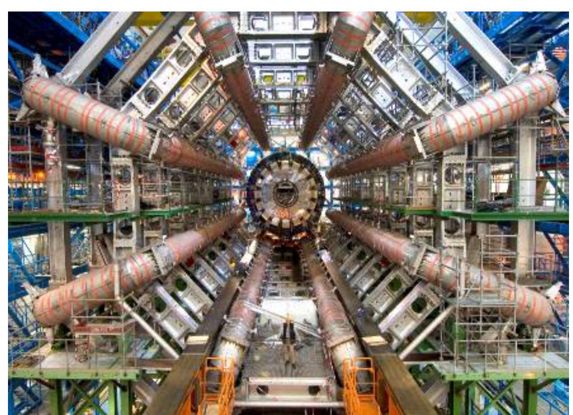
ORACLE use within the LCG

Eva Dafonte Pérez

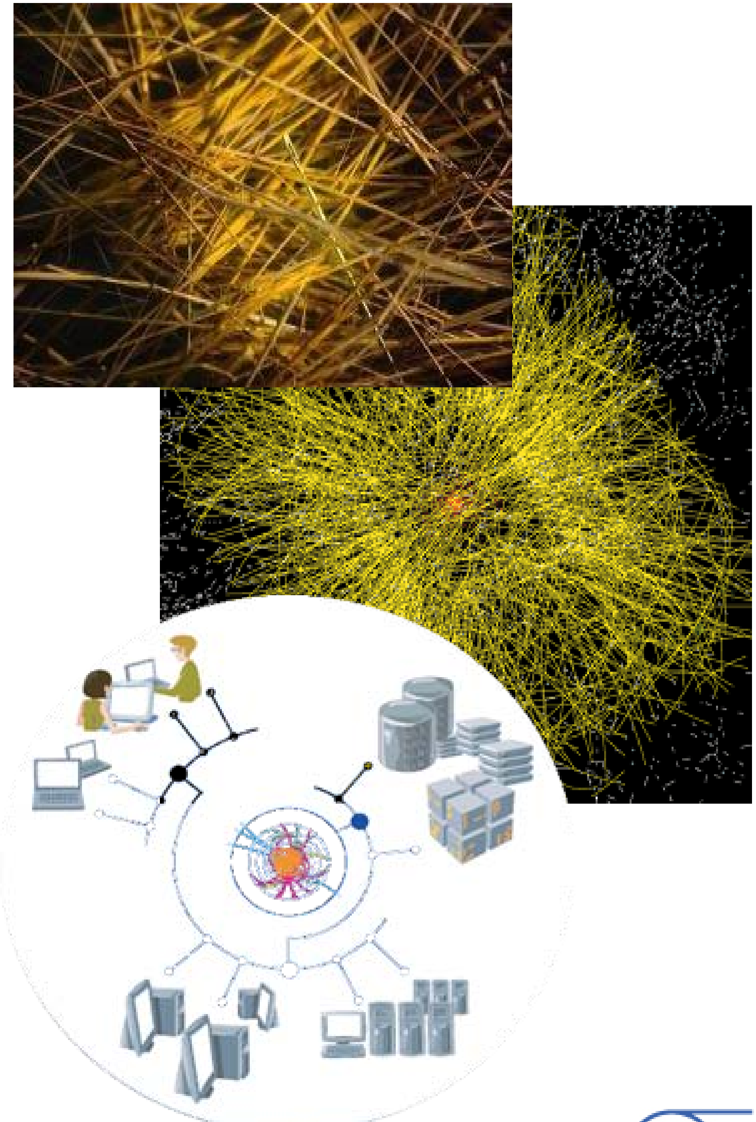
15 February 08

Outline

- CERN and LHC
- Role of databases in LHC data management
- Oracle database technologies and deployment models
- Oracle Streams replication
- Monitoring
- Future improvements
- Summary


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**

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

WLCG Collaboration

- The Collaboration
  - 4 LHC experiments
  - ~200 computing centres
  - 12 large centers (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

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

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

Oracle point of view

Oracle Real Application Clusters 10g - Foundation for Grid Computing
http://www.oracle.com/technology/products/database/clustering/index.html

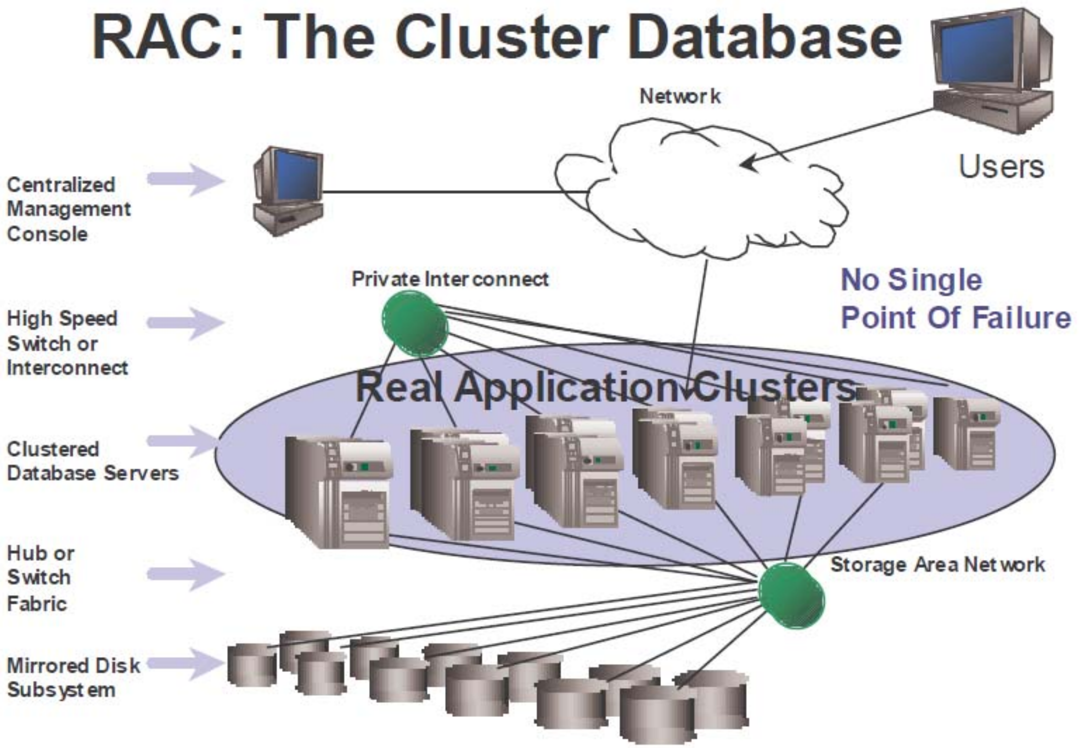
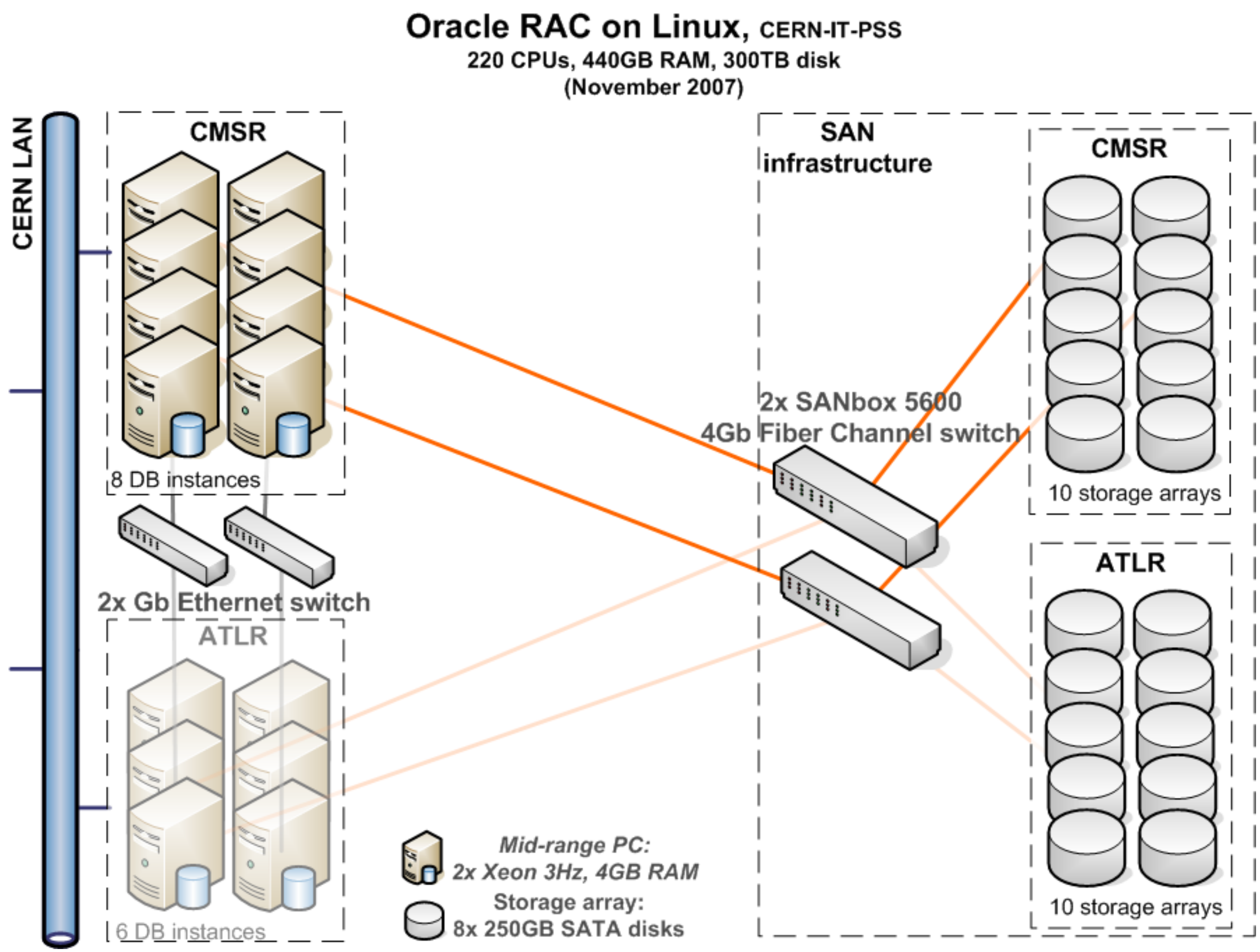


Figure 1: Oracle RAC –clustering database servers – foundation for Enterprise Grid Computing delivering high availability, scalability and flexibility.

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

Tier0 Database Architecture for Physics

- Applications consolidated on large clusters, per experiment
- Redundant and homogeneous HW across each RAC



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

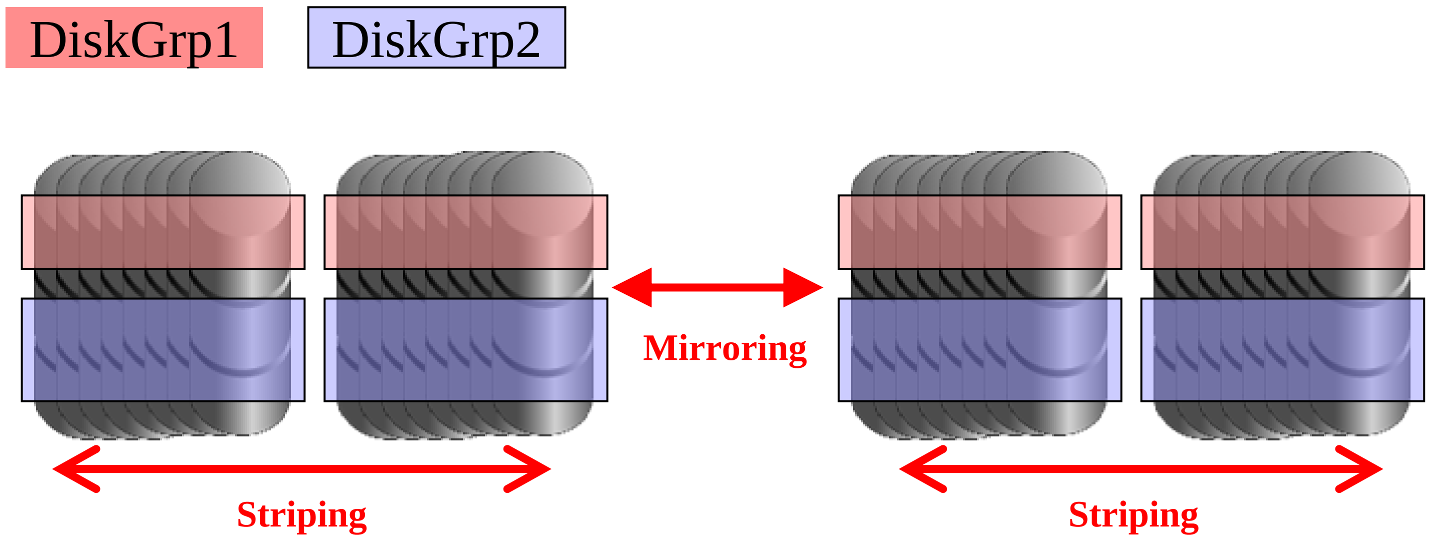
Architecture

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

Architecture (storage)

- Following SAME concept:
  - Oracle ASM for mirroring across arrays and striping
- Two diskgroups per database (‘data’, ‘recovery’)
- Destroking: most accessed data on external part of disk

- Example:



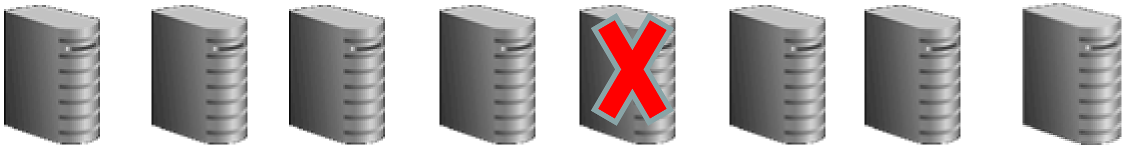
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Architecture (services)

- Resources distributed among Oracle services
  - Applications assigned to dedicated service
  - Applications components might have different services
- Service reallocation not always completely transparent

CMS_COND	Preferred	A1	A2	A3	A4	A5	A6	A7
CMS_C2K	Preferred	A2	A3	A4	A5	A6	A7	A1
CMS_DBS	A5	A3	A1	A2	Preferred	Preferred	Preferred	A4
CMS_DBS_W	A4	A5	A6	A7	Preferred	A1	A2	A3
CMS_SSTRACKER	Preferred	Preferred	Preferred	Preferred	Preferred	Preferred	Preferred	Preferred
CMS_TRANSFERMGMT	A2	Preferred	Preferred	Preferred	A1	A3	A4	A5
CMS RAC Node #	1	2	3	4	5	6	7	8

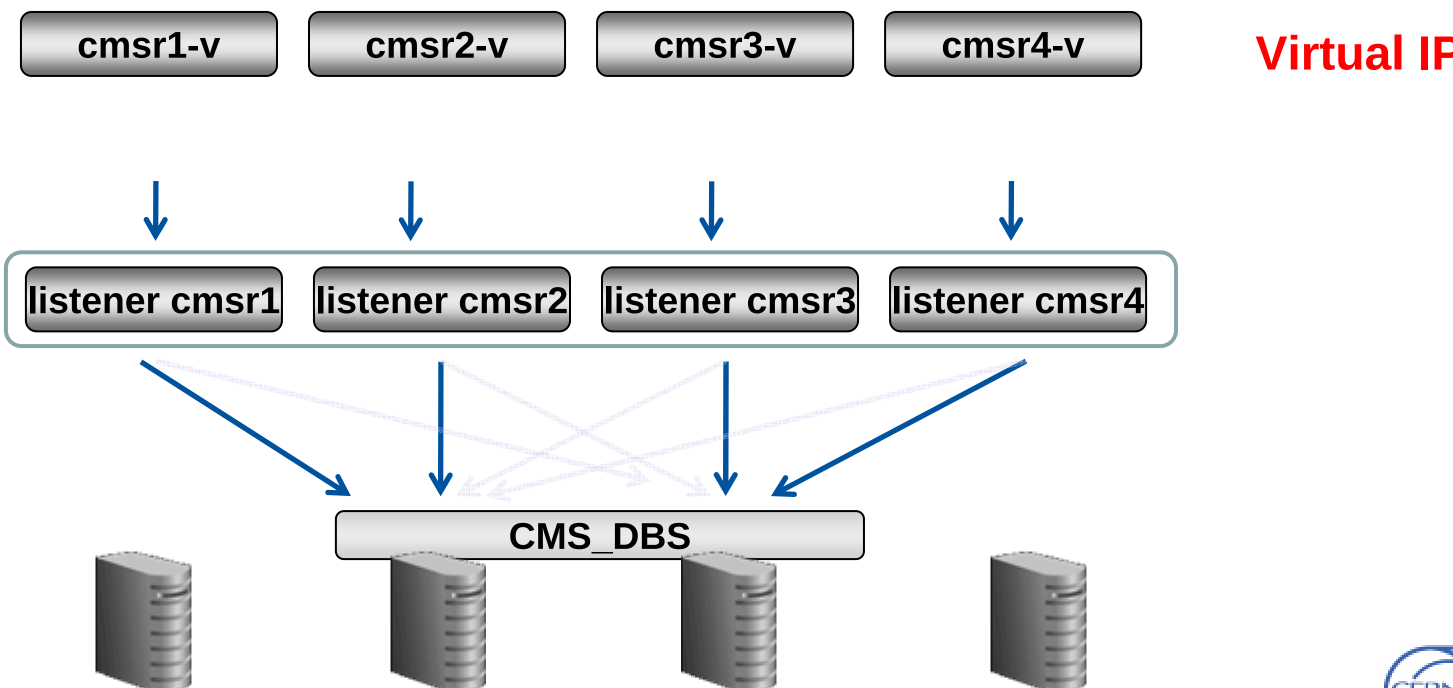
CMS_COND	Preferred	A1	A2	A3		A4	A5	A6
CMS_C2K	Preferred	A2	A3	A4		A5	A6	A1
CMS_DBS	A4	A2	Preferred	A1		Preferred	Preferred	A3
CMS_DBS_W	A3	A4	A5	A6		Preferred	A1	A2
CMS_SSTRACKER	Preferred	Preferred	Preferred	Preferred		Preferred	Preferred	Preferred
CMS_TRANSFERMGMT	A1	Preferred	Preferred	Preferred		A2	A3	A4



Architecture (load balancing)

- Service's *connection string* mentions **ALL** virtual IPs
- It connects to a random virtual IP (client load balance)
- Listener sends connection to least loaded node where service runs (server load balance)

```
$ sqlplus cms_dbs@cms_dbs
```

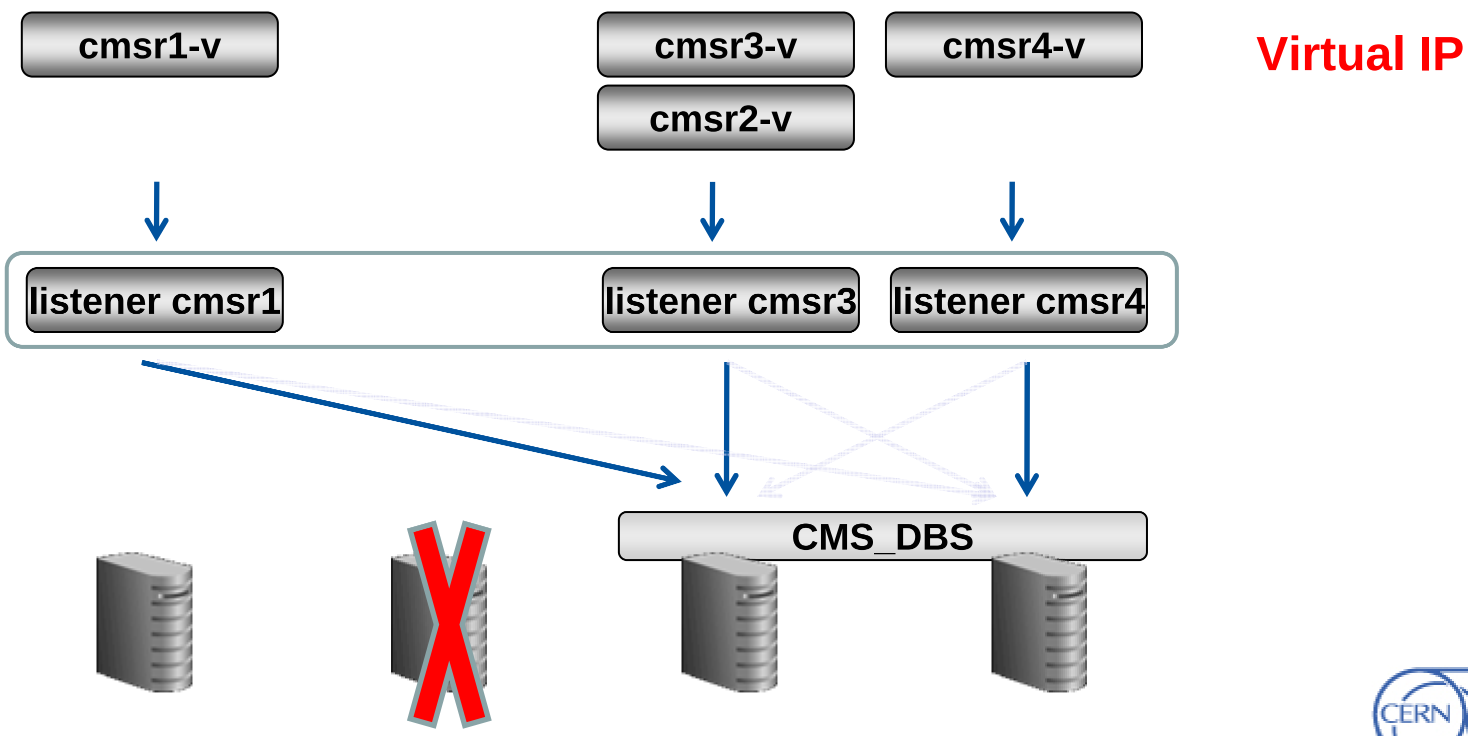


CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Architecture (load balancing)

- Used also for rolling upgrades (patch applied node by node)
- Small glitches might happen during VIP move
  - no response / timeout / error
  - applications need to be ready for this → catch errors, retry, not hang

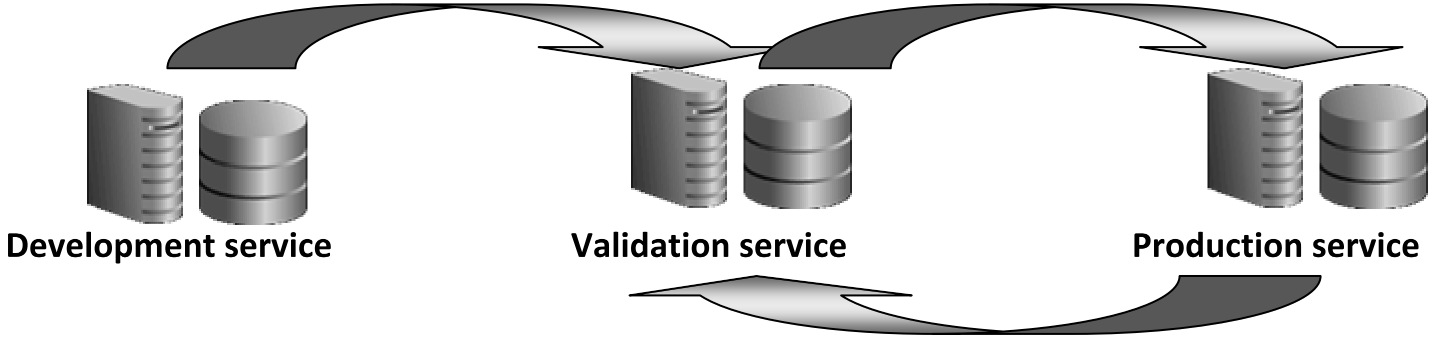
```
$ sqlplus cms_dbs@cms_dbs
```



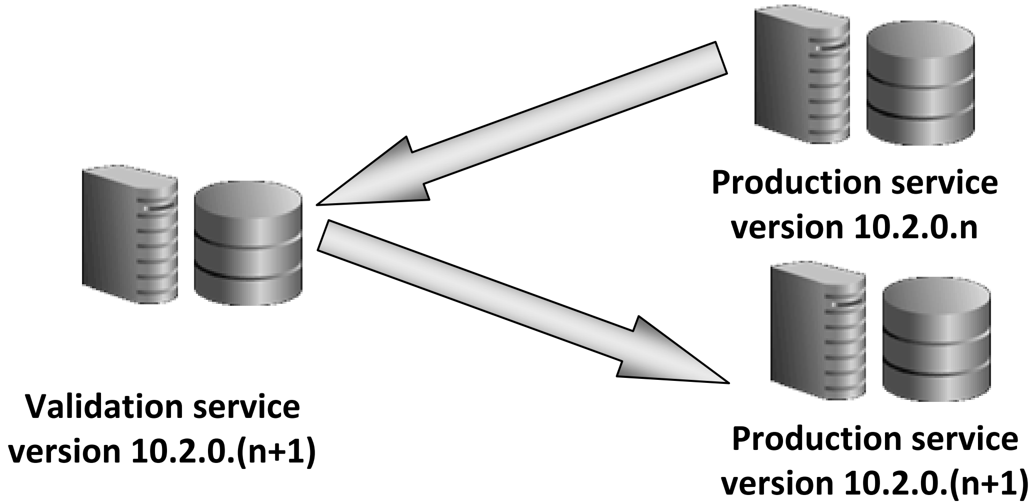
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Apps and database release cycle

- Applications’ release cycle



- Database software release cycle



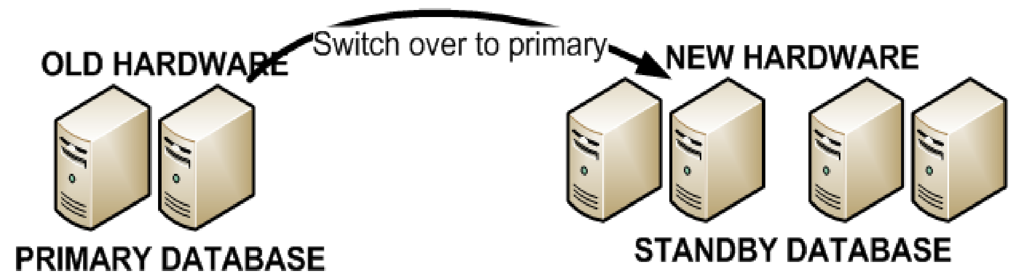
Backup strategy used at Tier0

- Both Oracle-recommended strategies implemented for all production systems using RMAN
- Incremental backup strategy:
  - backups go to tapes
  - weekly or biweekly level 0 backups (depending on the DB size)
  - level 1 cumulative backup inbetween
  - daily incremental level 1 differential backups
  - archivelog backup every 30 minutes
- Incrementally updated DB copy strategy:
  - daily incremental differential backups applied with 2 days of delay
  - copies, incremental backups and archived redo logs stored in the Flash Recovery Area
- Central machine to schedule and run all the backups
- Central RMAN catalog exported on regular basis

Oracle DataGuard for RAC migration

CERN IT Department

- Commodity hardware has small warranty periods
- Hardware specifications progress very fast
- Minimal downtime required independent of database size
- Easy to fallback in case of error



- Can include
  - version change
  - migration to 64bit
  - hardware upgrades
- Our use cases: migrate hardware (storage + servers) and
  - upgrade 10.2.0.2 to 10.2.0.3
  - upgrade 32bit to 64bit OS+RDBMS

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

Physics databases numbers

CERN IT Department

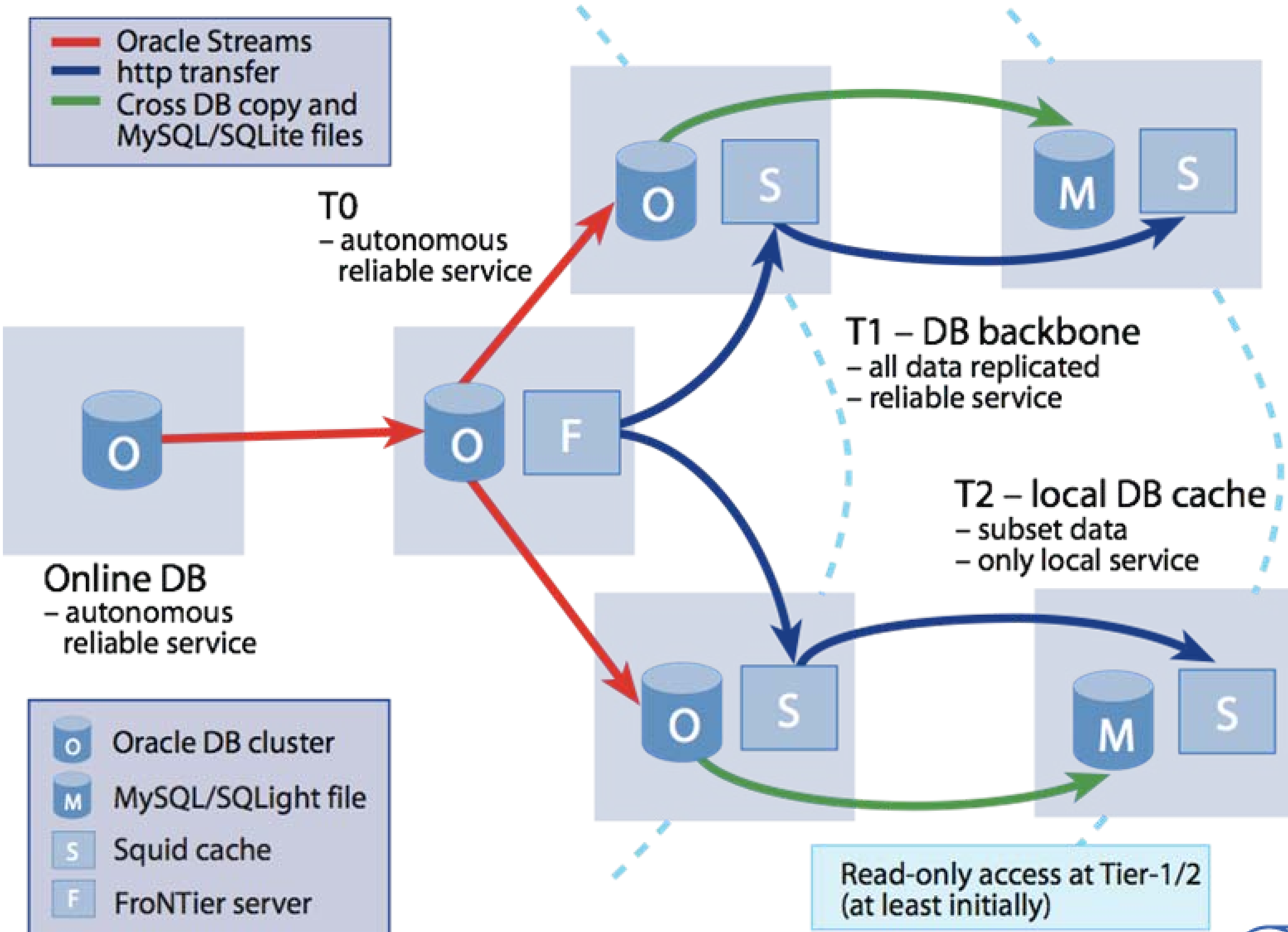
- ~20 RACs
  - 6-8 nodes
  - ALICE 1TB, ATLAS and CMS 4TB, LHCb and WLCG 3TB
- 220 CPUs, 440GB RAM, 300TB disk
  - soon quad-cores (3-4 node RAC)
- 345 production schemas (feb07)
- 9.625 GB production data
  - largest table: 3.02 billion rows (IOT, non partitioned)
- Workload
  - 5.2 million sessions/week (week 3/2008)
  - 128 MB/s (average for week 3/2008)
  - 995 CPU hours/week (CPU time avg for week 3/2008)

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

DM

Distributed Service Architecture

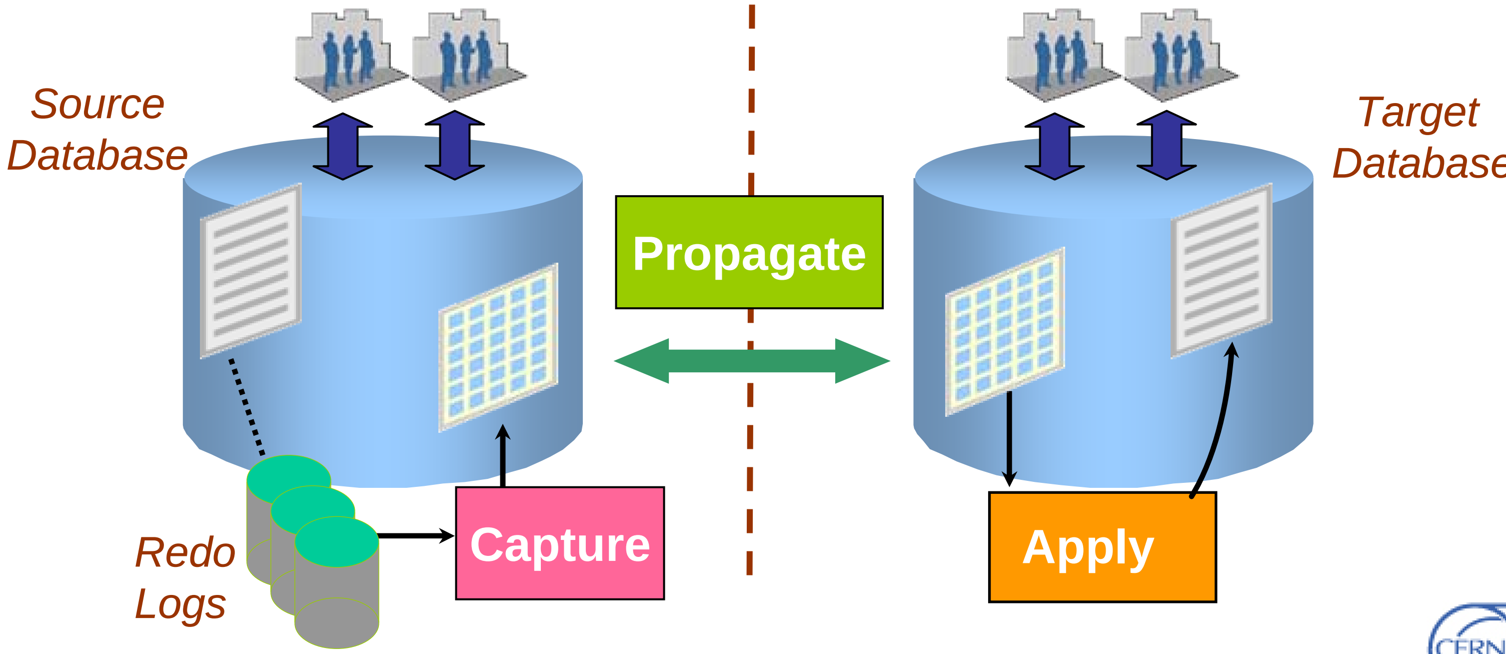
CERN IT Department



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

Oracle Streams replication

- Technology for sharing information between databases
- Database changes captured from the redo-log and propagated asynchronously as Logical Change Records (LCRs)



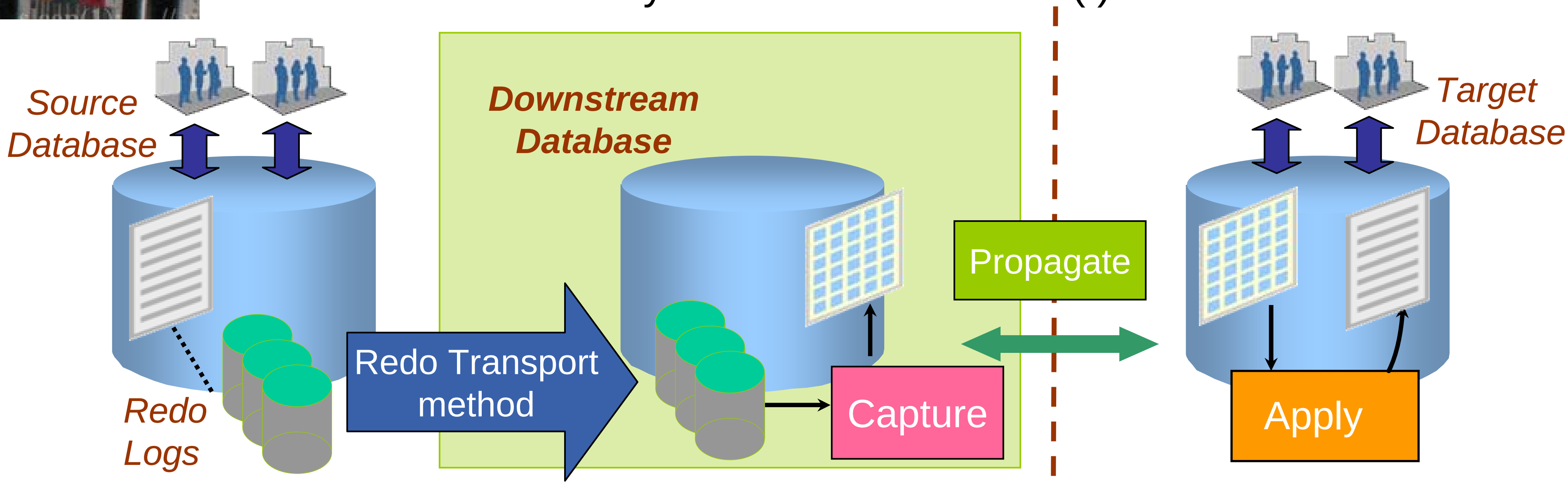
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

DM

Downstream capture & Network optimizations

CERN IT Department

- Downstream capture to de-couple Tier 0 production databases from destination or network problems
  - source database availability is highest priority
- Optimizing redo log retention on downstream database to allow for sufficient re-synchronisation window
  - we use 5 days retention to avoid tape access
- TCP and Oracle protocol optimisations yielded significant throughput improvements (factor 10)
  - network latency to some sites 300 ms(!)



Streams lessons learned

- Filtering - capture vs. propagation rules
  - performance difference significant
  - obtained a factor 5 in apply speed
- Be aware of row-id based operations
  - apply side can be significantly less efficient
- SQL bulk operations (at the source db)
  - may map to many elementary operations at the destination side
  - need to control source rates to avoid overloading
- Streams fail-over in case of site problems
  - naïve set-up can run into problems with spilling
  - proposed set-up with dedicated “problem” stream proven to work during site/network problems

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

Streams enhancements in 11g

ORACLE DATABASE 11g

- Performance improvements
  - minimizes disk I/O and reduce capture latency
  - reduces CPU consumption
- Combined Capture and Apply
  - direct communication between capture and apply processes
  - observed significant gain in replication throughput for several applications
    - e.g. 12,000 lcr/s (11g) instead of 5,000 lcr/s (10g)
  - this will help us to
    - increase the replicated data volume for 1-to-1 replication set-ups
    - decrease the time for tier sites to catch-up after outages/interventions
- Data comparison, performance advisor, automatic split and merge procedures, ….

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

Monitoring

- 24x7 reactive monitoring
  - Lemon Alarms, Operators, SysAdmins
    - HW failures, OS problems, High load
  - Host and, instance and service availability
    - home grown monitoring
- Active monitoring
  - Oracle Enterprise Manager
    - execution plans, resource usage per service
  - 3D monitoring included into experiments dashboards
  - Lemon
  - Weekly reports (sent to experiment DBAs/links/3D)
    - SQL changes, service usage, bad connection management, bad indexes

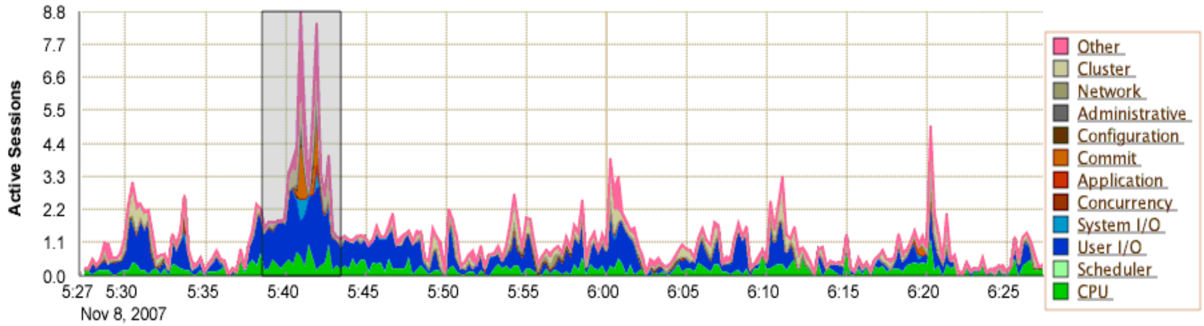
Oracle Enterprise Manager

Top Activity

Switch Database Instance atlr.cern.ch_atlr1 Go

Drag the shaded box to change the time period for the detail section below.

View Data Real Time: 15 Second Refresh



Detail for Selected 5 Minute Interval

Start Time **Nov 8, 2007 5:38:34 PM MET**

Run ASH Report

Top SQL

Schedule SQL Tuning Advisor | Create SQL Tuning Set

Select All | Select None

Select	Activity (%)	SQL ID	SQL Type
☐	42.78	3t4zyft1ycmp7	SELECT
☐	11.30	aq8utg96sc8kq	SELECT
☐	8.30	0pcmq7hryn2x0	SELECT
☐	5.87	aahvgmuar9z07	SELECT
☐	4.15	7yhvgrsw7x8d7	SELECT
☐	2.72	7grmrsp8dy3yd	SELECT
☐	1.72	2f3ddpjygkxhv	SELECT
☐	1.72	9scc29h43m5ap	INSERT
☐	1.29	7hzn20h90y9rk	SELECT
☐	1.00	6khrc3pvyw9j6	SELECT

Schedule SQL Tuning Advisor | Create SQL Tuning Set

Total Sample Count: 699

Top Sessions

View Top Sessions

Activity (%)	Session ID	User Name	Program
29.72	635	ATLAS_PVSS_READER	root.exe@roata01 (TNS V1-V3)
5.77	579	ATLAS_PS_W3	httpd@ccwbsn01.in2p3.fr (TNS V1-V3)
4.77	675	ATLAS_DASHBOARD_DM_WRITER	data.stats.collection@lxarda11.cern.ch (TNS V1-
4.08	673	ATLAS_PS_W3	
3.48	661	ATLAS_PS_W1	python@atlas002.uta.edu (TNS V1-V3)
3.38	874	SYS	oracle@itrac21.cern.ch (LGWR)
3.28	545	ATLAS_DQ2_R	httpd.worker@lxb7239.cern.ch (TNS V1-V3)
3.28	859	SYS	oracle@itrac21.cern.ch (ARC0)
3.18	858	SYS	oracle@itrac21.cern.ch (ARC1)
2.78	636	ATLAS_DQ2_W	httpd.worker@lxb7238.cern.ch (TNS V1-V3)

Total Sample Count: 1,006

RACMon - monitoring

CERN IT Department

Overview | ASM | Prod Services | All Services | Unavailability | Performance | Backup | CDB State | StreamMon

atlr ⊠ atonr ⊠ cmsonr ⊠ cmsr ⊠ compr ⊠ d3r ⊠ int11r ⊠ int12r ⊠ int2r ⊠ int6r ⊠ int8r ⊠ int9r ⊠ intr ⊠ lcgr ⊠ lhcbr ⊠ pdbr ⊠ t11g64 ⊠ test1 ⊠

Last update: 2008-02-13 11:33 (4 minutes ago)

Cluster: atlr - RAC for ATLAS (monitoring enabled) last update: 2008-02-13 11:33

	DB instances	Production services	ASM	CDB	Backup
Availability last 7 days	100%	100%	100%		
Status	OK	OK	OK	OK	OK

6 Nodes: atlr1 (itrac21) ✓ , atlr2 (itrac22) ✓ , atlr3 (itrac23) ✓ , atlr4 (itrac24) ✓ , atlr5 (itrac25) ✓ , atlr6 (itrac26) ✓
10 Storages: itstor30, itstor31, itstor32, itstor33, itstor34, itstor35, itstor36, itstor37, itstor38, itstor39
Sessions (per node): atlr1 (36), atlr2 (43), atlr3 (102), atlr4 (108), atlr5 (123), atlr6 (12)

Cluster: atonr - RAC for Atlas Online (monitoring enabled) last update: 2008-02-13 11:33

	DB instances	Production services	ASM	CDB	Backup
Availability last 7 days	100%	100%	100%		
Status	OK	OK	OK	OK	OK

6 Nodes: atonr1 (itrac45) ✓ , atonr2 (itrac46) ✓ , atonr3 (itrac47) ✓ , atonr4 (itrac48) ✓ , atonr5 (itrac49) ✓ , atonr6 (itrac50) ✓
10 Storages: itstor42, itstor43, itstor44, itstor45, itstor54, itstor55, itstor56, itstor57, itstor58, itstor59
Sessions (per node): atonr1 (42), atonr2 (58), atonr3 (47), atonr4 (63), atonr5 (50), atonr6 (4)

Cluster: cmsonr - RAC for CMS Online (monitoring disabled) last update: 2008-02-13 11:33

	DB instances	ASM	CDB	Backup
Availability last 7 days	98.88%	99.08%		
Status	BAD	BAD	BAD	BAD

6 Nodes: cmsonr1 (cms-srv-c2c01-11) , cmsonr2 (cms-srv-c2c01-12) , cmsonr3 (cms-srv-c2c01-13) , cmsonr4 (cms-srv-c2c01-14) , cmsonr5 (cms-srv-c2c01-15) , cmsonr6 (cms-srv-c2c01-16)
0 Storages:

Cluster: cmsr - RAC for CMS (monitoring enabled) last update: 2008-02-13 11:33

	DB instances	Production services	ASM	CDB	Backup
Availability last 7 days	100%	100%	100%		
Status	OK	OK	OK	OK	OK

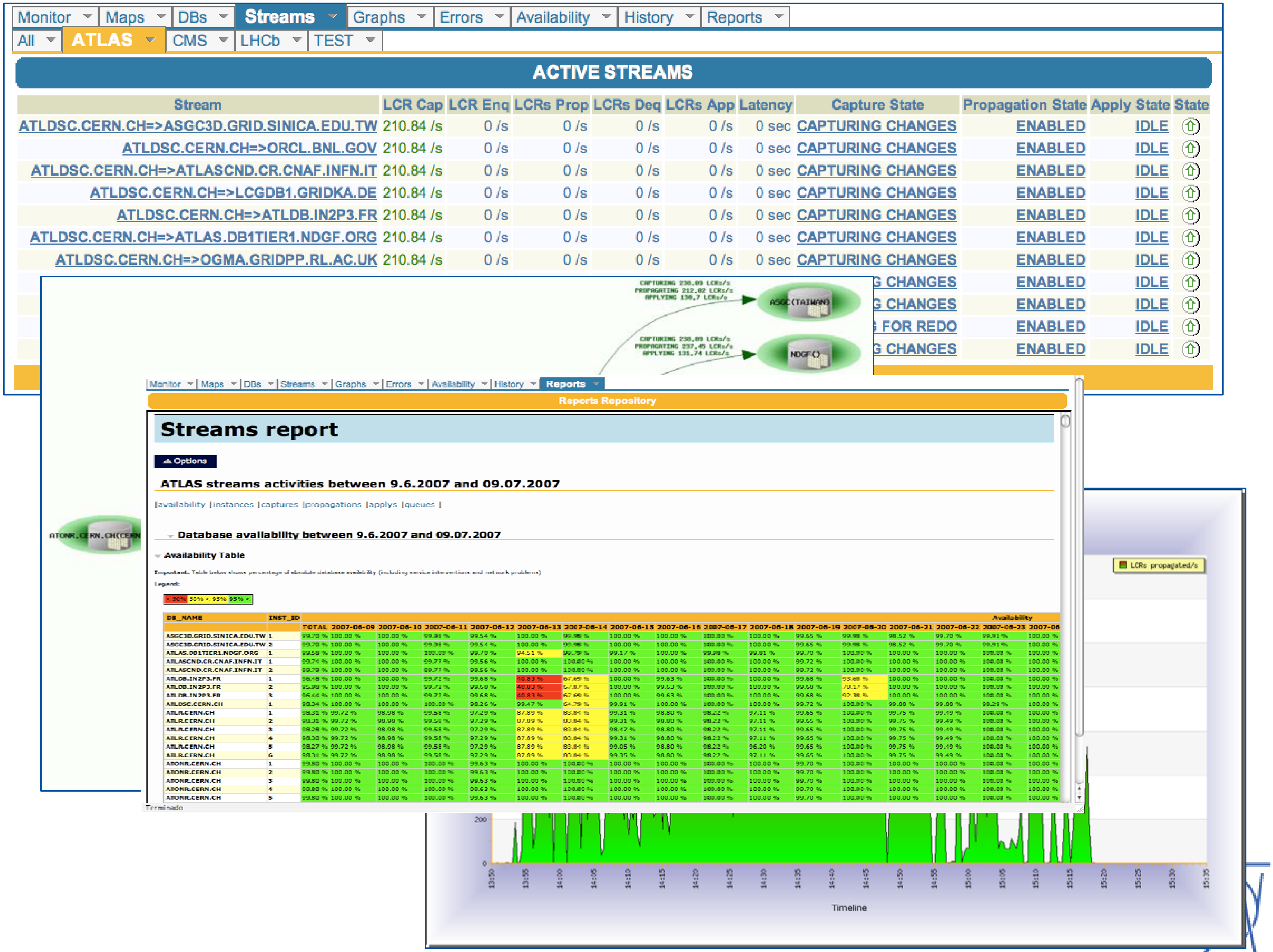
8 Nodes: cmsr1 (itrac301) ✓ , cmsr2 (itrac302) ✓ , cmsr3 (itrac303) ✓ , cmsr4 (itrac304) ✓ , cmsr5 (itrac305) ✓ , cmsr6 (itrac306) ✓ , cmsr7 (itrac307) ✓ , cmsr8 (itrac308) ✓
10 Storages: itstor301, itstor302, itstor303, itstor304, itstor305, itstor306, itstor307, itstor308, itstor323, itstor324
Sessions (per node): cmsr1 (41), cmsr2 (86), cmsr3 (166), cmsr4 (49), cmsr5 (94), cmsr6 (69), cmsr7 (69), cmsr8 (42)

Cluster: compr - RAC for COMPASS (monitoring enabled) last update: 2008-02-13 11:33

	DB instances	Production services	ASM	CDB	Backup

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

3D Streams monitoring



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

Future improvements

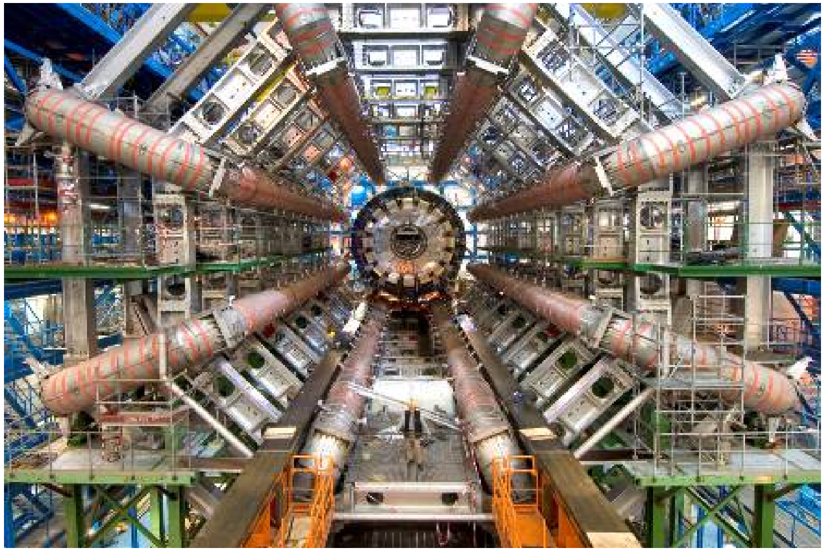
- Quad-core servers
  - smaller RACs, 1 quad-core better than 6 instance RAC
- Data guard
  - parallel RAC with small lag (few hours)
  - fast disaster recovery
  - can be open read-only to recover from human errors
- Streams replication
  - add redundancy for downstream and streams monitoring
  - automation of the split-merge (procedure used when one site needs to be dropped/re-synchronized)
- Oracle 11g new features
  - "SQL Replay" to have load on validation RAC
  - "Data Guard - Standby snapshot" allows make a snapshot of production DB
  - "SQL Plan Management" to stabilize optimizer
  - "Result Cache" for faster results

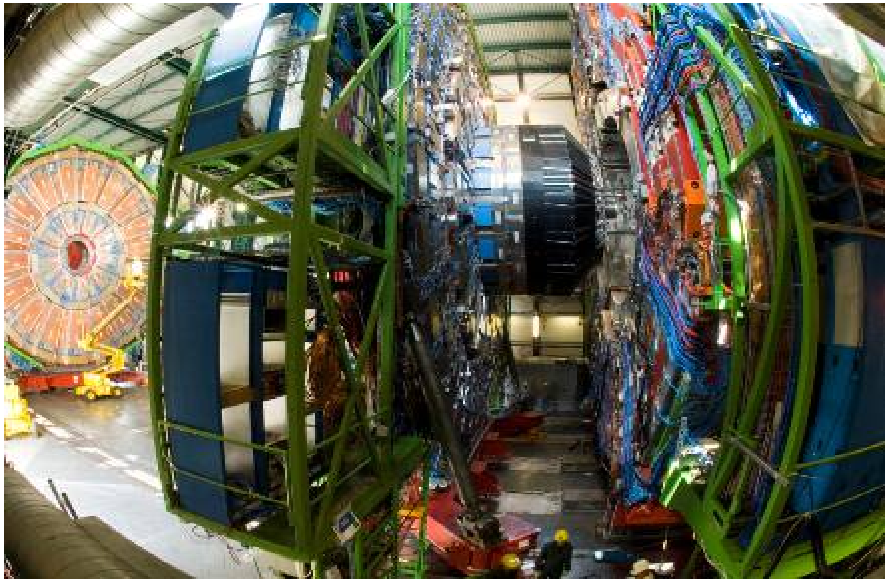
Summary

- 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

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

Questions?





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