Databases for the CERN LHC: Techniques and Lessons Learned

2nd XLDB Workshop, SLAC, 29-30 Sept 2008

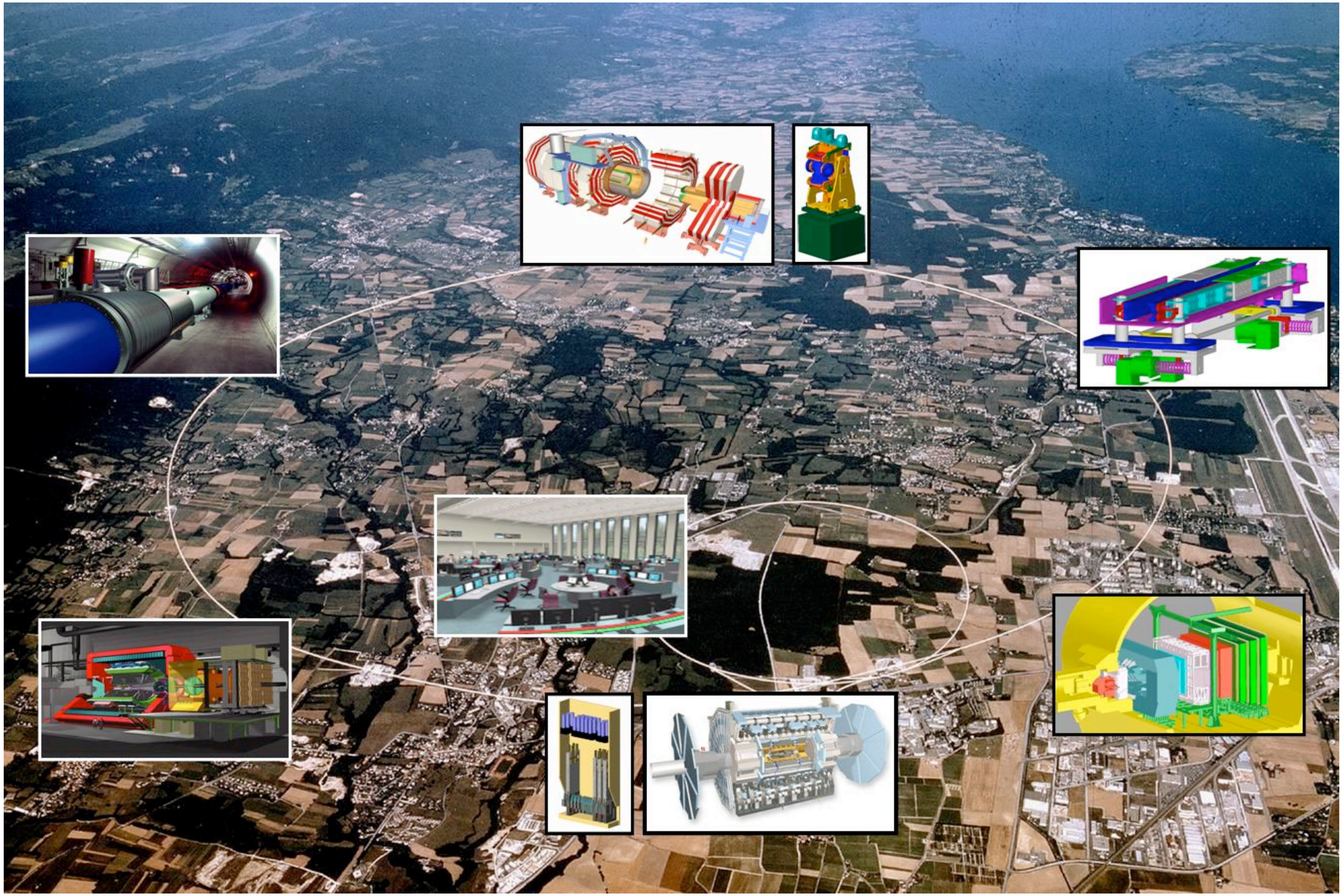
Maria Girone, CERN - IT

Outline

- Databases in the LHC Computing Grid
- Technologies behind
- 10 Lessons Learned
  - Lessons Learned in deploying and operating the distributed WLCG service(s) may also be relevant, but not covered by this talk

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

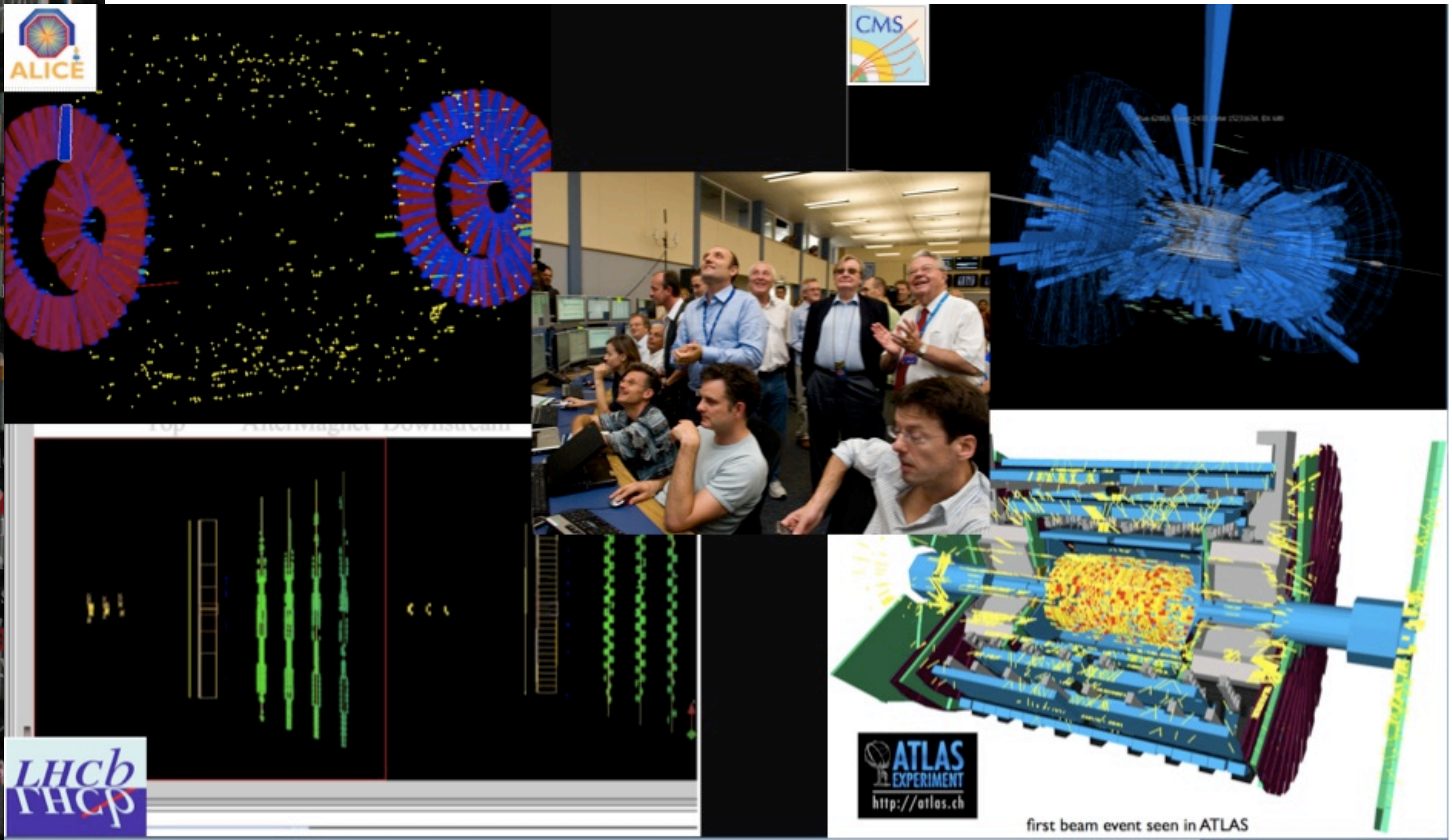
LHC accelerator and exeriments



DM

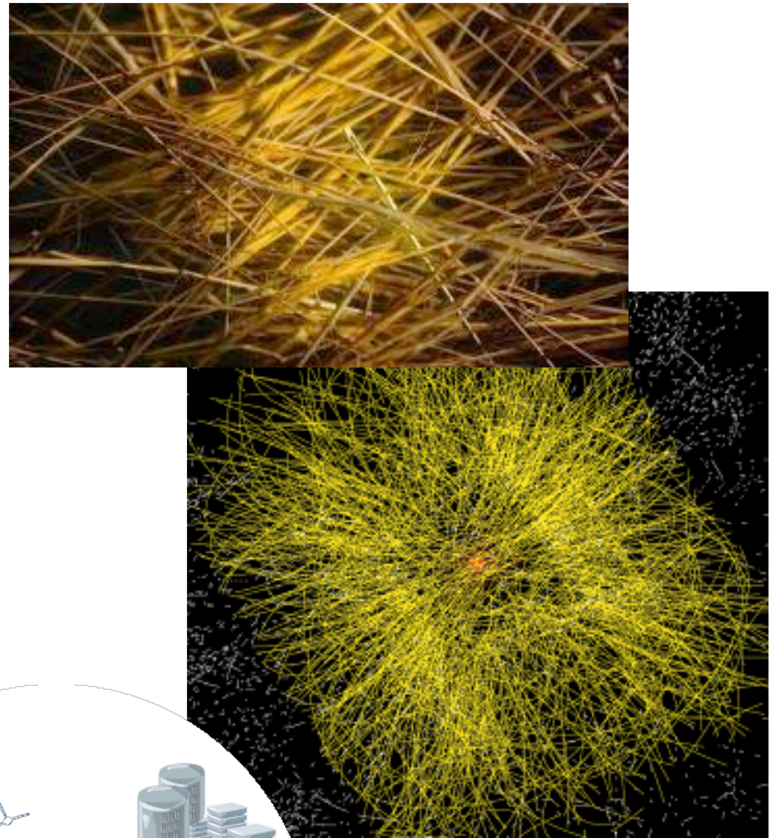
LHC first beams: 10th Sept 2008

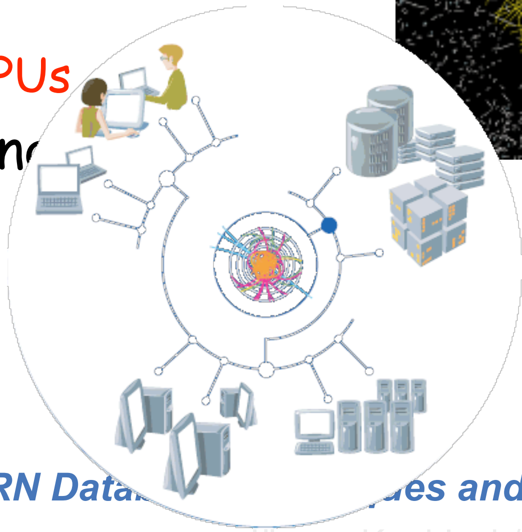
CERN IT Department



The LHC Computing Challenge

- Data volume
  - High rate x large number of channels x 4 experiments
  - ➔15 PetaBytes of new data each year stored
  - ➔ Much more data discarded during multi-level filtering before storage
- Compute power
  - Event complexity x Nb. events x thousands users
  - ➔ 100 k of today's fastest CPUs
- Worldwide analysis & funding
  - Computing funding locally in major regions & countries
  - Efficient analysis everywhere
  - ➔ GRID technology





Databases and LHC

Relational databases are used by a wide-range of mission-critical applications that are part of the Grid infrastructure:

- middleware and storage related services (CASTOR, DPM, FTS, LFC, SRM)
- key infrastructure and operations services (dashboards, SAM, GridView, …)
- LHC experiments' conditions, geometry, alignment, calibration, meta-data book-keeping..
  (COOL, PVSS, …)

Connected to 10 Tier-1 sites for synchronized Databases. Sharing policies and procedures

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Key Technologies Behind

CERN IT Department

- Oracle Real Application Clusters (RAC) with Automatic Storage Management (ASM) : database engine
- Oracle Streams: for sharing information between databases
- Oracle Data Guard: for additional protection against failures (human errors, disaster recoveries,)

10 Lessons Learned

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

1. Communication among DBA teams

- Communication with a very large end-users community and with 11 DBA teams from large centers (Tier0, 10 Tier1) is a challenge
  - Emphasis on homogeneity
  - Sharing policies & procedures
  - Regular meetings and workshops
- Different time zones may delay coordination and problem resolution

2. Patching and Upgrades

CERN IT Department

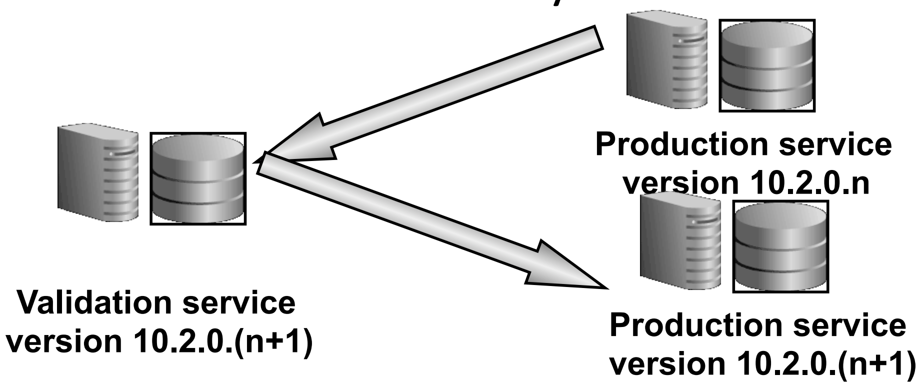
- Databases are used by a world-wide community: arranging for scheduled interventions (s/w and h/w upgrades) requires quite some effort
- Rolling upgrades and use of stand-by databases help somewhat
  - **0.04% services unavailability = 3.5 hours/year**
  - **0.22% server unavailability = 19 hours/year (Patch deployment, hardware)**
- Interventions typically shorten than the time it takes to arrange for them

3. Testing and Validation

CERN IT Department

- Introduced strict policies for hardware, DB versions, applications testing
  - Application release cycle

  - Database software release cycle



- Proven key to smooth production

4. Monitoring

- Comprehensive monitoring hard to achieve but essential for smooth operation
- Out of the box ORACLE tools (such as Grid Control) do not fully cover:
  - Streams
  - Storage
  - End-users database availability and performance

  > In-house tools developed and fed back to Oracle development
- Coherent status board of distributed database services for all the 11 Tiers still under development

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

5. Backup & Recovery

CERN IT Department

- On-tape backups: fundamental for protecting data, but recoveries run at ~40MB/s (70 hours for LHC DB size of 10TB)
  - Very painful for an experiment in data-taking
- Put in place on-disk image copies of the DBs: able to recover to any point in time of the last 48 hours activities
  - Recovery time independent of DB size
- Use of Oracle Data Guard (physical stand -by) gives additional protection
  - Disasters, multi-point failures data corruption

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

6. Streams Replication

- Connected to 10 Tier1 sites for synchronized databases:
  - Operations involve source (Tier0) and destination (Tier1) databases
    - Limited Streams knowledge at Tier1 sites
    - Based on Tier0 expertise

- Several bugs affecting Streams
  - Problem debugging takes time
  - Fixes are not always produced in time
  - Workarounds cause more manual work

- Unique design due to CERN Stream's setup particularities (topology and performance needs)

7. Execution Plans and Statistics

CERN IT Department

- Execution plans not stable in time
  - Performance differences often of a order magnitude
- May change with s/w upgrades or with more data
- Use of explicit hints can only be a short term workaround
- For some applications the main DBA concern is to stabilize the execution plan

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

8. Scalability within a cluster

- Oracle RAC well proven with our – mostly read-only – applications
  - I/O with ASM scales well adding more disk spindles

- But, some key write applications need to be optimized to scale
  - Important application changes maybe required
  - Move to multi-core hardware can help
  - We had a major upgrade to 8-core servers before the LHC start-up

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

9. Resource allocation

- Assigning resources to users is done
  - Without clear resource plan from the community
  - With a long hardware acquisition cycle (8-9 months)
- Difficult to provide and maintain a service due to “last minute” changes
  - Often requires re-prioritization within the available hardware budget
  - Spare hardware can help somewhat

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

10. Resource Throttling

- Users workload driven by external factors (start-up, conferences, re -processing, discoveries?)
- Databases can become unstable under high-load
- Service throttling is key and implemented via Oracle Services for each large application (connection, CPU, memory)

Conclusions

- Recognizing the importance DB services to the experiments' activities, we have focused on robustness, scalability and flexibility
- Testing and validation – hardware, DB versions, applications – proven key to smooth production
  - close cooperation between application developers and database administrators
- Extra complexity comes from distributed operations in the LHC Computing Grid
- Several data-challenges but data-taking starts only now

More Details

- Questions?
- References:
  - CERN Physics Databases wiki:
    - General advice
    - Connection management
    - http://cern.ch/phydb/wiki
    - Support: phydb.support@cern.ch
  - LCG 3D wiki
    - interventions, performance pages
      - http://lcg3d.cern.ch