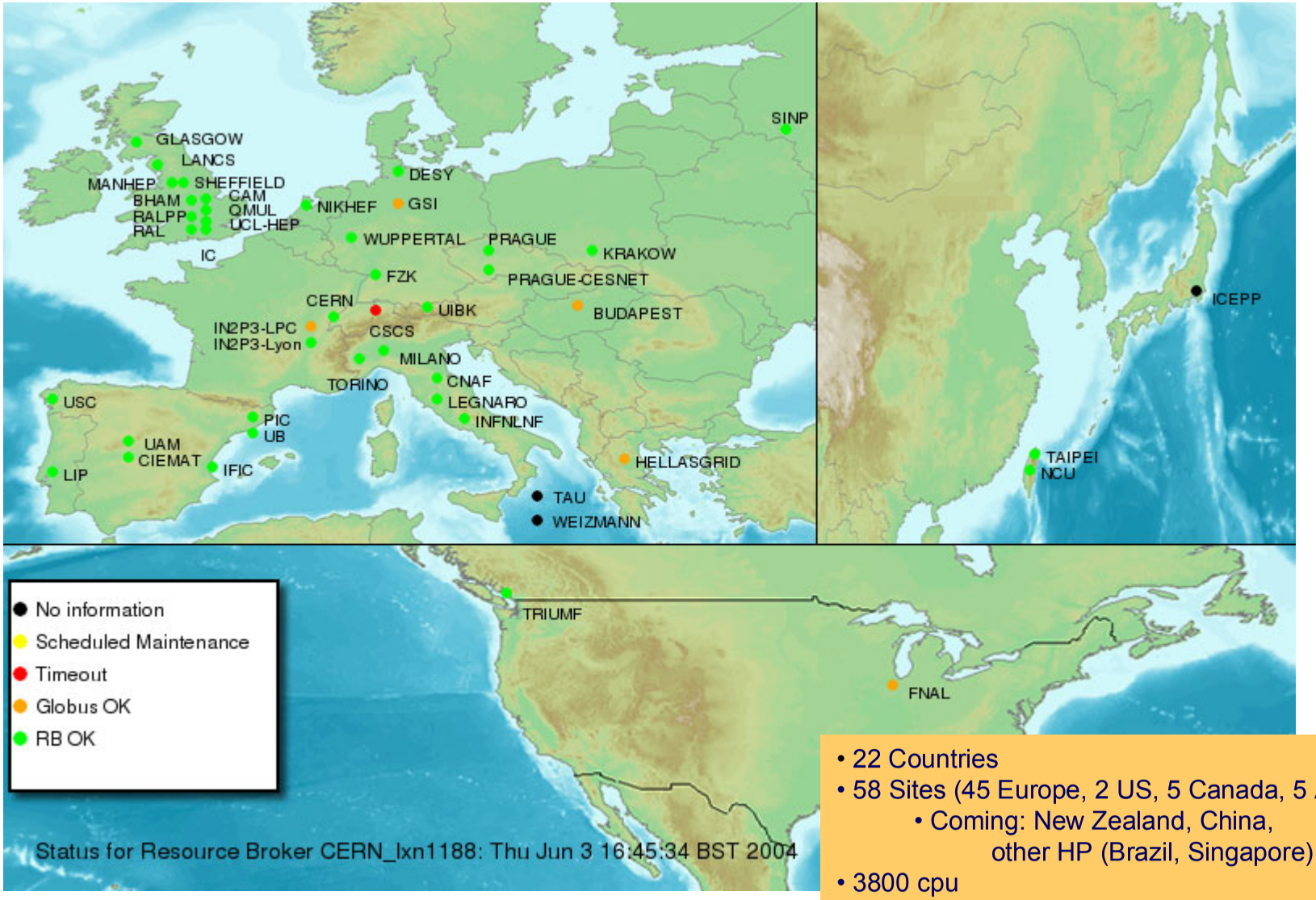
Ian Bird
LCG, CERN

GDB Meeting
CERN
15th June 2004

Overview

- Current Status/Issues with LCG-2
- Timeline & Possible medium-term work
- Interoperability with Grid3/Nordugrid
- Priorities

Sites in LCG-2/EGEE-0 : June 4 2004



- 22 Countries
- 58 Sites (45 Europe, 2 US, 5 Canada, 5 Asia, 1 HP)
  - Coming: New Zealand, China, other HP (Brazil, Singapore)
- 3800 cpu

Current status

- Middleware release end May
  - Consolidation release – see details
- Near future releases (mid/end June) and subsequently
  - Provide functional upgrades (add-on's)
  - Not major releases
  - Continue to provide bug fixes – but basic system is stable
  - Ports to other Linux flavours

LCG-2 May 31 release

- Consolidation/maintenance activities - ~ 50 bugs fixed
  - New VDT 1.1.14 (Globus 2.4.3)
  - Workload Management maintenance
  - Data Management maintenance and features
  - GridICE monitoring improvements
- New features
  - Data Management
    - lcg-utils - tools requested by experiments based on C++ API
    - GFAL integration
    - “long names” Castor client
- And: (already deployed)
  - Updated BDII – information system now expected to scale to 500 sites
  - RB performance much faster – BDII removes need to query sites during ranking – submission time now ~ 1s

Data Management maintenance

The main changes in this release are:

- Upgrade of GSoap runtime to 2.3 for all C++ clients (needed by ATLAS)
- Addition of extra methods into catalogs for bulk operations (requested by CMS/POOL)
- Integration of EDG-SE StorageResource for ADS interaction at RAL
- Refactoring of info system interaction and printInfo command in Replica Manager (internal request to rewrite a buggy component that caused many error reports)
- LRC/RMC C++ clients v2.3.0
  - edg-rm-* replaced now by lcg-rm-*

Timeline

- The context for setting priorities for LCG-2 is the timeline of LCG-2 vs EGEE middleware development and maturity
  - Will be constantly re-evaluated, but have not yet seen first release
- Still expect to be running LCG-2 until Spring 2005
  - Both for LCG experiment work and service challenges
  - EGEE applications (non-LCG) have requirements
  - Many possible areas of improvement
    - Some are basic infrastructure – not necessarily dependent on exact middleware
    - Some are problems/issues with existing middleware – how much effort is put here needs:
      - Input from experiments (priorities),
      - Consideration of what could be learned
      - Effort involved in implementation
  - … and hope/expect components/services coming from EGEE prototype

Future roadmap – what could be done?

- LCG-2 – there might be effort invested in:
  - Data management –
    - see list of possible work – short term vs longer term
    - dCache
  - R-GMA for monitoring applications
  - VOMS
  - Other Infrastructure issues:
    - Accounting
    - User and operational support
    - Security (incident response, auditing, etc)
    - VO Management
    - IP connectivity issues
    - …

Add-on services in June

- R-GMA
  - This is ready to be deployed – as basic service that can be used by accounting system, experiment monitoring, etc.
- VOMS
  - Basic implementation
    - VOMS servers, generation of gridmap files has been tested – will be deployed imminently
      - Will run LDAP and VOMS servers in parallel until stable
    - Needs work on management interface (VOX/VOM RS or VOMS)
- Also in progress:
  - Integration of ROOT with GFAL
    - Will allow POOL and other clients to use GFAL

Data management

- What is missing:
  - Fully functional storage element, with components:
    - Disk pool manager (dCache, …)
    - File transfer service (layer over gridftp)
    - SRM interface (exists)
    - POSIX I/O interface (GFAL and rfio, dcap, etc)
    - Security service (…)
  - Acceptable Replica Location Service
    - Adequate performance, scalability, distribution/replication
  - Higher level tools
    - Many have been added, more simple tools can be provided
    - Replica Manager service

Data management – 1 : dCache as DPM

- dCache was proposed as a disk-pool manager with existing SRM
  - Was in production use by CMS
- LCG has been working with dCache developers at FNAL and DESY for several months
  - We found and resolved several significant performance issues
  - Many iterations on packaging,
  - SRM (in)compatibilities caused by Java vs C implementations of GSI, different Globus versions
  - Simple tests with RB could hang service (similar problems seen at FNAL – they limited to 15 clients)
    - Now resolved (this weekend)? – restarts needed only 1/week?
  - dCache is clearly a much more complex solution than the problem
- Now (!) expect we can do some test deployments
- Clarify support commitments from DESY
  - This week – phone conf on Thursday
- ..and FNAL… (for SRM, grdiftp)

Data management – 2 : Potential RLS work

- Focus on low level services
  - Essential catalogs, storage, data movement
  - Experiments have higher level services (TMDB, Don Quijote)
  - Provide lightweight, fast tools (lcg_utils)
- Catalogs:
  - Split into 3:
    - Replica Catalog
      - Stores GUID -> PFN mappings. It DOES NOT store any attributes
      - Allow Bulk inserts of GUID->PFN mappings
      - Provide Bulk queries (with either a list of PFNs or a list of GUIDs as argument)
      - OPTIONAL: Provide a SQL query operation (probably only for admin usage e.g. Give me all the PFNs on this site X
      - OPTIONAL: Provide cursors on queries to allow for very large result sets (needed if SQL queries are provided)
    - File Catalog
      - Stores LFN -> GUID mappings and GUID attributes (NOTE: There are no LFN attributes)
      - Make the LFN space be hierarchical. Provide hierarchical operations (rm, mv, ls, …)
      - Provide GUID attributes. Also provide a 'stat' like object for middleware information (file size, creator, creation time, checksum, …) This will allow for consistency checking and efficient 'ls –l' operations
      - This will supply bulk inserts (of all attributes and stat infor for a guid, along with optional LFN)
      - This will supply efficient queries, along with cursors for dealing with large results set
    - Metadata Catalog
      - We assume that most of the metadata will be in experiment catalogs.
  - Query is always a 2 stage operation; BUT Replica and File catalogs physically in same db

Other RLS/RM issues

- We have ideas on how these issues can also be addressed:
  - Naming
    - Finalise naming schemes for LFNs, GUIDs and PFNs
  - Transactions
    - Supported in db but not exposed – several options.
  - Cursors
    - Currently queries are not consistent, and also are rerun every time as you page through them. This leads to tools breaking and also to high db load. Proper cursors from the db would help us make these 'transactional' as well.
  - WAN interaction
    - RM as a service (or use RRS). Could buffer/cache requests. IP connectivity
  - Client Side interaction
    - RM service also allows timeouts, retries, connection management etc. Very thin clients.
  - GSI
    - Authentication – allows accounting and auditing
  - Management
    - Logging and log analysis tools, to make it easier to debug problems (tied in with GSI in order to get identity)
    - GUI + Management. Need simple (web) view of contents.
    - Accounting. Need GSI in order to get user identity
  - DB Replication
    - Some ideas to use DB tools

Data management – 3 : longer term

- Four broad areas:
  - Basic file transfer service
    - A basic service that is required – collaborate with EGEE/JRA1
  - File and replica catalogs
    - See previous discussion, priorities need to be set
  - Lightweight disk pool manager
    - Recognised as a need – dCache will be suitable for large Tier 2 sites who have large disk pools, or for MSS front-ends, but still miss a simple dpm
      - Some research to do, before a concrete proposal
  - Posix I/O
    - Based on GFAL, integration with ROOT in progress

Portability – What after RH 7.3?

- In the absence of consensus, we have chosen to port LCG-2 to:
  - RH Enterprise Server 3.0 IA32, the CERN variant
    - RH, recompiled by CERN + CERN mods
    - It should be freely downloadable when certified by CERN
    - Already well integrated in autobuild
    - We have started to install a small testbed which we will later connect to bi C&T for interoperability testing
    - we have a WN working (manual installation)
  - RH Enterprise Server 3.0 IA64
    - Needed to support OpenLab
    - External OpenLab partners involved (HP, IBM)
    - Most of the work has already been done manually by OpenLab people, work is progressing to integrate it into the autobuild system
- This work is directly usable for Scientific Linux
  - Will become primary platform we expect

After RH 7.3

- The porting to other than RH 7.3 has become much higher priority
- We are working on some ports ourselves
- Collaboration
  - with TCD (they have some experience with non-Linux systems such as IRIX)
  - Start collaborations with QMUL and LeSC who offered their resources
- We provide anonymous access to our CVS server and will advise on how to setup the build process
- We will then introduce all changes into the CVS server for all LCG C&T tested architectures
- If we do not have a necessary hardware, we will solicit help in providing necessary access to such resources and help in certification

- Start with Worker Nodes, leaving service nodes on IA32 (probably RH 7.3 for now) and adding CE, SE, and others services later

Interoperability

- Observations:
  - BDII is now very reliable, RB can make use of information at top level
  - LCG-2, Grid3, NorduGrid all use MDS (BDII is MDS)
  - LCG-2 and Grid3 use GLUE schema
    - But semantics, and extensions may differ
- Idea to build a BDII filter, that
  - Plugs Grid3 IS into LCG-2 and would allow RB to submit jobs to all
  - Single (central) LCG RLS would still work in this case (does not solve integration of catalogs)
  - Needs some work to understand GLUE schema differences
  - Subsequently could do the same with NorduGrid
    - But IS schema is not GLUE – how easy would the filter be?
- This is not full interoperability, but would be a good first step
  - Agreed last week to work with Grid3 people to investigate this

Priorities

- Several things are in progress
  - Tools etc as part of DC support efforts, more will come as actual usage continues to become apparent (!)
  - R-GMA, VOMS, GFAL/ROOT, etc.
  - Basic infrastructure improvements (accounting, monitoring, etc)
- Several potential improvements possible
  - Reliable data transfer service
  - Potential RLS/RM improvements
  - Lightweight dpm
- Understanding with EGEE middleware group
  - Not duplication – leverage each others efforts
- Missing – clear priorities from experiments

Summary

- Basic LCG-2 is stable
  - Continue to stabilise, bug fixes,
  - Add-on services, tools to make use easier
- Other LCG infrastructure work will continue
  - Monitoring, accounting, tools, …
- Data management – large potential scope of work
  - Reliable file transfer
  - Alternative dpm
  - RLS/RM improvements
  - DB replication strategies
  - Some of this should be longer term developments with EGEE and Grid3, but some can be done soon
- Clarify priorities
- Ideas for interoperability - to be investigated