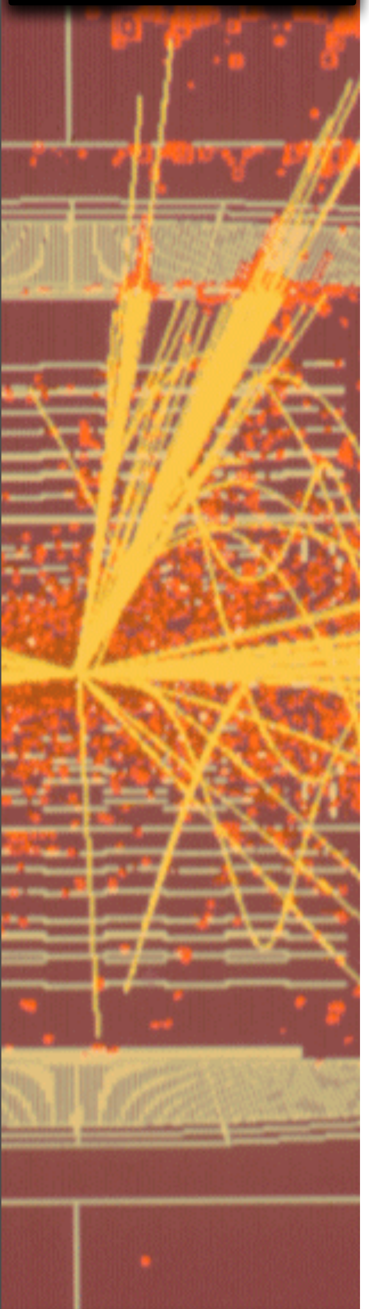


Databases for the Large Hadron Collider at CERN

Dirk Duellmann, CERN IT

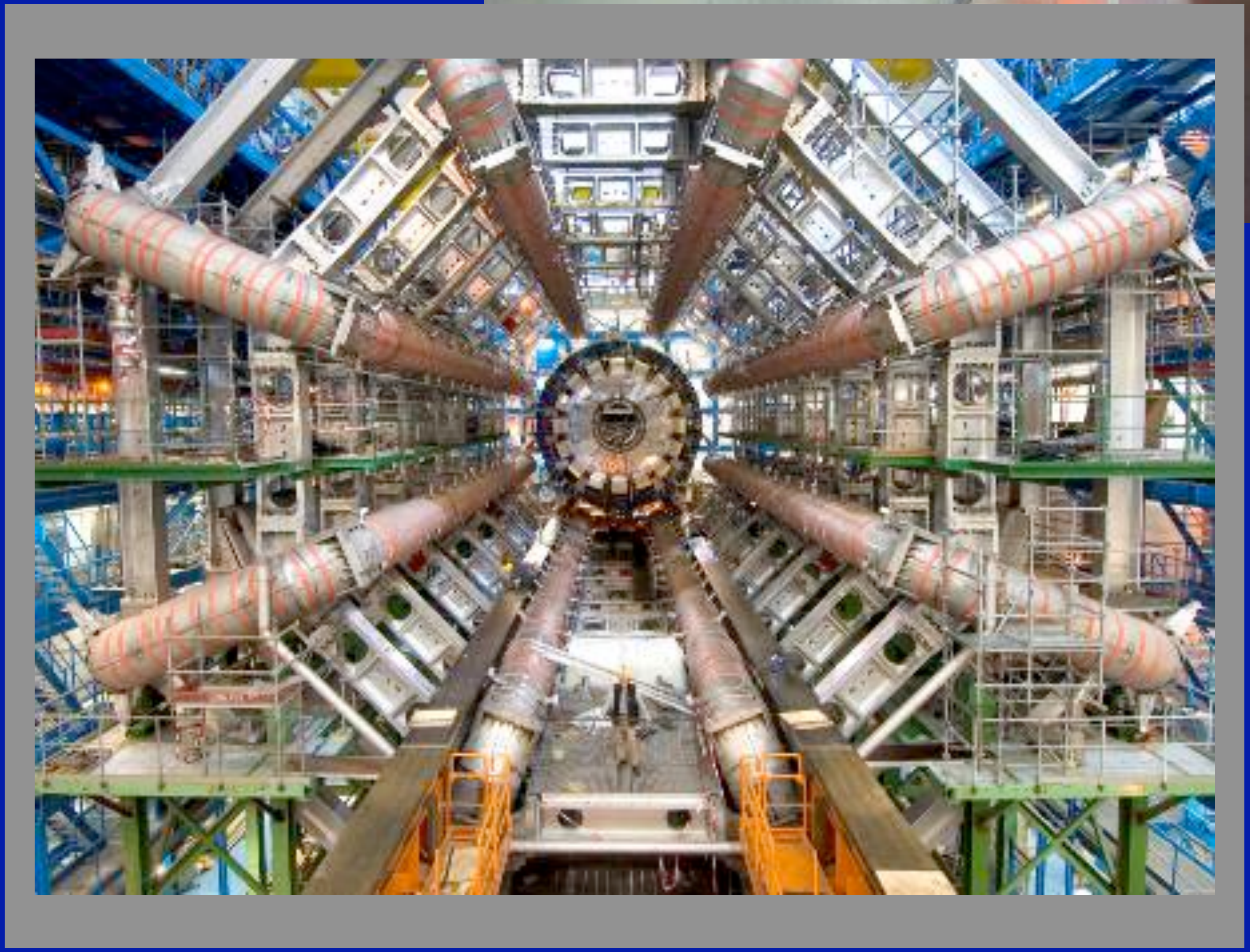
XLDB Workshop @ SLAC

25. October 2007

Outline

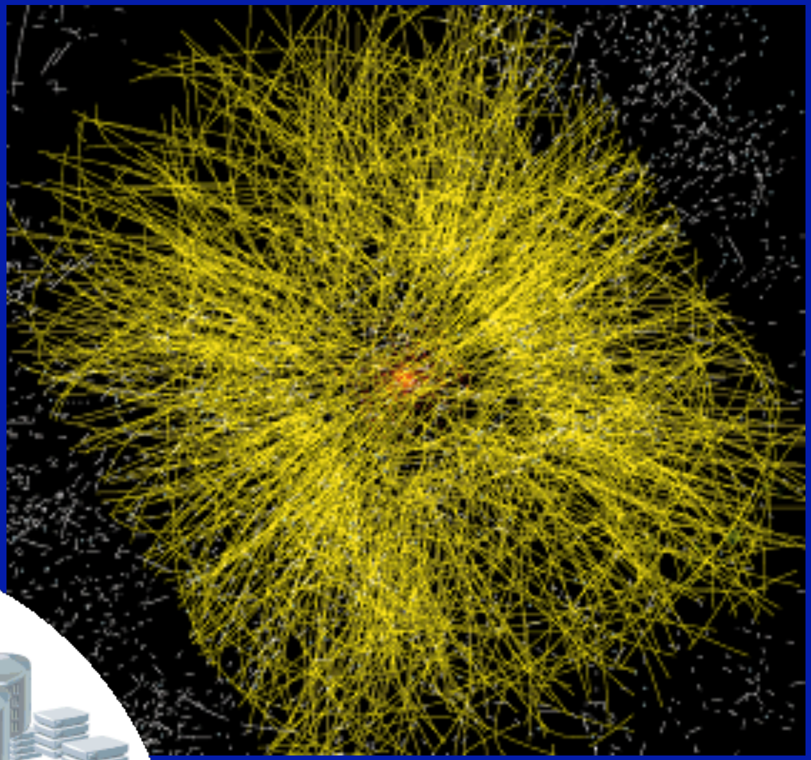
- CERN and LHC
  - Intro - project goals and schedule
- Role of databases in LHC data management
  - Key applications and use cases
- Physics software and databases
  - Integration with physics code & development model
- Database technologies and deployment models
  - Scalability, availability, replication
- Remaining questions / issues / concerns
  - Areas for future improvement
- Conclusions

LHC gets ready …



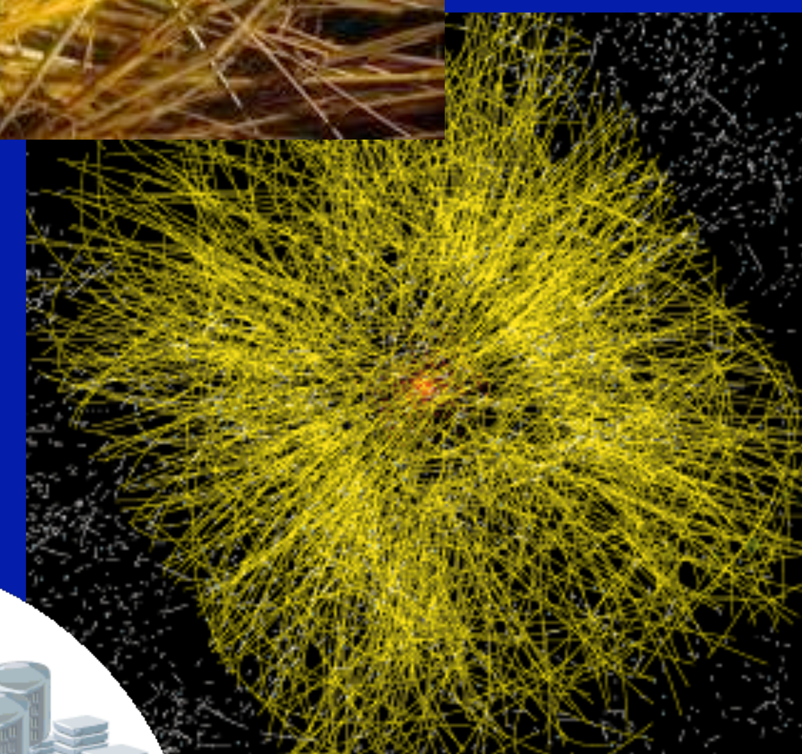
The LHC Computing Challenge

- **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**



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**

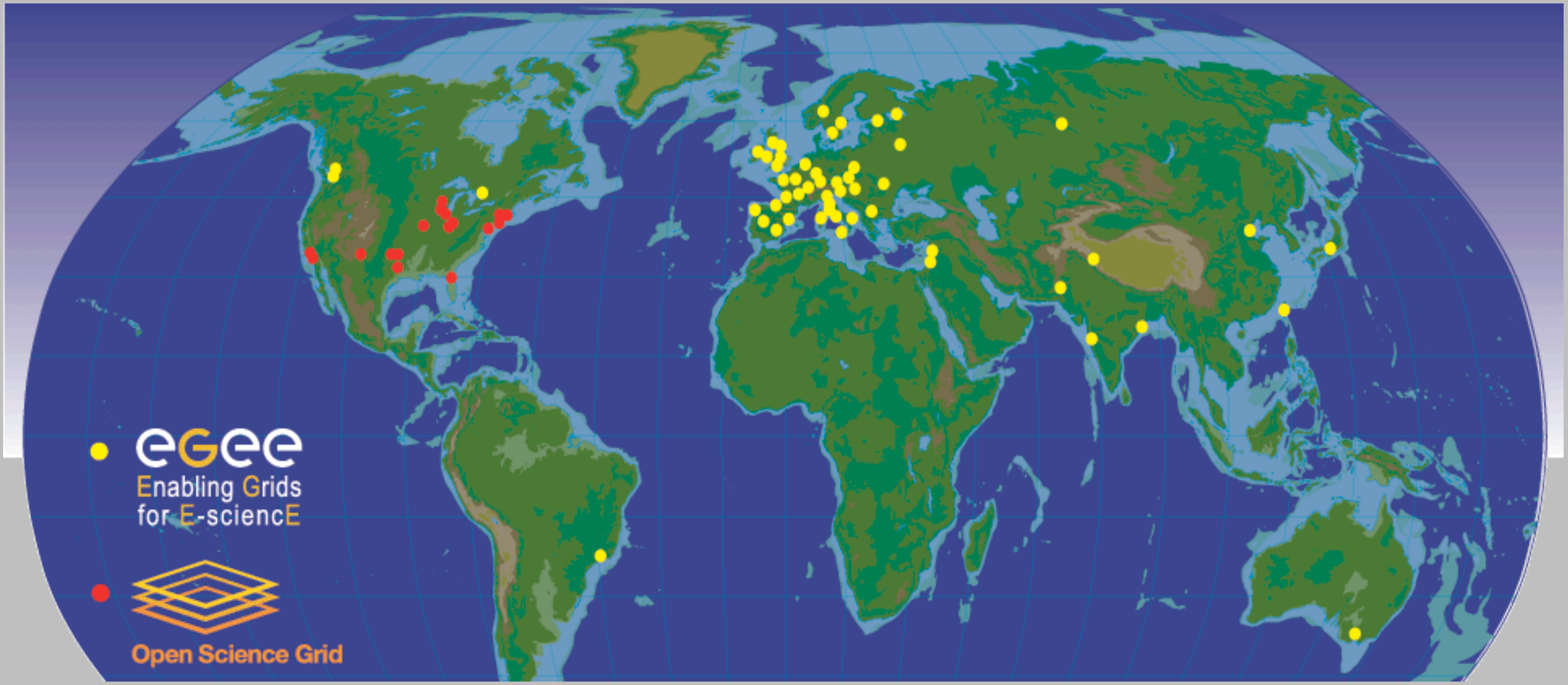


WLCG Collaboration

- **The Collaboration**
  - 4 LHC experiments
  - ~200 computing centres
  - 12 large centres (Tier-0, Tier-1)
  - 38 *federations* of smaller “Tier-2” centres
  - Growing to ~40 countries
  - Grids: EGEE, OSG, Nordugrid
- **Technical Design Reports**
  - WLCG, 4 Experiments: June 2005
- **Memorandum of Understanding**
  - Agreed in October 2005
- **Resources**
  - 5-year forward look

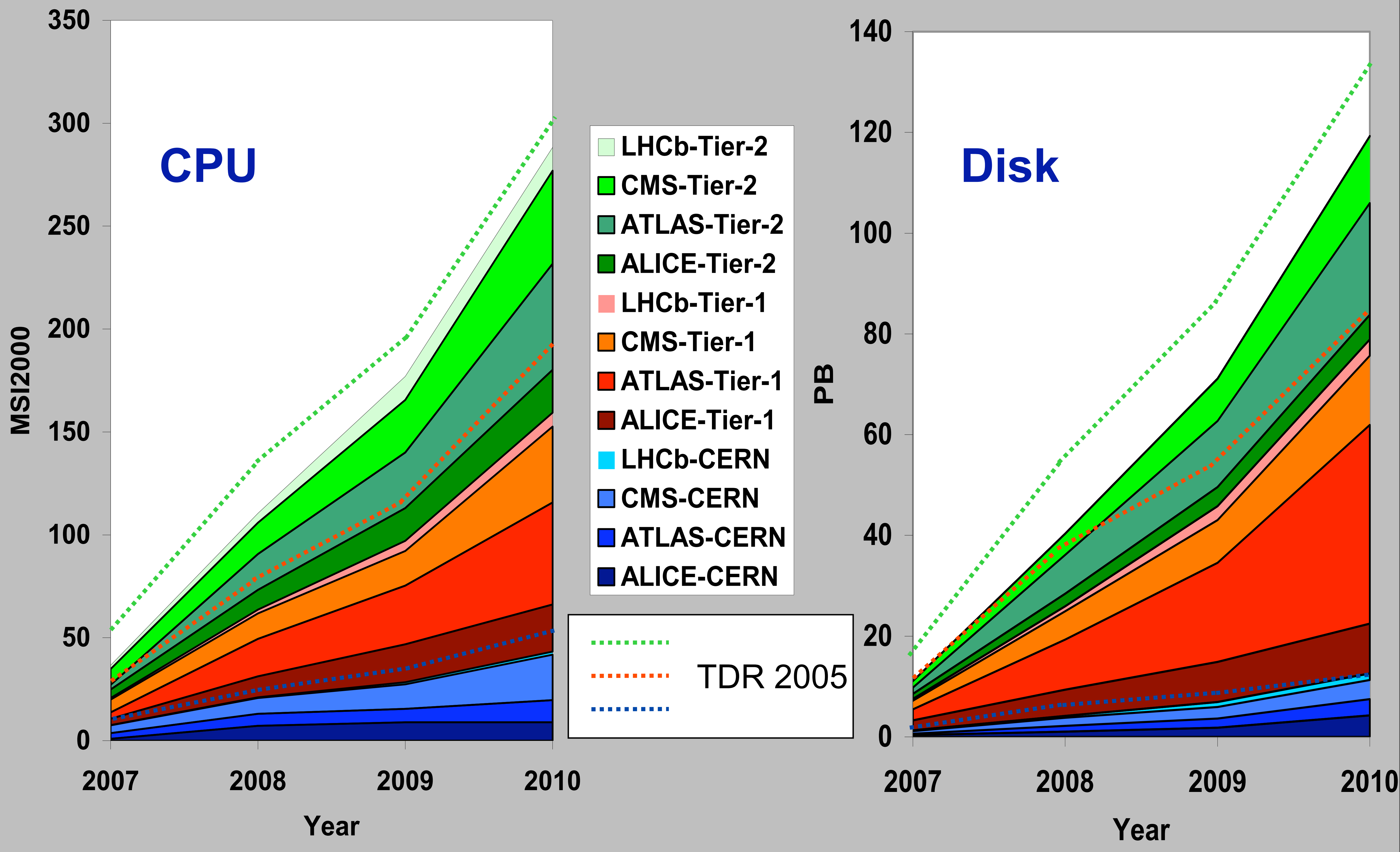
Centers around the world form a Supercomputer

- The EGEE and OSG projects are the basis of the Worldwide LHC Computing Grid Project WLCG



Inter-operation between Grids is working!

CPU & Disk Requirements 2006



HEP's interpretation of "Divide and Conquer"

- Physics community uses a well established selection & storage cascade
  - RAW
    - pure detector measurements, simple structure
  - AOD
    - Analysis Object Data, complex objects describing full reconstruction detail
  - ESD
    - Event Summary Data, combined high level description across several detector components
  - TAG
    - Event selection tag, highly condensed and abstracted key features of a reconstructed collision
- Each step includes
  - further filtering (often by orders of magnitude)
  - data reclustering to suite next processing step

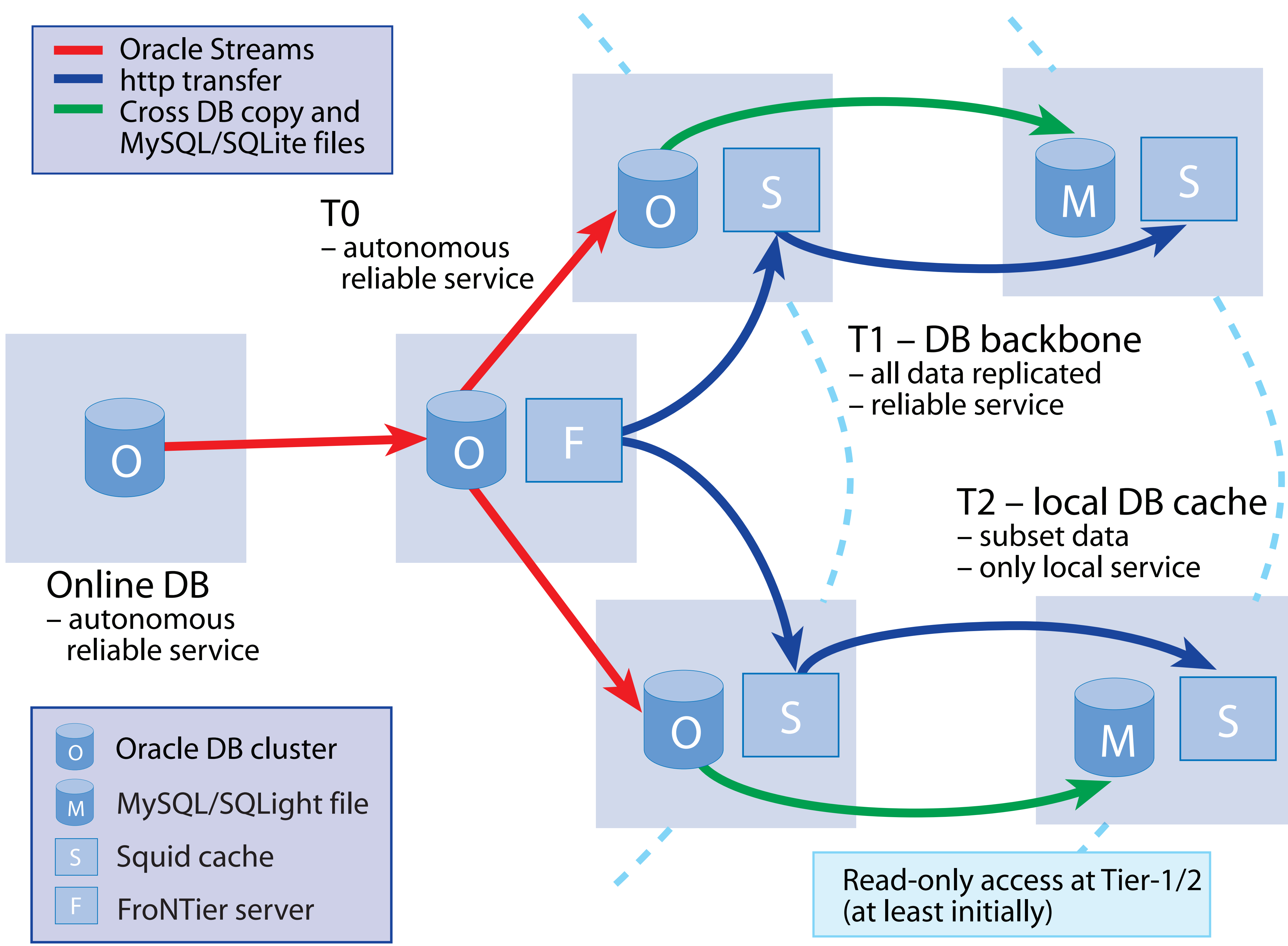
Databases? Sure - but which ones?

- What happened so far from the high energy physics point of view:
- ~1995 - Object Databases
  - good match with complex physics data models and programming languages (OO)
- ~2000 - OODB stagnating market
  - In-house OODB or RDBMS or ORDBMS?
  - Eg pure RDBMS
    - difficulties matching complex data models
    - cost of consistency, which is not always required
- Since 2001 - RDBMS + files
  - Idea of consistent storage of all data in databases was dropped
  - Hybrid model
    - Bulk data in files (largely read-only)
    - Only key meta-data in RDBMS

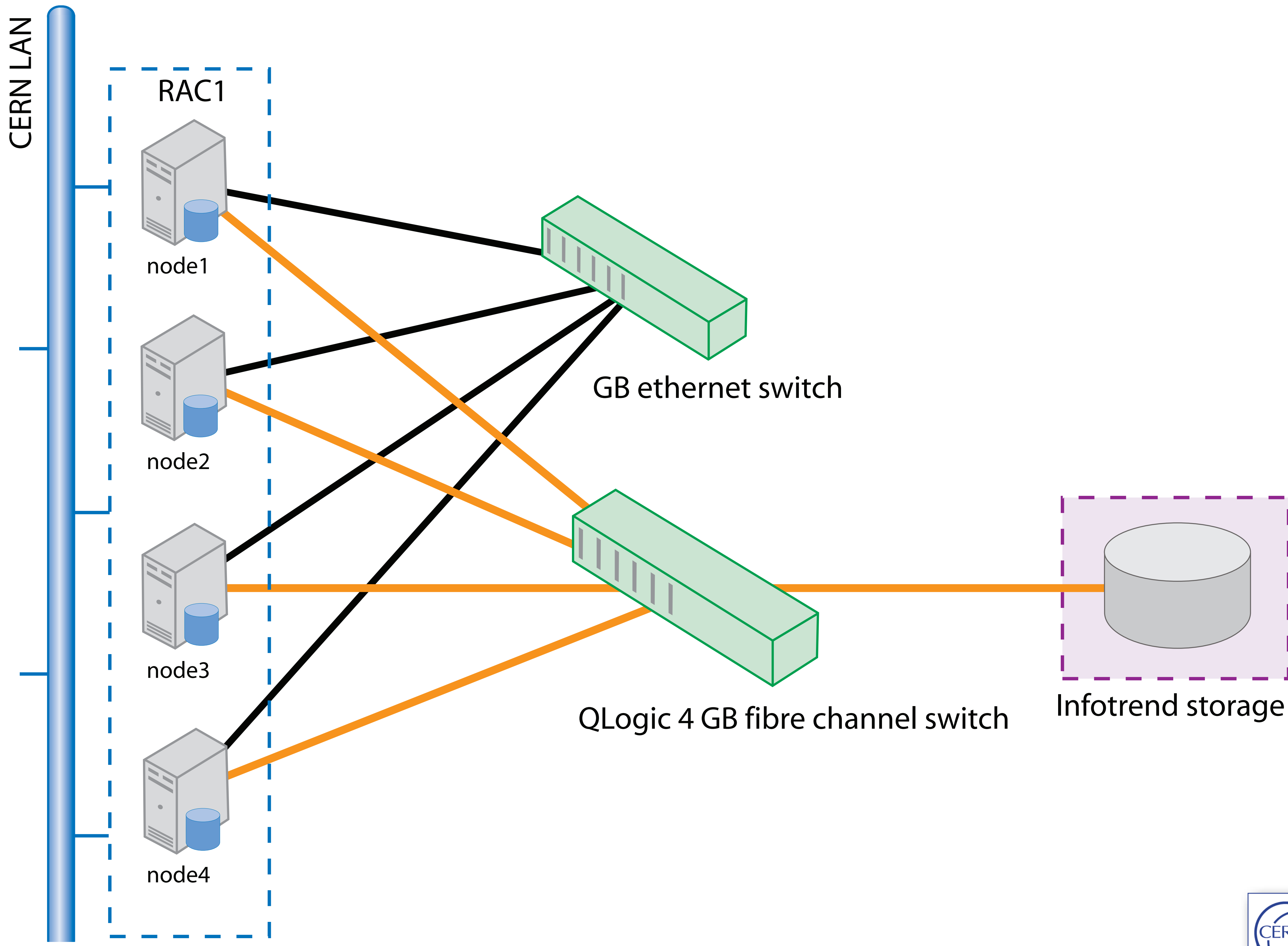
High Energy Physics - User Community

- Very many application developers
  - with varying levels of DB training
- A large number of different applications
  - Detector geometry, conditions, calibration, configuration, production workflow, analysis data
  - Grid services: file catalogs, transfer workflow
- Very different operational environments
  - online systems:
    - HA required, controlled environment
  - data production:
    - coordinated batch access by production managers, grid computing
  - data analysis:
    - chaotic access by a large number of users

Distributed Service Architecture



Building Block for LHC DBs- Database Clusters



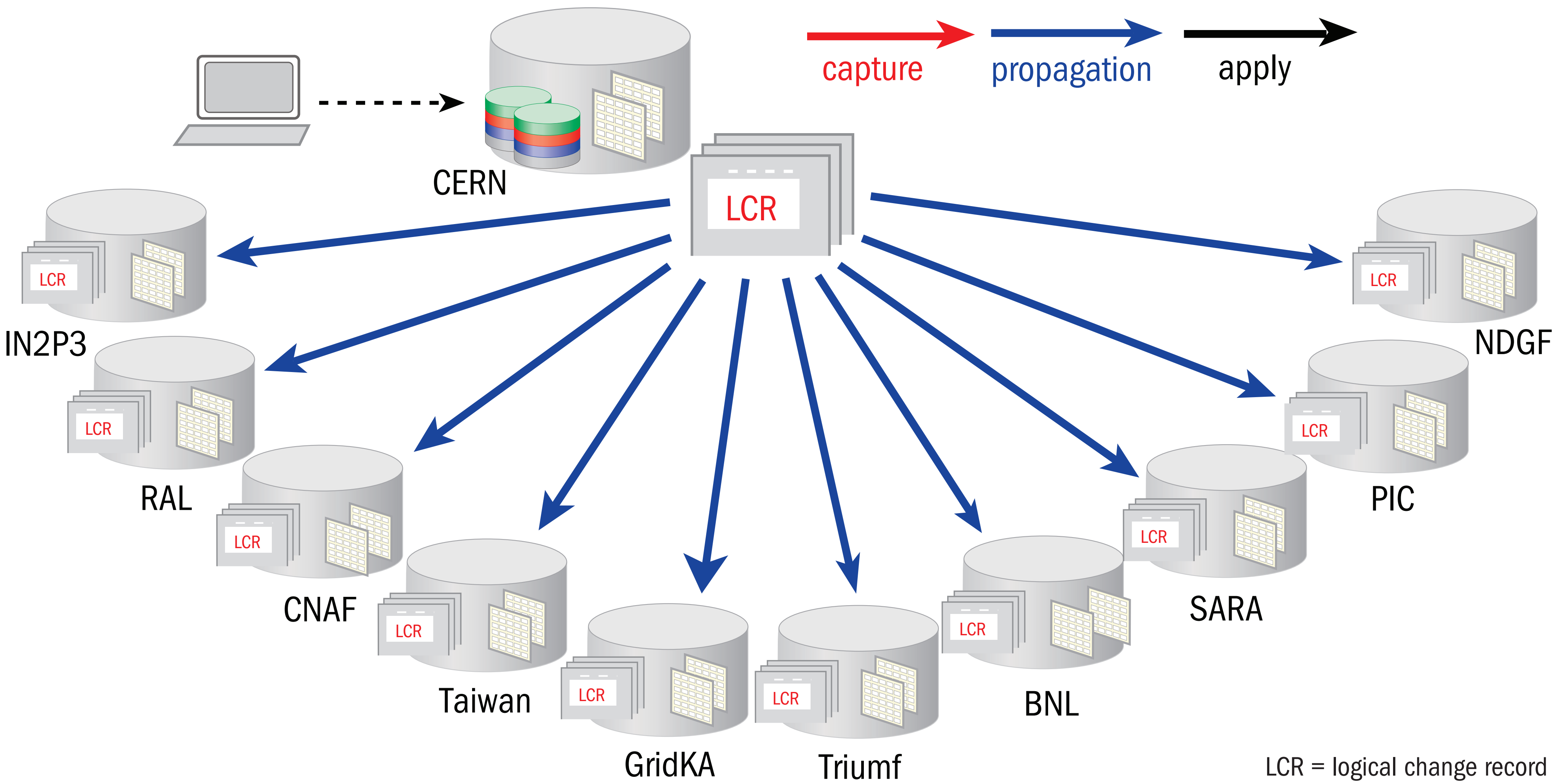
Physics Database Service Deployment Numbers (CERN only)

- 110 server nodes - RHEL 5, Oracle 10g
  - 0.5 TB of RAM
  - spread over some 15 clusters
  - service levels: development, validation, production
- 112 disk arrays, 300TB total (single DB in few tens TB range)
  - SATA disks attached via fibre channel controller
- Some average numbers (before LHC running)
  - 3 Million sessions /week
  - 100 MB/s physical I/Os (per cluster)
- Moving to quad-core and 64-bit OS & Oracle
  - significant gains and promising scaling with increasing available CPU power
  - will add 32 QC CPU nodes + 60 disk arrays before LHC start
- 5 Database Administrators
  - OS & box level support from other CERN teams
  - Reliability today around 99.98%

Deployment Issues

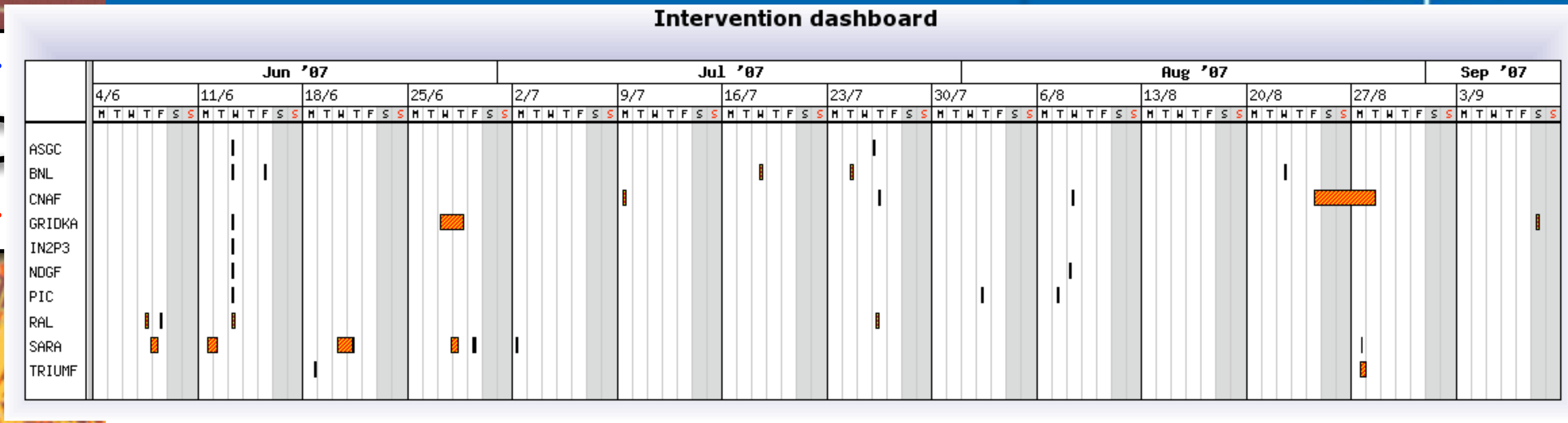
- Power and UPS
  - Both are limited as the CERN computing center evolves with LHC requirements
  - A/C and power problem cause significant h/w loss and require precious DBA time
- Increasing CPU power per box needs more and more disk spindles per box
  - JBOD & ASM approach -> many devices on linux level
- Bulk orders of inexpensive h/w
  - Exposed to bulk h/w problems
- Disks and CPU nodes do fail
  - That's ok - our normal mode of operation!
- Oracle (security) patches - not always 'rolling'
  - big improvement recently

Oracle Streams Replication

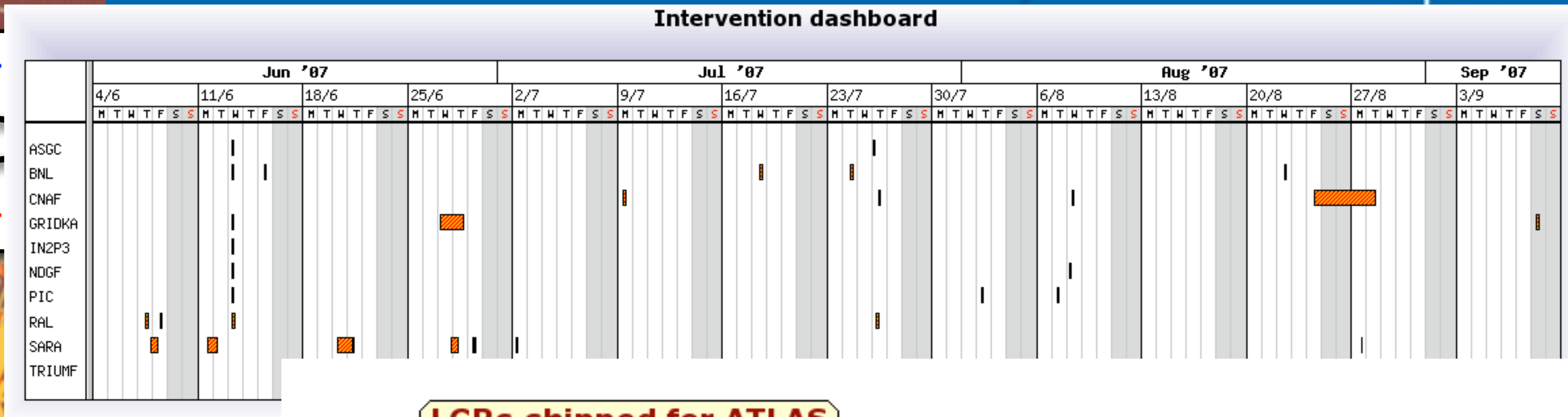


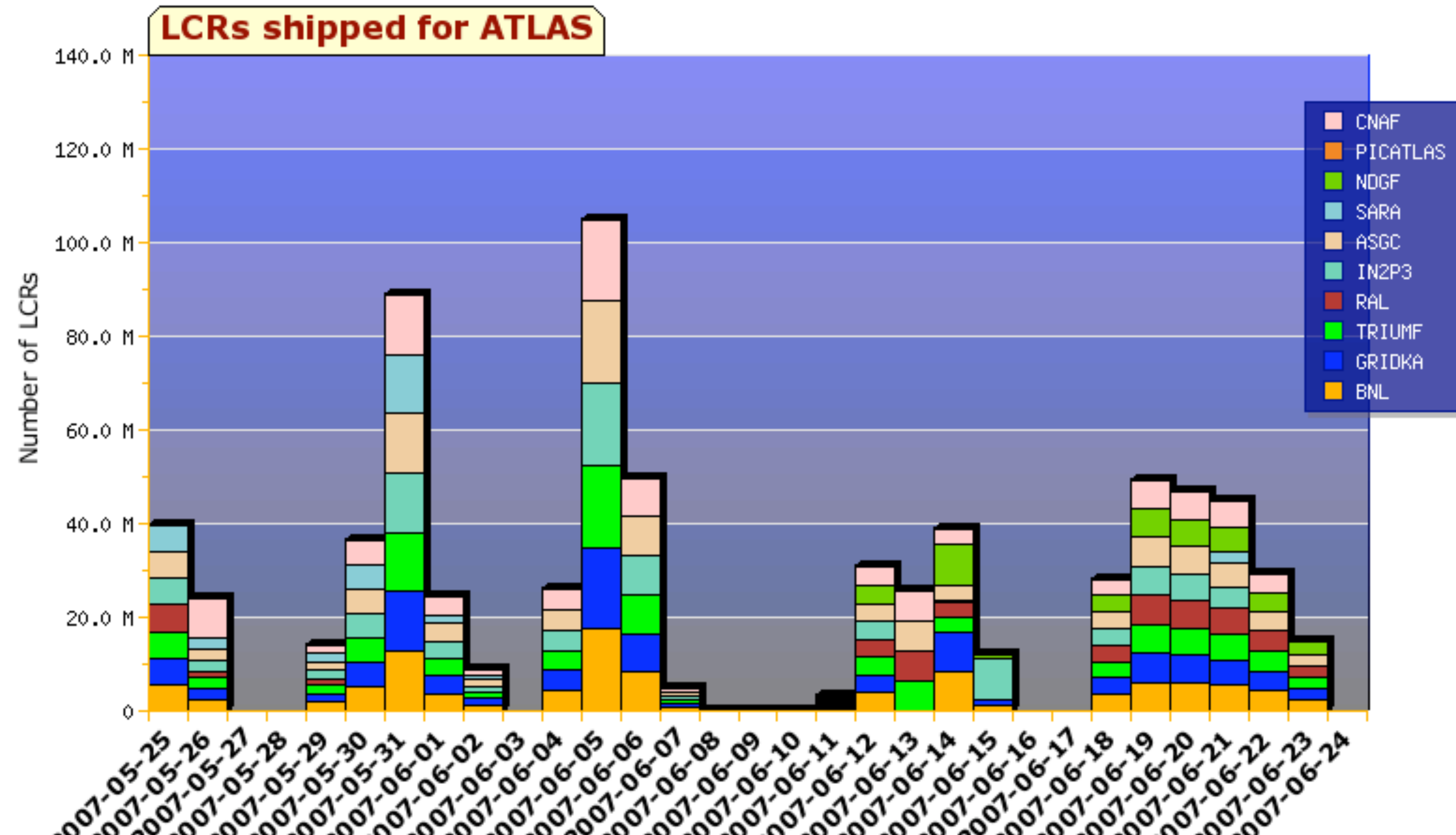
- Database changes captured from the redo-log and propagated asynchronously as Logical Change Records (LCRs)
- All changes are queued until successful application at all destinations
  - need to control change rate at the source in order to minimise the replication latency
  - 2GB/day user data to Tier 1 can be sustained with the current DB setups
- significant overheads between user data and redo-log volume apply

Intervention & Streams Reports



Intervention & Streams Reports





Distributed Deployment Issues

- Data consistency - distributed recoveries exercising the integration between local recovery and global syncronisation
  - Joint training with all DBA teams is essential
- Database software licenses, versions and updates
  - Not all sites have the same schedule and security policies
- Monitoring and Diagnostics
  - Global system monitoring had to be developed
- Database services exposed to the internet
  - Firewalls and security closure procedures
- Application side retry and failover among accessible replicas, db replica catalog
  - Handled via common application s/w layer (CORAL)
- Grid (remote batch) processing and security
  - Shipping access credentials with applications
    - User / password approach, certificates, proxy-cert's

Queries with many dimensions

CERN IT Department

- Physics analysis needs multi-dimensional interval queries on very large input data sets (>10**9)
  - select .. where v1 > 4 and v2 > 5 … v99 > 3
    - B-Trees implementation of limited use
    - large input data sets (10**9 or more)
    - Bitmap indices for continuous variables
      - long standing research topic
      - significant space and maintenance overhead
- Today
  - implemented via column-wise clustered files and specialised analysis programs (eg ROOT)
- Tomorrow
  - Petabyte flash-RAM and in-memory databases?

Questions for discussion?

- Filesystems add meta-data queries - Database add file storage (and remove file systems)
  - signs of conversion - or just expansion of each camp?
- Databases as
  - transactional system?
  - efficient query implementation?
  - highly available shared storage?
  - Not all applications need all of the above
    - but service costs to provide above qualities are very different (and usually significantly higher than for files)
- Are hybrid systems unavoidable ($-wise) for very large stores?

PSS

Questions for discussion?

CERN IT Department

- Will we see more hybrid (hierarchical) structures as the available memory increases wrt active data?
  - proxy-caches, in memory databases, solid-state disks
- Is the disk volume a good metric to characterise the scaling of database systems?
  - active data fraction, write/update fraction, IOops/TB, IOops/SPECINT
  - many DB apps are limited by CPU or cluster interconnect traffic
- Shared everything or shared nothing ?
  - which architecture will win the scaling race with a typical(?) application mix?

Summary

- High Energy Physics and Astronomy produce unprecedented amounts of data
  - Databases are a key component of the data handling with an increasing scope in all areas of data handling & analysis
- Joint work between database vendors and science community (eg in CERN openlab) has been extremely beneficial for both sides
  - Allowed to construct one of the worlds largest distributed database deployments world-wide for LHC
- Many of the technology and deployment issues are/will soon be relevant also for larger commercial data management systems
  - The open environment of science is an ideal place to push the limits of current technology further
  - Also to the benefit of non-science applications