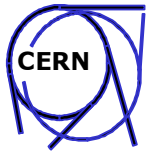



# LCG-1 Deployment Status

Ian Bird
IT Division, CERN

GDB Meeting, RAL
8 July 2003

# Outline

- Middleware status
- Deployment schedule
- Adding resources
- Priorities for 2003

# Status of Middleware – July 4

Running system on certification test-bed (~ 1 week – still very unstable)
- (original expectation was to have this all working on April 30)

- RB - seems to work, certainly better than before
  - 50 + 50 jobs (2 streams in parallel) – no problem
  - 200 + 200 jobs (in parallel) – 1 job died, still under investigation
  - But:
    - RB hangs (no job submission possible)
    - Service does not start after reboot
    - Workload manager dies
    - Need fix to allow non-standard job managers (LCG provides to workaround shared file systems)
  - Expect fixes this weekend

- RLS – still many problems
  - we can transfer files but a lot needs to be done, it is possible lots of problems are configuration problems but are very hard to determine
  - Currently (4/7/03) some 9 outstanding issues

Just a representative list – but is this the end or are we still only part-way through debugging?

- MDS
  - Only part of the information available sometimes (Nikhef provided solution)
  - Real MDS code bug found by LCG this week – fixed and submitted to VDT/Globus

- R-GMA – single registry version
  - Being tested by developers in UK
  - System runs for several hours – not ready for deployment yet
  - This week a complete re-write of GIN was done

R-GMA: Clearly not ready for deployment

# Middleware - Comments

- VDT – 1.1.8-9
  - Excellent collaboration – very responsive to our needs – fast turnaround

- LCG contributions to VDT and EDG
  - added accounting and security auditing (connection between the process, the batch system and the submitter)
  - fixed gatekeeper log file growing infinitely by implementing rotating log schema, avoiding gatekeeper restarts
  - incorporated MDS bug fixes from NorduGrid, improving at the same time the timeout handling, this allows LCG to deploy MDS on bigger scale; found and fixed bugs in GRIS
  - new LCG versions of job managers that do not need shared home directories
    - issue raised by deploying LCG-0
    - solves scalability issues that prevented use of more than a few tens of worker nodes
  - gass-cache inode leak, "dead" inodes were never removed - filling up the disk, eventually causing the service to crash
  - and much more…
  - Many bugs found by trying "simple cases"

# Testing

- Progress in the last few weeks with new Russian people involved
- We have the following tests – to define when LCG-1 can be deployed
  - Sequential & parallel job submission (RB functionality)
  - Job storms (parametrizable)
    - normal
    - replica manager
    - copy (gridFTP)
    - checksum (big files through the sandbox with verification)
  - MDS testing suite
  - Replica Manager (simulates Monte Carlo production)
  - Globus functionality tests (through VDT test suite)
  - Service nodes functionality tests
    - MDS x BDII coherence tests
    - LRC x MRC coherence tests
- Many of these based on EDG "stress-tests" of Atlas and CMS

- We still need to define a "site verification test suite",
  - to be run as a validation of a site installation, before the site connects to the grid

# State of roll-out

## Two activities in parallel

### Certification

- Week 1 (June 30)
  - Finalise integration of candidate release
  - Invite Loose Canons to test
- Week 2 & 3
  - Loose Canon testing
  - Problem fixing
  - Iteration with EDG of significant problems
  - Take updates from EDG as needed
- July 18
  - Release of LCG-1

### Deployment

- Week 1 (June 30)
  - Preparation of sites, organise machines, etc.
  - Finish set up of VO, registration servers
- Week 2 & 3
  - Prepare configurations for each site
  - Deploy and install middleware at each Tier 1
  - Update packages from certification/testing outcomes
- July 21
  - Loose canons work on LCG-1 system
  - Invite experiment production managers to test

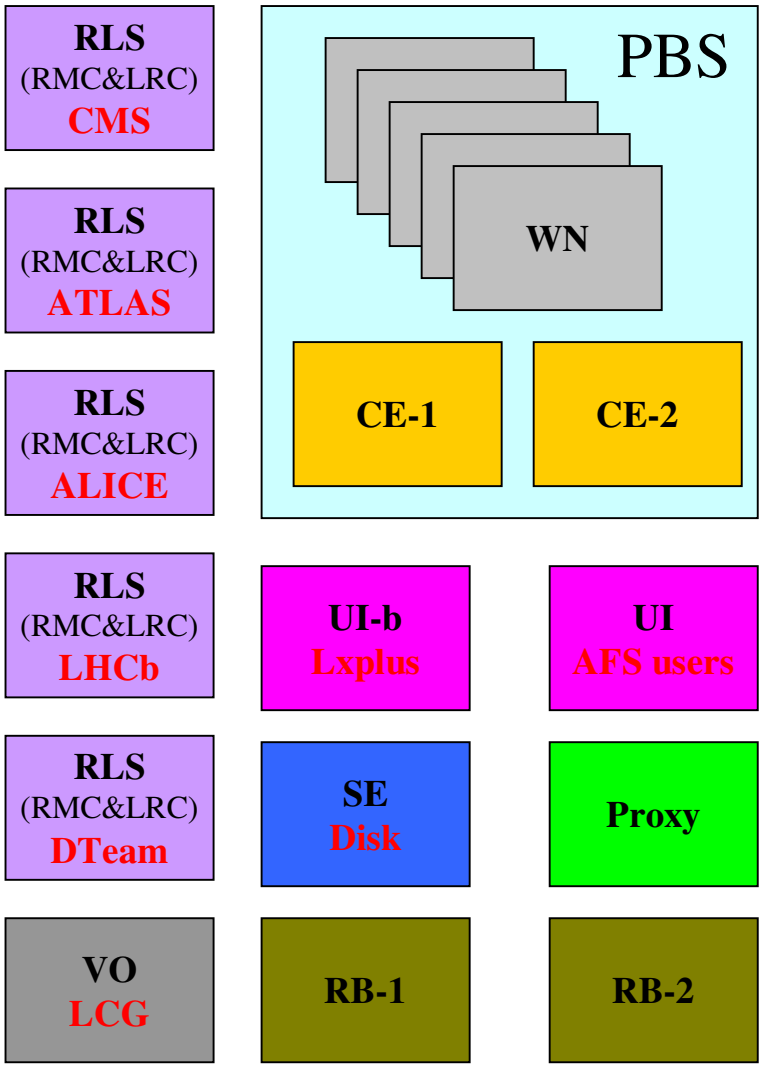
# LCG-0 Deployment Status

**These sites deployed the LCG-0 pilot system and will be the first sites to deploy LCG-1**

| | Site | Scheduled | Status |
|---|---|---|---|
| **Tier 1** | | | |
| 0 | CERN | 15/2/03 | Done |
| 1 | CNAF | 28/2/03 | Done |
| 2 | RAL | 28/2/03 | Done |
| 3 | FNAL | 30/3/03 | Done |
| 4 | Taipei | 15/4/03 | Done |
| 5 | FZK | 30/4/03 | Done |
| 6 | IN2P3 | 7/5/03 | In prep. ?? |
| 7 | BNL | 15/5/03 | Done ? |
| 8 | Russia (Moscow) | 21/5/03 | In prep. |
| 9 | Tokyo | 21/5/03 | Done |
| **Tier 2** | | | |
| 10 | Legnaro (INFN) | After CNAF | Done |

# CERN LCG-1 First Launch

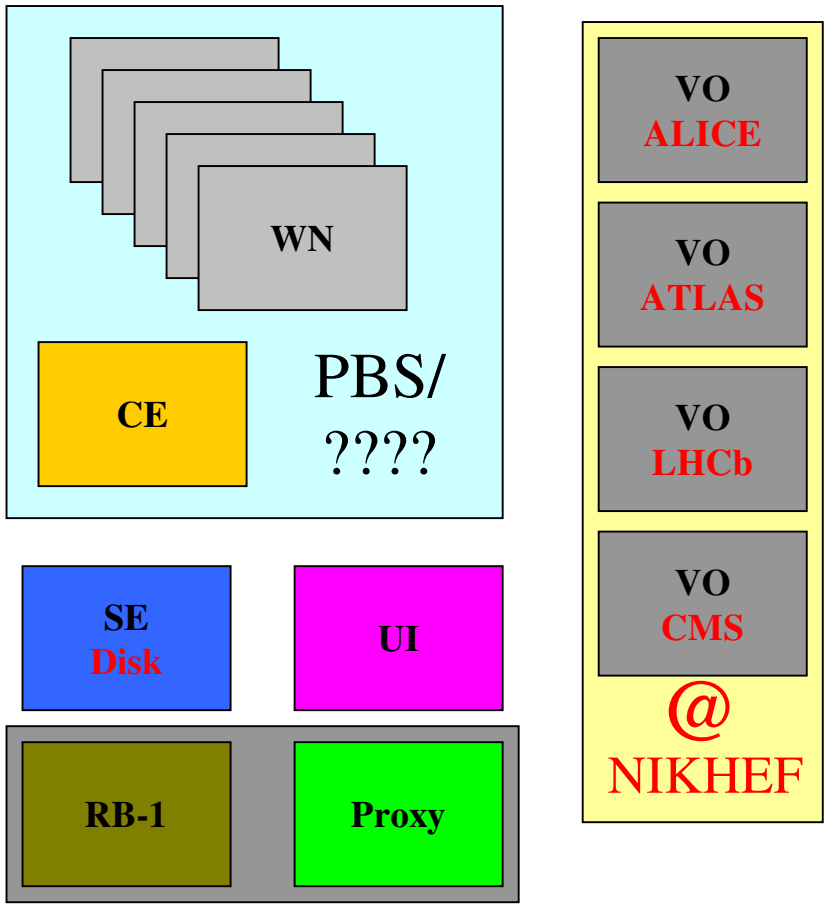
## Information System excluded




- One RLS server needed for every VO supported
  - Run at CERN
- The VO servers for LCGauthorisation and DTeam will be run at CERN
  - Other VOs at NIKHEF
- Registration Server at CERN
- CVS + RPM repository at CERN
- Order of 10 WNs to allow some computing.
- 2 CE to test config for multiple CEs
  - Initially PBS, add CE for LSF and migrate
- Disk (gridFTP) SE to avoid problems with MSS interface.
  - Test MSS separately and add when stable
- Proxy server to allow long running jobs
- UI installed on a CERN type node. User accounts through AFS (LCFG lite)
- UI-b by installing the UI software and default configuration in an AFS directory (later)
- WNs have outbound connectivity

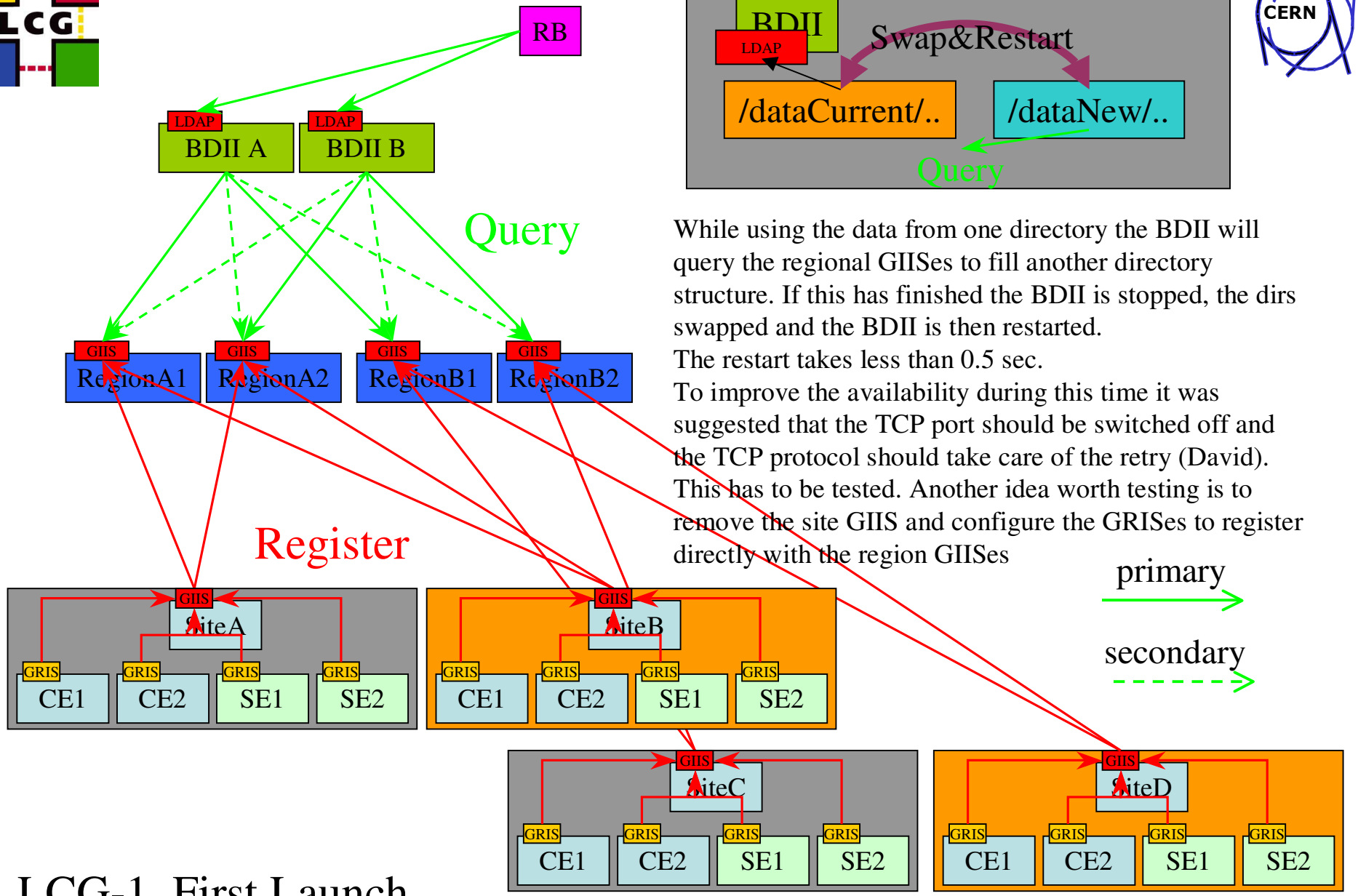
# Other Sites LCG-1 First Launch

## Information System excluded




- The sites decide how they want to setup their UI.
  - In case they have no idea, local user management (YP) is suggested.
- RB and Proxy server are optional.
  - At least one additional site in Europe and one in the US will run RB(s).
  - Multiple RBs and UIs are essential to allow tests with asynchronous chaotic requests to the services.
- Default local resource management system is PBS, but the sites have the freedom to use their own system.
  - For the initial deployment will start with a minimal system using PBS
- We should aim to get additional RBs at RAL, FNAL and CNAF at day 5.
- LYON should start to try setting up by their own means. The WP1 documentation is very detailed.

Using multiple BDIIs requires RB changes







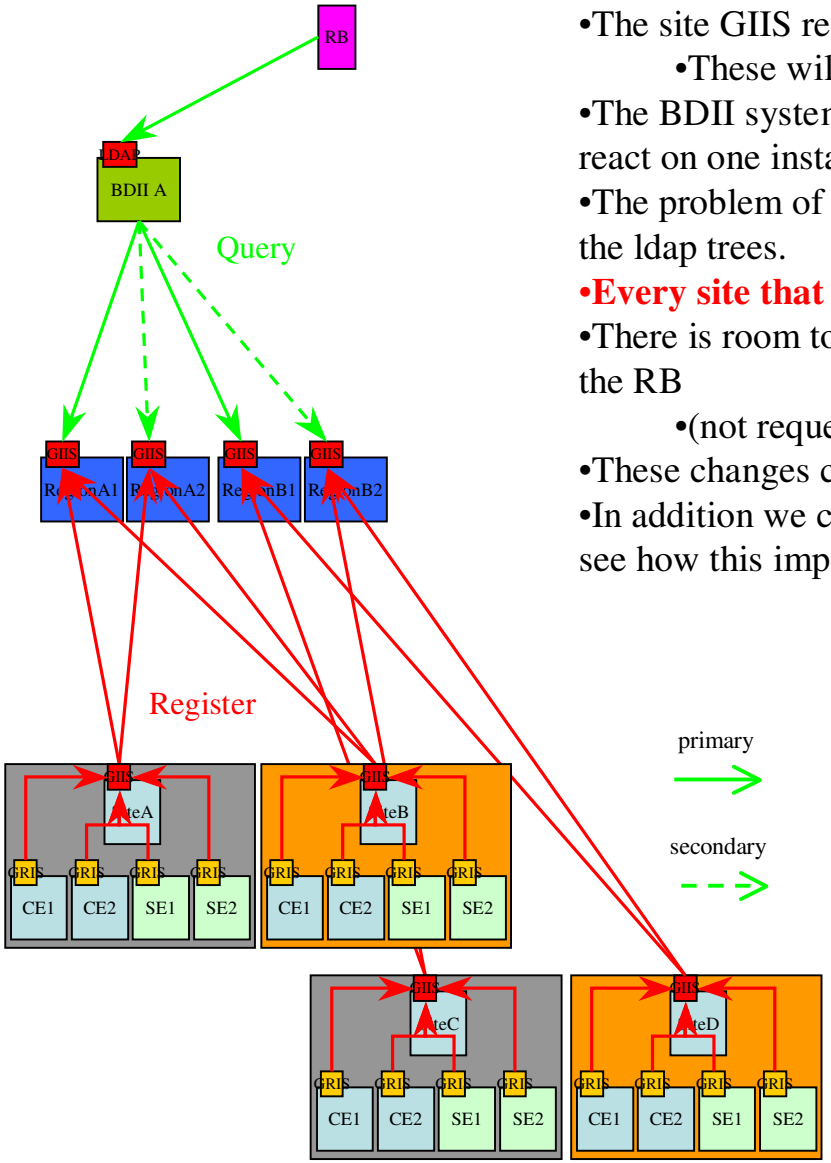
While using the data from one directory the BDII will query the regional GIISes to fill another directory structure. If this has finished the BDII is stopped, the dirs swapped and the BDII is then restarted.
The restart takes less than 0.5 sec.
To improve the availability during this time it was suggested that the TCP port should be switched off and the TCP protocol should take care of the retry (David). This has to be tested. Another idea worth testing is to remove the site GIIS and configure the GRISes to register directly with the region GIISes

# LCG-1 First Launch

## Information System Overview

# LCG-1 First Launch

## Information System Structure

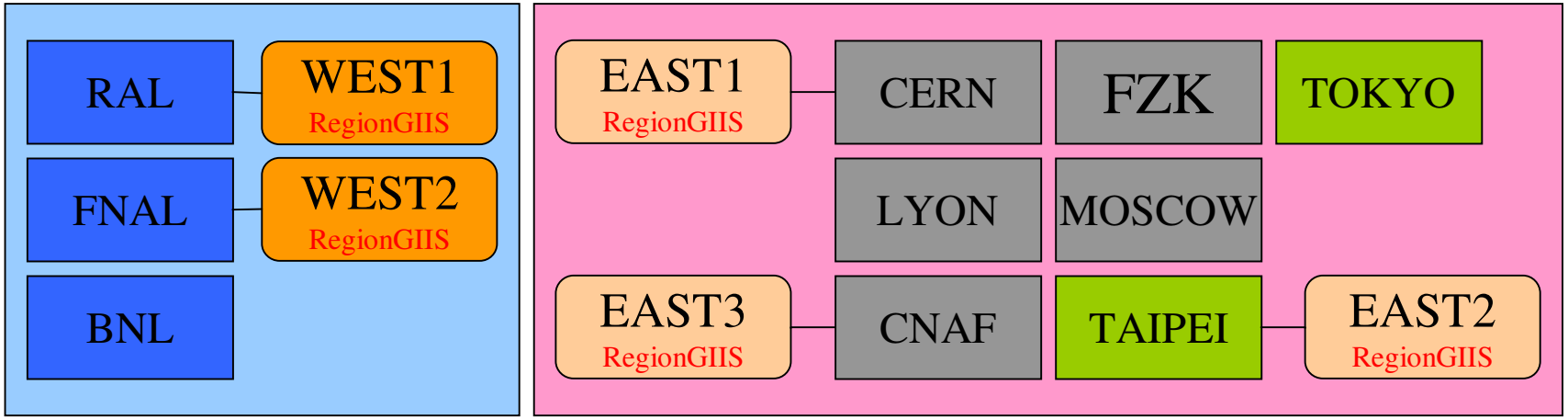



- Each site will run a site GIIS. On day one this will be run on one of CEs.
- The site GIIS registers to two or more regional GIISes.
  - These will be well known and in the configuration that we distribute.
- The BDII system has been modified by LCG to handle multiple regions and to react on one instance of the regional GIIS failing.
- The problem of stale information has been limited by repopulating and swapping the ldap trees.
- **Every site that runs an RB should run its own BDII.**
- There is room to improve the way this system works via small modifications to the RB
  - (not requesting DNs, using alternate multiple BDIIs…)
- These changes can be handled after we gain experience with the first release(s)
- In addition we can try to register the GRISes directly with the region GIISes to see how this improves the reliability.

# LCG-1 First Launch

## Information System Sites and Regions

•A Region should not contain too many sites since we have observed problems with MDS if a large number of sites are involved.
•To allow for future expansion, but not to make the system too complex we suggest starting with two regions and if needed split later to smaller regions.
    •The regions are: West and East of 0 degrees
•The idea is to have a large region and a small one and see how they work
•For the West 2 regional GIISes, and for the East 3 will be setup at the beginning,



# LCG-1 First Launch

## The first day(s)

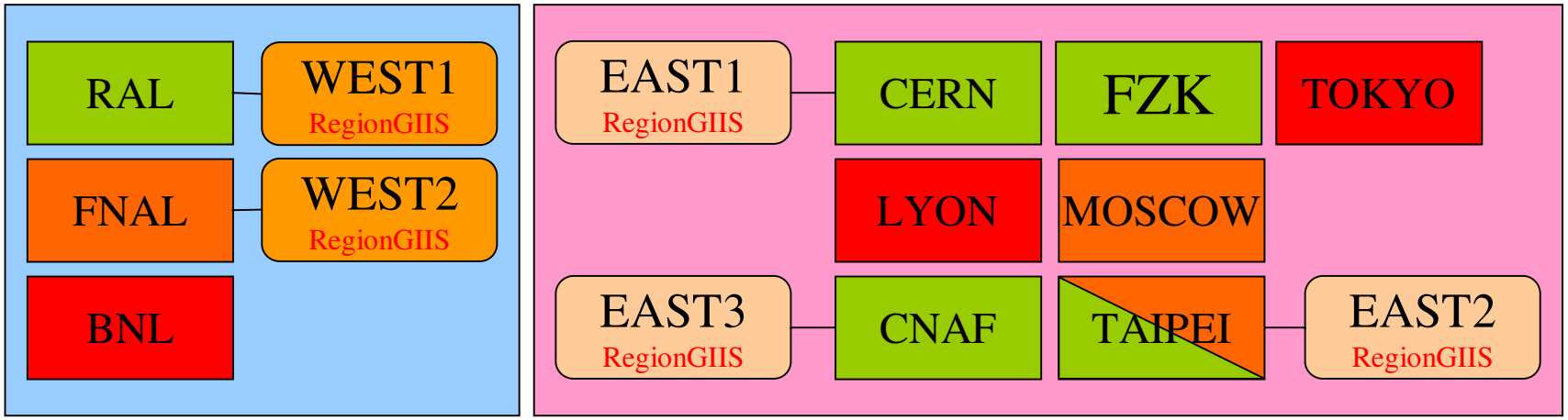
•It is very important that we get an initial integrated system during the first three days.

•To get this we leverage the experience of the sites that use LCFGng already. Those can be grouped as follows:

| | |
|---|---|
| **Solid Knowledge**: | RAL, CNAF, CERN |
| **Basic Knowledge**: | FZK, FNAL, Taipei, Moscow |
| **Non LCFG sites:** | BNL, Tokyo, Lyon |

•To accelerate the first deployment we sent a questionnaire to the LCFGng sites and instruct them to get the right host certificates for their nodes. Then we will write the site configuration files for the ORANGE sites and make them available through CVS.

•CERN-RAL-CNAF will launch their services and integrate first. After initial tests we add the ORANGE sites. Since we haven't tried a manual configuration of the service nodes up to now it can take a while until the RED sites join. The lcfg configuration files of the sites will be kept on the LCG CVS server.



# LCG-1 First Launch

## Questionnaire to the sites

To prepare the initial config and create the site config files for the ORANGE sites questions have been sent to the site managers to allow the team at CERN to prepare a set of configuration files to start with. These files and an initial installation guide are what we start from.

**All Nodes**
DNS servers
Network Time Server (ntp)
Domain
Subnet Mask
IP address
Router
Time Zone
In case you use a firewall:
Port range for grid services
Configure your firewall
Name of the LCFGng server

**Minimal Sites**
Hostnames of:CE,SE,UI,WN(s)(<10 Nodes)
Batch system(PBS,LSF,Condor..)
Is it possible to start with PBS?
Description of the WNs:
CPU(s), clock speed, memory, disk
WNs have outgoing network connectivity
Storage of your SE
Description of the SE
Protocols supported between SE and WNs
VOs supported by your site
**Create host certificates for CEs and Ses**
Which CA are you using for host certs?

will change

**Communication**
Contact email for site
Contact phone for site
Emergency phone for site
Name(s) + email of staff to participate in IM

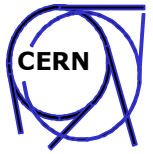
**Sites providing RBs**
Hostnames of: RB, Proxy Server, BDII
**Create host certificates for RBs and BDIIs**

**Information System**
CERN, CNAF, TAIPEI, FNAL and RAL:
Hostname for the regional GIIS
Every site should choose on which CE the site GIIS will run (otherwise we will pick one)

NFS: For the first release /home is shared between CE and WNs. In case you want to use a dedicated file server: hostname, exported directories
For the disk based SE a fileserver can be used too:
Hostname and directories exported, size of the partitions

# LCG-1 First Launch

## Prerequisites: Communication




**Before we can start to deploy we need a few things in place at the ten core sites**

- IM server for rapid communication,
  - Located at RAL
  - The contents must be archived and restricted access via a web interface should be provided (to avoid security exploits of the records)
- Mailing list setup that is used ONLY for this initial deployment.
  - In case the IRC is silent send mail, or if more detailed info needs to be exchanged.
  - Located at the future GOC site
  - Needs to be archived
  - Access only by the group that does the setup of LCG-1
- Template web page to report status of the sites.
  - Primitive, will be used only for the very first deployment.
  - Central one with links to the others at one place (RAL? CERN?)
  - The information that will be provided (see example page)
    - Contact mail address of site
    - Message of the day
    - Table with install and config status of the nodes

# What happens next?

- Bug fixing, bug fixing, bug fixing, …

- Middleware:
  - incremental features
  - Support for gcc x.x
  - Support for RH x.x
- Sites:
  - Aim for 15-20 sites by end 2003
  - New countries
  - Tier 2's in US, UK, IT, etc.
- Expand resources available to LCG in each site
  - Have to demonstrate that LCG-1 is useful
  - Solve integration issues

# Middleware expectations

Middleware developments LCG would like in 2003

- Cut-off for this is October in order to have system ready for 2004
- EDG further developments will finish in September

- R-GMA:
  - Multiple registries: are essential for a full production service – this work still has to be done, may not be simple
  - Will not be started until initial R-GMA is stable
  - If not delivered have a single point of failure/bottleneck, or MDS (**)
- RLS:
  - Need the RLI component to have distributed LRCs
  - Code is ready; being stress tested by developers
  - Fallback is single global catalog – but is single point of failure/bottleneck
- RLS: proxy service for Replica Management
  - Essential to remove requirement on WN IP connectivity
  - If not delivered limits sites that can run LCG, and resources that can be included (**)
- VOMS (VO management service)
  - Service itself is ready and tested
  - Needs integration with CE and Storage – must be done by developers of those services
  - Fallback is to continue as now with basic authorization via grid-map
- GFAL (Grid File Access Library)
  - LCG development
  - Prototype available, expect production version in August

# Schedule for adding sites

- LCG-1 will be 10 sites initially
- Aim for end 2003 to include 16 countries that commit resources for 1Q04
  - Czech. Republic, Hungary, Netherlands, Poland, Switzerland, Spain, Sweden
- Additional Tier 2's:
  - At least in UK, USA, IT (others that are ready?)
  - Expect that Tier 1s support the Tier 2 deployments, with LCG team as second line of support

Request from GDB members:

- Proposed sites (for Tier 2s)
- Date when site is ready to join LCG-1
- Contacts at site: system manager and security contact
  - For Tier 2 need also name of responsible Tier 1 person who will coordinate

# LCG Resource Commitments – 1Q04

| | CPU (kSI2K) | Disk TB | Support FTE | Tape TB |
|---|---|---|---|---|
| CERN | 700 | 160 | 10.0 | 1000 |
| Czech Republic | 60 | 5 | 2.5 | 5 |
| France | 420 | 81 | 10.2 | 540 |
| Germany | 207 | 40 | 9.0 | 62 |
| Holland | 124 | 3 | 4.0 | 12 |
| Italy | 507 | 60 | 16.0 | 100 |
| Japan | 220 | 45 | 5.0 | 100 |
| Poland | 86 | 9 | 5.0 | 28 |
| Russia | 120 | 30 | 10.0 | 40 |
| Taiwan | 220 | 30 | 4.0 | 120 |
| Spain | 150 | 30 | 4.0 | 100 |
| Sweden | 179 | 40 | 2.0 | 40 |
| Switzerland | 26 | 5 | 2.0 | 40 |
| UK | 1656 | 226 | 17.3 | 295 |
| USA | 801 | 176 | 15.5 | 1741 |
| Total | 5600 | 1169 | 120.0 | 4223 |

# Expanding resources

- Requires:
- LCG demonstrate that LCG-1 is useful:
  - Experiments coordinate with site managers migration of resources into LCG-1
- Resolution of some issues:
  - Shared file systems
    - LCG has alternative job managers that do not require shared fs
  - Worker node connectivity
    - Replica management and file transfers must be done by services on behalf of the WN
      - Needs Replica Management Service
  - MSS systems – must deploy SRM interfaces

# Summary

- We are late having integrated release of LCG-1
- Testing is much curtailed
- Aim to deploy the basic system before end July
  - Very much depends on what bugs are still to be found

- The initial system will be basic – but we *must* start to run a service or we cannot make progress

- Hope we can get experiments involved in August

- Must start to plan developments for the rest of 2003