Databases for the CERN LHC

Techniques and Challenges

Maria Girone, CERN

Leader of the Database Services for Physics

Oracle Open World, San Francisco, 13th Oct 2009

Outline

- CERN and LHC
- Databases and LHC Computing Grid
- Techniques & Challenges
- CERN and Oracle 11gR2 testing

What is CERN?

CERN is:
-~ 2500 staff scientists (physicists, engineers, …)
- Some 6500 visiting scientists (half of the world's particle physicists)

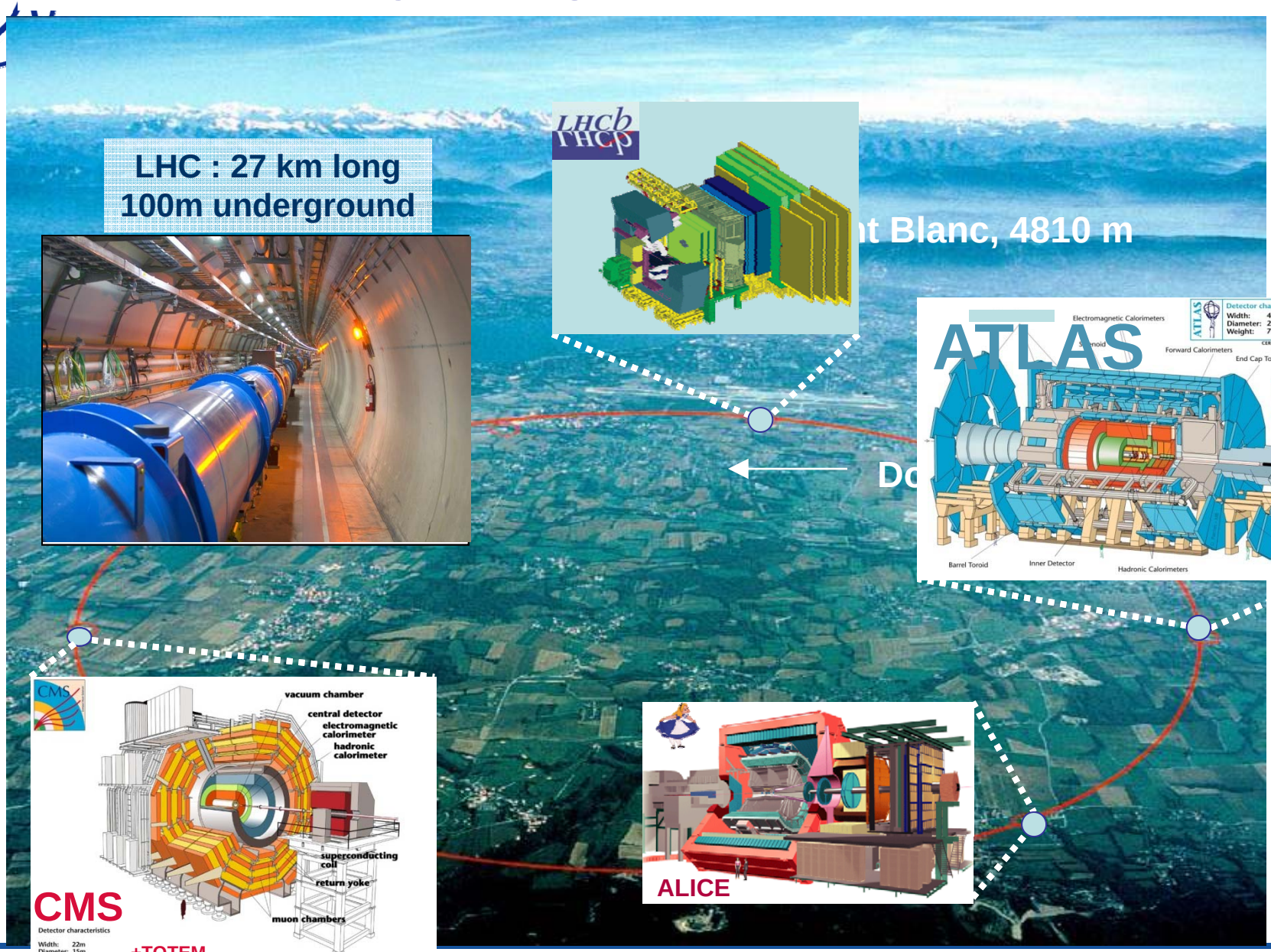
They come from 500 universities representing 80 nationalities.

- CERN is the world's largest particle physics centre

- Particle physics is about:
  - elementary particles and fundamental forces
- Particles physics requires special tools to create and study new particles

  - ACCELERATORS, huge machines able to speed up particles to very high energies before colliding them into other particles
  - DETECTORS, massive instruments which register the particles produced when the accelerated particles collide

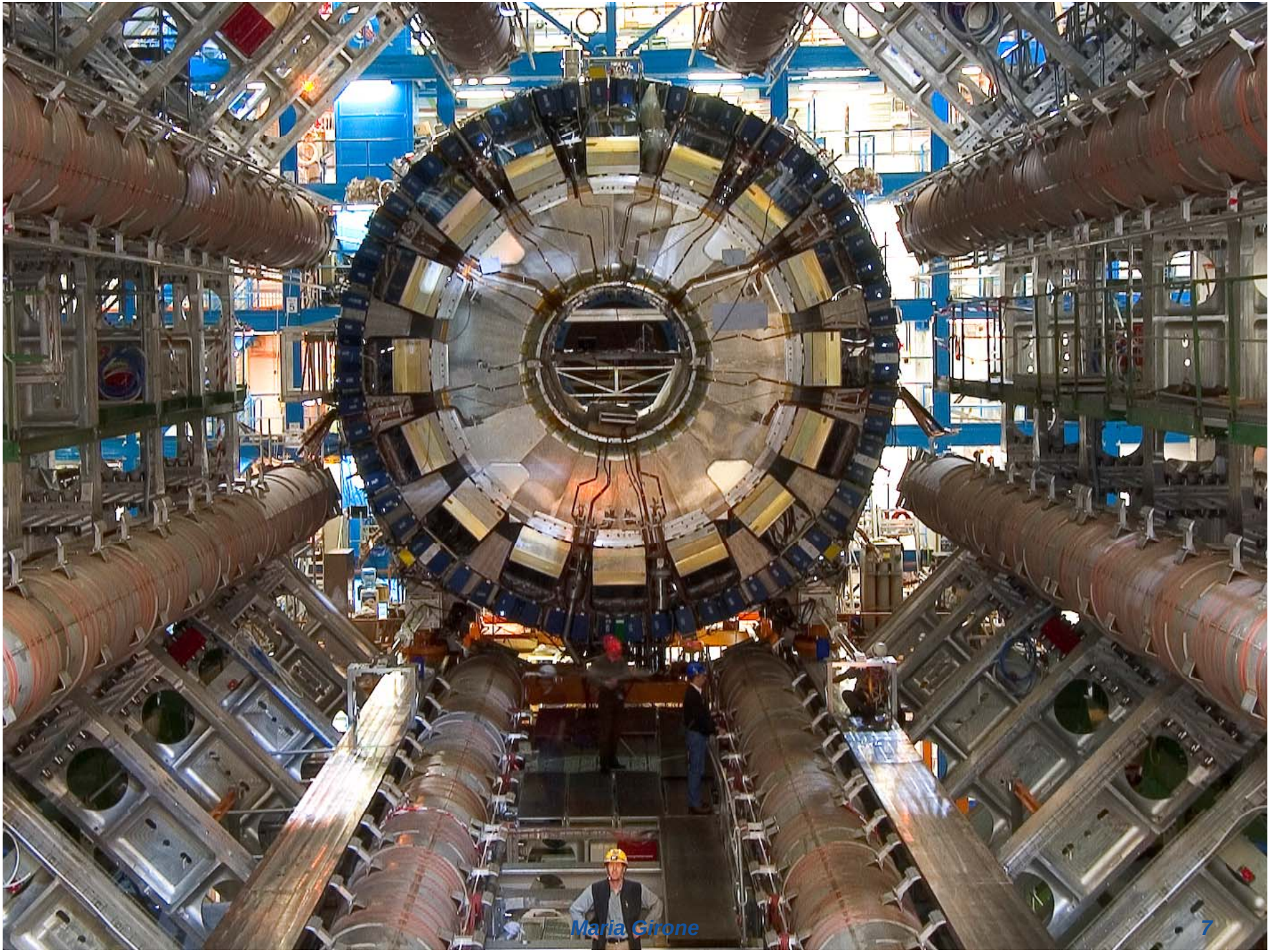
LHC: a Very Large Scientific Instrument



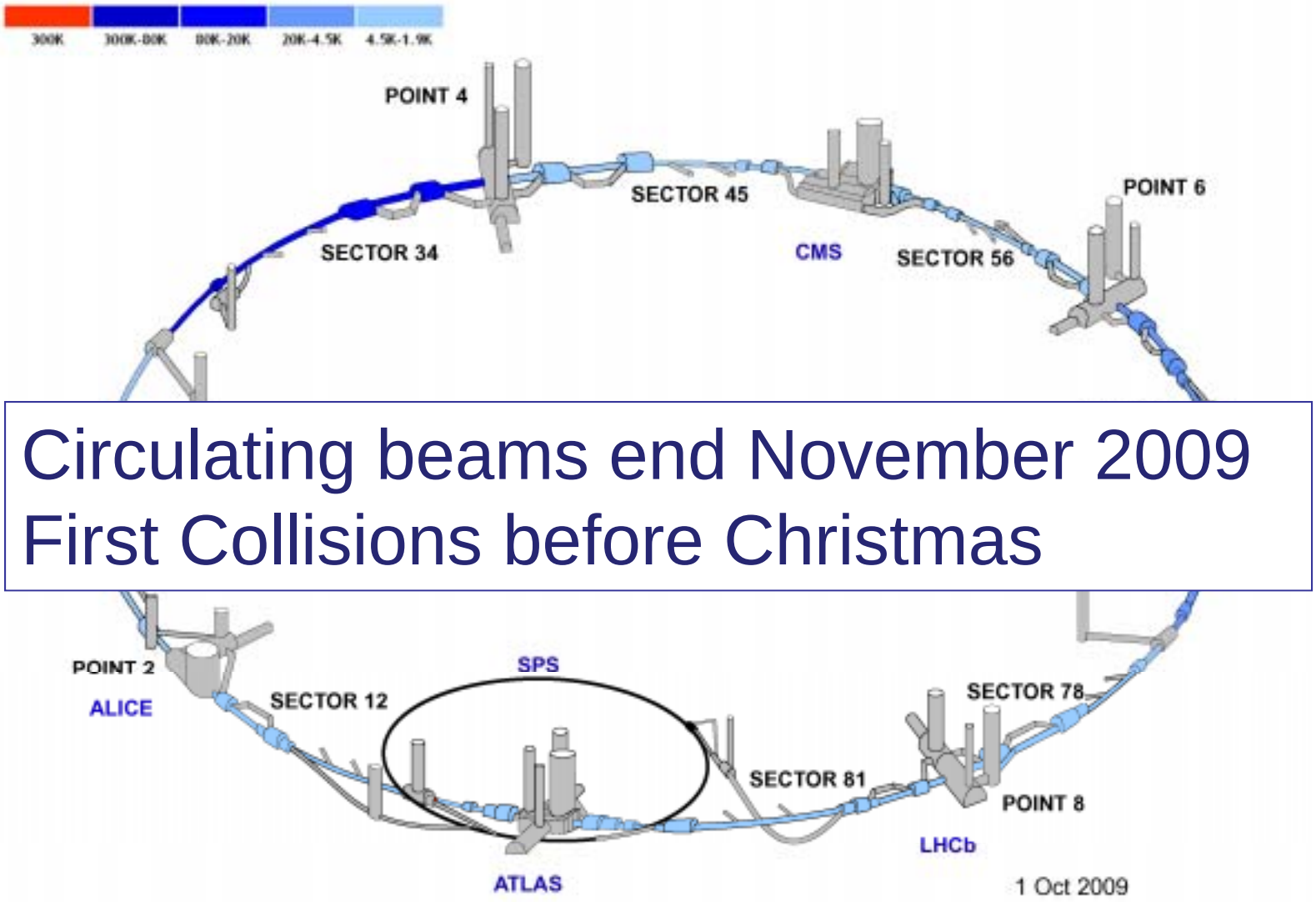
… Based on Advanced Technology

27 km of superconducting magnets
cooled in superfluid helium at 1.9 K

CMS Closed & Ready for First Beam



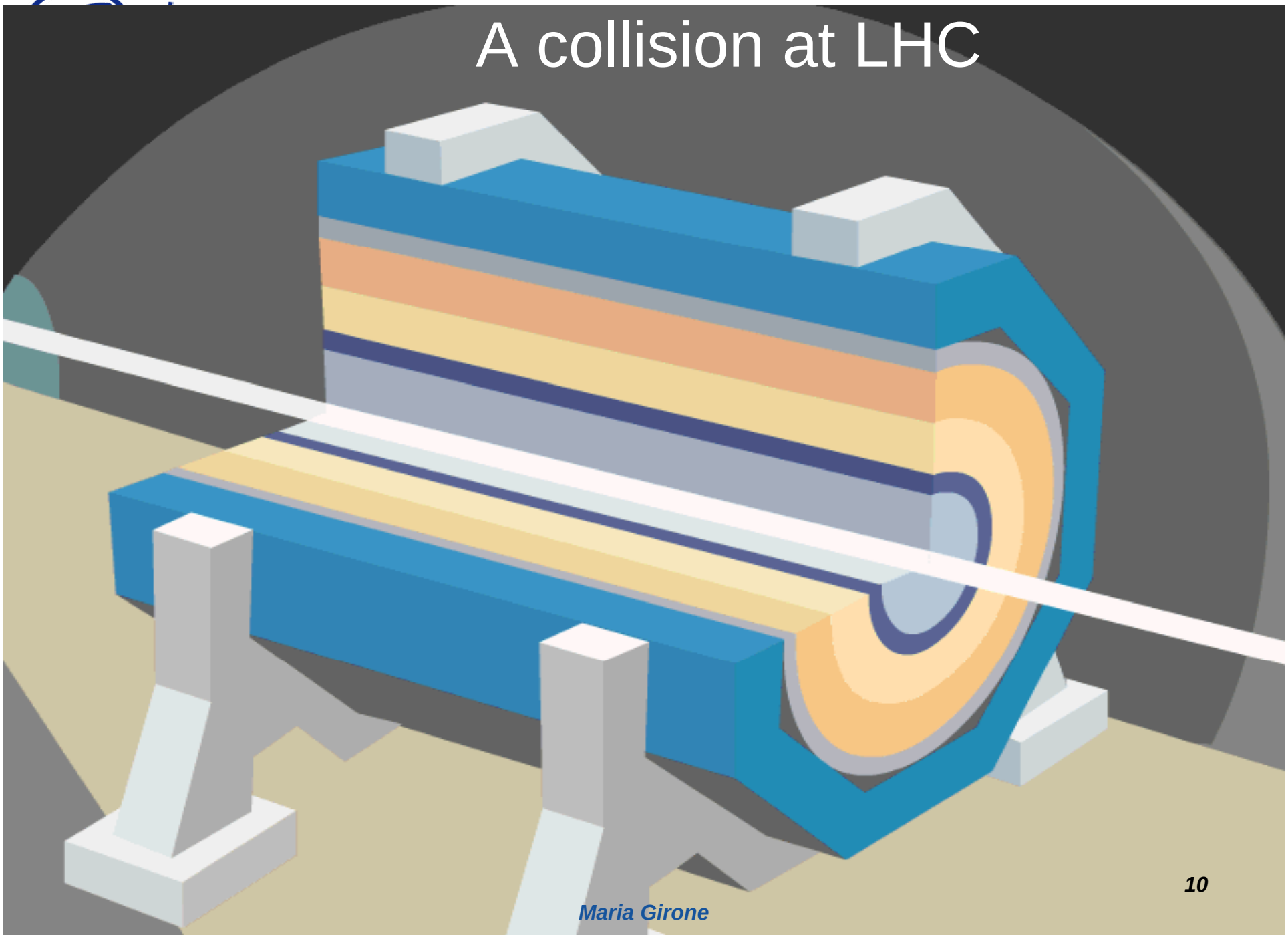
LHC Cooldown Status



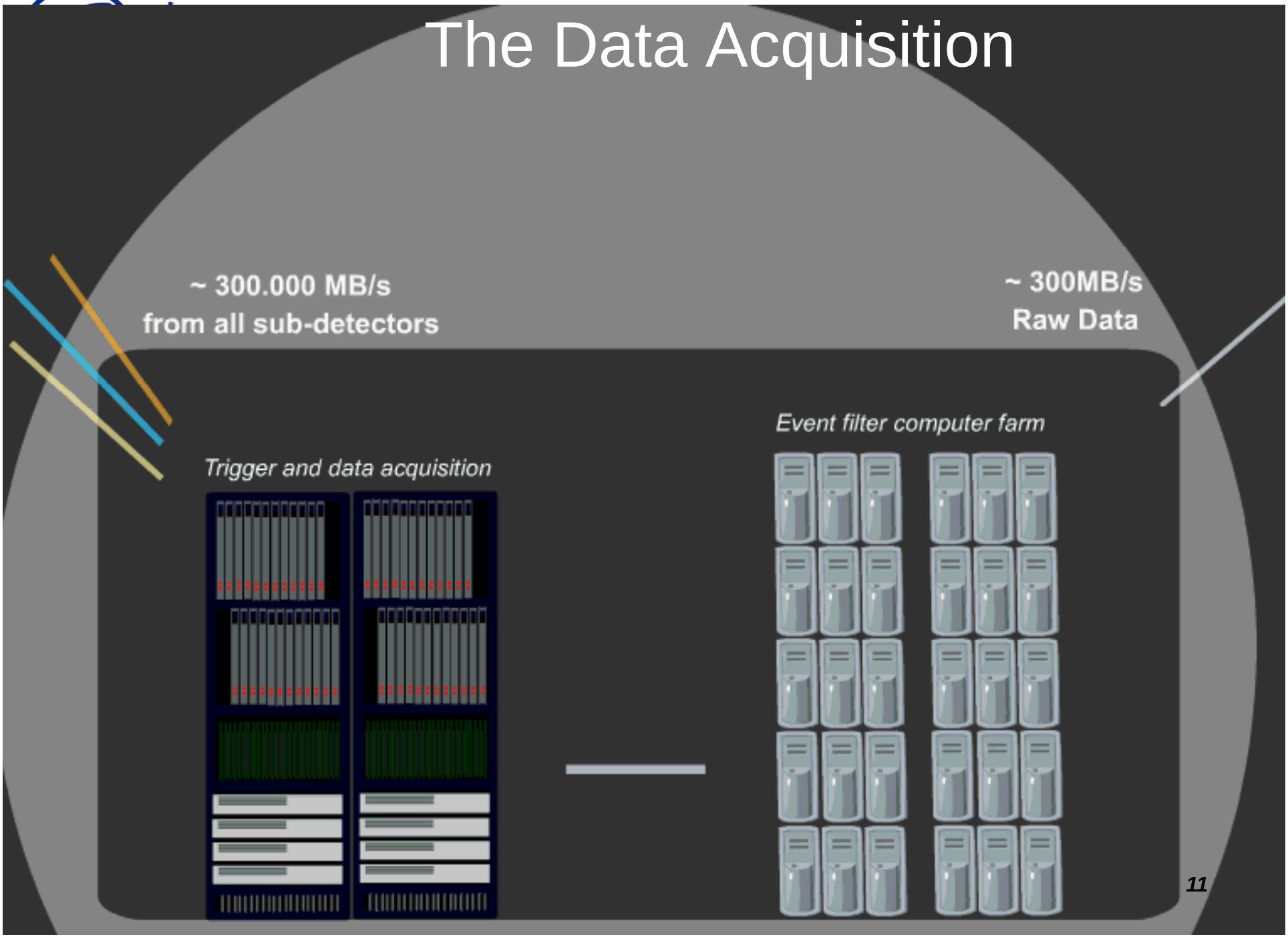
Circulating beams end November 2009
First Collisions before Christmas

The LHC Computing Grid

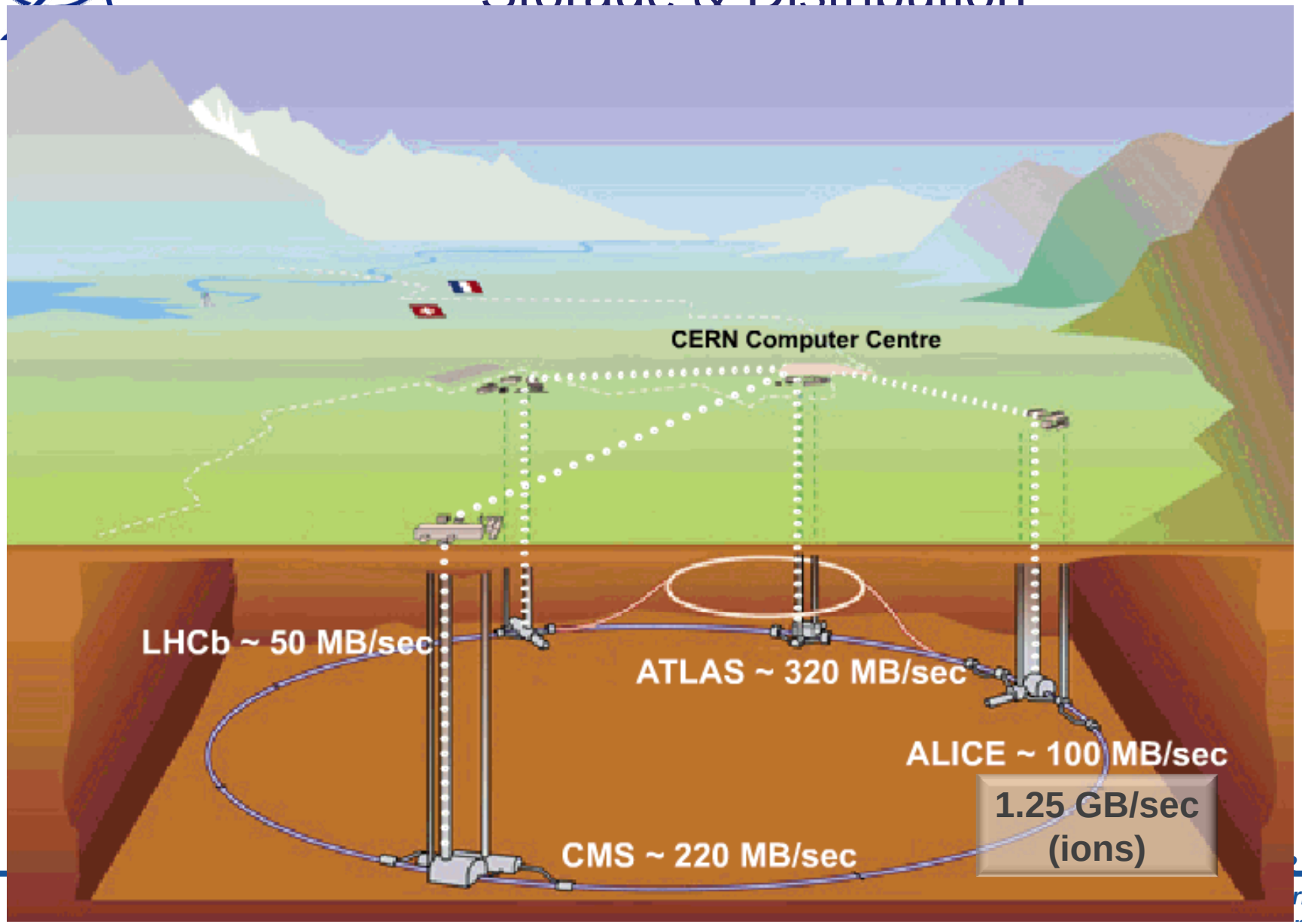
A collision at LHC



The Data Acquisition



Tier 0 at CERN: Acquisition, First pass processing
Storage & Distribution



The LHC Computing Challenge

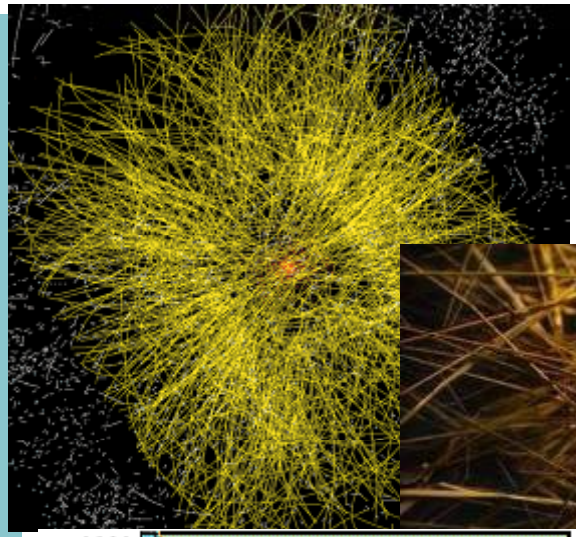
- Signal/Noise: 10^{-9}
- Data volume
  - High rate * large number of channels * 4 experiments

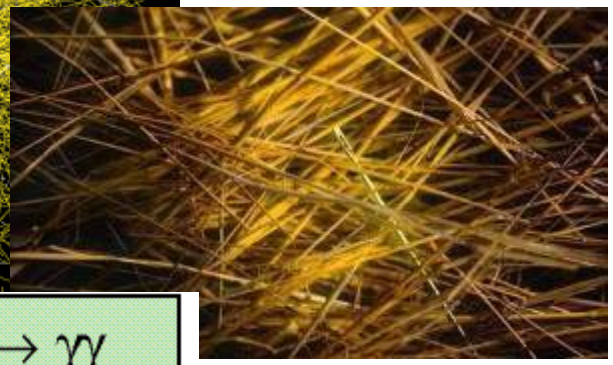
  ➔ **15 PetaBytes of new data each year**
- Compute power
  - Event complexity * Nb. events * thousands users

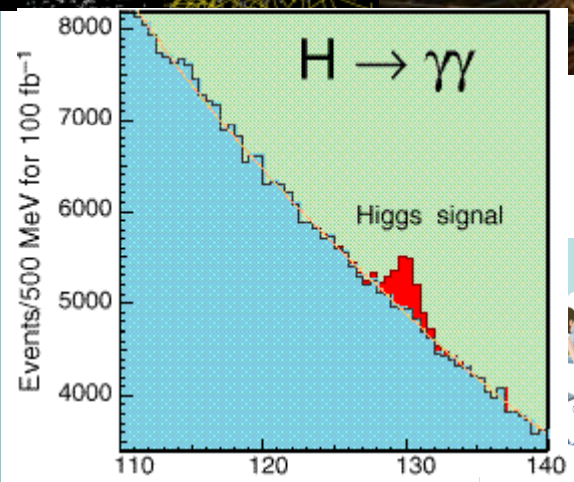
  ➔ **100 k of (today's) fastest CPUs**

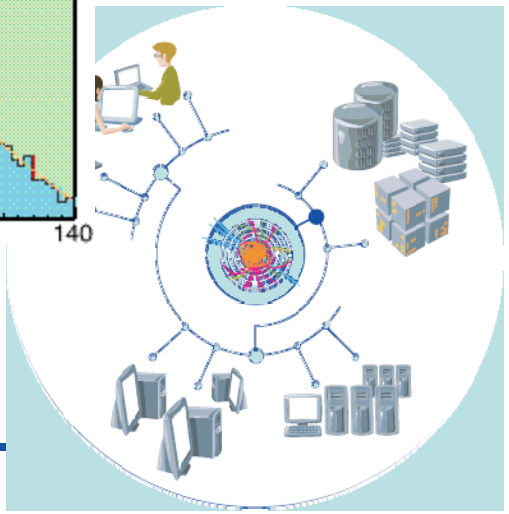
  ➔ **45 PB of disk storage**
- Worldwide analysis & funding
  - Computing funding locally in major regions & countries
  - Efficient analysis everywhere

  ➔ **GRID technology**
- Bulk of data stored in files, a fraction of it in databases (~30TB/year)





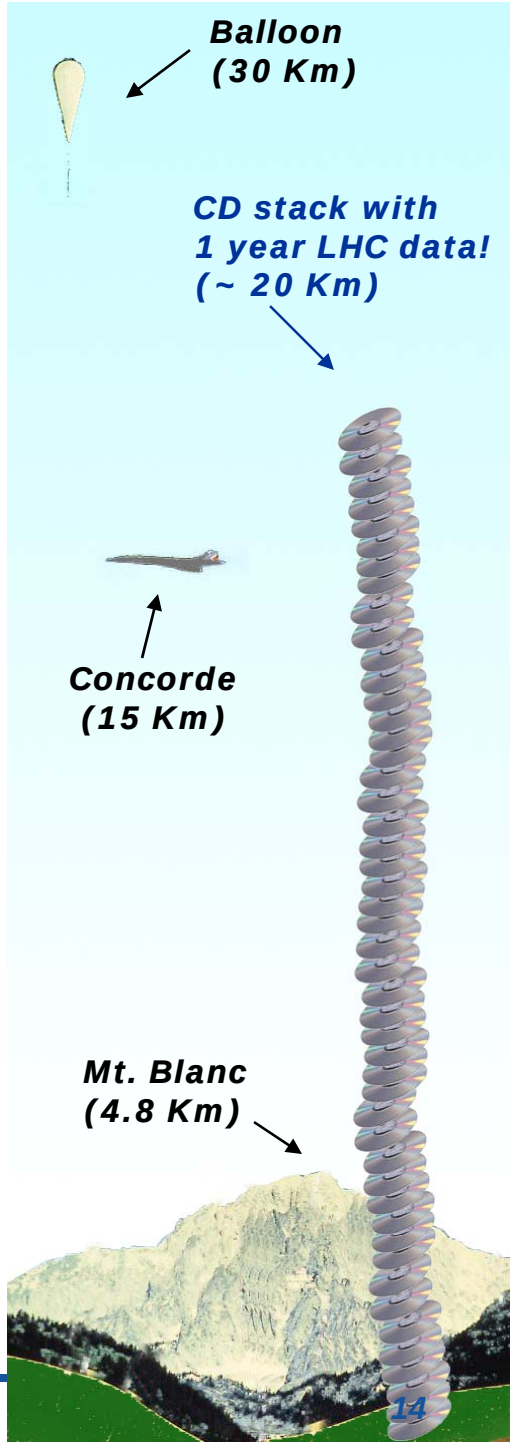




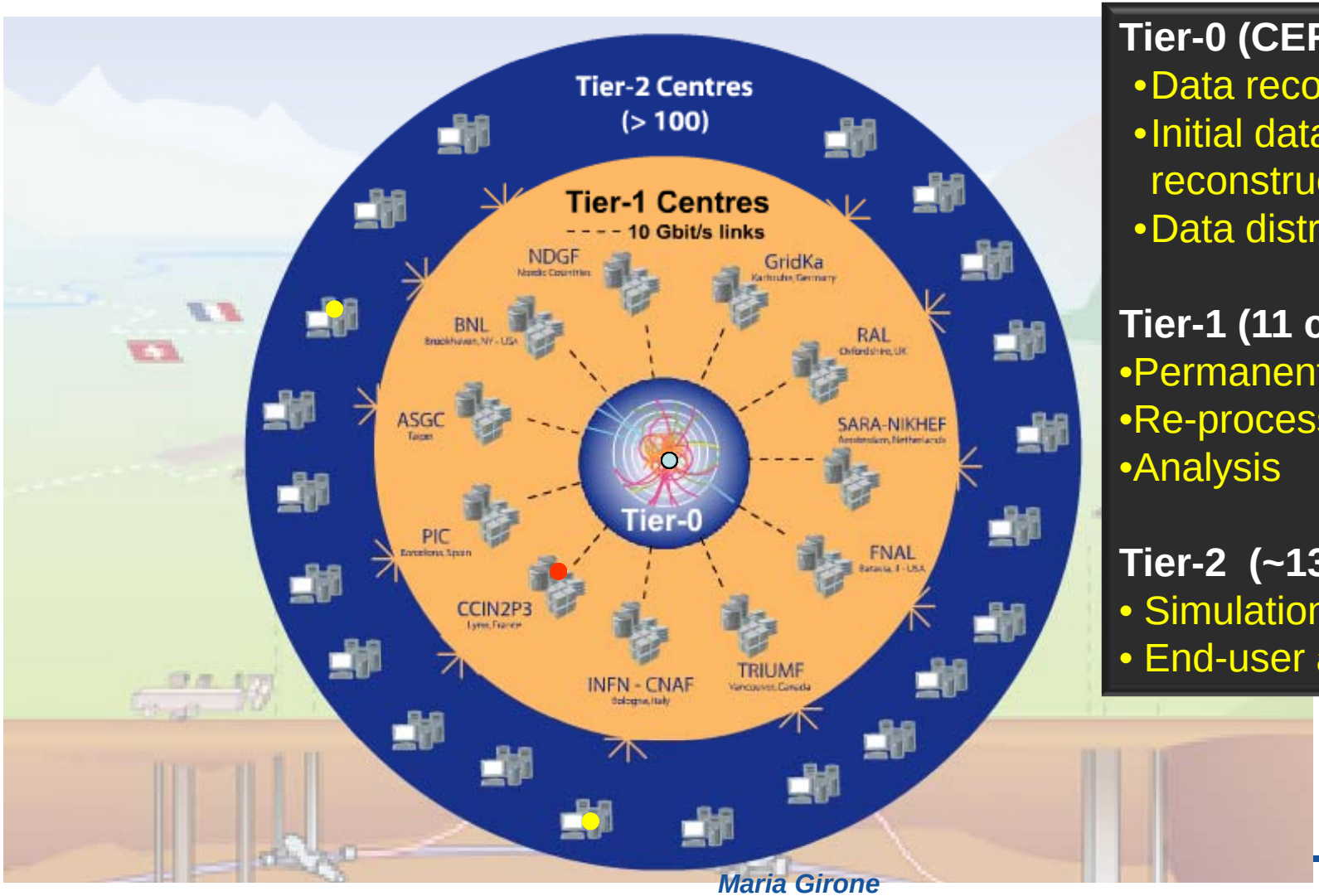
LHC data

LHC data correspond to about 20 million CDs each year!

Where will the experiments store all of these data?



Tier 0 – Tier 1 – Tier 2



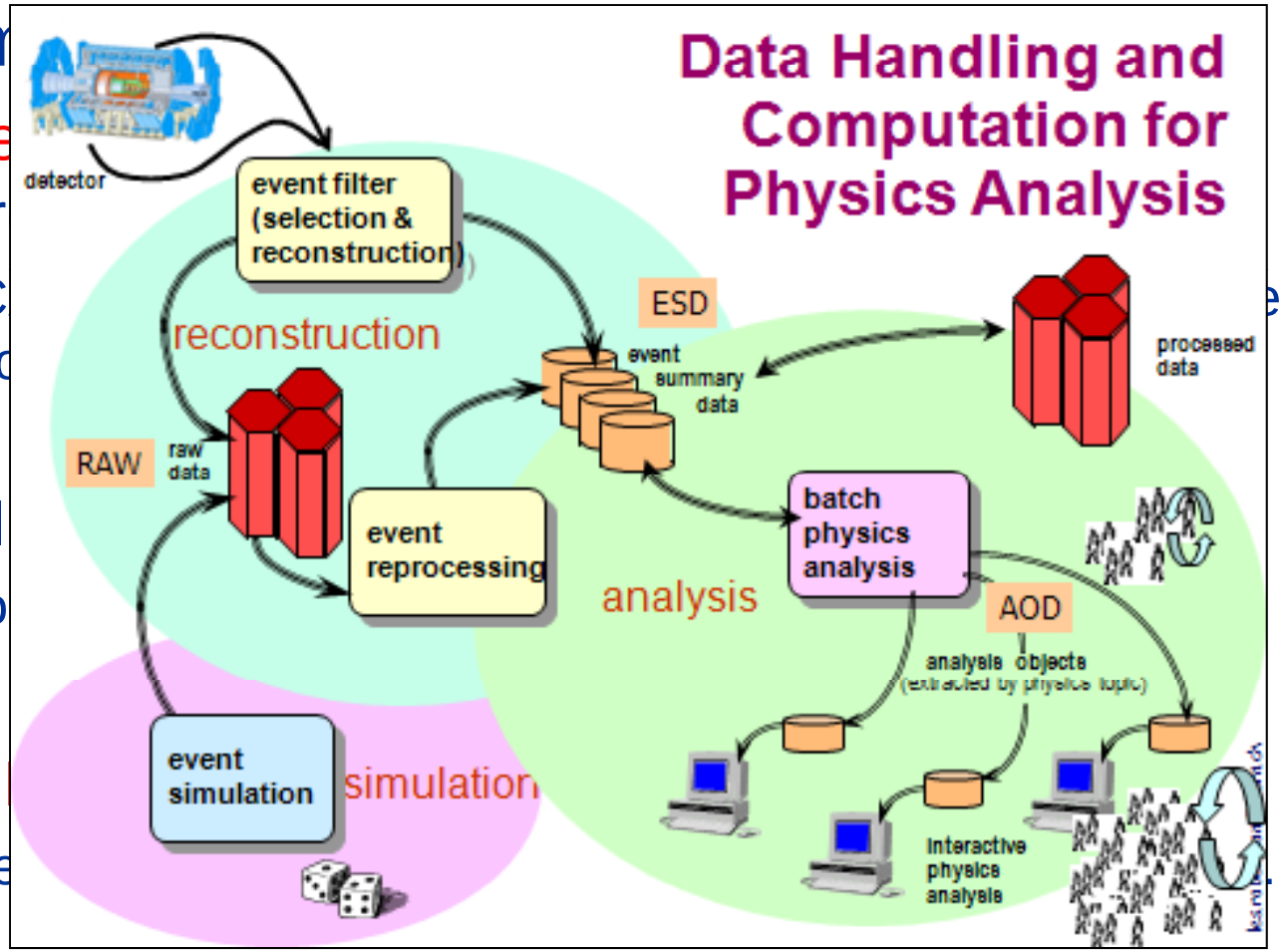
Tier-0 (CERN):
- Data recording
- Initial data reconstruction
- Data distribution

Tier-1 (11 centres):
- Permanent storage
- Re-processing
- Analysis

Tier-2 (~130 centres):
- Simulation
- End-user analysis

Databases and LHC

- Relational DBs play today a key role in the experim
  - online
    (re)pr
    - SC
      boo
  - Grid I
    - Mo
  - Data
    - File



Techniques…

Service Key Requirements

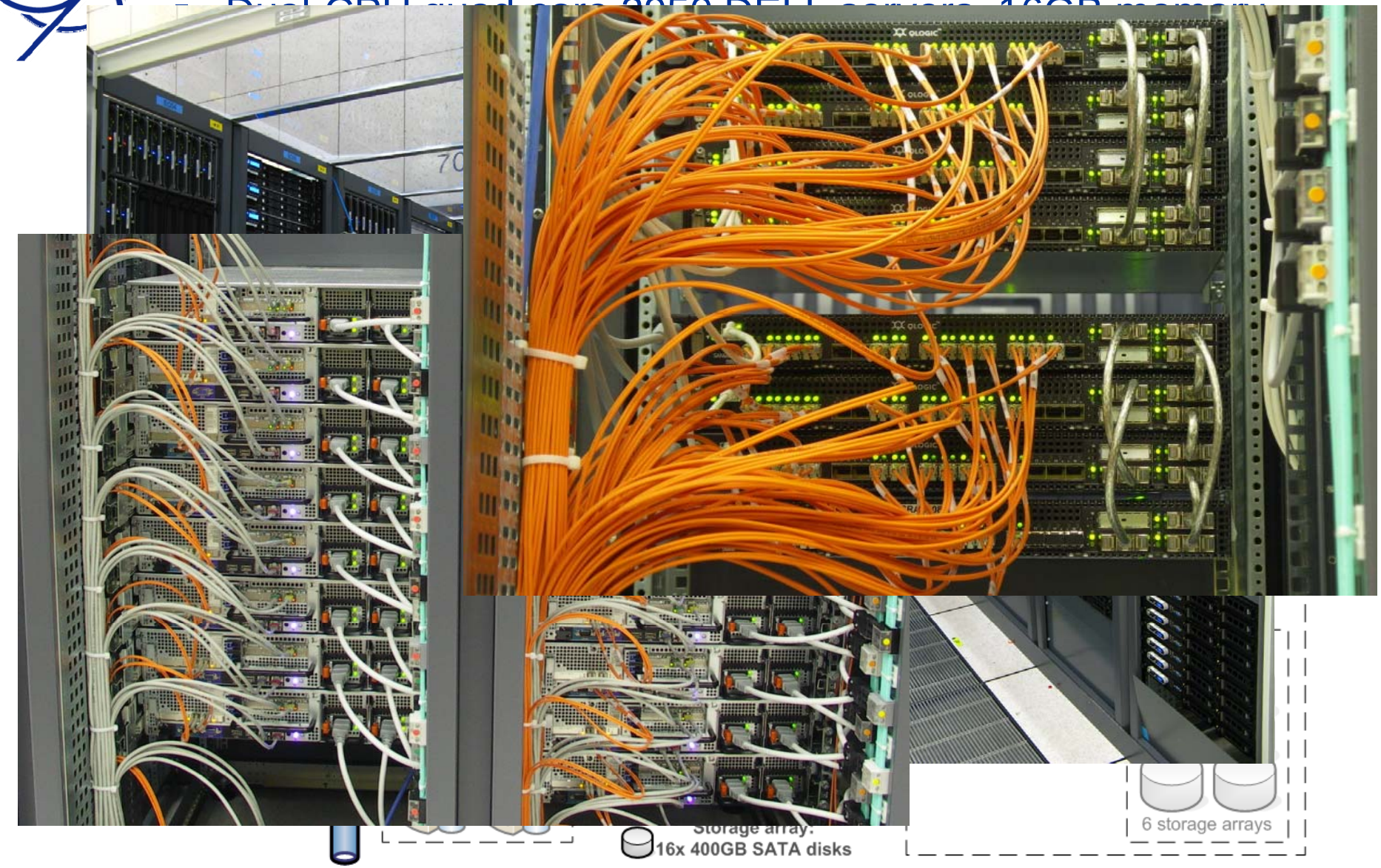
- Data Availability, Scalability, Performance and Manageability
  - Oracle RAC with ASM: building-block architecture for CERN and Tier1 sites
  - Rolling upgrade and failover capabilities essential for service continuity
- Data Distribution
  - Oracle Streams: for sharing information between databases at CERN and 10 Tier1 sites
- Data Protection
  - Oracle RMAN on TSM for backups
  - Oracle Data Guard: for additional protection against failures (data corruption, disaster recoveries,...)

Physics DB Services in a Nutshell

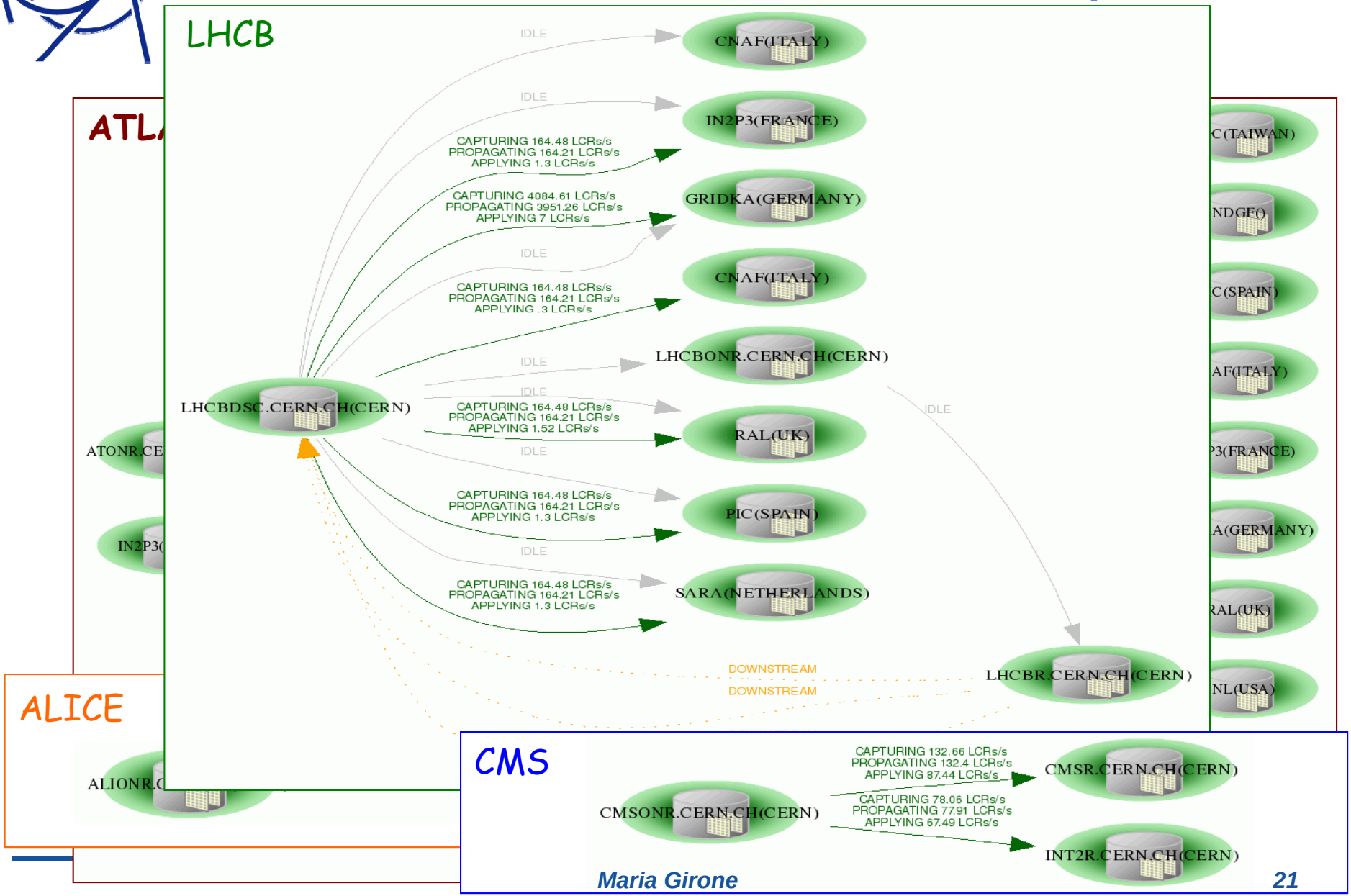
- About 30 RAC with ASM databases (clusters up to 6 nodes). Oracle 10.2.0.4, 64 bit
  - 150 servers, 200 disk arrays (2300 spindles)
  - 650 CPU cores, 1300GB of RAM, 850TB raw disk space
  - More than 1000 deployed schemas
- Connected via Oracle Streams replication to 10 Tier1 sites
- Team of 6 DBAs supporting mission-critical DBs on a 24x7 rota

CERN Set-up

- Dual CPU quad core 2950 DELL servers, 16GB memory



LHC Streams Set-up

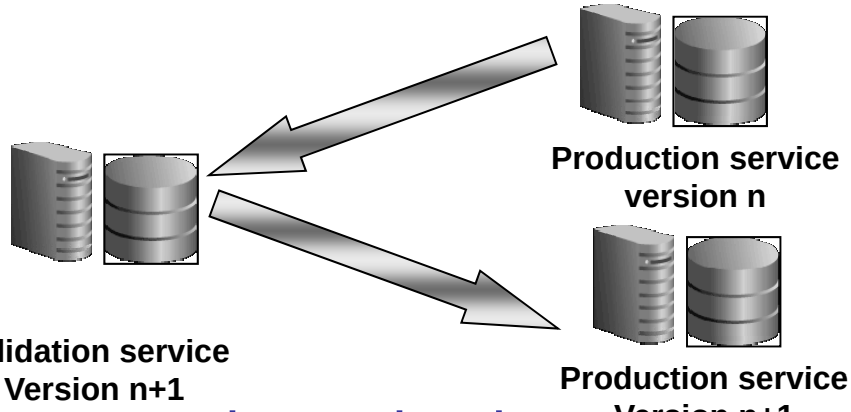


… and Challenges

Application Deployment Policy

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

  - Database software release cycle

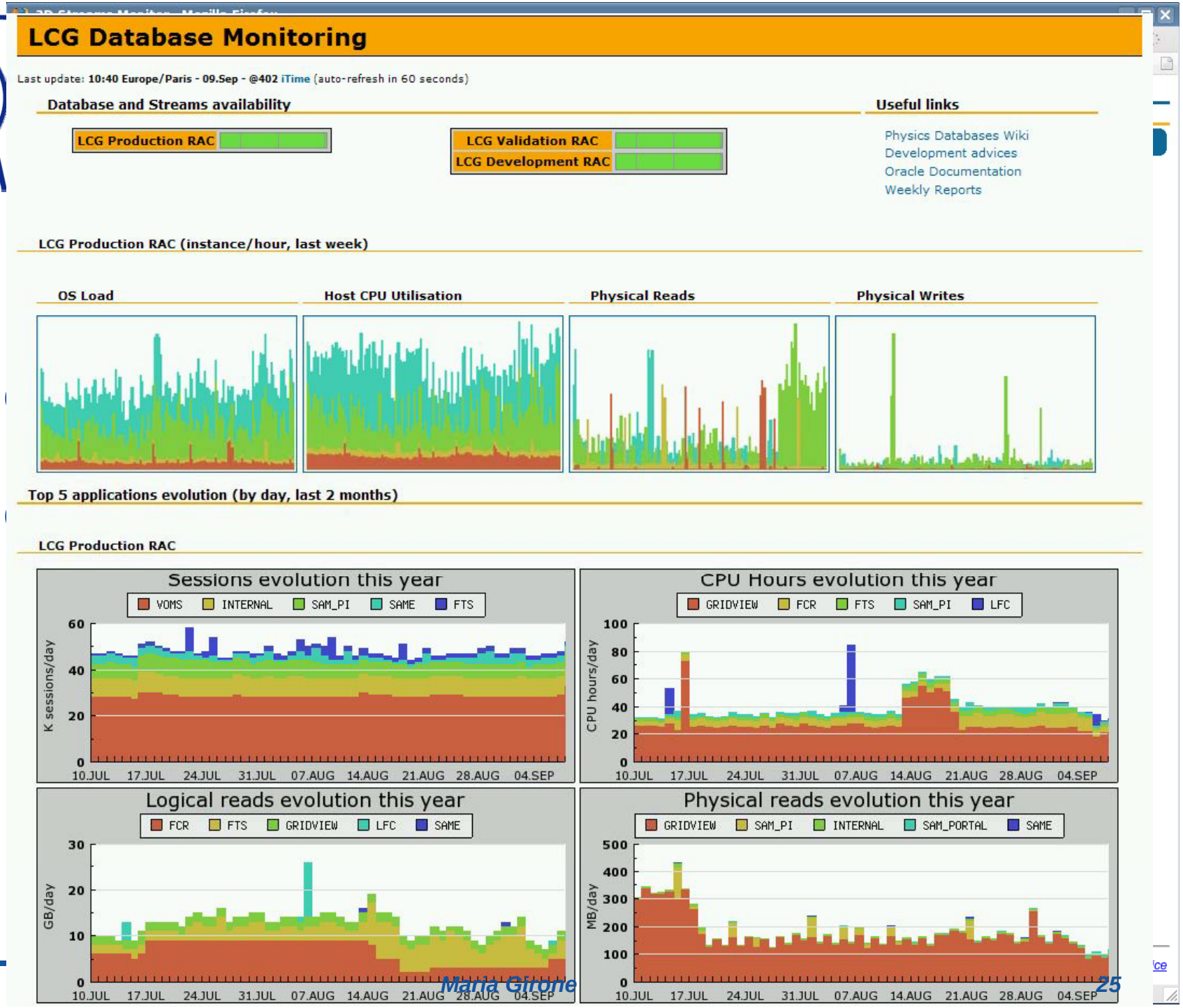


- Proven key to smooth production

Patching and Upgrades

- Databases are used by a world-wide community: arranging for scheduled interventions (s/w and h/w upgrades) requires quite some effort
  - Services need to be operational 24x7

- Minimize service downtime with rolling upgrades and use of stand-by databases
  - **0.04% services unavailability = 3.5 hours/year**
  - **0.12% server unavailability = 9.5 hours/year (Patch deployment, hardware)**


LCG Database Monitoring
Last update: 10:40 Europe/Paris - 09.Sep - @402 iTime (auto-refresh in 60 seconds)
Database and Streams availability
LCG Production RAC
LCG Validation RAC
LCG Development RAC
Useful links
Physics Databases Wiki
Development advices
Oracle Documentation
Weekly Reports
LCG Production RAC (instance/hour, last week)
OS Load
Host CPU Utilisation
Physical Reads
Physical Writes
Top 5 applications evolution (by day, last 2 months)
LCG Production RAC
Sessions evolution this year
VOMS INTERNAL SAM_PI SAME FTS
K sessions/day
CPU Hours evolution this year
GRIDVIEW FCR FTS SAM_PI LFC
CPU hours/day
Logical reads evolution this year
FCR FTS GRIDVIEW LFC SAME
GB/day
Physical reads evolution this year
GRIDVIEW SAM_PI INTERNAL SAM_PORTAL SAME
MB/day
10.JUL 17.JUL 24.JUL 31.JUL 07.AUG 14.AUG 21.AUG 28.AUG 04.SEP

Backup & Recovery

- On-tape backups: fundamental for protecting data, but recoveries run at ~100MB/s (~30 hours to restore datafiles of a DB 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

Security

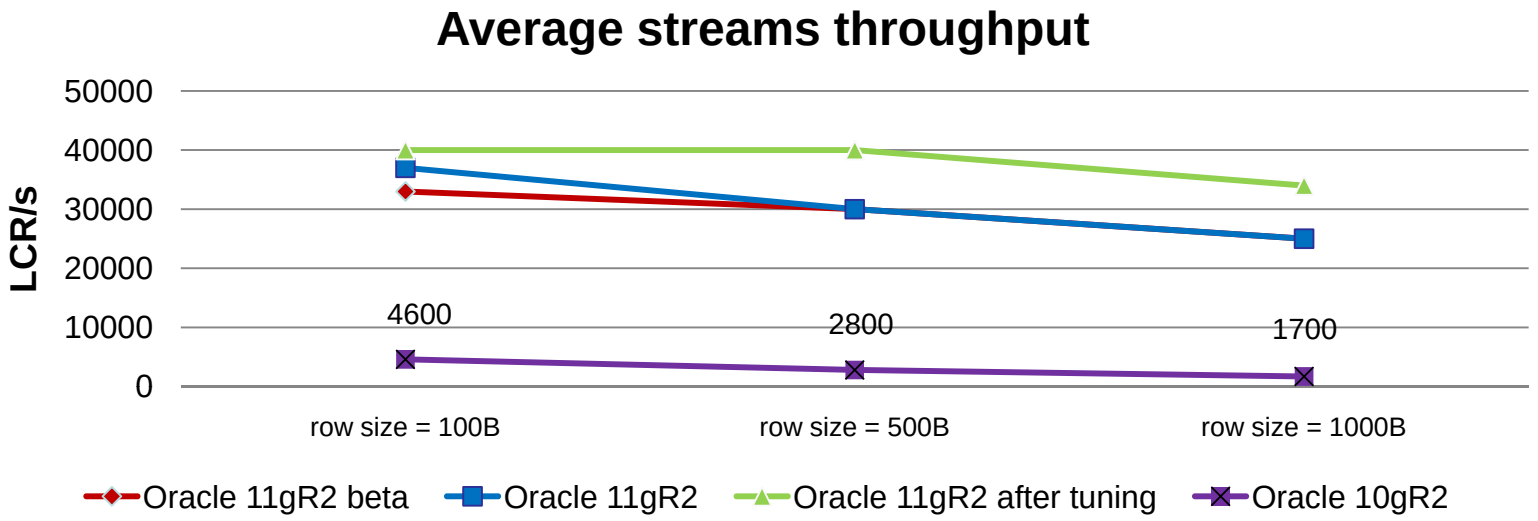
- Schemas setup with ‘least required privileges’
  - account owner only used for application upgrades
  - reader and writer accounts used by applications
  - password verification function to enforce strong passwords
- Firewall to filter DB connectivity
  - CERN firewall and local iptables firewall
- Oracle CPU patches
  - Production up-to-date after validation period
  - Policy agreed with users
- Custom development
  - Audit-based log analysis
  - Automatic pass cracker to check password weakness

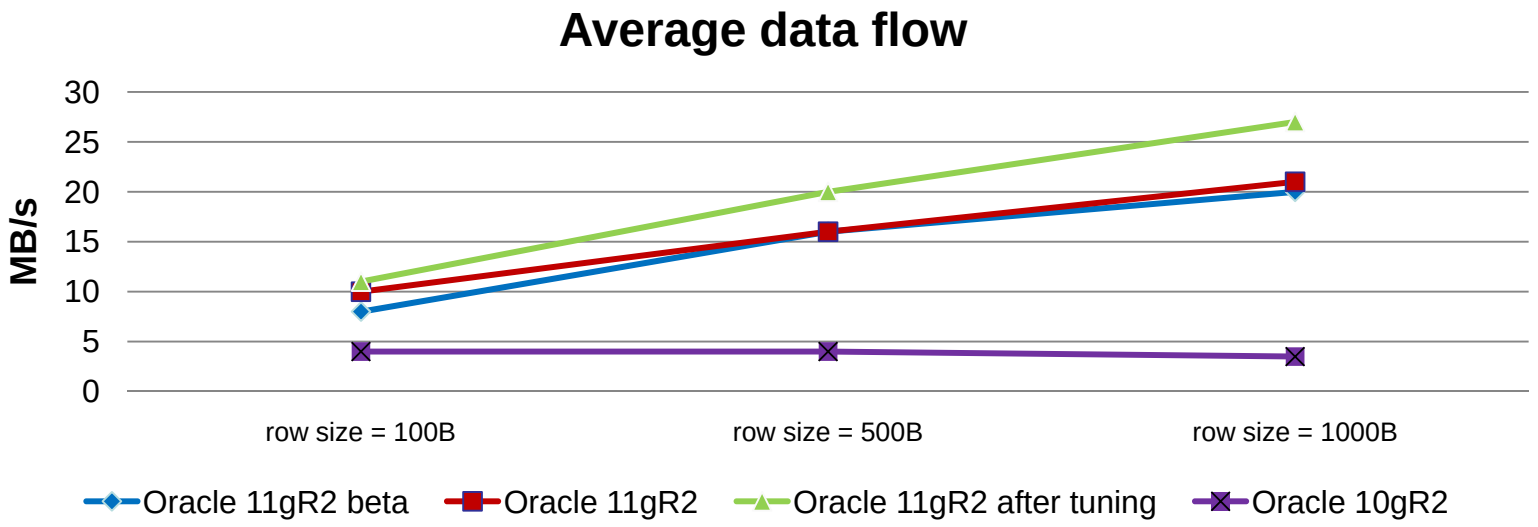
CERN and Oracle 11gR2

Streams 11gR2 tests

- ✓ Tests coverage
  - ✓ Combined Capture and Apply (beta and production)
    - ✓ Functional and performance tests
  - ✓ Automatic Split and Merge (beta)
    - ✓ Functional test
  - ✓ Compare and Converge (beta)
    - ✓ Functional and performance tests

Streams Throughput & Data Flow





ASM – Test Scope

11g R2 beta1 and beta2 (on Linux x64_86)

- ✓ Functionality
  - ✓ Placing clusterware files on ASM
  - ✓ asmcmd tool new features
- ✓ Stability
  - ✓ Different HW failure scenarios while accessing a disk group configured with normal redundancy
  - ✓ Re-location/replacement of clusterware files
  - ✓ ASM fast mirror resync (will be very useful in production)
    - ✓ Already in 11gR1
- ✓ Performance
  - ✓ Disk group rebalancing (comparing 11gR2 behaviour with measured values for our current 10g and 11gR1 setups)
  - ✓ Intelligent Data Placement
    - ✓ important data stays on the external part of the disk

ASM – Test Results

- Functionality and stability tests did not reveal any obvious issues
  - Possibility to store clusterware files on ASM is very important for our service
- Sequential and random IO tests confirmed that ASM offers performance close to raw devices
- Rebalancing tests showed big performance improvements (a factor four gain)
  - Excessive re-partnering and incorrect estimates of the data to be relocated seem to be fixed

(Test results reported to ASM product manager)

ASM Cluster File System – Test Scope

- ACFS tests performed using beta1 software mainly

- ✓ Functionality
  - ✓ Command-line tools: asmcmd, acfsutil and standard Linux file system interfaces with ACFS
  - ✓ Creation and usage of a huge ADVM volume
  - ✓ ACFS snapshots
- ✓ Performance
  - ✓ ACFS vs ext3

ACFS – Test Results

- ACFS is a very interesting solution for storing Oracle logs, trace and export files in a RAC environment
- Comparative performance tests between ACFS and ext3 were very encouraging:
  - ACFS offers much better read and write performance
  - Some performance issues noticed during file deletion test
    - Looked at by ACFS development team
- ACFS proved to be robust and mature

(Test results reported to ACFS product manager)

Active Data Guard

- Data Guard: foundation of Oracle Maximum Availability Architecture best practices (together with RAC and Streams)
  - Software to create and keep in sync one (or more) standby databases
- All critical DBs in Physics Databases have a standby DB (but not available for continuous read-only)
  - LHC experiments online and offline RACs
- In 11G Active Data Guard can be opened for continuous read-only access
  - Requested by all experiment's online groups as all monitoring/analysis can be run there

Active Data Guard

- ✓ Tests
  - ✓ Functional tests
  - ✓ Long term stability
  - ✓ Performance
- ✓ Smooth running over months achieved. No performance issues encountered
- ✓ Failover tests and human error recoveries scenarios also tested

Data Life Cycle

- Several Physics applications generate very large data sets and have the need to archive data
  - Performance-based: online data more frequently accessed
  - Capacity based: Old data can be read-only, rarely accessed, in some cases can be put online ‘on demand’

- Technologies:
  - Oracle Partitioning: mainly range partitioning by time
  - Application-centric: tables split and metadata maintained by the application
  - Archive DB initiative: offline old partitions/chunks of data in a separate ‘archive DB’
    - Archive DB HW is focused on storage capacity more than throughput

Oracle Compression for Physics Applications

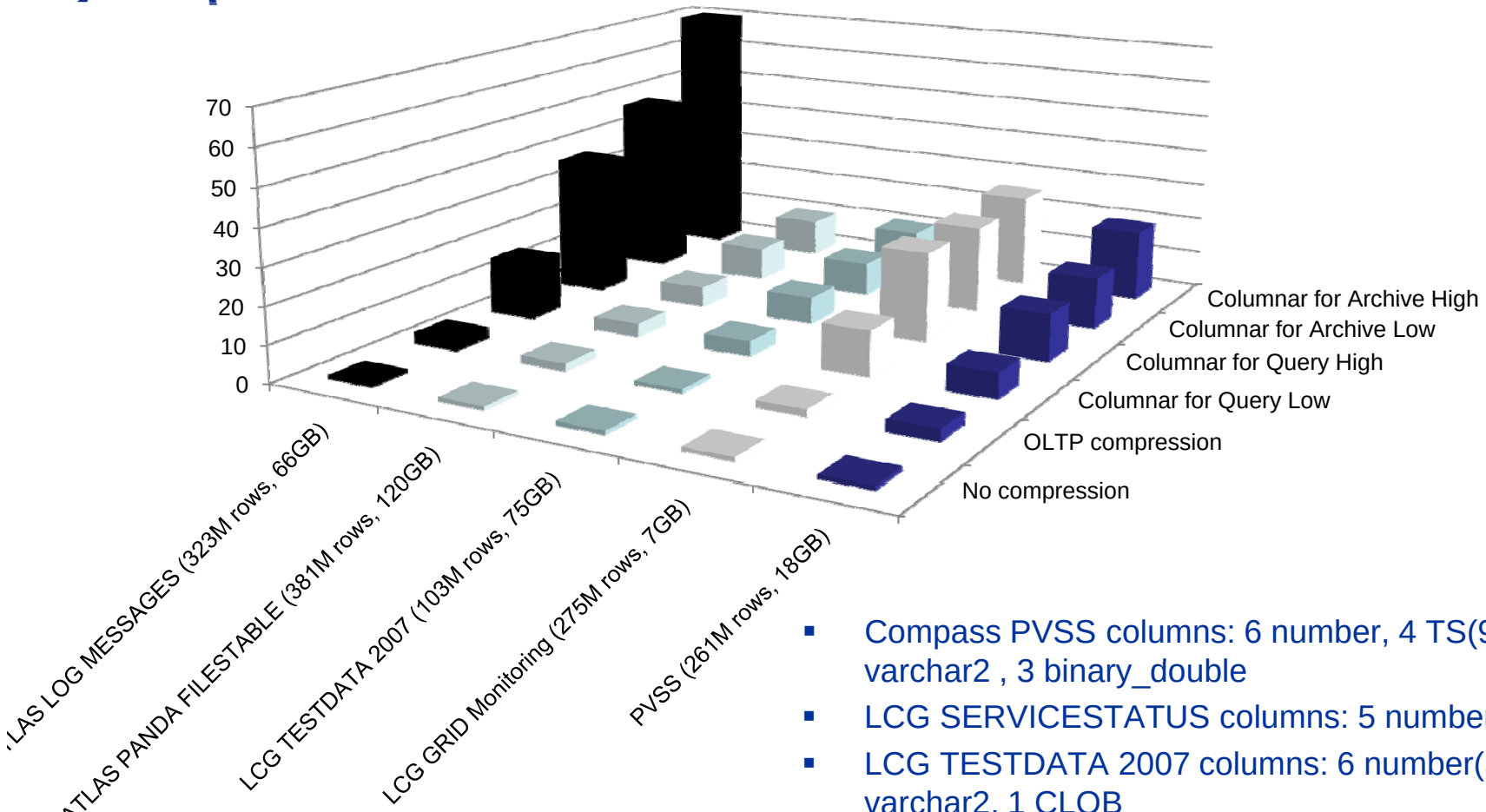
- Compression for Physics DBs:
  - DB volumes expected to grow ~30 TB/year for Oracle data
  - Large volumes of data becomes ‘read only’
    - In some cases datasets kept ‘silent’ for a long time
- “10g compression” (for direct load) already used at CERN
  - Advanced compression has higher compression factors
  - Hybrid columnar compression 10-50 times compression
  - Compression for OLTP considered for transactional applications
  - Compression of Secure files also very interesting
- Some datawarehouse-like applications being developed
  - Although currently most applications are OLTP-like

Advanced Compression Tests

- Representative subsets of data from production exported to Exadata V2 Machine:
  - Applications: PVSS (slow control system for the detector and accelerator)
  - GRID monitoring applications
  - File transfer applications (PANDA)
  - Log application for ATLAS
  - Exadata machine accessed remotely to Reading, UK for a 2-week test
- Tests focused on :
  - OLTP and Hybrid columnar compression factors
  - Query speedup

Advanced Compression Option 11gR2

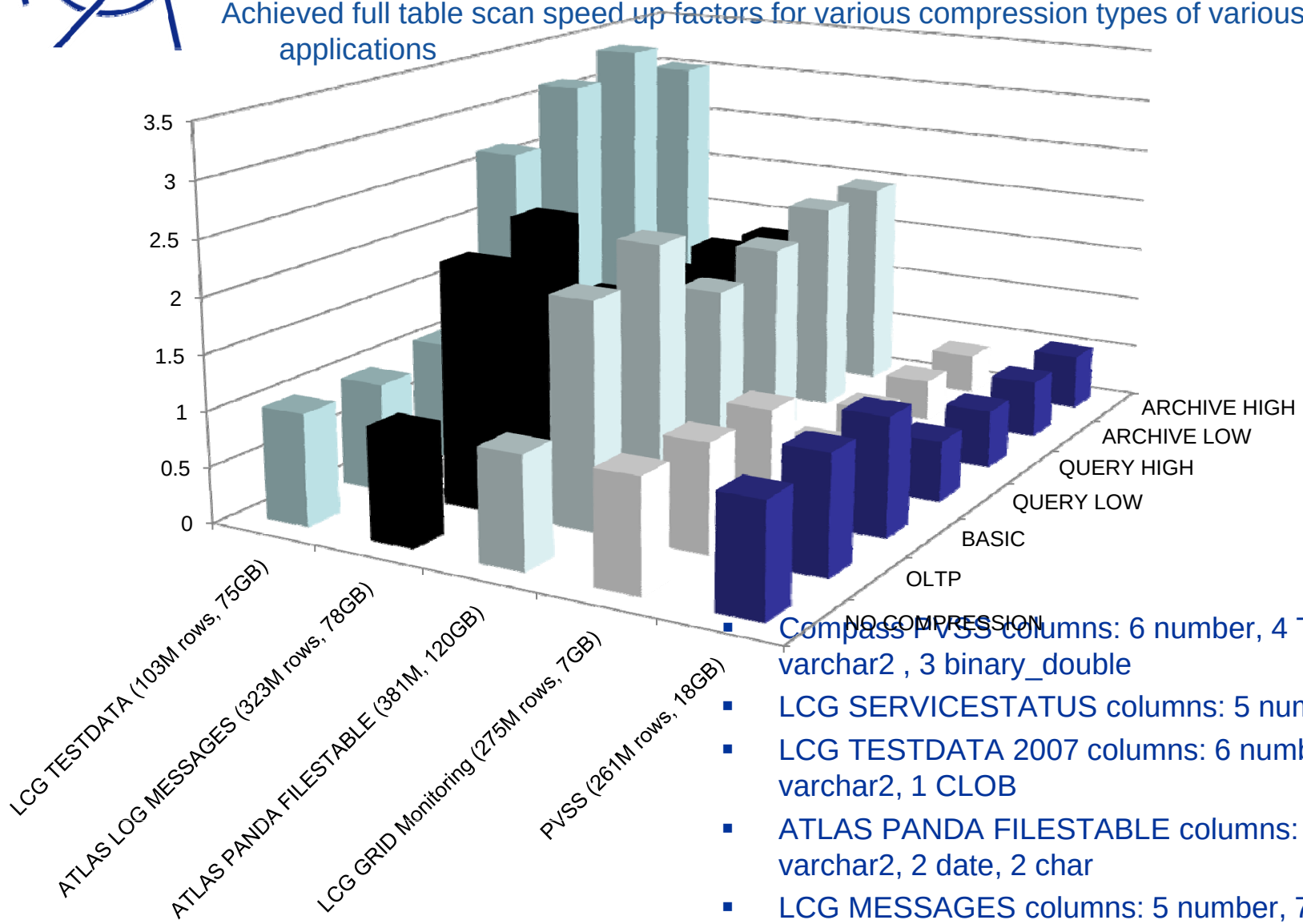
Achieved compression factors for various compression types of various physics applications



- Compass PVSS columns: 6 number, 4 TS(9) , 5 varchar2 , 3 binary_double
- LCG SERVICESTATUS columns: 5 number
- LCG TESTDATA 2007 columns: 6 number(38), 1 varchar2, 1 CLOB
- ATLAS PANDA FILESTABLE columns: 3 number, 12 varchar2, 2 date, 2 char
- LCG MESSAGES columns: 5 number, 7 varchar2, 1 TS

Advanced Compression Option 11gR2

Achieved full table scan speed up factors for various compression types of various physics applications



- Compass PVSS columns: 6 number, 4 TS(9) , 5 varchar2 , 3 binary_double
- LCG SERVICESTATUS columns: 5 number
- LCG TESTDATA 2007 columns: 6 number(38), 1 varchar2, 1 CLOB
- ATLAS PANDA FILESTABLE columns: 3 number, 12 varchar2, 2 date, 2 char
- LCG MESSAGES columns: 5 number, 7 varchar2, 1 TS

Summary

- Physics database services run on Oracle 10g RAC and ASM
- DB service architecture follows Oracle Maximum Availability Architecture best practices
- Beta testing has allowed us to study and test new 11g features of interest for the services
  - ASM, ACFS and clusterware improvements
  - Streams new features
  - Active Data Guard
  - Advanced Compression Option
  - Real Application Testing
- Need to complete all tests on production release and study the integration of 11gR2 in our set-up
  - Priority: top-quality service for the LHC

Conclusions

- We have set up a world-wide distributed database infrastructure for LHC Computing Grid
  - Oracle RAC with ASM key DB services at Tier0 & Tier1s on 10.2.0.4
  - Oracle Streams for detector conditions: key for data (re-)processing
  - Oracle Data Guard for data protection: critical databases

- The enormous challenges of providing robust, flexible and scalable DB services to the LHC experiments have been met using a combination of Oracle technology and operating procedures

- Close collaboration with Oracle at different levels has been essential