LCG2 Administrators Course
Oxford University, 19th- 21st July 2004
LCG Grid Elements

Steve Traylen
s.traylen@rl.ac.uk
Rutherford Appleton Lab

Grid Elements - Outline

- Grid Layout, Node Requirements.
- Middleware Components In LCG.
- Installation Guide.
- Registration Procedure.
  - LCG Central
  - GOC
- Site Requirements.
- Network (firewall) requirements.
- LCFGng.
  - Configuration Theory
  - Installation Process
- LCG installation with LCFGng overview.

Grid Layout

- All machines and services within LCG are provided or owned by one of three entities on the grid.
  - Grid Wide Services.
    - Grid wide level services, e.g. there is only one information system.
  - VO Services.
    - VOs will run VO wide services, e.g. user management and meta data catalogues.
  - Site Services.
    - Sites provide the physical CPU power and storage.
- Of course the VO and Grid level services, many of which are distributed, are physically located at particular sites.

Providing Services

- How does a site provide a service or resource?
  - It publishes the existence of the resource into an agreed schema.
- Within LCG and the whole EGEE this is the GLUE schema. A short example of GLUE, it is in LDIF.

```
# lcgce02.gridpp.rl.ac.uk/siteinfo, local, grid
dn: in=lcgce02.gridpp.rl.ac.uk/siteinfo,Mds-Vo-name=local,o=grid
objectClass: SiteInfo
objectClass: DataGridTop
objectClass: DynamicObject
siteName: RAL-LCG2
sysAdminContact: lcg-support@gridpp.rl.ac.uk
dataGridVersion: LCG-2_1_0
installationDate: 20040116115700Z
```

Grid Wide Services(1)

- LCG grid wide services are all VO and site neutral.
  - Authorisation service.
    - Currently unique and runs at CERN.
    - EGEE could create another, e.g. a stronger one for biomedical applications.
    - Importantly though users individually use this service when they sign up for LCG.
  - Top levels of the information system, i.e. BDII.
    - These are duplicated at many of the Tier1s in LCG and offer a view of all resources (services, CPUs, storage.).
    - A VOs information may flow through a site that does not support the VO.

Grid Wide Services(2)

- "User Interface"
  - The user interface is VO neutral, it supports a user not a VO and can access all of LCG.
  - In LCG it belongs here more than as a member of a site service or VO service.
  - With customisation it may be a VO service, e.g. GANGA.
- Grid middleware repositories and recommended deployment.

- Within EGEE responsibility for running grid wide services is made by the Core Infrastructure Centres (CIC).

Virtual Organisation Services(1)

- The VO maintains a list of its members.
  - Could be in an LDAP server.
  - A VOMS service.
- The VO will run catalogues and high level application systems. All can be VO specific and may not be a grid service.
  - Within LCG the EDG Replica Location Service is very common.
  - LHCb runs it's DIRAC service for interfacing LCG to it's application.

Virtual Organisation Services(2)

- Within EGEE Regional Operation Centres (ROCs) can support a VO.
- Often a VO hosting institution is obvious.
  - DESY Lab hosts the H1 VO.
  - CERN hosts the LHC experiments.

Virtual Organisation Services(3)

- A VO providing a replica catalogue.
- Replica manager clients locate their services with the following information.

```
# http://rlscms.cern.ch:7777/cms/v2.2/edg-local-replica-catalog/services/edg-local-replica-catalog, local, grid
dn: GlueServiceURI=http://rlscms.cern.ch:7777/cms/v2.2/edg-local-replica-catalog/services/edg-local
                                        -replica-catalog,mds-vo-name=local,o=grid
GlueServiceURI: http://rlscms.cern.ch:7777/cms/v2.2/edg-local-replica-catalog/services/edg
                                        -local-replica-catalog
GlueServiceType: edg-local-replica-catalog
GlueServicePrimaryOwnerName: LCG
GlueServiceHostingOrganization: CERN
GlueServiceAccessControlRule: cms
GlueServiceInformationServiceURL: MDS2GRIS:ldap://lxn1194.cern.ch:2135/mds-vo-name=local,o=grid
```

Site Services(1)

- Within LCG there are exactly two physical resources that a site may provide.
  - CPU.
    - All computing is presented as a single Computing Element (CE).
    - A GlueCE in the schema is actually a queue on the batch system. Multiple queues results in multiple CEs.
  - Storage.
    - All storage is presented as a single Storage Element (SE) in the schema, a GlueSE. You have one Storage Element published per interface per blob of storage……
- Terms CE and SE are "usually misused" to describe the head node of the resources.
- Realistically one node is needed as a CE, one as an SE per site.

Site Service(2) GlueCE

- A GlueCE publishes how many CPUs there are and queues lengths amongst other things.
- Queues support one or more VOs. Supported VOs are published within the GlueCE.
- GlueCE information cached in the information system.
- The Resource Broker (workload manager) uses this information to match a user's submitted job requirements to a suitable site.

Site Sevices(3) GlueCE

```
# lcgce02.gridpp.rl.ac.uk:2119/jobmanager-lcgpbs-long, local, grid
dn: GlueCEUniqueID=lcgce02.gridpp.rl.ac.uk:2119/jobmanager-lcgpbs-long, mds-vo-name
                =local, o=grid
GlueCEName: long
GlueCEUniqueID: lcgce02.gridpp.rl.ac.uk:2119/jobmanager-lcgpbs-long
GlueCEInfoHostName: lcgce02.gridpp.rl.ac.uk
GlueCEInfoLRMSType: pbs
GlueCEInfoLRMSVersion: torque_1.0.1p5
GlueCEInfoTotalCPUs: 144
GlueCEStateEstimatedResponseTime: 0
GlueCEStateFreeCPUs: 4
GlueCEStateRunningJobs: 70
GlueCEStateStatus: Production
GlueCEPolicyMaxCPUTime: 4800
GlueCEPolicyPriority: 1
GlueCEAccessControlBaseRule: VO:alice
GlueCEAccessControlBaseRule: VO:atlas
GlueInformationServiceURL: ldap://lcgce02.gridpp.rl.ac.uk:2135/mds-vo-name=local,o=grid
```

Site Sevices(4) GlueSE

- The GlueSE publishes what type of management interface is available for the storage of files.
- Within LCG there are 3 interfaces in use today.
  - Disk - No management, put and get possible.
  - SRM_V1 - More management, reserve, advisory delete, ….
  - EDG-SE - Some where between the above two.
- GlueSE is used by data management tools, the edg replica manager in particular, to learn how to shift files around.
- Today we only consider disk.

Site Sevices(3) GlueSE

```
# castorgrid.cern.ch, local, grid
dn: GlueSEUniqueID=castorgrid.cern.ch,mds-vo-name=local,o=grid
objectClass: GlueSETop
objectClass: GlueSE
objectClass: GlueInformationService
objectClass: Gluekey
objectClass: GlueSchemaVersion
GlueSEUniqueID: castorgrid.cern.ch
GlueSEName: CERN-TEST2:disk
GlueSEPort: 2811
GlueInformationServiceURL: ldap://lxn1194.cern.ch:2135/Mds-Vo-name=local,o=grid
GlueForeignKey: GlueSLUniqueID=castorgrid.cern.ch
GlueSchemaVersionMajor: 1
GlueSchemaVersionMinor: 1
```

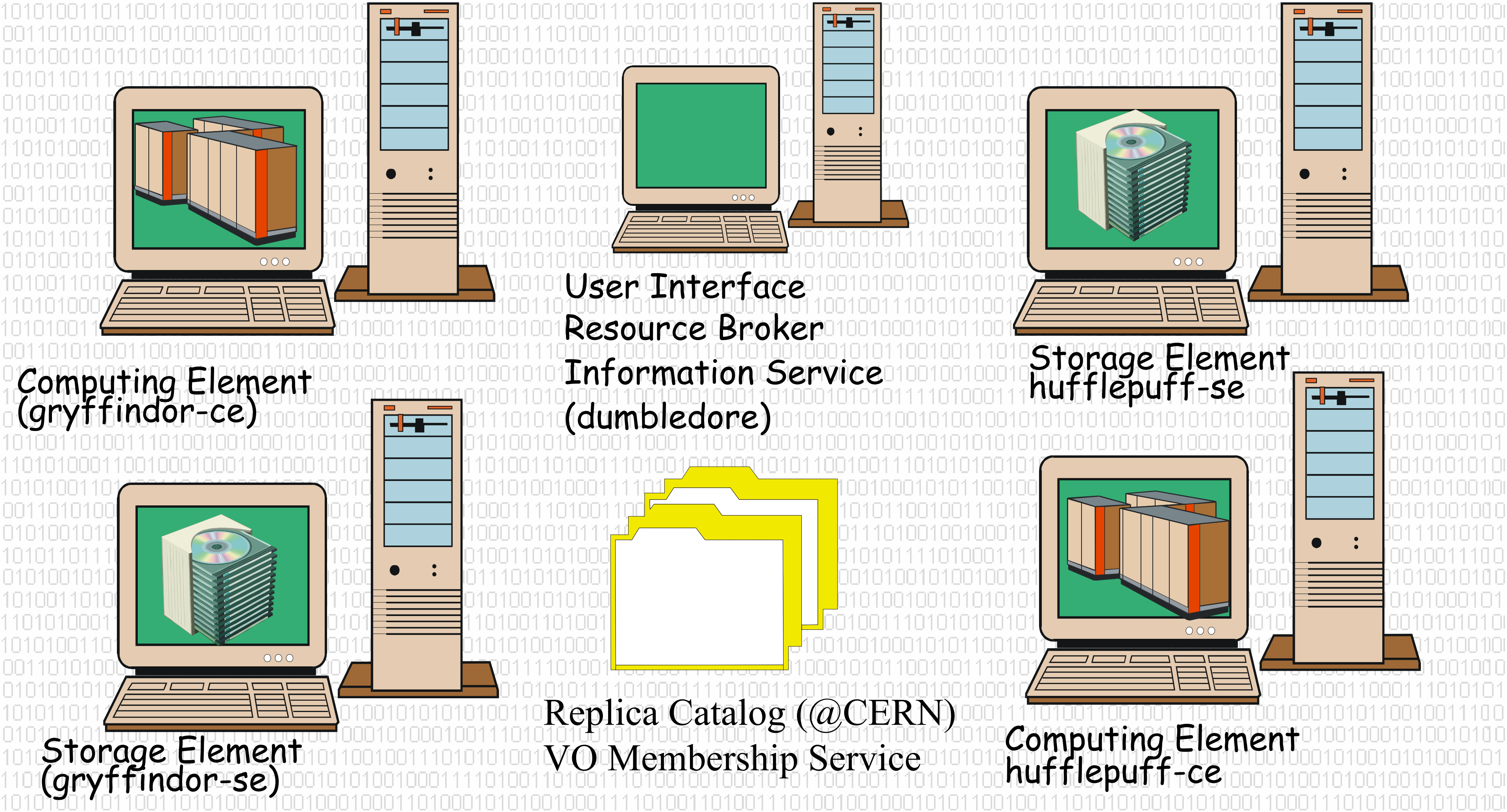
Site Sevices(3)

- In addition to the GlueCEs and GlueSEs other services offered by a site that describe how the site is set up rather than physical resources.
- The most interesting are:
  - GlueSARoot and GlueSAControlAccessBase
    - Publishes a Storage Area that a VO is permitted to write to.
  - GlueCESEBind
    - Publishes how a CE and SE are connected at a local site.
- Replica management tools require this information to be correct and valid.

Resource Broker

- Provided by sites.
- Matches a job requirements against the user's submitted job.
- Only particular VOs are permitted to use a particular RB.
- Currently RBs are located from a static configuration file on the user interface.
  - This is starting to change and they now do publish their existance.

Course Site Configuration



Computing Element
(gryffindor-ce)

User Interface
Resource Broker
Information Service
(dumbledore)

Storage Element
hufflepuff-se

Storage Element
(gryffindor-se)

Replica Catalog (@CERN)
VO Membership Service

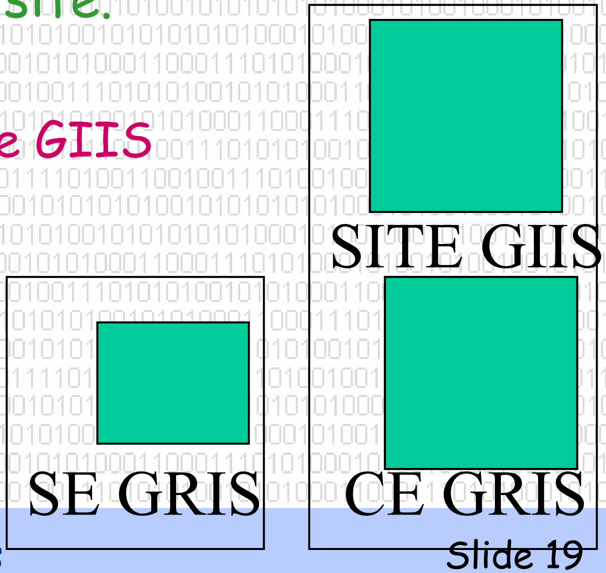
Computing Element
hufflepuff-ce

LCG Middleware Components at Sites

- Information (Provider Scripts, MDS).
- Information (BDII)
- MkGridMap and Pool Accounts.
- GridFTP.
- Gatekeeper.
- LCG PBS Job Manager.
- Software manager areas.
- EDG replica manager.
- Job submission and job brokering.
  - From end user command to result.
  - Our aim is to help problem isolation at later stages.

Information Service in LCG

- Example , Find all the CEs (queues) at RAL
  - ldapsearch -x -H ldap://lcgce02.gridpp.rl.ac.uk:2135 -b 'Mds-vo-name=rallcg2,o=Grid' '(objectClass=GlueCE)'
  - Displays all the CEs at RAL.
- What happended.
  - We queried an ldap server (site GIIS) on lcgce02 which queried two GRISes. One GRIS per node at each site.
    - 1st on the CE, 2nd on the SE.
    - The GRISes pre soft state register to the GIIS /etc/globus.conf on CE and SE.
  - Traditionally the SITE GIIS is on the CE node in LCG.



Information Provider

- At the back of the GRIS are various scripts that collect required information.
- Backend scripts are configured in the grid-info-resource-ldif.conf file.

```
dn: Mds-Vo-name=local,o=grid
objectclass: GlobusTop
type: exec
path: /opt/edg/libexec
base: ceinfo-wrapper.sh
args:
cachetime: 30
timelimit: 20
sizelimit: 50
```

Information Provider(2)

- This configuration defined that:
  /opt/edg/libexec/ceinfo-wrapper.sh
  should be run to fetch some information.
- Assuming PBS is the LRMS this script will call qstat, pbsnodes,… and print LDIF to STDOUT to fill in the GlueCE object that we saw earlier.
- Running the info provider by hand is an extremely good way to catch errors that are otherwise lost in the information system.

Information System (BDII)

- Introduced by NIKHEF during EDG to avoid the problem of site GIISes hanging an entire in hierarchy of GIISes within MDS.
- Scripts polls site GIISes and populate a standard OpenLDAP tree of the data.
- This is served out and used by the Resource Broker and the replica management clients.
- Within LCG there is no soft state registration from site to higher level - a configuration file references all the resources to check for.

Information System (BDII)(2)

- With BDII's having resources manually added to their configuration they can each be looking at different grids.
- Each of these grids is referred to as a Zone within LCG.
- The obvious examples being:
  - Testing Zone.
  - Core Zone.
- Also VOs have their own views as well.

MkgridMap and PoolAccounts

- The combination of these utilities allows sites to be able to support remote VO servers.
- mkgridmap configuration.
  - The file edg-mkgridmap.conf contains a list of URIs pointing to VO databases.
  - It also contains a URI pointing to the LCG auth server.
- Running mkgridmap
  - When run it looks at VO members who have also signed the auth guidelines.
  - These users are placed in the grid-mapfile.

PoolAccounts

- Example generated grid-mapfile.
  "/C=UK/O=eScience/OU=CLRC/L=RAL/CN=steve traylen" .dteam
- Users with a '.' in are illegal in UNIX. The '.' is a signal to the globus libraries to allocate a pool account, eg dteam005.
- A permanent mapping is maintained in the gridmapdir directory as below.
- Within LCG user accounts have a life time of 10 days before being recycled. A check is made first to see if it is active.

```
# ls –irt /etc/grid-security/gridmapdir/
2020015 dteam005
2028015 %2fc%3duk%2fo%3descience%2fou%3dclrc%2fl%3dral%2fcn%3dsteve%20traylen
2027976 atlas012
2027976 %2fc%3duk%2fo%3descience%2fou%3dclrc%2fl%3dral%2fcn%3dsteve%20burke
```

GridFTP

- Not much to say…
  - It runs on the disk SE for moving data in and out.
  - It runs in the CE so that end users can tag sites with software environments.
  - It runs on the RB so i/o sandboxes can be moved around.
  - It uses the pool account techniques.
  - The FTP is a passive FTP.
  - It logs to syslog.
  - Supplied by VDT.

Gatekeeper

- Globus Gatekeeper modified by EDG.
  - Supports dynamically loadable modules – LCMAPS
  - LCMAPS is configured currently to do nothing in LCG.
  - Becomes useful with addition of VOMS credentials.
- The default job manager is the fork job manager.
  - A requirement for the Resource Broker to work I believe.
- LRMS in use within LCG today or soon.
  - OpenPBS, PBSPro, Torque, LSF, BQS, Condor, SGE (ongoing).
  - Anything else and you need to write a job manager!

LCG PBS Job Manager(1)

- It is LCG (PBS JobManager) and not the (LCGPBS) JobManager.
- Big differences over globus provided PBS JobManager.
  - Does not require a shared /home across nodes.
    - Removes an NFS dependence ☺
  - Persistent process on the CE node reduced to one per user instead of one per job.
    - Less processes ☺.
  - Lots more to go wrong.
    - Multiple protocols and security infrastructures (GSI and SSH) used to move files around farm. ☹.
    - Harder to debug (though the original was plenty bad enough anyway), errors get lost.

LCG PBS JobManager(2)

- Reading the Perl this is what happens… I think.
  1. GRAM job arrives from gram client (Globus or Resource Broker).
  2. User's job script and associated files are placed in a tar.gz file.
  3. A simple identical job is submitted to PBS.

```
#!/bin/sh
#PBS –W stagein x509_proxy_file.
globus-url-copy tar from CE.
Unpack and run script.
```

LCG PBS Job Manager(3)

- Two files are transferred from the CE to the WN.
  - Proxy file with scp.
  - Tar ball with globus-url-copy.
- The first of these requires that pool users on the WN can SSH to the CE unchallenged.
- This is achieved through ssh shost.equiv configuration.
- Wrong or degraded configuration in this area is common reason for sporadic job failure.

Software Manager Areas

- LCG has the concept of a software managers in a VO.
- They are able to install software at sites in a special NFS/AFS shared area available on all WNs.
- Having validated a site they are able to tag a site as suitable for other people and production jobs.
  - They drop a file into a directory on the CE.
- As system managers this shared area must be provided for the SW managers to use.
- We will create this area when installing machines later on today and tomorrow.

EDG Replica Manager(1)

- Example I want to copy a file to a SE and register it with a LFN.
- Command: (cr = copy and register)
  - edg-rm --vo dteam cr file:/etc/group –l lfn:stevesFile.lfn –d lcgads01.gridpp.rl.ac.uk
- What happens:
  - The edg-rm has a configuration file that points to the top level of the information system.
  - It queries the information system (BDII) to discover.
    - Where the dteam replica catalogue is located.
    - Does lcgads01.gridpp.rl.ac.uk exist as an SE.
    - Am I permitted to use it and if so how should I use it.
  - A consequence of this is that you cannot test your own SE until your site is visible in the wider grid.

EDG Replica Manager(2)

- Having registered the file with a LFN it can be copied again back.
  - edg-rm –vo dteam cp lfn:stevesFile.lfn file:/tmp/myfile
- Many more operations possible, delete, replicate,…
  - Replica catalogue will only be consistent if replica manager tools are used.
  - As a site you have no way of enforcing use of replica management.
- Currently with unmanaged storage there is a problem for sites, remote pointers into sites can not be redirected by the sites. DCache is the proposed solution to this.

Resource Broker(1)

- Resource broker processes JDL and finds a suitable execution site. It should consider:
  - Sites that have my software installed.
  - Sites that I am permitted to run at.
  - Site that have some physical file present.
  - If my job fails what should I do?
  - How should I choose my execution site if multiple sites match?
- The RB has only two sources of information.
  - The information system (IS) for matching resources, locating services. The IS is found using a config file.
  - The replica catalogue API for locating files. The RC API uses the IS as well.

Resource Broker(2)

- A job is submitted as JDL. (Job Description Language)

```
Executable=“/bin/sh";
StdOutput=“aliroot.out";
StdError=“aliroot.err";
InputSandbox={“start_aliroot.sh", “rootrc”};
OutputSandbox={``aliroot.out", ``aliroot.err"};
RetryCount=7;
InputData={“lfn:ALICE-hits"};
Arguments=“start_aliroot.sh 3.02.04 3.07.01";
Environments={“ROOT_ALICE = $ALICE_ROOT_DIR/root/"};
Requirements=Member(“ALICE3.07.01",
        other.GlueHostApplicationSoftwareRunTimeEnvironment);
```

Resource Broker(3)

- All the previous attributes concerned matching. The result is Boolean at each site.
- The fact that the BDII has old (15 minutes) information is acceptable.
- For a JDL entry such as

  ```
      rank = other.FreeCPUs;
  ```

  old information was thought to be less suitable.
- Currently the RB queries all the matching site GIISes individually to make this decision….
- I believe this will change shortly to improve scalability.

Installation Guide

- Details for latest LCG release are here:
  - http://grid-deployment.web.cern.ch/grid-deployment/cgi-bin/index.cgi?var=releases
- Question: LCG2_0_0 or LCG2_1_0 what do you want to do?.... If LCG2_1_0.
- Some hands on at last!!
- CVSROOT should be set for you.

```
echo $CVSROOT
cvs login
cvs checkout LCG-2_1_0 lcg2
```

- The following slides are all just extracts and summaries from the new **HTML** install guide.

Site Registration

- Your local ROC should help, advise and support you through this process.
- Join LCG-ROLLOUT mailing list.
  - Announcements of new releases.
  - This goes to a lot of busy people and should be considered the last point for getting help. Your local ROC/Tier1 or within GridPP your Tier2 coordinator should be the first line of support.
- There is small questionnaire in Appendix G to fill in and pass onto both your ROC and LCG Central at CERN.
- Site admins are entitled to join the dteam VO and should do so thus enabling them to use their own site.☺

GOC Registration

- The LCG Grid Operations Centre is responsible for monitoring and so creating a reliable and usable grid system on LCG.
- To register just visit the URL http://goc.grid-support.ac.uk/gridsite/db-auth-request/
- More details of the GOC's operations will be in a later talk.
- Note for UK non-HEP people, the LCG GOC partially hosted in the UK is not the same as the E-Science GOC though overlap does exists.

Site Requirements

- Currently there are no formal site requirements.
- You do not sign anything as a site to join LCG.
- A service level agreement is in preparation with draft versions linked from a GRIDPP FAQ. http://www.gridpp.ac.uk/tb-support/faq/response.html

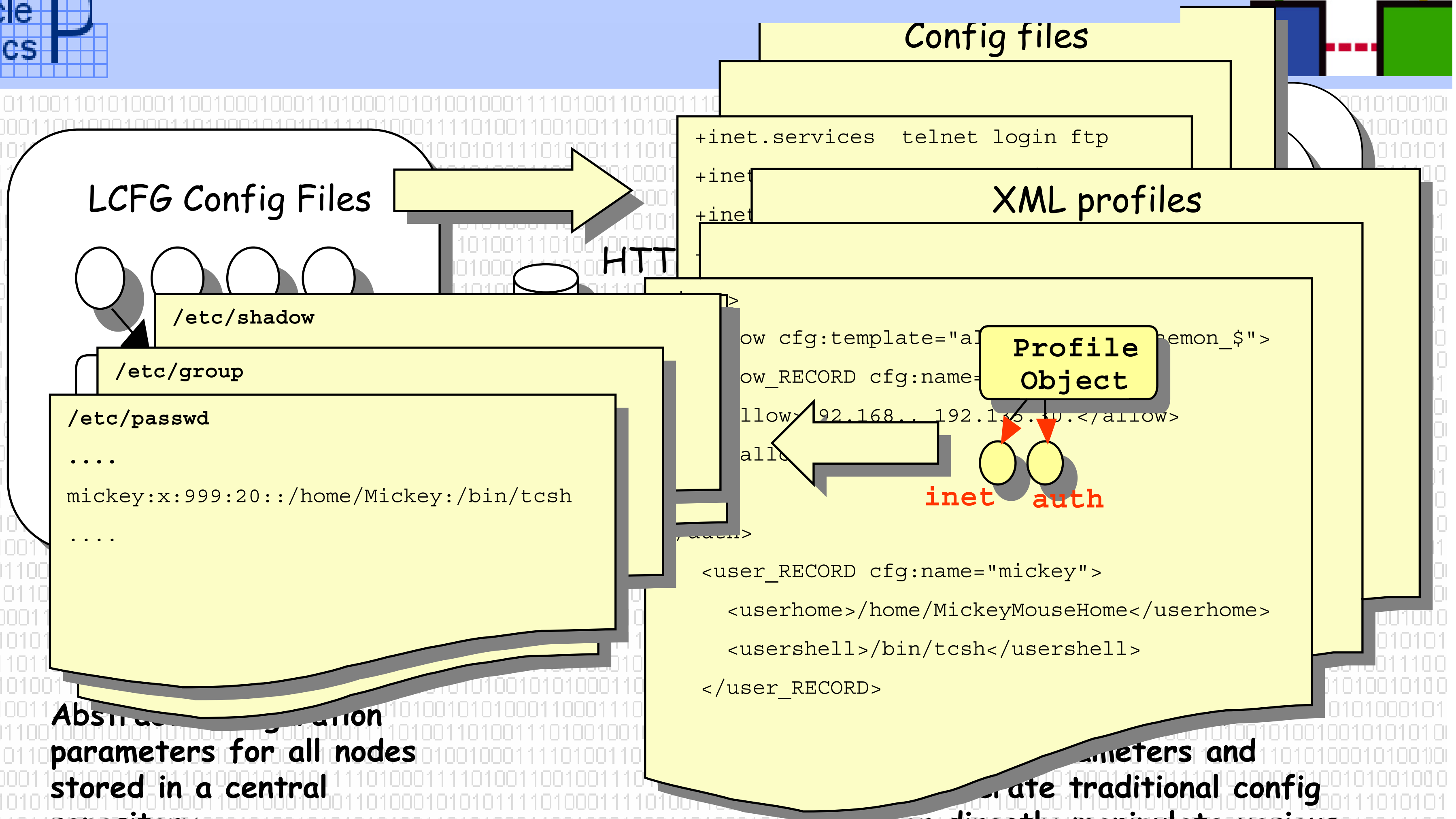
Network (Firewall) Requirements

- All services and ports required are outlined in a document. http://lcgdeploy.cvs.cern.ch/cgi-bin/lcgdeploy.cgi/lcg2/docs/lcg-port-table.pdf
- Most services are simple and firewall friendly, e.g. to access your site GIIS you must open port 2135.
- The exception to this simple case are services and clients which make use of the GLOBUS_TCP_PORT_RANGE.
- This range is set on a host that needs to accept connections back on random ports and makes the choice of port far less random.

LCFGng

- LCFGng is the local configuration tool, next generation. Developed in Edinburgh and then within EDG. It was by the end of EDG and into the start of LCG the only way being used to install an EDG/LCG grid site.
- LCG now has manual instructions that have allowed many sites to avoid using LCFG completely.
- Both LCFG and manual installations are very much supported by LCG.
- We will be using LCFG.
  - It is the fastest way to do things if you can dedicate hardware.

How LCFG Works in One Slide.



```
+inet.services  telnet login ftp
+inet
+inet
```

```
/etc/shadow
/etc/group
/etc/passwd
....
mickey:x:999:20::/home/Mickey:/bin/tcsh
....
```

```
>
ow cfg:template="al          emon_$">
ow_RECORD cfg:name=
llow>192.168.. 192.135.30.</allow>
all
/auth>
<user_RECORD cfg:name="mickey">
  <userhome>/home/MickeyMouseHome</userhome>
  <usershell>/bin/tcsh</usershell>
</user_RECORD>
```

Absi   ation parameters for all nodes stored in a central repository

ameters and rate traditional config files or directly manipulate various services

How LCFG installs Nodes(1)

- AN XML profile is constructed on the LCFGng to completely define the LCFG client node.
- LCG provides the sources files from which to generate these XML files for all nodes of the LCG service.
- LCFG client installation process.
  - Client boots a kernel from a floppy, a CD or via PXE.
  - Client mounts a root filesystem over NFS. (root=/dev/nfs)
  - This file system is called the LiveOS from now on and is also in the install guide.
  - The kernel is passed the option init=/etc/rc_install.

How LCFG installs Nodes(2)

- The rc_install scripts downloads the XML profile and installs the machines from the node specification.
- Disks are partitioned, formatted and mounted in /root.
- An rpm repository is mounted onto the booted LiveOS from the LCFGng server.
- A base set of RPMs are installed with "rpm --root /root".
- LCFG objects configure the network and other OS services. (At this point all installs are identical)
- Node reboots from its own harddisk and continues to configure itself to be a CE, SE, ….
- Node reboots once again to leave a running grid node.

Installing LCG with LCFGng(1)

- LCG provide LCFGng profiles to be checked out from CVS. These are copied by hand into the correct location.
- A utility updaterep is used to download all the required RPMs.
  - RPMs for the LCFG server itself.
  - RPMs for all the LCFGng clients, the grid nodes.
- A utility lcfgng_update_server.pl is used to install the required RPMS on the LCFGng server.
- A utility called lcfgng_installroot.pl installs the LiveOS.

Installing LCG with LCFGng(2)

- The LCG profiles must be customised with local values.
  - DNS servers.
  - IP Addresses.
- The profiles are compiled with do_mkprof.sh utility.
- Clients are now ready to be installed.

Next Part

- We next plan to finish the installation of the LCFGng which has been partly done for you.