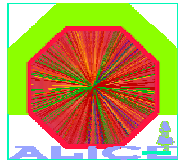


Detailed ALICE Goals & Plans for Service Challenge 3

P. Cerello (INFN – Torino)

LCG-SC Workshop

CERN

Jun 13th, 2005

ALICE 2005 Physics Data Challenge

- Number of events (preliminary, they will increase)
  - Simulation
    - 30,000 Pb-Pb (equivalent to: - 24,000 central)
    - 100,000 Pb-Pb (equivqlent to: - 60,000 central)
    - 100,000 p-p (equivqlent to: - 1,000 central)

      6-12h x 85,000 central

      -> 0.5 - 1 Mh
  - Reconstruction
    - Much quicker -> 15 Kh
  - Assume 1,000 CPUs:
    - Reconstruction: 1 day
    - Simulation: 0.5-1 Kh each (25-50 days occupancy)

ALICE 2005 Physics Data Challenge

- Physics Data Challenge
  - Until September 2005, simulate MC events on available resources
  - Register them in the ALICE File Catalogue and store them at CERN-CASTOR (for SC3)
- Coordinate with SC3 to run our Physics Data Challenge in the SC3 framework

ALICE & LCG Service Challenge 3

- Primary Goals:
  - Use of the deployed LCG SC3 infrastructure for the ALICE Data Challenge 2005
  - Test of data transfer and storage services (SC3)
  - Test of distributed reconstruction and calibration model (ALICE)
  - Integrate the use of LCG resources with other resources available to ALICE within one single VO interfaced to different “grids”
  - Analysis of reconstructed data

ALICE & LCG Service Challenge 3

- Secondary Goals:
  - Interactive Analysis of reconstructed data with PROOF

ALICE & LCG Service Challenge 3

- How do we define:
  - Success: meet primary and secondary goals
  - Partial Success: meet primary goals only
  - Failure: miss primary goals (any of them)
  - Metrics:
    - Let's first describe our plans and requirements

ALICE & LCG Service Challenge 3

- Use case 1: RECONSTRUCTION
  - (Get "RAW" events stored at T0 from our Catalogue)
  - First Reconstruct pass at T0
  - Ship from T0 to T1's (goal: 500 MB/S out of T0)
  - Reconstruct at T1 with calibration data
  - Store/Catalogue the output

ALICE & LCG Service Challenge 3

- Use Case 2: SIMULATION
  - Simulate events at T2's
  - Transfer Data to supporting T1's

ALICE & LCG SC3: Possible Data Flows

- A) Reconstruction: input at T0, run at T0 + push to T1s
  - Central Pb-Pb events, 300 CPUs at T0
  - 1 Job = 1 Input event
  - Input: 0.8 GB on T0 SE
  - Output: 22 MB on T0 SE
  - Job duration: 10 min -> 1800 Jobs/h

    -> **1.4 TB/h (400 MB/s)** from T0 -> T1s
- B) Reconstruction: input at T0, push to T1s + run at T1s
  - Central Pb-Pb events, 600 CPUs at T1s
  - 1 Job = 1 Input event
  - Input: 0.8 GB on T0 SE
  - Output: 22 MB on T1 SE
  - Job duration: 10 min -> 3600 Jobs/h

    -> **2.8 TB/h (800 MB/s)** from T0 -> T1s

ALICE & LCG SC3: Possible Data Flows

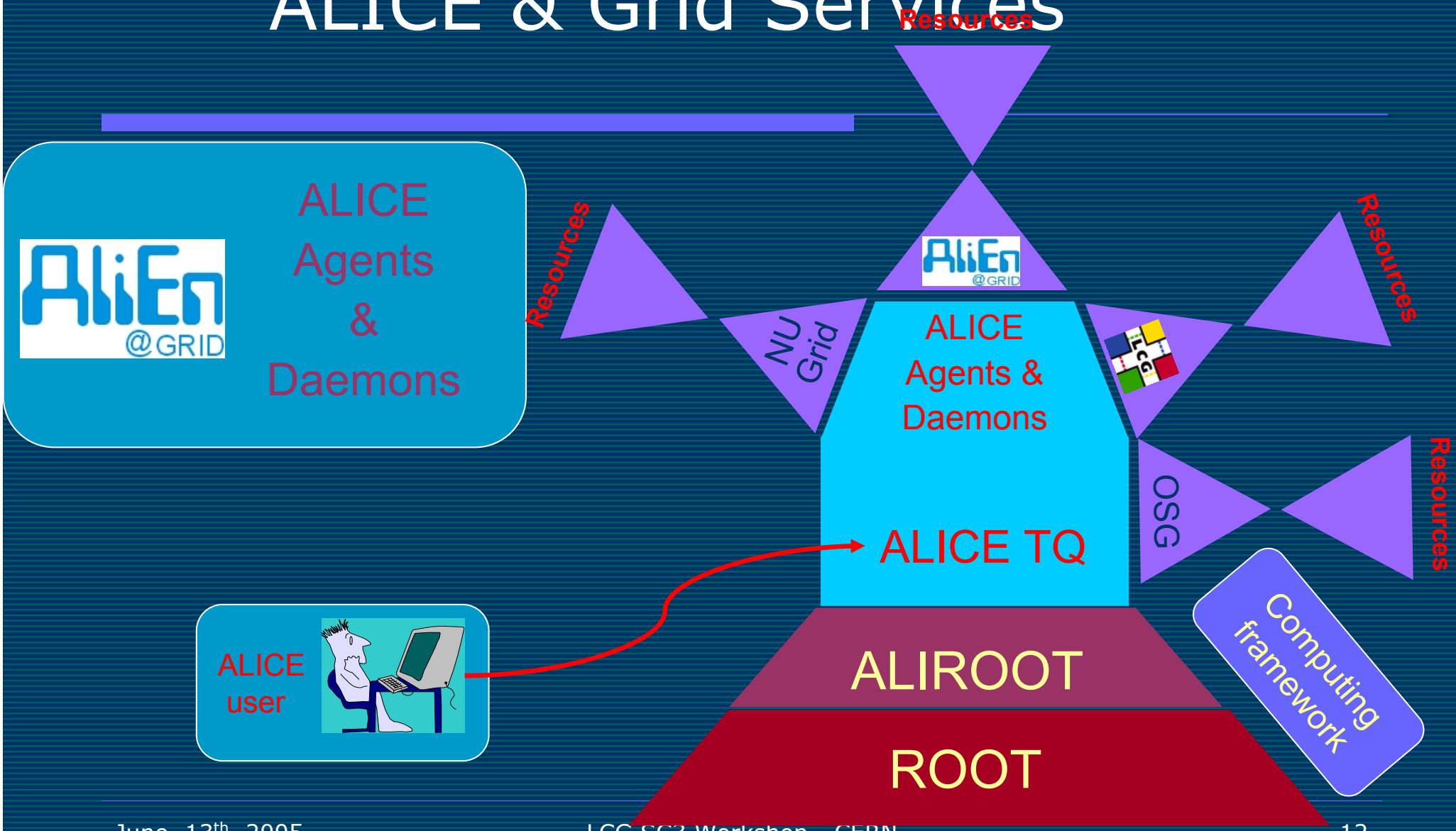
- Simulation
  - Assume 1000 CPUs availability at T2s and central Pb-Pb event
  - 1 Job = 1 Event
  - Input: few KB of configuration files
  - Output: 0.8 GB on T1 SE
  - Job duration: 6 h -> 4000 Jobs/day
    -> **3.2 TB/day (40 MB/s)** from T2s -> T1s
- Remark
  - in case T2 resources were not sufficient, we could obviously simulate at T1 as well

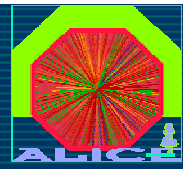
ALICE and LCG Service Challenges

- In other words:
  - Mimic our data (raw-like + simulated) flow
  - test the reconstruction and the prototype calibration framework
  - measure the performance of the SC3 services/components:
    - Data transfer efficiency
    - Storage Element efficiency

  - Start in september 2005
  - As new sites keep coming in, increase the scale of the exercise
  - As new middleware comes in, test it & add more functionality

ALICE & Grid Services





Baseline services

- Storage management services
  - Based on SRM as the interface
- Basic transfer services
  - gridFTP, srmCopy
- Reliable file transfer service
- Grid catalogue services
- Catalogue and data management tools
- Database services
  - Required at Tier1,2
- Compute Resource Services
- Workload management
- VO management services
  - Clear need for VOMS: roles, groups, subgroups
- POSIX-like I/O service
  - local files, and include links to catalogues
- Grid monitoring tools and services
  - Focussed on job monitoring
- VO agent framework
- Applications software installation service
- Reliable messaging service
- Information system

FTS summary – cont.

...

- ALICE:
  - See fts layer as service that underlies data placement. Have used FTD (with aiod as protocol) for this in DC04.
  - Expect gLite FTS to be tested with other data management service in SC3 – ALICE will participate.
  - Expect implementation to allow for experiment-specific choices of higher level components like file catalogues

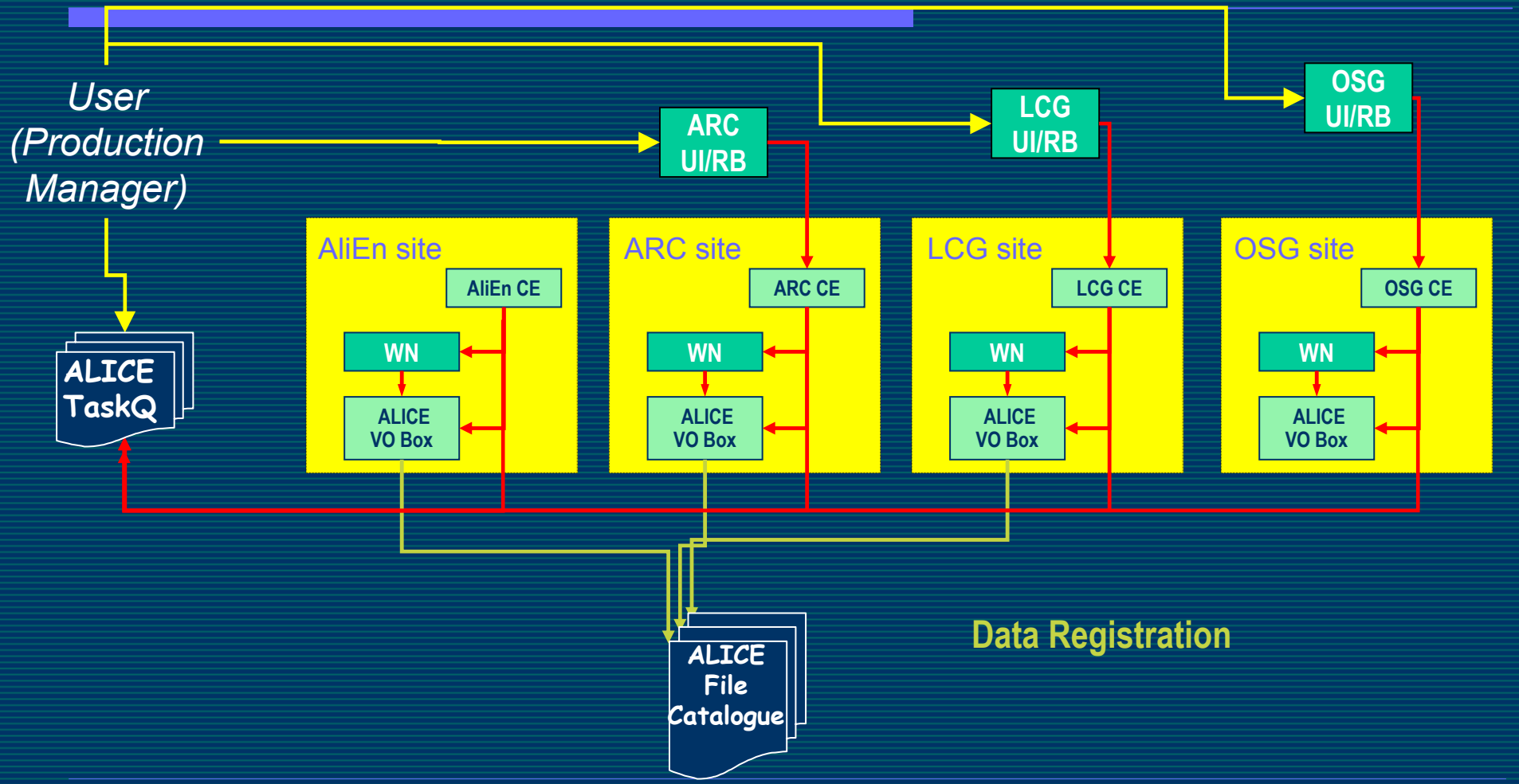
Summary of catalogue needs

- ALICE:
  - Central (Alien) file catalogue.
    - + local LFN - PFN mapping
  - No requirement for replication
  - will use the Alien FC
  - testing of LCG-LFC as a possible alternative to AliEn-LFC
  - if time & manpower available, have a look at Fireman

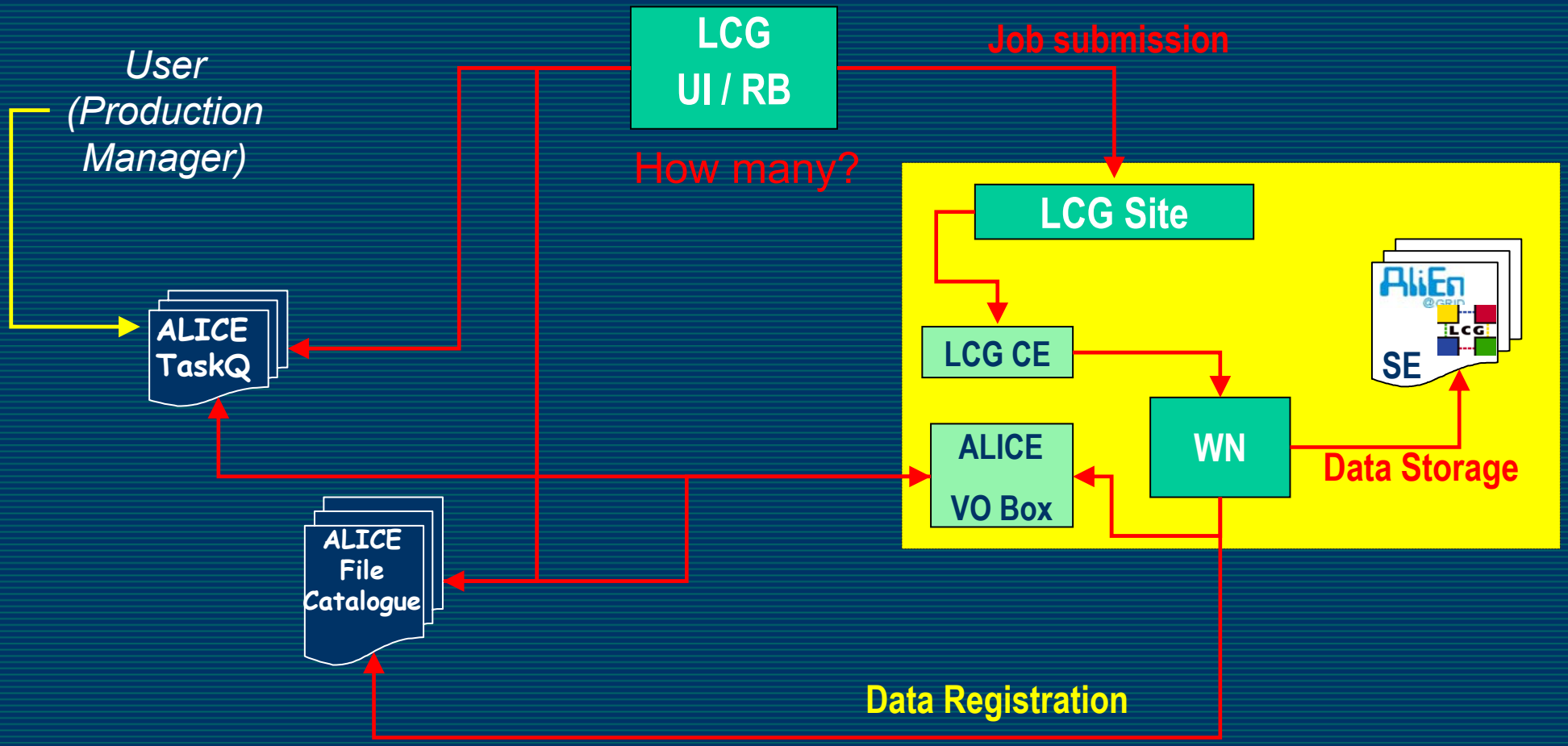
VO "Agents & Daemons"

- VO-specific services/agents
  - Appeared in the discussions of fts, catalogs, etc.
  - …. – all experiments need the ability to run "long-lived agents" on a site
  - At Tier 1 and at Tier 2
  - → how do they get machines for this, who runs it, can we make a generic service framework

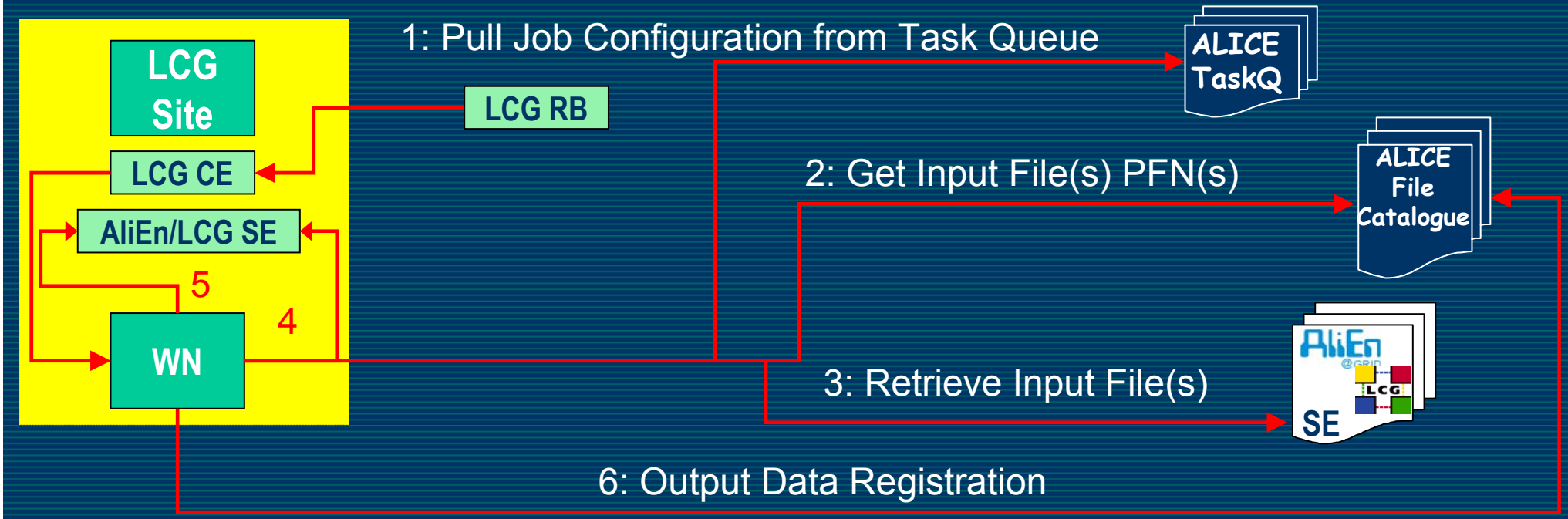
ALICE SC3 layout



ALICE SC3 layout



ALICE SC layout - seen from LCG-WN



ALICE & LCG Service Challenge 3

- Goal:
  - ALICE would like to implement a single VO operating across all flavours of Grid, with the possibility to specify roles and groups via VOMS

ALICE & LCG Service Challenge 3

- Requirements:
  - ALICE requires one VO-managed node at each site on which to deploy and manage long-lived agents.
  - ALICE would take the responsibility to deploy and support the agents/services on the dedicated node at each site

ALICE & LCG Service Challenge 3

- Expectations - Storage:
  - Standard interface to SE's
    - it is envisaged that this will be provided by SRM (functionality as specified for V2) and that the SE will present a single, SURL namespace.

ALICE & LCG Service Challenge 3

- Expectations - Data Management:
  - ALICE will have a *distributed* LFN-PFN mapping & a central file catalogue containing the file LFN, GUID, Storage Index (SI) and corresponding metadata.
  - In addition, we expect that there will be a local site service that will map the LFN to a PFN
    - If unavailable, we can provide it as AliEn service

ALICE & LCG Service Challenge 3

- Expectations - Data Replication/Transfer:
  - For data replication, ALICE expects a reliable file transfer service(s) (FTS) to be provided and deployed by the LCG project.

ALICE & LCG Service Challenge 3

- Expectations - Workload Management:
  - Job submission to LCG will go through the RB service
  - We envisage different ways to use the RB service:
    - Jobs will retrieve their configuration information from the ALICE Task Queue through the ClusterMonitor service as soon as they start on the WN
    - The number of needed RBs is undefined (yet) as it depends on their speed and on the ratio:

      average job duration/submission time
      - Option A) "infinite" speed: 1 RB
      - Option B) "slow" RB: 1 RB/CE in pull mode

ALICE & LCG Service Challenge 3

- Expectations - Compute Element:
  - Common interface to the local CE which hides the particular implementation of the deployed scheduling system.
  - Outbound network connection from the Worker Nodes (or an appropriate tunnelling mechanism)

ALICE & LCG Service Challenge 3

- VO node general requirements:
  - at least one normal user account (no super-user privileges) with access via ssh
  - optimal configuration: two accounts, belonging to the same group
  - account (home) directory shared among the WNs, min 5 GB disk space
  - outbound connectivity
  - inbound connectivity from CERN on one fixed network port
  - inbound connectivity from World on two fixed network ports
  - local tactical data buffer (local disk, LCG deployed Disk Pool Manager, NFS mounted disk) for intermediate input and output data storage of jobs. The buffer size is at least number of jobs slots on the site * 3GB. This buffer is not necessary if xrootd is running on the site storage element.
  - Linux kernel 2.4 or higher, any Linux flavour on i386, ia64 or Opteron
  - hardware: min. PIII 2GHz, 1024 MB RAM

ALICE & LCG Service Challenge 3

- AliEn and monitoring agents and services running on the VO node:
  - Storage Element Service (SES)
    – interface to local storage (via SRM or directly)
  - File Transfer Daemon (FTD)
    – scheduled file transfers agent (possibly using FTS implementation)
  - xrootd – application file access
  - Cluster Monitor (CM) – local queue monitoring
  - MonALISA – general monitoring agent
  - PackMan (PM) – application software distribution and management
  - Computing Element (CE)

ALICE & LCG Service Challenge 3

- What would we need for SC3?
  - AliRoot/ROOT etc. deployment on SC3 sites - ALICE
  - AliEn Top Level Services(ongoing) - ALICE
  - UI(s) for submission to LCG/SC3 - ALICE/LCG
  - WMS (n RBs) + CE/SE Services on SC3 - LCG
  - SC3 resources for ALICE (Computing/Storage) - LCG
  - Appropriate JDL files for the different tasks - ALICE

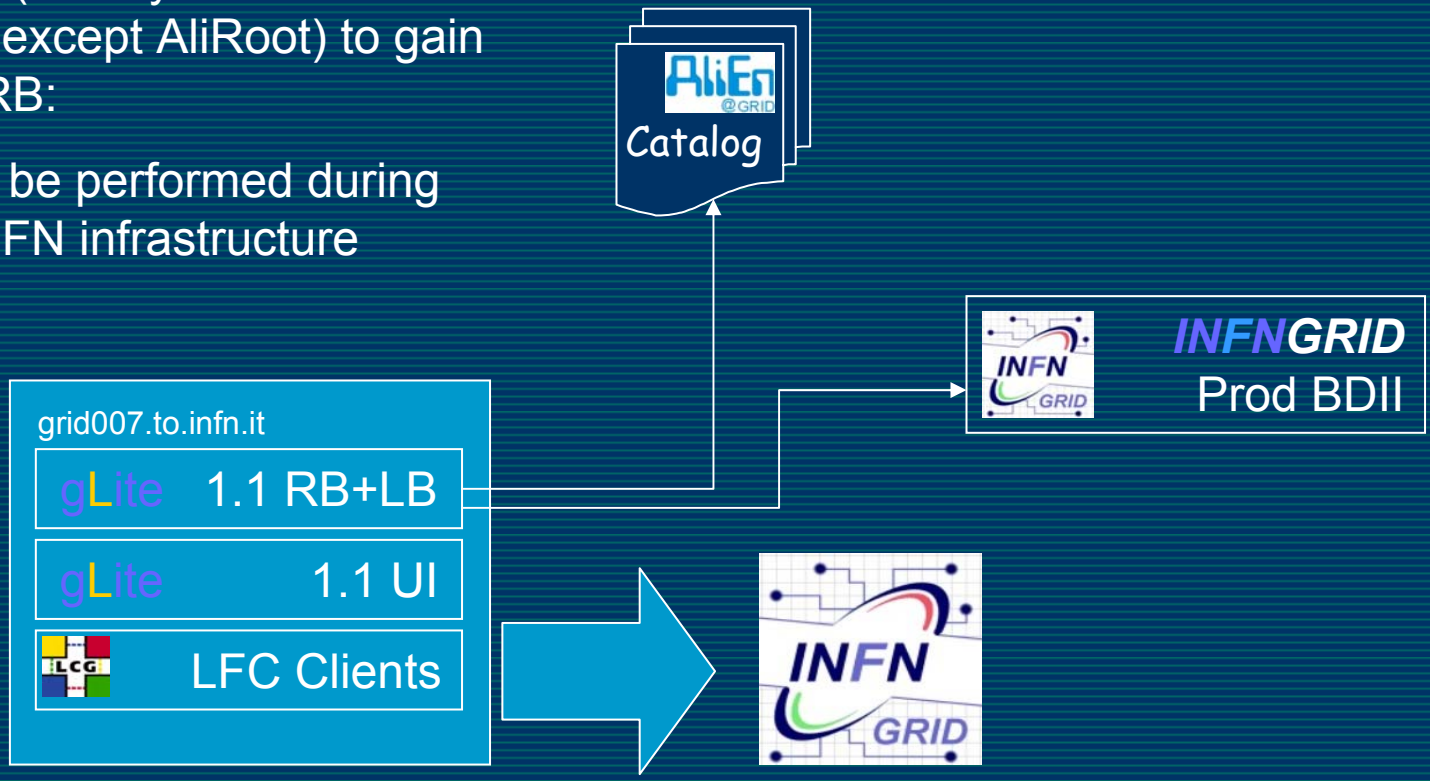
AliEn - gLite integration

- Set up of a test gLite RB & CE in Torino
  - test job submission, interaction with the ALICE file catalogue and, gradually, other pieces of the framework
- Tests of storage and data management components in Bari
  - dCache+SRM, FTS
  - To be integrated with the Torino setup to build a full testbed

AliEn - gLite integration

Small MC production (initially no ALICE-specific components except AliRoot) to gain confidence with the RB:

20k events, ~6TB, to be performed during next weeks on the INFN infrastructure



AliEn - gLite integration

- **Problems:**
  - The gLite UI command does not interact correctly with the VOMS.
    - known problem, fixed but the fix did not get through to the release (not even 1.1)
  - Submission to the gLite RB fails with certificates mapped to a SGM (software manager) account (Savannah bug #8616)
- **gLite RB interacts correctly with LCG 2.4.0 CEs but:**
  - 650 jobs submitted to INFNGRID just after the upgrade to 2.4.0 showed the same teething problems of last year, e.g.:
    - Hanging NFSs make software area inaccessible (this is a nasty one – remember the "Black Hole Effect"!)
    - Problems with environment configuration on WNs
  - Success rate: 423 (65%), not uniform on different sites
  - The support responsiveness definitely improved
    - Problem generally solved within an hour of submitting the ticket

AliEn - gLite integration: ongoing gLite production

Jobs sent, via gLite RB (grid007.to.infn.it), to LCG 2.4.0 on INFN-Grid CEs: 1000

Scheduled:	26
Running:	11
Completed:	661 (68%)
Aborted:	72 (8%)
Error:	230 (24%)
Location:	33 SNS, 70 Le, 35 unknown, 43 CNAF, 25 Pd: 206
AliROOT crash:	28
NFS crash:	143
WN disk space < 4GB:	55
Other (not understood yet)	4

RB problem: 100 (1 bunch) job destination lost

ALICE & LCG Service Challenge 3

- Sites/Resources/Planning: see
  - http://lcg.web.cern.ch/LCG/PEB/Planning/deployment/Grid%20Deployment%20Schedule.htm

  for latest schedule…
- Sites:
  - all ALICE T1s
  - T2s (at the beginning): GSI, Torino, RDIG,…
  - As more T2s join SC3, same approach as for T1s…

ALICE & LCG Service Challenge 3

- PDC2005 is also aimed at evaluating the performance of SC3 services in realistic conditions
- We need support from LCG and we are putting some of our manpower in for SC3
- We are willing to start as soon as possible so as to be ready for…
- …PDC2006 & SC4, where the MW components of the "LCG Production Service" will be tested