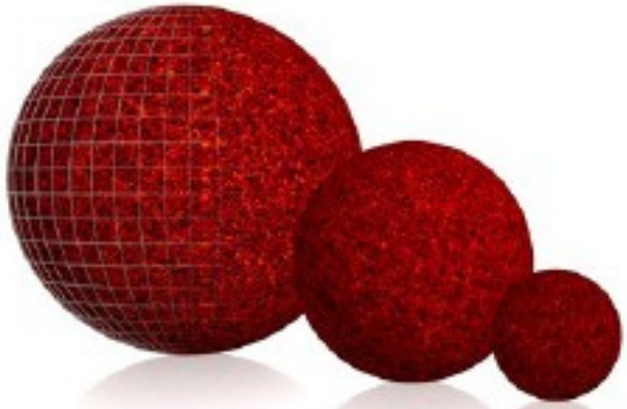
Overview of Database Technologies

Computing and Astroparticle Physics
2nd ASPERA Workshop



Eric Grancher
eric.grancher@cern.ch
CERN IT department

https://edms.cern.ch/document/1146858/1

Outline

- Experience at CERN
  - Current usage and deployment
  - Replication
  - Lessons from last 15 years
- Oracle and MySQL
- Trends
  - Flash
  - Compression
  - Open source “relational” databases
  - NoSQL databases

http://www.wordle.net/

Database SERVICES

At the heart of CERN, LHC and Experiment Operations

CERN IT Department

http://cern.ch/it-dep/db/

IT/DB GROUP

Experiment Offline Databases
9 Production DBs, 7 Integration DBs
5 Tests DBs, 7 Standby DBs

Streams

Experiment Online Databases
4 Production DBs

Data

CASTOR
CERN Advanced STORage manager
21 Production DBs
4 Development DBs

RAW Data

Middleware
73 Application Servers

Administrative/IT/ Engineering Databases
17 Production DBs
11 Development DBs
4 Ref/Test DBs

Accelerators ACC
12 Production DBs

LHC Operations

Streams

Tier-1 Centres
dedicated 10 Gbit/s links
NDGF
GridKa
BNL
RAL
ASGC
SARA-NIKHEF
PIC
FNAL
CCIN2P3
INFN - CNAF
TRIUMF

LHC Experiments
LHCb
CMS
ATLAS
ALICE

PARTNERS:
CERN openlab
ORACLE

CERN

Current status on February 2010

CERN databases in numbers

- CERN databases services
  - ~130 databases, most of them database clusters (Oracle RAC technology RAC, 2 – 6 nodes)
  - Currently over 3000 disk spindles providing more than ~3PB raw disk space (NAS and SAN)
  - MySQL service started (for Drupal)
- Some notable databases at CERN
  - Experiments' databases – 14 production databases
    - Currently between 1 and 12 TB in size
    - Expected growth between 1 and 10 TB / year
  - LHC accelerator logging database (ACCLOG) – ~70 TB, >2.10^{12} rows, expected growth up to 35(+35) TB / year
  - ... Several more DBs in the 1-2 TB range

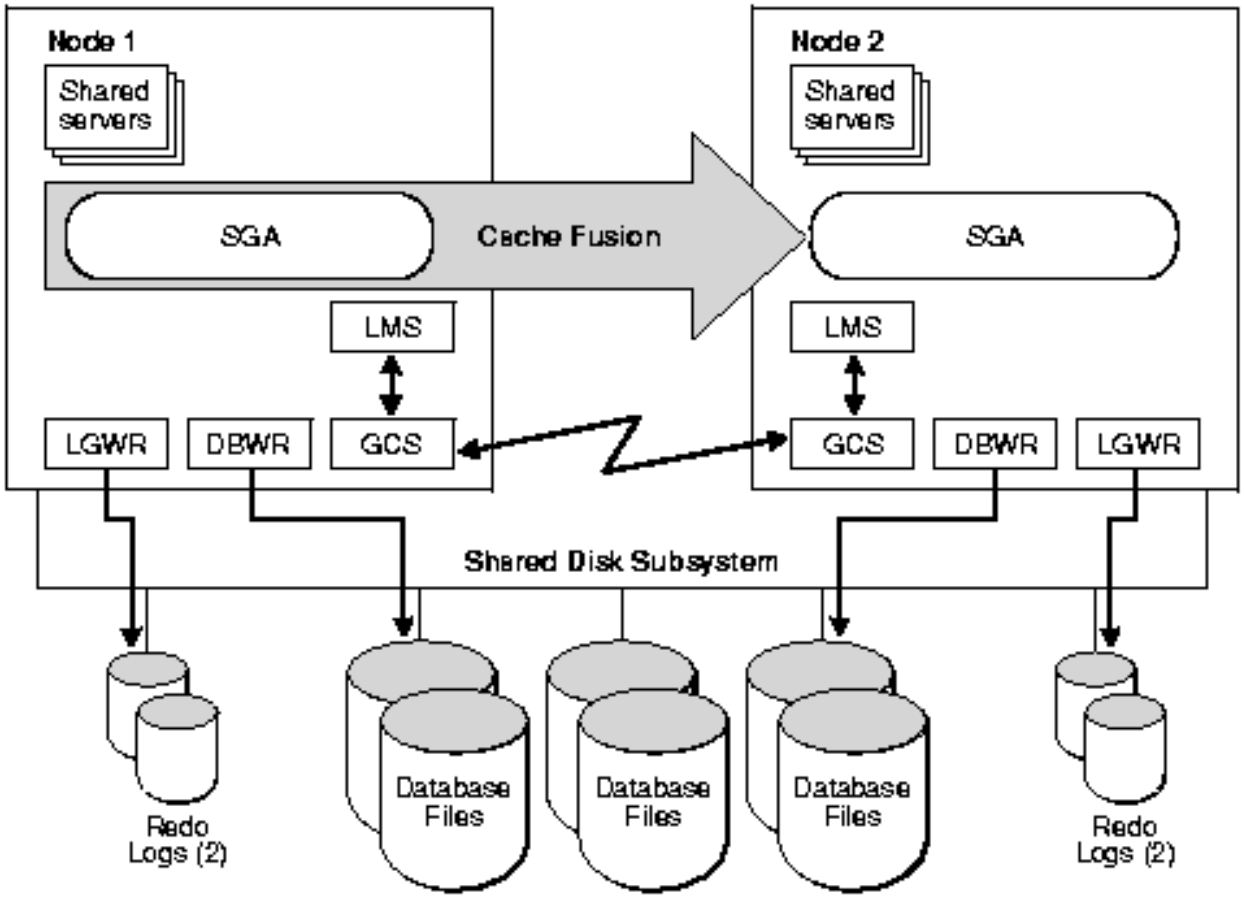
original slide by Luca Canali

Key role for LHC physics data processing

- Online acquisition, offline production, data (re)processing, data distribution, analysis
  - SCADA, conditions, geometry, alignment, calibration, file bookkeeping, file transfers, etc..
- Grid Infrastructure and Operation services
  - Monitoring, Dashboards, User-role management, ..
- Data Management Services
  - File catalogues, file transfers and storage management, …
- Metadata and transaction processing for custom tape-based storage system of physics data
- Accelerator control and logging systems
- AMS as well: data/mc production bookkeeping and slow control data

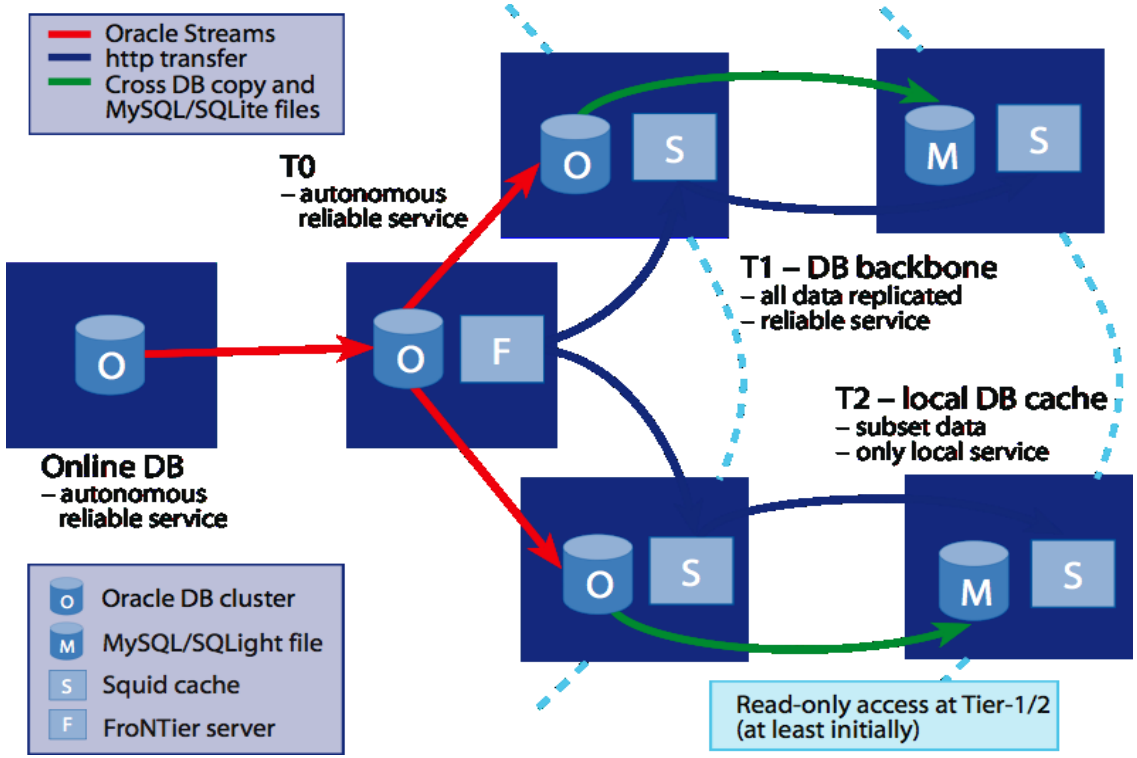
original slide by Luca Canali

Oracle Real Application Cluster

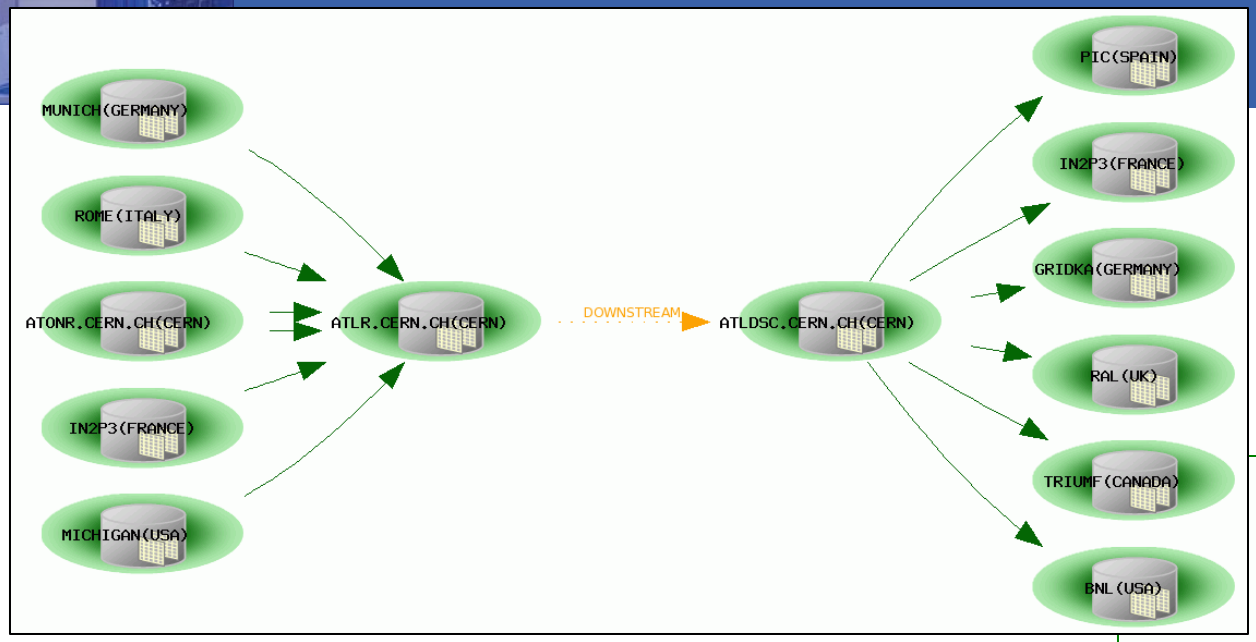


from Oracle9i Real Application Clusters Deployment and Performance

- Worldwide distribution of experimental physics data using Oracle Streams

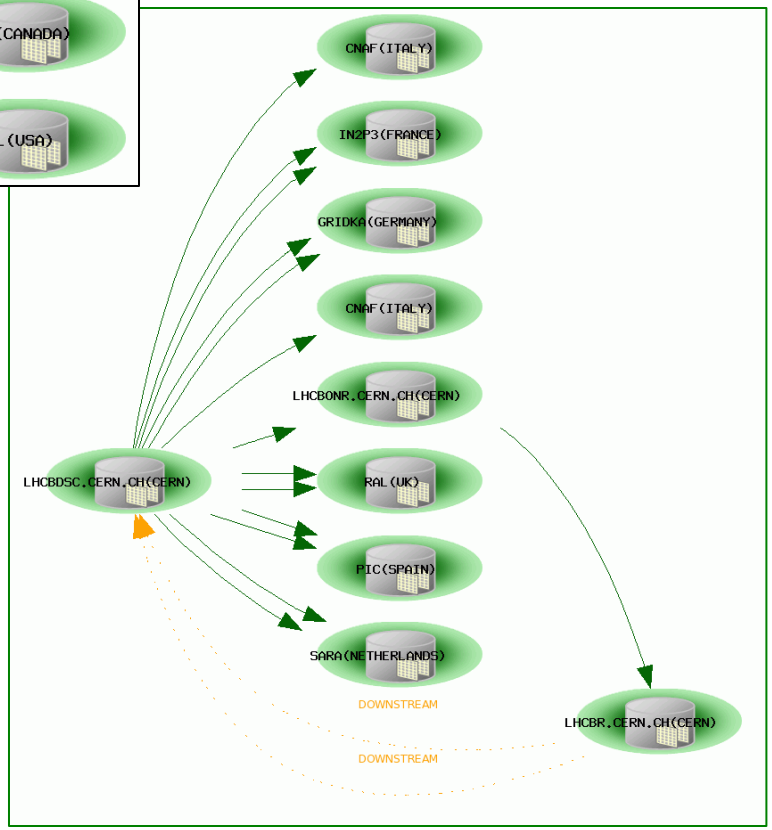


  - Huge effort, successful outcome

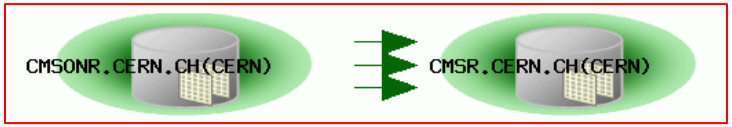


ATLAS

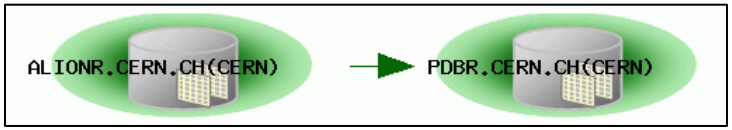
LHCb



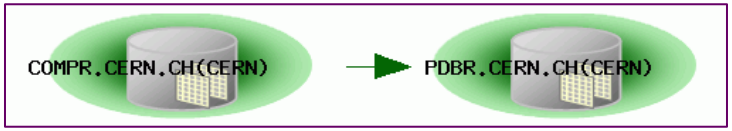
CMS



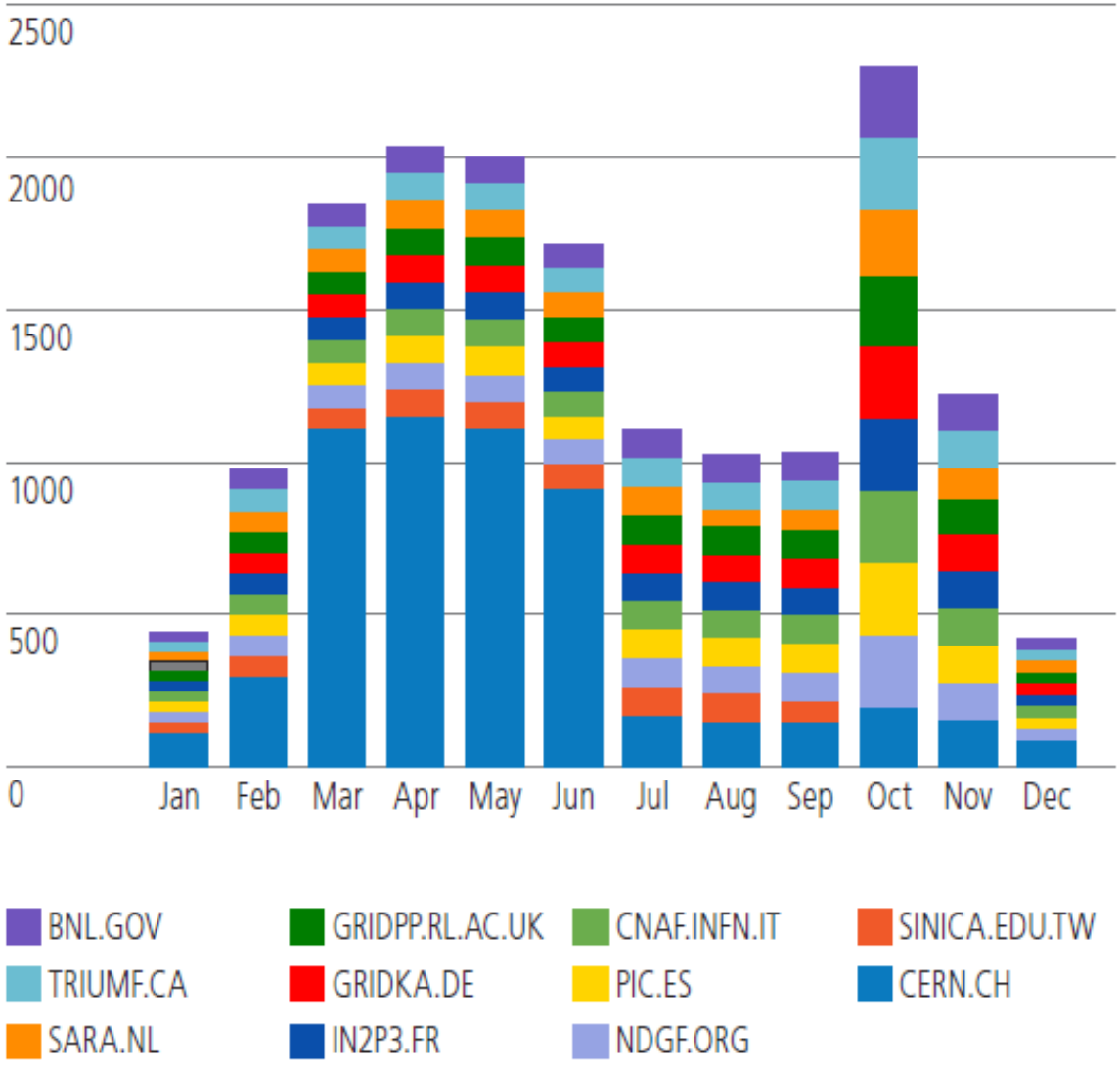
ALICE



COMPASS



Number of Logical Change Records (LCRs) in Millions, per Month, by Tier-1 Site



Replication rate for conditions data from the ATLAS experiment to the different WLCG Tier-1 sites in 2010

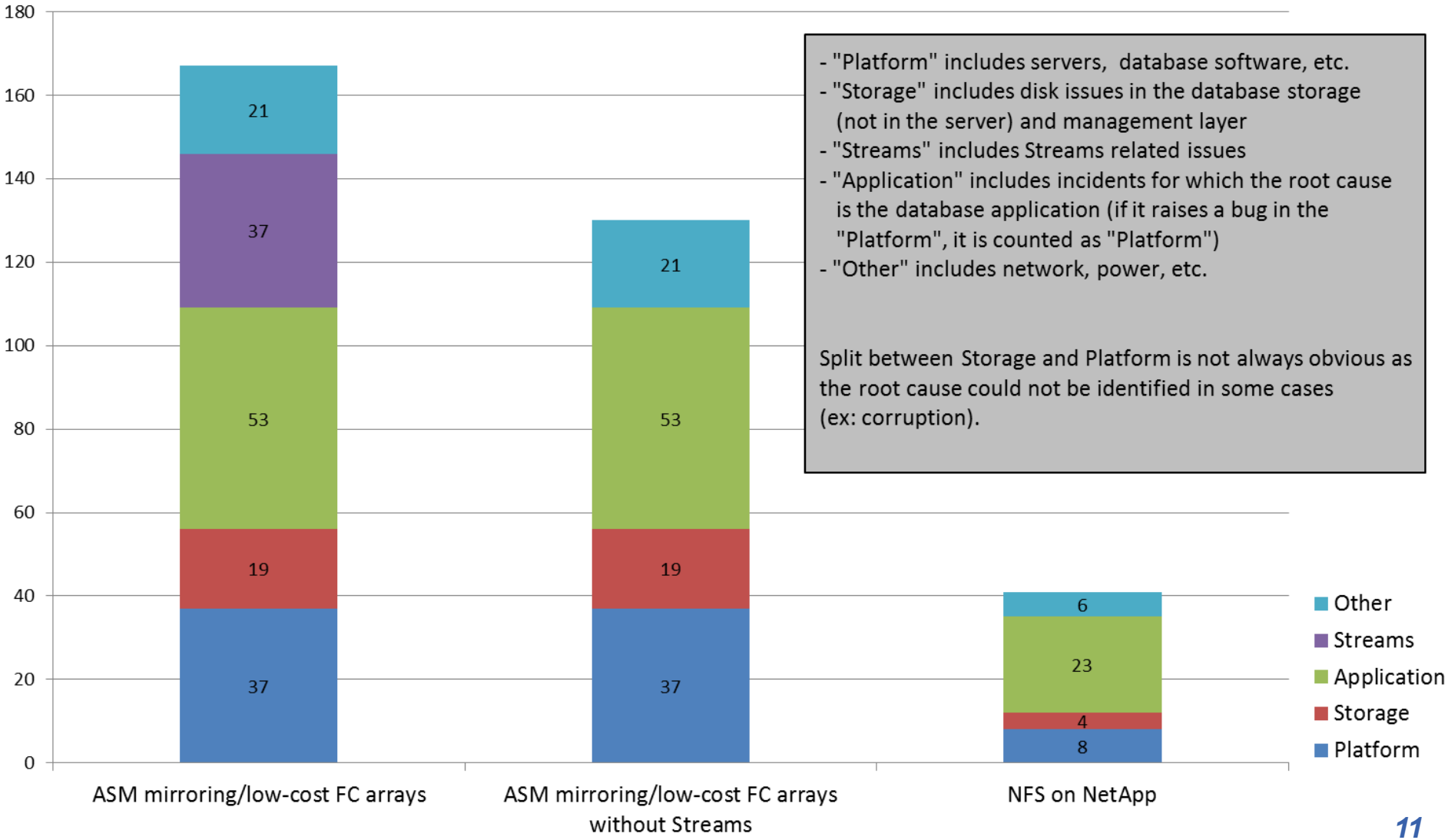
Lessons learnt

- Aiming for high-availability is (often) adding complexity… and complexity is the enemy of availability
- Scalability can be achieved with Oracle Real Application Cluster (150k entries/s for PVSS)
- Database / application instrumentation is key for understanding/improving performance
- NFS/D-NFS/pNFS are solutions to be considered for stability and scalability (very positive experience with NetApp, snapshots, scrubbing, etc.)
- Database independence is very complex if performance is required
- Hiding IO errors from the database leaves the database handle what it is best at (transactions, query optimisation, coherency, etc.)

Incidents review

CERN IT Department

Breakdown of incidents
(based on C5 incidents, January 1st 2010-March 4th 2011)



- "Platform" includes servers, database software, etc.
- "Storage" includes disk issues in the database storage (not in the server) and management layer
- "Streams" includes Streams related issues
- "Application" includes incidents for which the root cause is the database application (if it raises a bug in the "Platform", it is counted as "Platform")
- "Other" includes network, power, etc.

Split between Storage and Platform is not always obvious as the root cause could not be identified in some cases (ex: corruption).

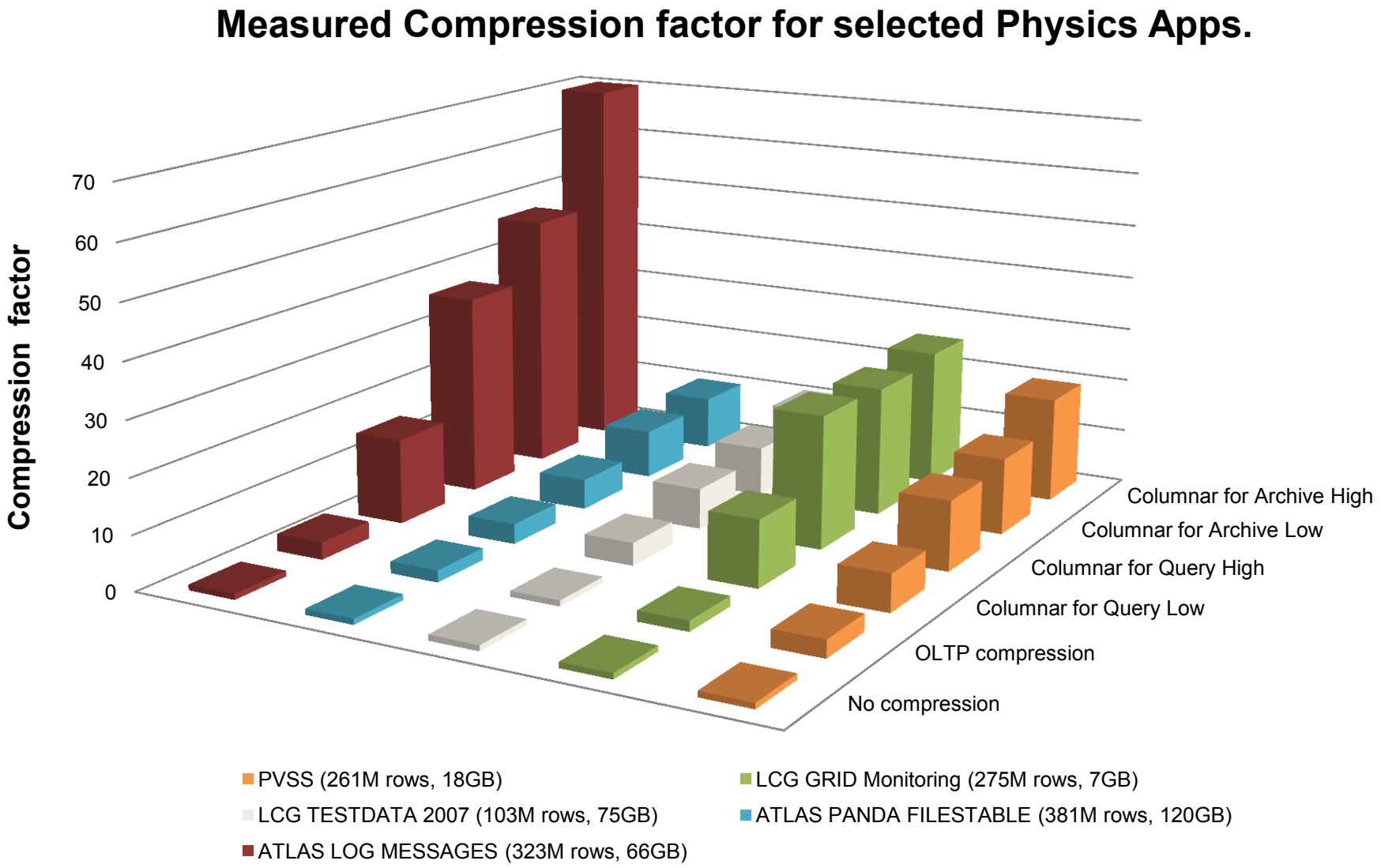
Evolution

- Flash
- Large memory systems
- Compression
- Open source “relational” databases
- NoSQL databases

Flash, memory and compression

- Flash changes the picture in the database area IO
  - Sizing for IO Operations Per Second
  - Usage of fast disks for high number of IOPS and latency
- Large amount of memory
  - Enables consolidation and virtualisation (less nodes)
  - Some databases fully in memory
- Compression is gaining momentum for databases
  - For example Oracle's hybrid columnar compression
  - Tiering of storage

Exadata Hybrid Columnar Compression on Oracle 11gR2



Local analysis of data

- “Big Data”, analysis of large amount of data in reasonable of time
- Goole MapReduce, Apache Hadoop implementation
- Oracle Exadata
  - Storage cells perform some of the operations locally (Smart Scans, storage index, column filtering, etc.)
- Greenplum
  - shared-nothing massively parallel processing architecture for Business Intelligence and analytical processing
- Important direction for *at least* the first level of selection

Open source relational databases

- MySQL and PostgreSQL
- Some components are not “free”, replacements exist (for example for hot backup Percona XtraBackup)
- MySQL default for many applications (OpenNebula, Drupal, etc.)
- PostgreSQL has a strong backend with Oracle-like features (stored procedures, write-ahead logging, “standby”, etc.)

NoSQL databases 1/2

- Not about SQL limitations/issues
  - about scalability
  - Big Data scale-out
- Feature-rich SQL engines are complex, lead to some unpredictability
- Workshop at CERN (needs and experience)
- ATLAS Computing Technical Interchange Meeting

NoSQL databases 2/2

- Lot of differences
  - Data models (Key-Value, Column Family, Key-Document, Graph, etc.)
  - Schema free storage
  - Queries complexity
  - Mostly not ACID (Atomicity, Consistency, Isolation and Durability), data durability relaxed compared to traditional SQL engines
  - Performance with sharding
  - Compromise on consistency to keep availability and partition tolerance
- -> application is at the core, evolution?

Summary

- Oracle
  - Critical component for LHC accelerator and physics data processing
  - Scalable and stable, including data replication
- CERN central services run on Oracle, for which we have components and experience to build high availability, guaranteed data, scalability
- MySQL is being introduced
  - Nice features and light, lacks some scalability and High-Availability features for solid service
- NoSQL is being considered
  - Ecosystem still in infancy (architecture, designs and interfaces subject to change!)

Oracle

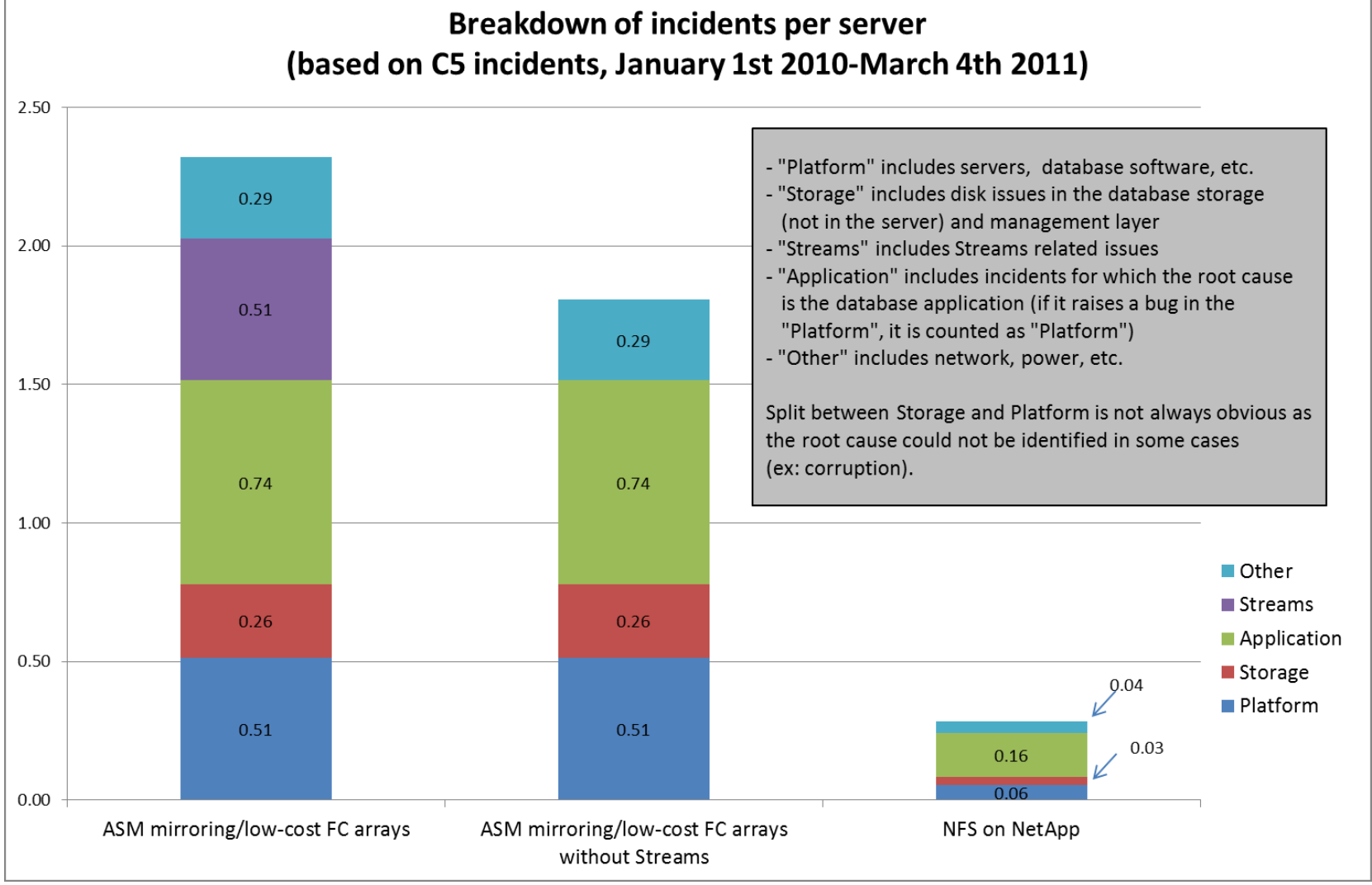
- EU director for research, Monica Marinucci
- Strong collaboration with CERN and universities
- Well known solution, easy to find database administrators and developers, training available
- Support and licensing

References

- NoSQL ecosystem, http://www.aosabook.org/en/nosql.html
- Database workshop at CERN https://indico.cern.ch/conferenceDisplay.py?confId=130874 and ATLAS Computing Technical Interchange Meeting https://indico.cern.ch/event/132486
- Eva Dafonte Pérez, UKOUG 2009 “Worldwide distribution of experimental physics data using Oracle Streams”
- Luca Canali, CERN IT-DB Deployment, Status, Outlook http://canali.web.cern.ch/canali/docs/CERN_IT-DB_deployment_GAIA_Workshop_March2011.pptx
- CERN openlab, http://cern.ch/openlab/
- CAP theorem, http://portal.acm.org/citation.cfm?id=564601
- ACID, http://portal.acm.org/citation.cfm?id=291

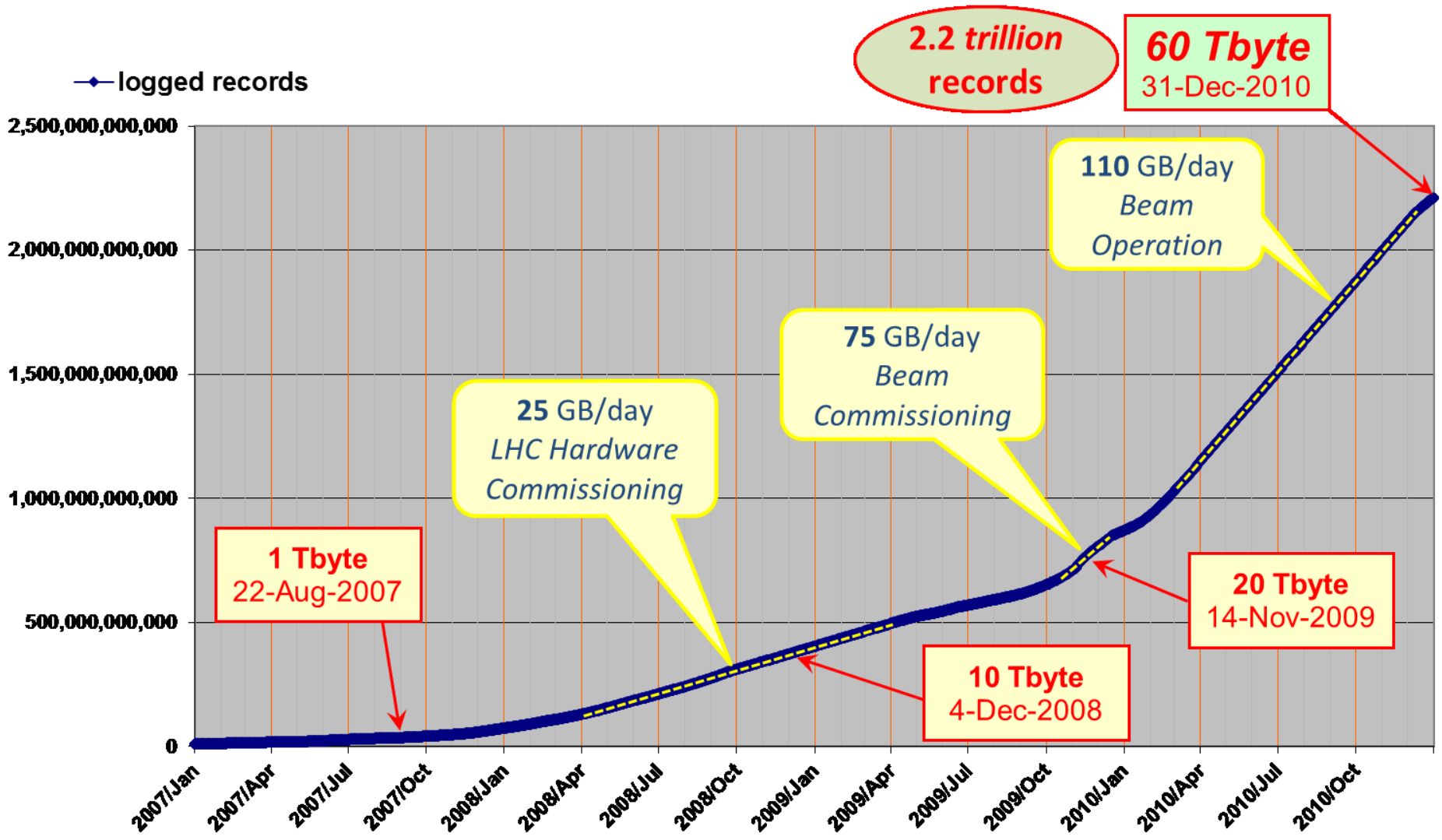
Backup slides

Incidents review



LHC logging service, >2.10^{12}

CERN IT Department



graph by Ronny Billen