



5 December 2013, Eric Grancher
CERN, head of database services

CERN and Oracle, a 30-year collaboration

Outlook

- CERN
- History of using Oracle
- Current usage
- Collaboration
- Why using Oracle in our research environment?

CERN

- **European Organization for Nuclear Research**
 - Founded in 1954
 - Research: Seeking and finding answers to questions about the Universe
 - Technology, International collaboration, Education



Twenty Member States

Austria, Belgium, Bulgaria, Czech Republic, Denmark, Finland, France, Germany, Greece, Italy, Hungary, Netherlands, Norway, Poland, Portugal, Slovakia, Spain, Sweden, Switzerland, United Kingdom

Seven Observer States

European Commission, USA, Russian Federation, India, Japan, Turkey, UNESCO

Associate Member States Candidate State

Israel, Serbia

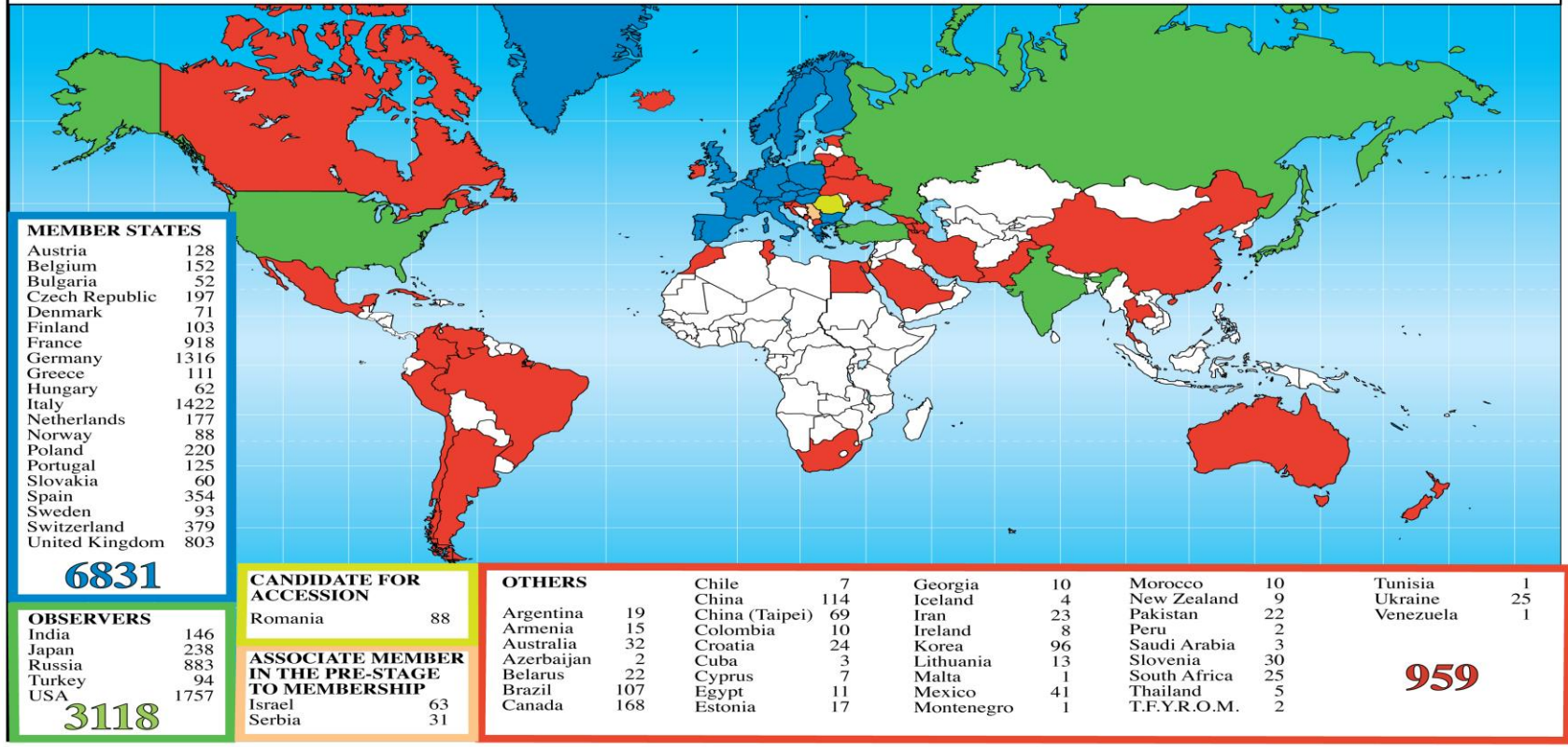
Romania

People

~2400 Staff, ~900 Students, post-docs and undergraduates, ~9000 Users, ~2000 Contractors

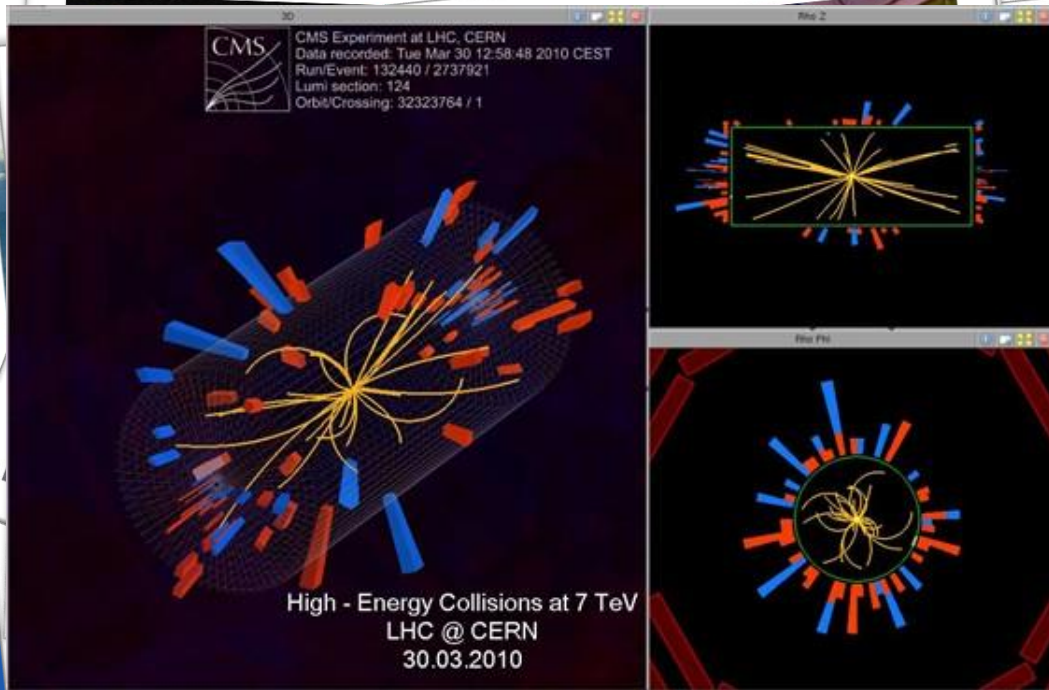
A European Laboratory with Global reach

Distribution of All CERN Users by Location of Institute on 14 January 2013



LHC

- The **largest** particle accelerator & detectors



17 miles (27km) long tunnel

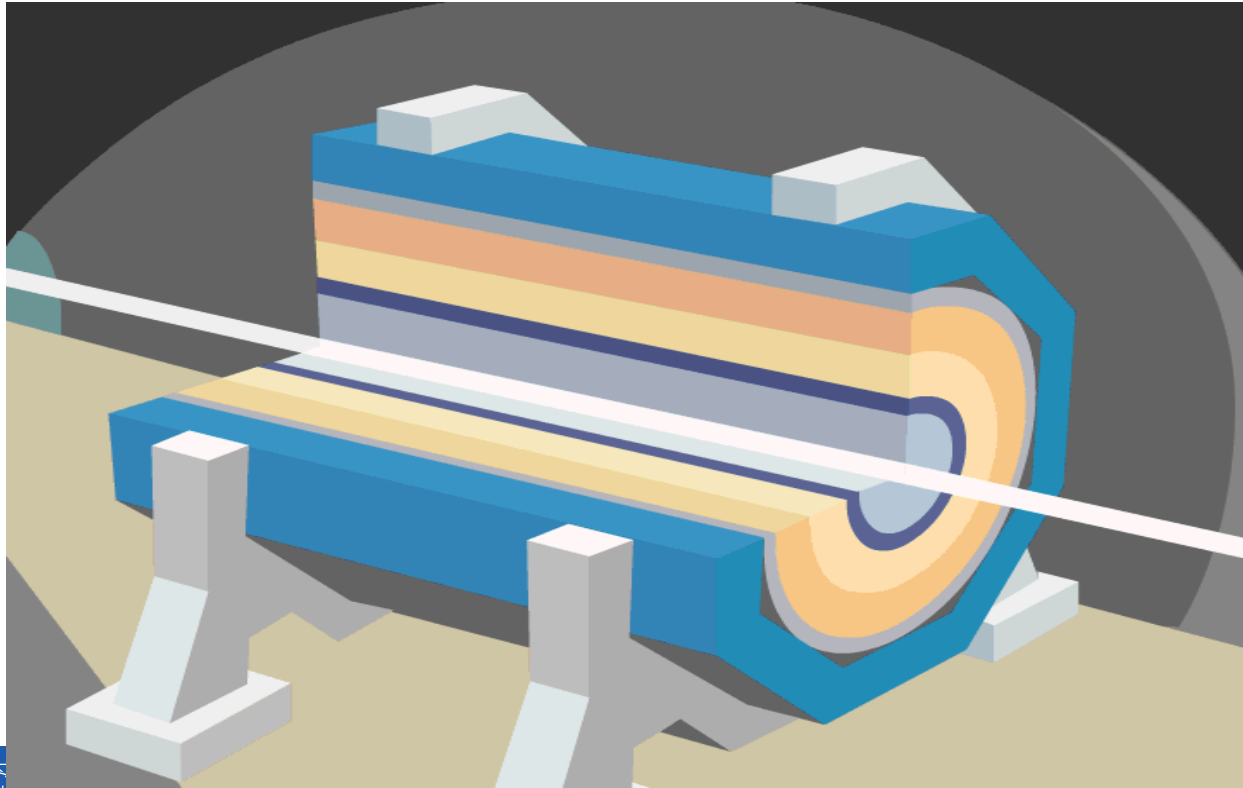
Thousands of superconducting magnets

Coldest place in the Universe: *1.9 degrees kelvin*

Ultra vacuum: 10x emptier than on the Moon

600 million collisions per second / analysis is like finding a needle in 20 million haystacks

Events at LHC

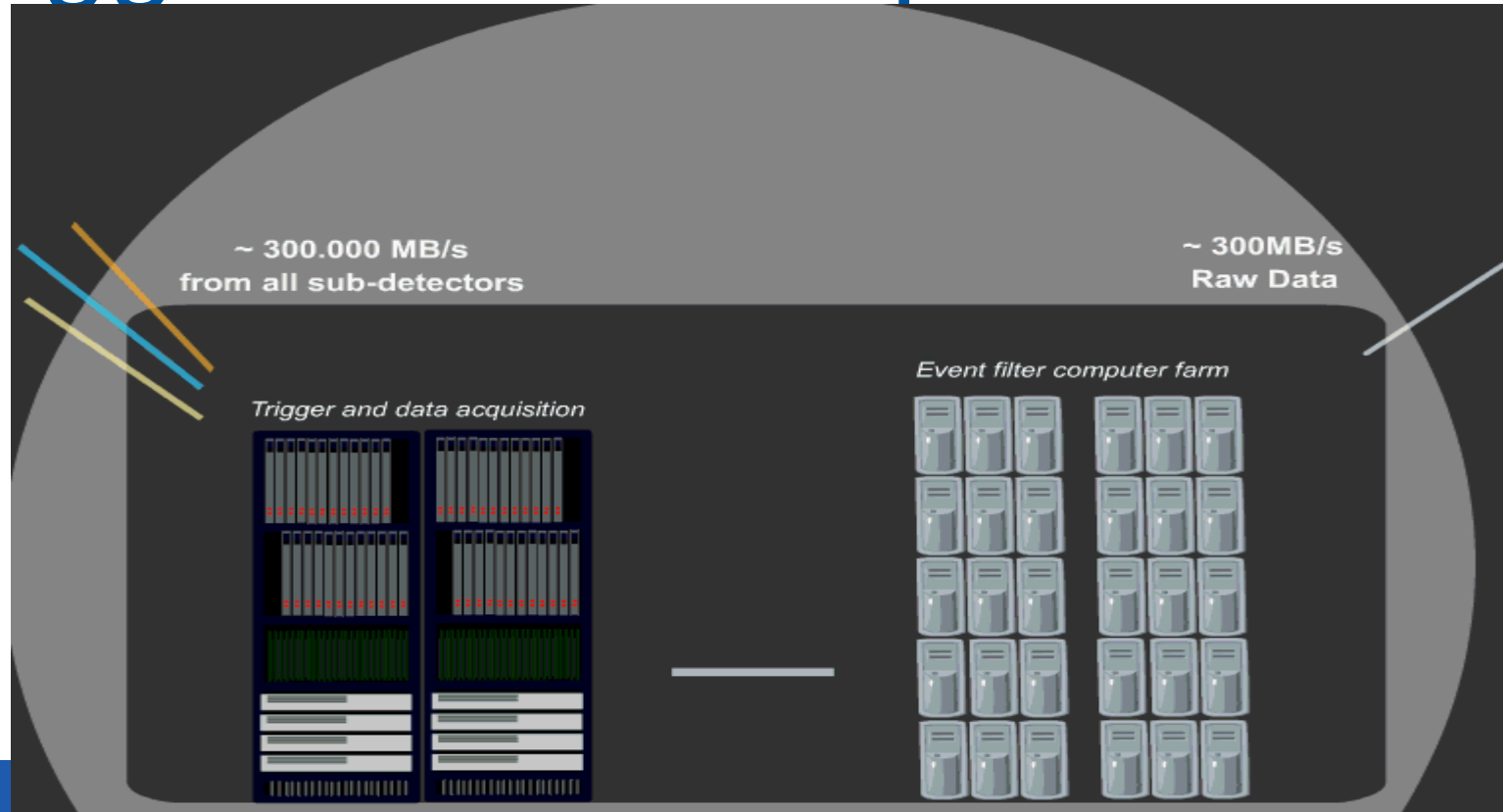


Luminosity :
 $10^{34} \text{cm}^{-2} \text{s}^{-1}$

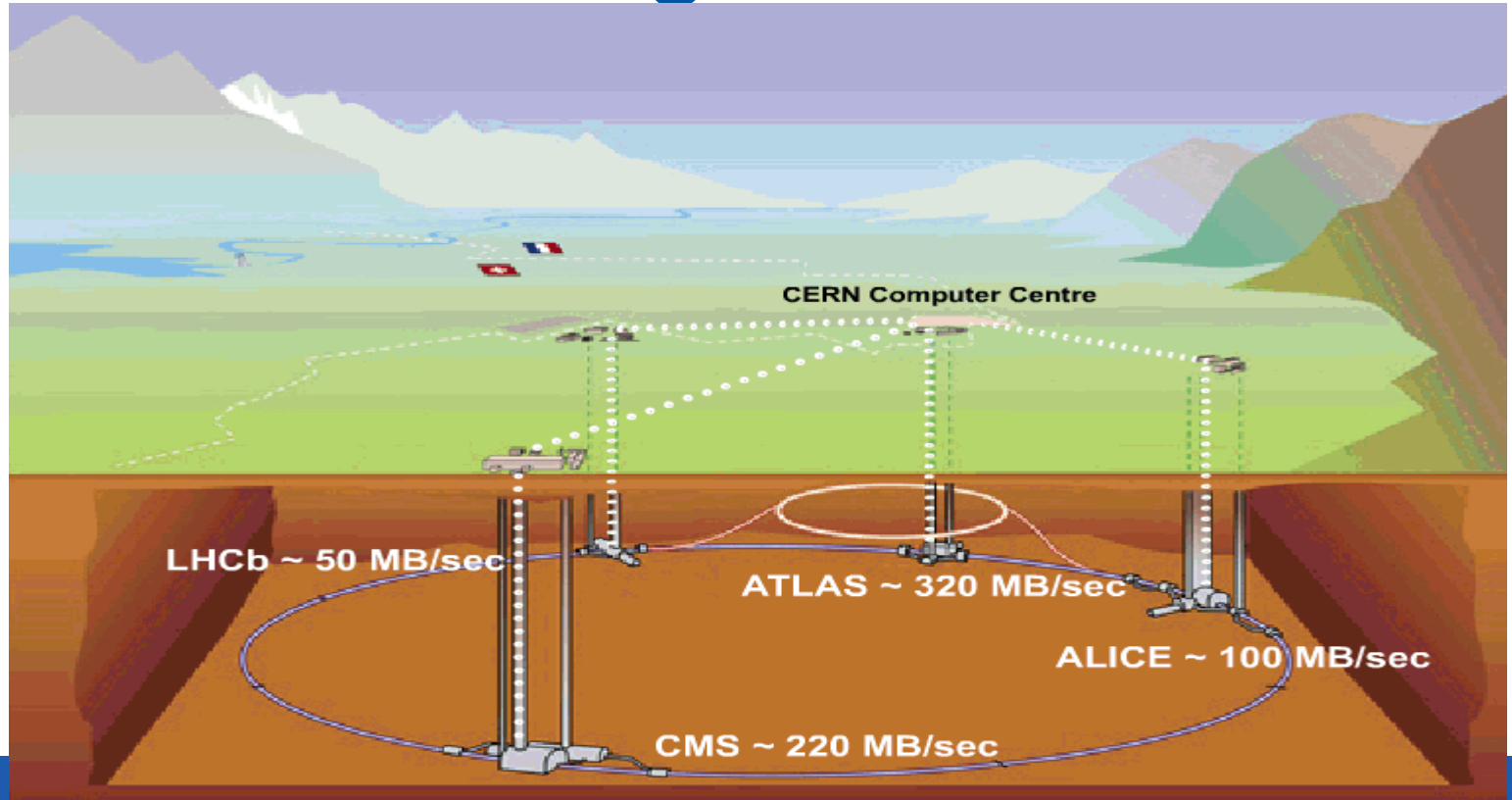
40 MHz – every 25 ns

20 events overlaying

Trigger & Data Acquisition



Data Recording



World's largest computing grid - WLCG



- 1 PB raw data per second before filtering / >20 PB of new data annually
- 68,889 physical CPUs / 305,935 logical CPUs
- 157 computer centres around the world

Oracle at CERN, 1982 accelerator control

<http://cds.cern.ch/record/443114?ln=en>

ISR LIBRARY
26.4.1982

CERN LIBRARIES, GENEVA

SCAN-0009042

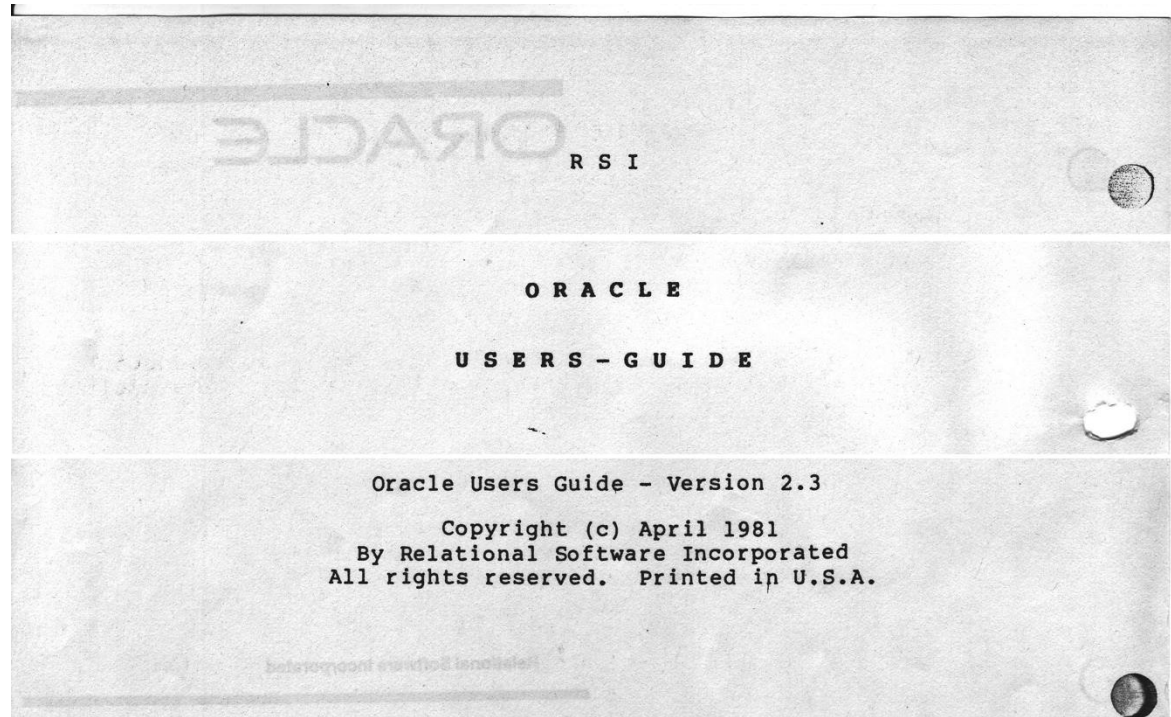
LEP NOTE 374
26.4.1982

ORACLE - the data base management system for LEP

J.Schinzel

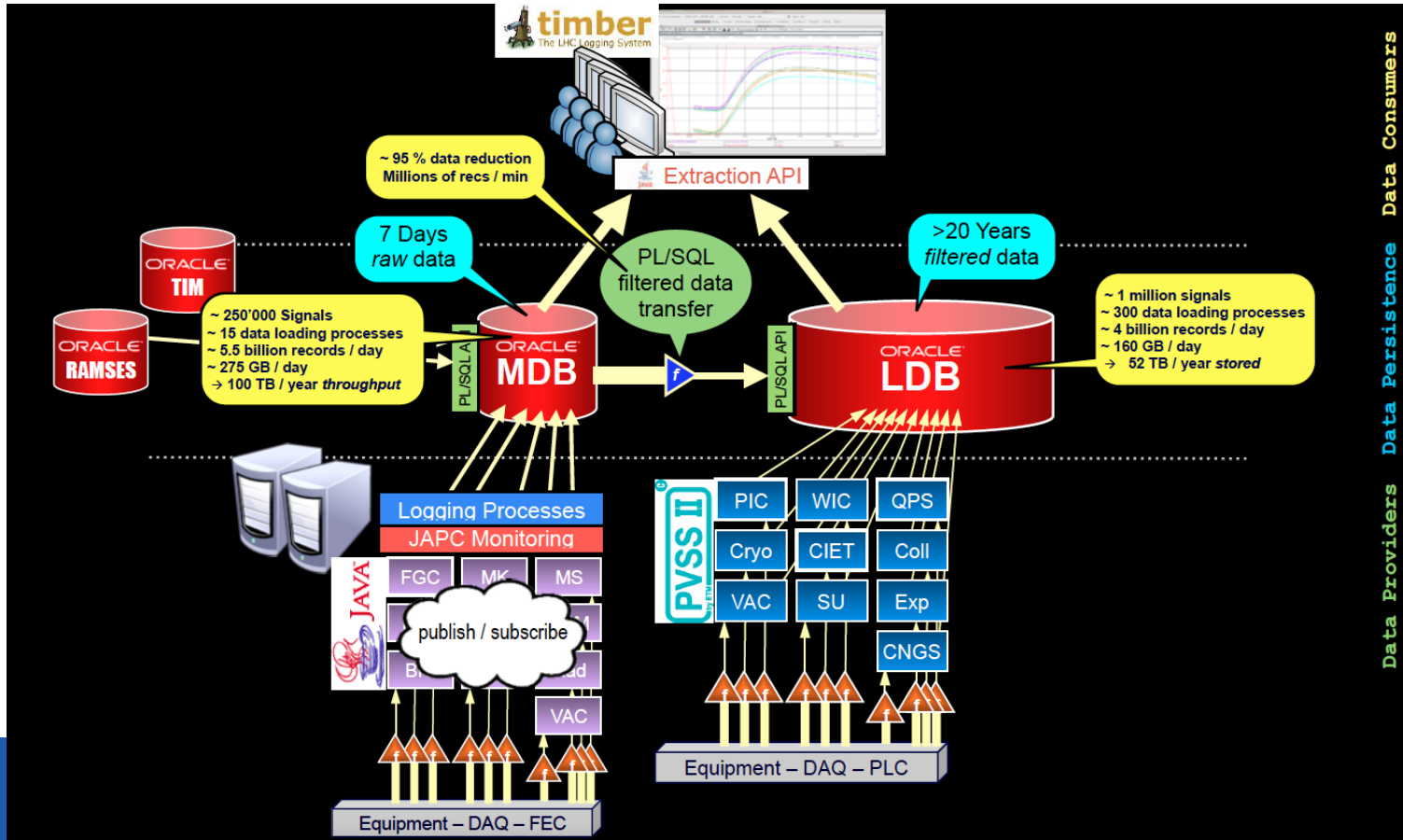
Following the decision that an efficient data base system is required for the LEP project and that the systems at present in use at CERN are not adequate, an enquiry into possible data base management systems on the market was launched early this year.

Oracle at CERN, version 2.3



Credit: N. Segura
Chinchilla

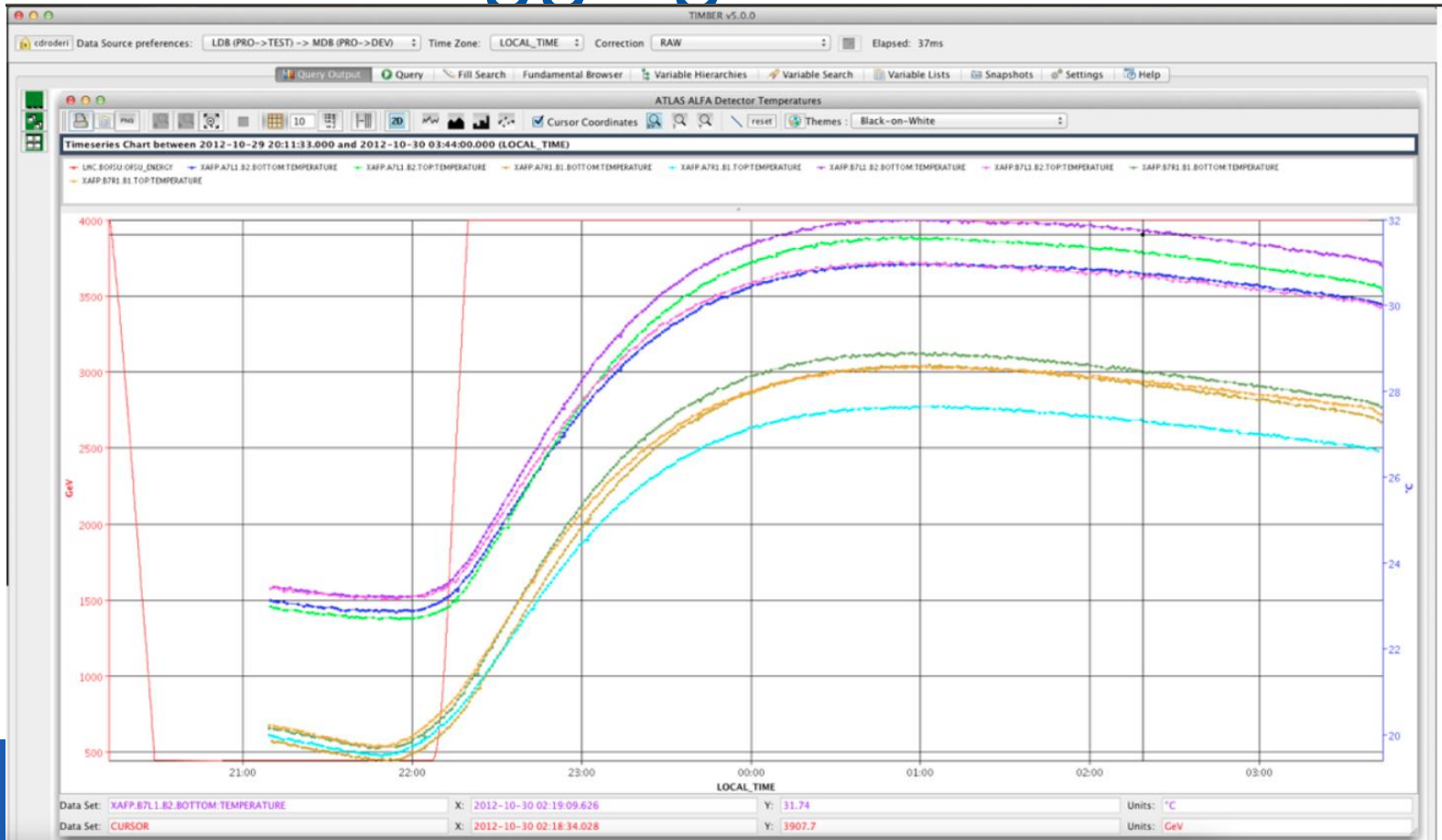
Accelerator logging



Credit:
C. Roderick



Accelerator logging



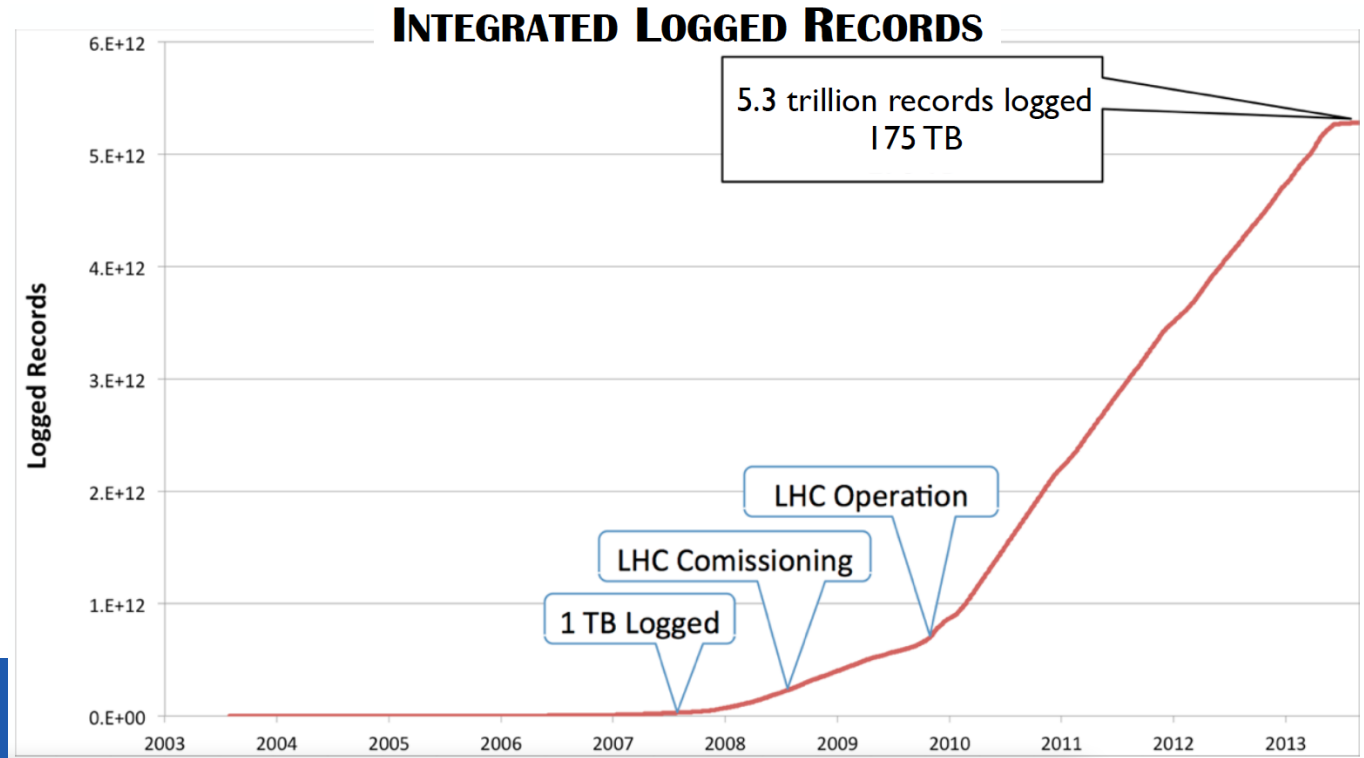
Credit:
C. Roderick



Accelerator logging

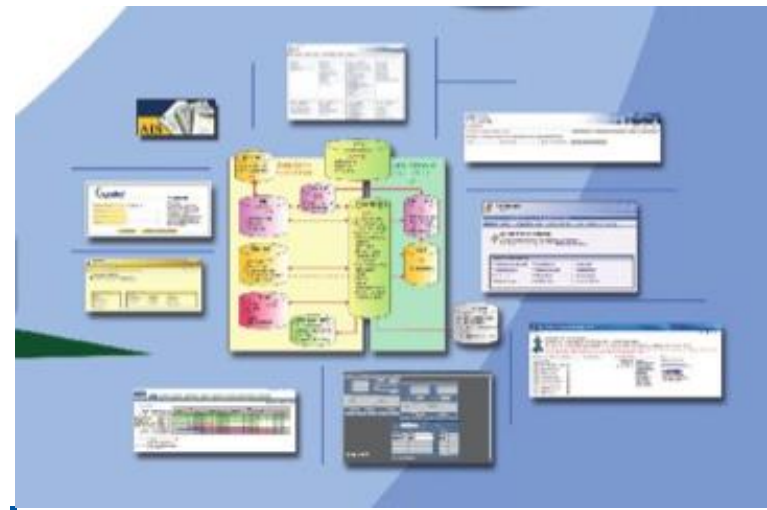
50TB/year,
rate to
increase to
100 – 150
TB in 2014
(Quench
Protection
System)

Credit:
C. Roderick



Administrative systems

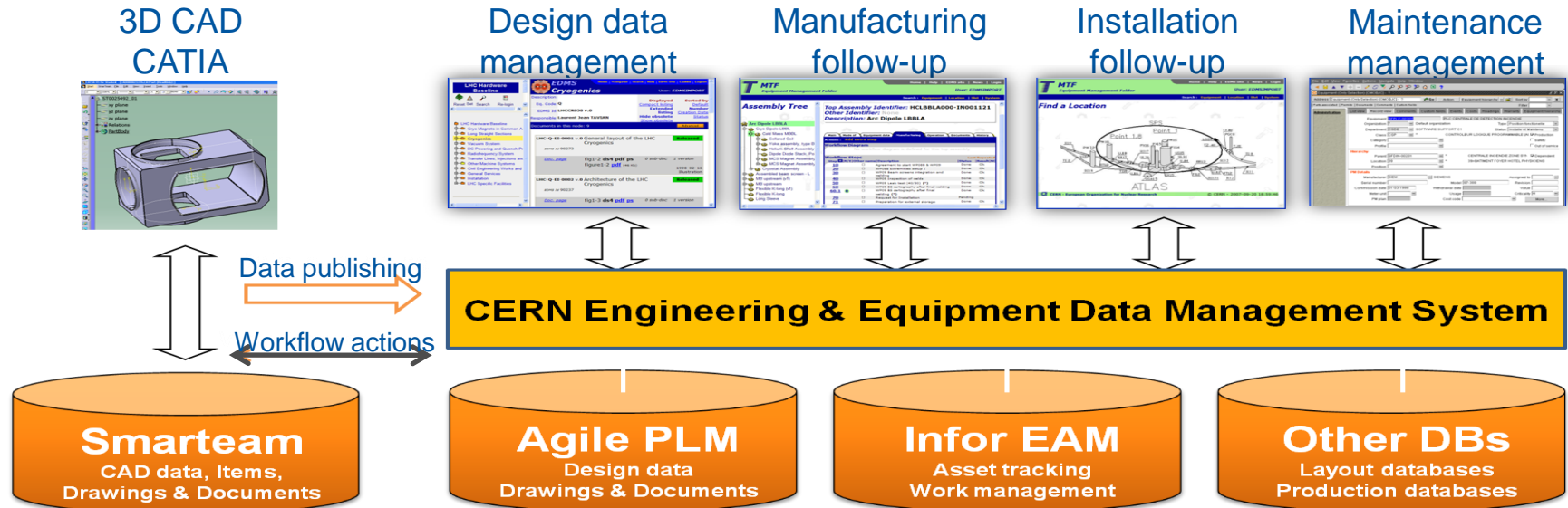
- AIS has standardized on Oracle as database and uses the Oracle database as interface between the tools
- Java EE and Apex, deployment with Weblogic
- Oracle E-Business HR, managed as Oracle Managed Cloud Services (“OnDemand”)



Engineering applications

Credit:
D. Widgren

- An integrated PLM platform based on commercial tools.
- Simplified web interfaces for precise tasks

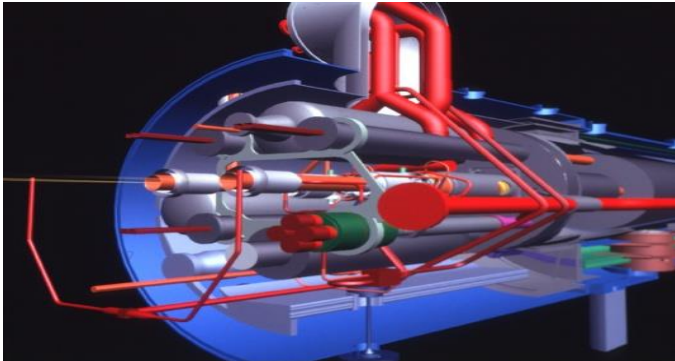


Design data management

Credit:
D. Widgren

Design baseline with full configuration management

- Workflows, versioning rules and access control based on project dependent contexts.
- Fully web-based and distributed approval processes.



✓	Thomas NICOL on 2005-08-24, 18:09 said: Ok as is.	Accept ✓
■	Christian BOCCARD on 2005-08-26, 08:54 said: Not directly concerned by Q3.	Seen ■
■	Davide BOZZINI on 2005-08-29, 09:44 said: For a more clear understanding of the instrumentation cabling the naming of the wires of table 7 should also appear in the figure 4. The names 'a' and 'b' should also be shown in figure 4.	Seen ■
■	Gilbert TRINQUART on 2005-08-31, 11:26 said: seen	Seen ■
■	Helene MAINAUD DURAND on 2005-09-05, 16:36 said: Seen.	Seen ■
✘	Raymond VENESS on 2005-09-05, 18:02 said: Tolerances for the V line extremities have been changed from +/-1 in all directions (v0.4) to +/-2 in X-Z and +/-6 in Y (v0.5). Beam vacuum interconnects have been designed and manufactured according to the values of v0.4 and cannot accept those proposed in this document. In addition, such large lateral offsets will lead to significant reduction in aperture, which should be discussed with AB/ABP experts.	Reject ✘

En

Pages 8 - **Show**

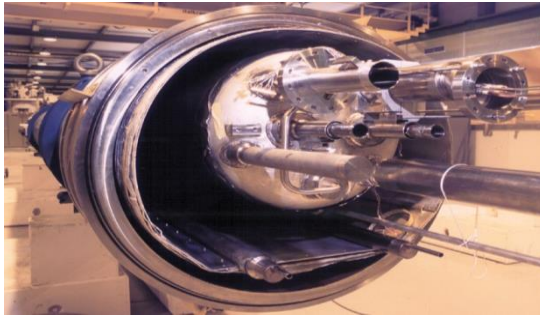
dd

Manufacturing follow-up

Credit:
D. Widgren

Follow-up of each manufactured component

- Manufacturing & test data captured at manufacturing sites.
- Predefined manufacturing workflows for each equipment type.



MTF-Travellers
Manufacturing and Test Folder

Home | Help | EDMS site | News | Login
User: WIDGREN
Quick access: Equipment search

Assembly Tree

- Cryo Dipole LBBR
 - Cold Mass MBBR**
 - Collared Coil
 - MCS Magnet Assembly (A1)
 - MCS Magnet Assembly (A2)
 - Yoke assembly, type B
 - Half Upper Yoke
 - Half Lower Yoke
 - Bus Bars Set Type B
 - Magnetic Insert Lamination
 - Austenitic Insert
 - Helium Shell Assembly Type
 - Dipole Diode Stack, Polarity 1
 - Cryostat Assembly

Assembly Folder: Manufacturing Workflow

Assembly Identifier: HCMBBRA001-03000003
Other Identifier: NB0105M
Description: Cold Mass MBBR

Main | Made of | Equipment data | **Workflow** | Documents

Actions: Add extra step

Workflow Diagram

No workflow diagram is defined for this assembly

Workflow Steps

Step	R/E	Other name	Description	Status	Result	INC
1	0		Geometrical measurement (ITP 20)	Done	Ok	
2	0		Warm magnetic measurements (ITP 21)	Done	Ok	
3	0		Electrical measurement (ITP 25)	Done	Ok	
4	0		Agreement for Shipment	Pending		
5	0		WP01 Arrival	Done	Ok	
6	0		WP01 Electrical Test	Done	Ok	
7	0		WP01 Mechanical Test	Done	Ok	
8	0		WP01 Optional Geometrical Test	Done	Ok	
9	0		WP01 Optional Magnetic Test	Done	Ok	

PLM @ CERN in numbers

Credit:
D. Widegren

Document & Drawings (incl. CAD):

~1,500.000 documents & drawings

~7,000 new documents & drawings created per month

Components:

~1,300,000 registered individually followed equipment

~3,000,000 equipment interventions/jobs logged

~ 15,000 equipment interventions/jobs logged per month



CASTOR and Oracle, tapes

Credit: German
Cancio Melia

- Home made mass storage system, relies on Oracle databases for name server, request handling and staging
- 4 libraries, SL8500
- 10088x4 = 40K slots (4500 free)
- Occupancy: 65PB worth of data
- Drives: 20 T10KB legacy drives; 40 T10KC drives (to be replaced by T10KD's)

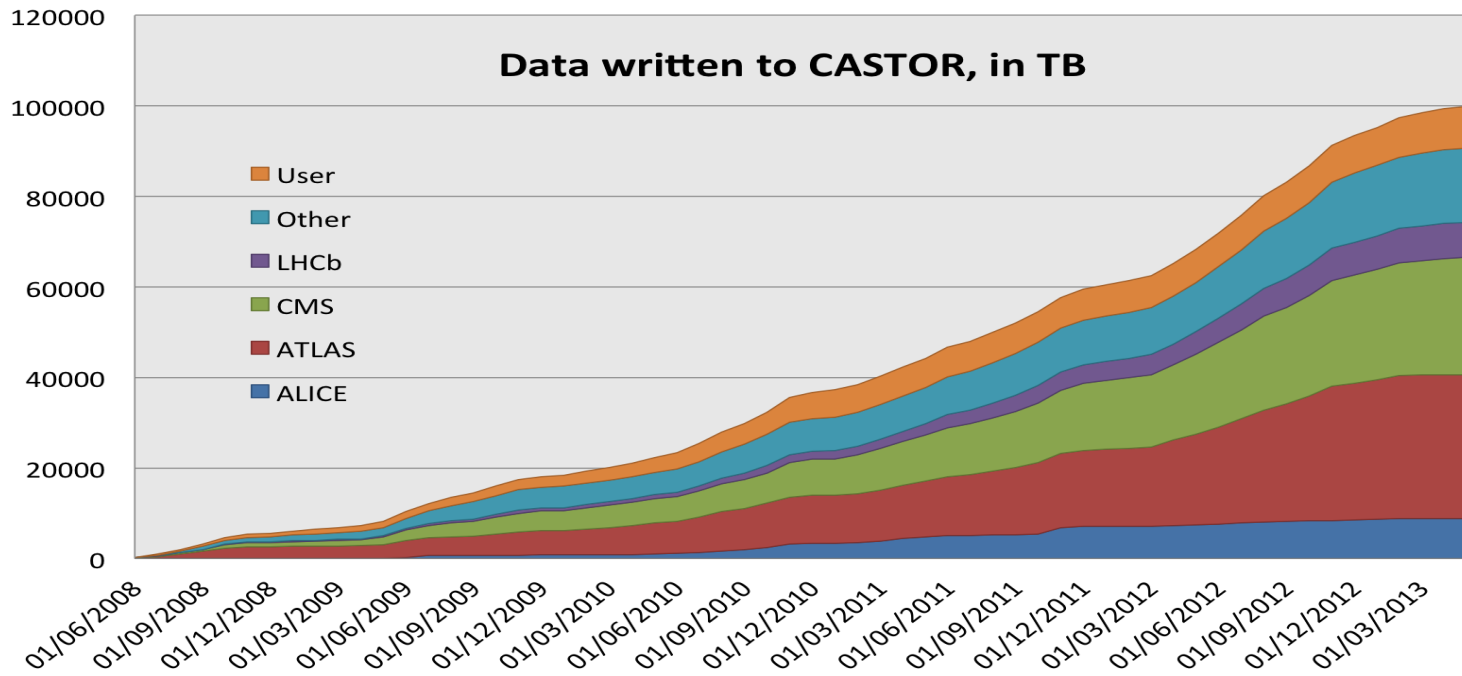


CASTOR Archive in Numbers

Credit: German Cancio Melia

Data:

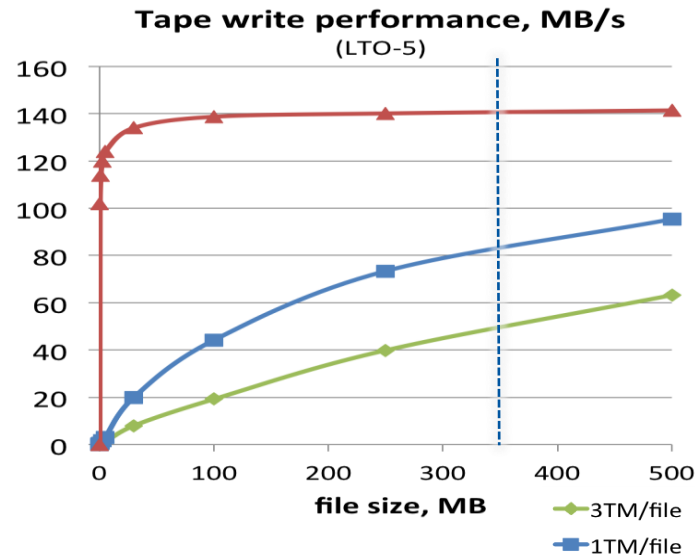
- ~90PB of data on tape; 250M files
- Up to 4.5 PB new data per month
- Over 10GB/s (R+W) peaks





Tape Writing

- Writing to tape is well managed and understood
 - Detector data can't wait for us!
- High *aggregate* transfer rates are not a problem
 - System designed to split the write stream onto several tape drives
- However, *low per-drive performance*
 - “Tape Mark” overhead per tape file up to 2s
 - CASTOR has 3 tape files per payload file
 - No significant overhead 14y ago on 20MB/s drives...!
 - Today with avg files of 350 MB: < 50% efficiency!



Credit: German Cancio Melia

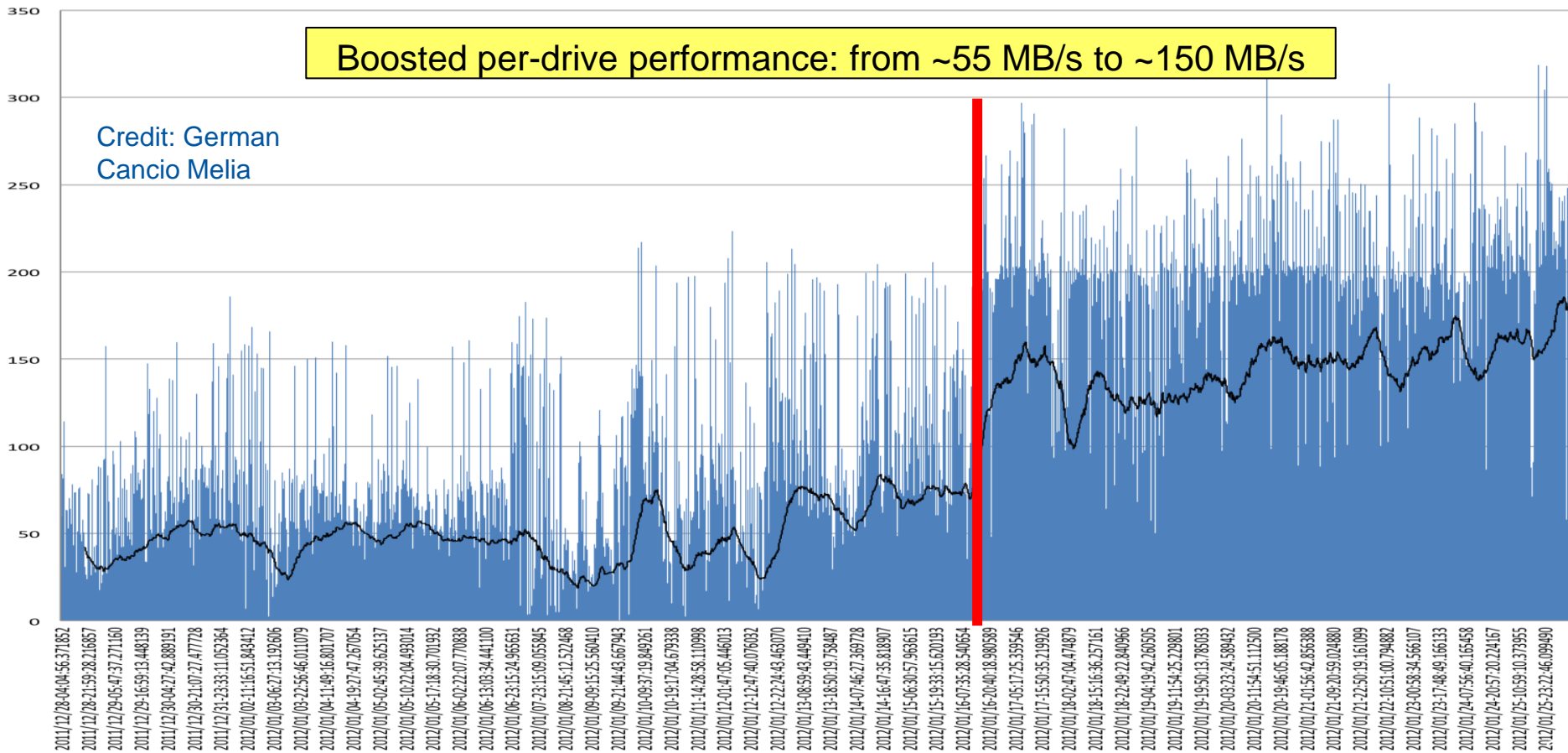
- Changed CASTOR tape layer to write “logical” (or “buffered”) tape marks between files (via CERN-prototyped Linux kernel extension)
- Now, achieving near-native drive speed by writing data in effective chunks of ~32GB
 - Significant savings in tape drive purchases



Data Write Rate Increase

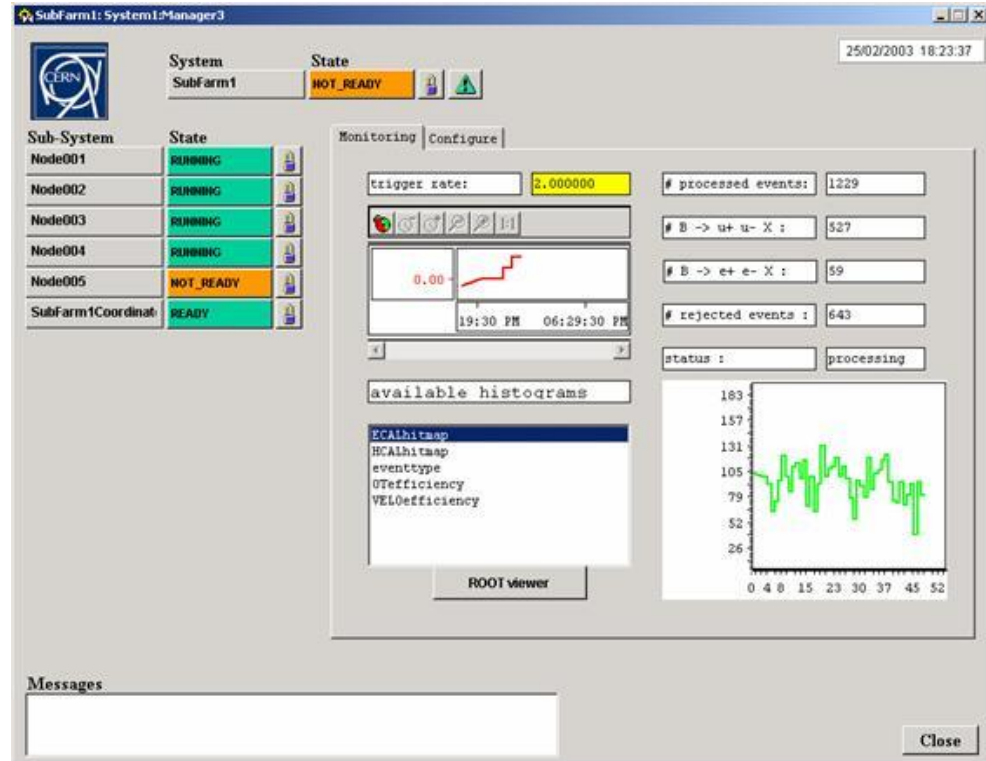
Boosted per-drive performance: from ~55 MB/s to ~150 MB/s

Credit: German
Cancio Melia



Experiment online systems

- Experiments rely on a SCADA system for their control
- Up to 150,000 changes / second stored in Oracle databases



Experiment offline systems

Credit:
Vakho Tsulaia

- **Geometry DB**
 - Relational database of **Primary Numbers** for the ATLAS Detector Description
 - All data for building **GeoModel** description in single place
 - Contains pointers to external files
 - **Identifier dictionaries**
 - **Magnetic field maps** (becoming obsolete)
 - All such files are shipped with the s/w release, no extra steps needed for getting them
 - Large relational database containing information about **Detector Status, Data-Taking Conditions, Calibrations, Alignment ...**
 - ATLAS Conditions DB is a **COOL Database**
 - COOL: one of 3 components of the **LCG Persistency Framework** (other two: **POOL, CORAL**)
- **Conditions DB**

Oracle at CERN

- From accelerator control to
 - accelerator logging,
 - administration,
 - engineering systems,
 - access control,
 - laboratory infrastructure (cabling, network configuration, etc.),
 - mass storage system,
 - experiment online systems,
 - experiment offline systems,
 - Etc.

openlab (1/3)

- Public-private partnership between CERN and leading ICT companies, currently in fourth phase (started in 2003)
- Its mission is to accelerate the development of cutting-edge solutions to be used by the worldwide LHC community
- Innovative ideas aligned between CERN and the partners, for products “you make it, we break it”

Partners



ORACLE

SIEMENS

openlab (2/3)

- Many successes:
 - RAC on Linux x86 (9.2 PoC and 10.1 production with ASM),
 - Additional required functionality (IEEE numbers, OCCl, instant client, etc.),
 - PVSS and RAC scalability,
 - Monitoring with Grid Control,
 - Streams world wide distribution,
 - Active DG, GoldenGate,
 - Analytics for accelerator, experiment and IT,
 - Etc.
- Regular feedback with joint selection of topics, some of the projects are common with more than one partner



openlab (3/3)

- Publications (web, paper) and presentations of results, visitors
- Maaïke Limper, best poster award at The International Conference on Computing in High Energy and Nuclear Physics 2013

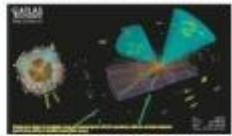
An SQL-based approach to Physics Analysis

M. Limper CERNopenlab

SELECT HIGGS FROM DATA_LHC;

Introduction: As part of the CERN openlab collaboration an investigation has been made into the use of an SQL-based approach for physics analysis. Currently, physics analysis is done using data stored in centrally produced root-staples that are accessible through the LHC computing grid. We'll present an alternative approach to physics analysis where analysis data is stored in a database. This would remove the need for customized staple production, and allows some of the calculations that are part of the analysis to be done on the database side.

Benchmarks: A simplified version of the search for the Higgs in association with a Z boson was implemented, both as a single root-macro and as an SQL-query. This analysis returns the invariant mass of the lepton- and jet-pair and uses 40 variables.



Dataset and database design: The benchmark analysis was tested using a subset of ATLAS experiment data from root-staples that were centrally produced for the ATLAS top-physics group. Root-staples store data columns wide, while Oracle groups all related attributes together by row. A database design was chosen where physics objects were stored in separate tables.

Column	Oracle	Oracle	Oracle	Oracle
EventNumber	100	100	100	100
JetNumber	100	100	100	100
LeptonNumber	100	100	100	100
...

The table on the right shows the volume of the test-data in our database, extracted from a subset of 127 staples containing a total of 7.3 million events, with 4000 analysis variables per event.

Physics Analysis in SQL: The SQL-version of the benchmark analysis is built through a series of select statements on each column table, each with a WHERE clause to apply selection criteria. Object-selection can be done via temporary tables using the WITH AS statement.

```
WITH goodhiggs AS (SELECT ... FROM meta WHERE ...)
```


or by explicitly creating a table holding the objects. Materialized views can be used to define common selection criteria. For example, the benchmarks used a materialized view to define the good lepton-jet selection. At the end of the query, JOIN statements are used to put information from the different selections together.

```
SELECT ... FROM good_higgs JOIN good_jet USING (JetNumber,EventNumber) WHERE (goodhiggs.M > 200)
```

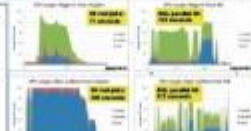
Test setups: Two types of test setup were used. The "test3"-setup used 2 machines with network-based file storage (NFS) accessible from all nodes. The 2nd test setup, "test2", was designed to run either Hadoop or Oracle RAC and was optimized for fast I/O using 5 machines connected to 5 disk arrays holding a total of 60 disks. On this test-setup the Oracle database used the Automatic Storage Management feature, and Hadoop used its HDFS filesystem, to spread the data evenly over all devices. For the comparison with root on the magnet-cluster, the staples were distributed evenly over all disks.

Test Setup	Time (min)	Time (min)	Time (min)
test2	2	1	1
test3	240	180	210
test4	32	42	42


Parallel execution: An SQL query can be executed in serial or in parallel and the degree of parallelism can be set on the table or by a hint inside the query. For the single analysis, parallelism was restricted by running multiple simultaneous root-jobs, each analyzing a subset of files. The root-version gained more from parallelism than the DB-version of the analysis. This is because the DB-version is limited by I/O speed as it needs to read many columns in the table to find the relevant variables.



CPU usage: The magnet-cluster was used to study CPU usage. Here, the staple version was executed using 40 root-jobs (1 per disk). The plots on the right show that the Higgs-Z benchmark was fastest with the single version and both versions were limited by I/O. The Higgs outflow benchmark was faster with the DB-version as the RootCore packages were limited by CPU.



Hadoop: On the magnet-cluster the test dataset was also stored as comma-delimited text files in the Hadoop filesystem (hdfs). The Hadoop system was configured to have 40 task slots (8 per node) to match the number of cores in the system. The Higgs-Z benchmark analysis was reimplemented using MapReduce-code written in Java. The Higgs-Z analysis in Hadoop used a relatively large amount of CPU and was slower than both the single and DB-versions.



Conclusion: Physics Analysis using SQL on data stored in a database can provide an alternative way to analyse the large datasets produced by the LHC experiments. Row-based storage in combination with wide tables limits performance by the I/O read speed of the system. Future studies will focus on columnar stores to improve performance.

Oracle in our research environment

- Even if computing is critical for HEP, it is not the goal, there is a lot to do using solutions from commercial vendors which are industry supported and scalable
- Oracle has provided solutions along the years
- We have worked with Oracle to improve the tools to our (and others') needs with success
- Good for staff to work on industry standards for their future career

Conclusion

- Not every day you build a 30+ years collaboration
- A long way since 1982, now very wide usage with applications, tape and database
- Oracle has proven to be reliable partner who cares and supports research
- Provide feedback and ideas for enhancements
- Helps focus on our core challenges
- A collaboration which works!



www.cern.ch