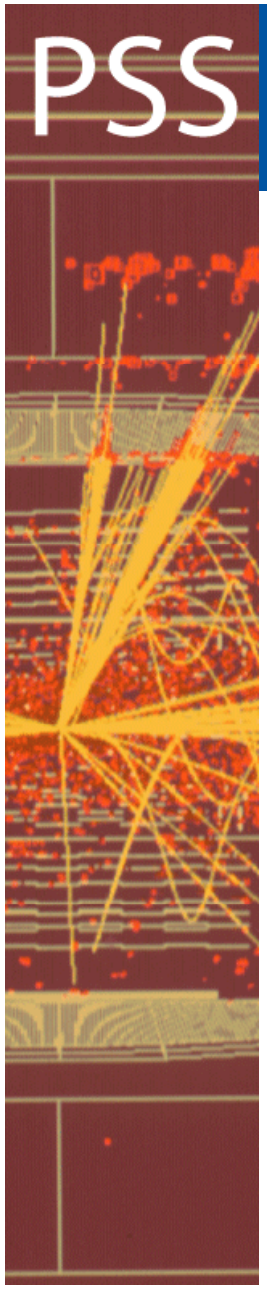


PSS Physics Services Support

CERN IT Department

Oracle Real Application Clusters (RAC)

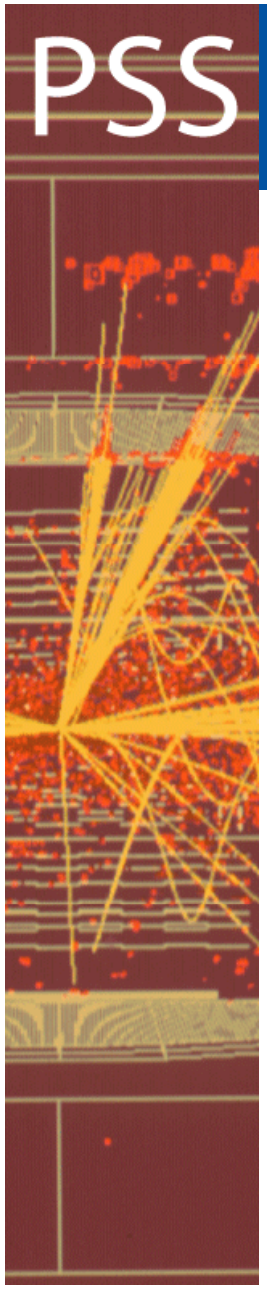
Techniques for implementing & running robust and reliable services

WLCG Service Reliability Workshop

26 November 2007

***Miguel Anjo**, CERN-Physics Databases*

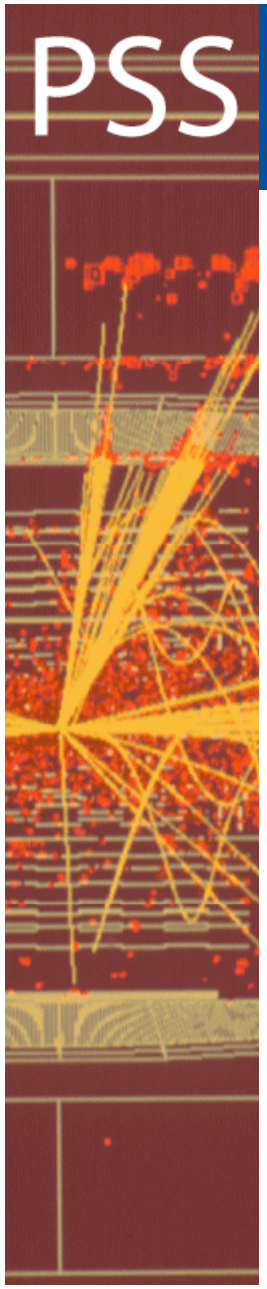
CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it



Outline

- Experiments Critical services
- Oracle RAC solution
  - Architecture
  - Monitoring
  - Administrative work
  - Current and future usage
  - Concerns
  - Future improvements
- Summary

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



Critical Services CMS

CERN IT Department

https://twiki.cern.ch/twiki/bin/view/CMS/SWIntCMSServices

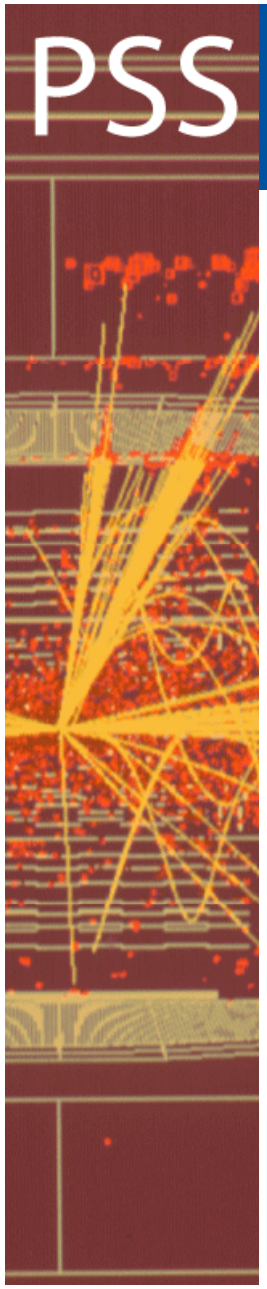
Rank	Definition	Max. downtime per incident (Hrs)	Comment
11	CMS Stops operating	0.5	Not covered (yet) here
10	CMS stops transferring data form Cessy		Cessy output buffer time
9	T0 Production stops		min(T0 input buffer/CESSY output buffer) or defined time to catch up
8	T1/T2 Production/analysis stops		defined time to catch up
7	Services critical when needed but not needed all the time (currently includes documentation)	0.5	
6	A service monitoring or documenting a critical service	8	
5	CMS development stops if service unavailable	24	
4	CMS development at CERN stops if service unavailable	24	
3	Services not critical CMS	24	
2	Services required for CMS	72	
1	Used by a significant fraction of CMS	72	
0	Not used or discouraged by CMS	forever	

The list of Services agreed by the task force in July 2007 is ...

Service	Description	IT/CMS	Rank	LB/HAS/SS ?	Notes	IT contact	CMS contact	Current Machines/resources
Oracle	Main Orcale back end. Serves a number of other services.	IT	10	HAS	There may be demand for mySQL			

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

CERN



Critical Services ATLAS

CERN IT Department

https://twiki.cern.ch/twiki//bin/view/Atlas/AtlasCriticalServices

This page contains a running list of the critical services that are needed for ATLAS computing operations at CERN and on the WLCG Grids.

The services are divided into 3 categories, according to their "criticality":

- **Very high**: interruption of these services affects online data-taking operations or stops any offline operations
- **High**: interruption of these services perturbs seriously offline computing operations
- **Moderate**: interruption of these services perturbs software development and part of computing operations

ATLAS_Computing_Services-Nov07.xls

Tier	Service	Criticality	Consequences of service interruption
0	Data transfer from Point1 to Castor	High	Short (<1 day): events buffered in SFO disks, backlog transfered as connection is resumed. Long (>1 day): loss of data.
0 (but is part of online system)	Oracle database RAC (online, ATONR)	Very high	Possible loss of DCS, Run Control, and Luminosity Block data while running. Run start needs configuration data from the online database. Buffering possibilities being investigated.
0	Online-offline database connectivity	High	No export of conditions data, this delays offline calibration and reconstruction. Import of calibration and configuration data proceed via files and does not require this connection.
0	Castor internal data movement	High	Slow down or interruption of Tier-0 processing. Buffers will fill up after 5 days.
0	Tier-0 processing farm	High	Slow down or interruption of Tier-0 processing. Buffers will fill up after 5 days.
0	Oracle database RAC	High	Slow down or interruption of Tier-0 processing. Buffers will fill up after 5

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

Critical Services Alice

http://indico.cern.ch/conferenceOtherViews.py?view=standard&confId=20080

Critical services list

- WLCG WMS (hybrid mode OK)
  - LCG RB
  - gLite WMS (gLite VO-box suite a must)
- FTS for T0->T1 data replications
  - SRM v.2.2 @ T0+T1s
- CASTOR2 + xrootd @ T0
- MSS with xrootd (dCache, CASTOR2) @ T1
- PROOF@CAF @ T0

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

Critical Services LHCb

https://twiki.cern.ch/twiki/bin/view/LHCb/CCRC08

Rank	Definition	Max downtime (hrs)	Comment
10	Critical	0.5	
7	Serious disruption	8	
5	Major reduction in effectiveness	8	
3	Reduced effectiveness	24	
1	not critical	72	

Service	Rank	Comment
CERN VO boxes	10	
CERN LFC service	10	
VOMS proxy service	10	
T0 SE	7	
T1 VOboxes	7	
SE access from WN	7	
FTS channel	7	both CERN to/from T1 & inter T1
WN misconfig	7	
CE access	7	
Conditions DB access	7	
LHCb Bookkeeping service	7	AMGA service hosted on ORACLE@CERN
Oracle streaming from CERN	7	
SAM service	7	We should rely on this to OK a site for use in RB?
LHCb RB	5	both at CERN and T1
T1 LFC service	3	
Dashboard	3	

-- NickBrook - 13 Nov 2007

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

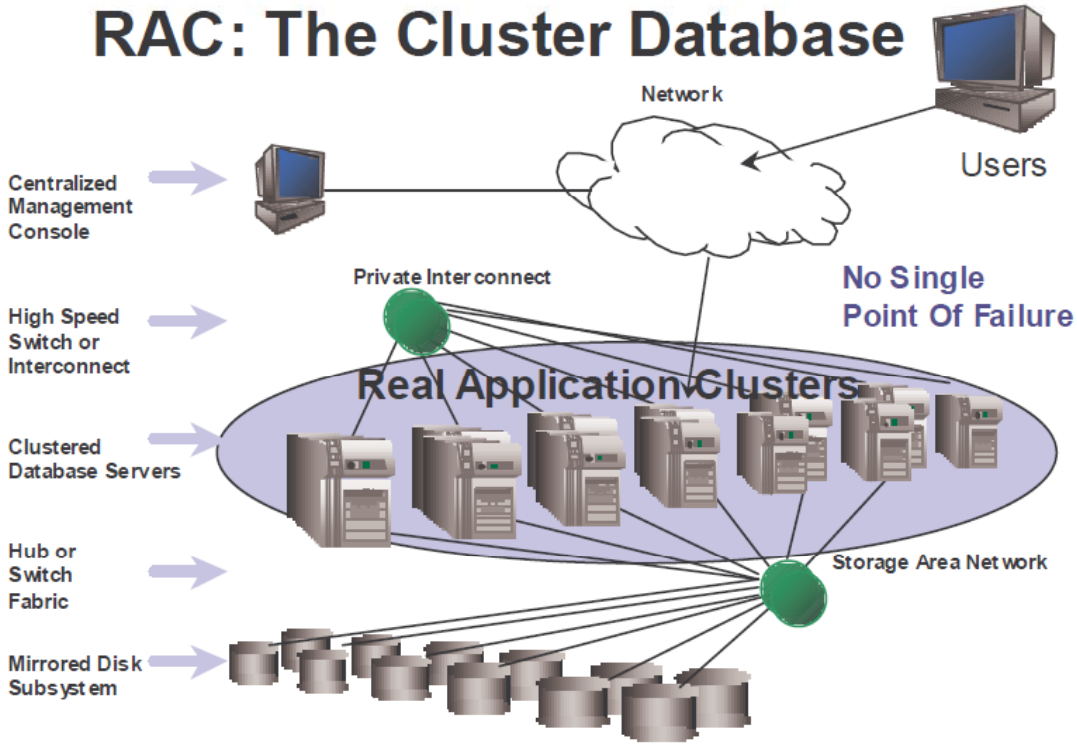
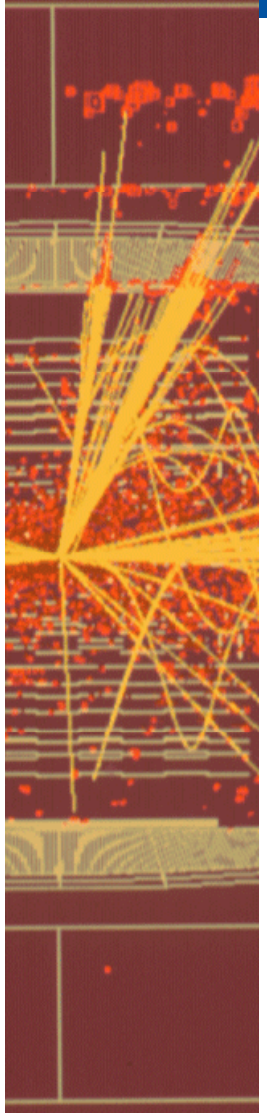


Figure 1: Oracle RAC –clustering database servers – foundation for Enterprise Grid Computing delivering high availability, scalability and flexibility.

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

The reality at CERN (Oct07)

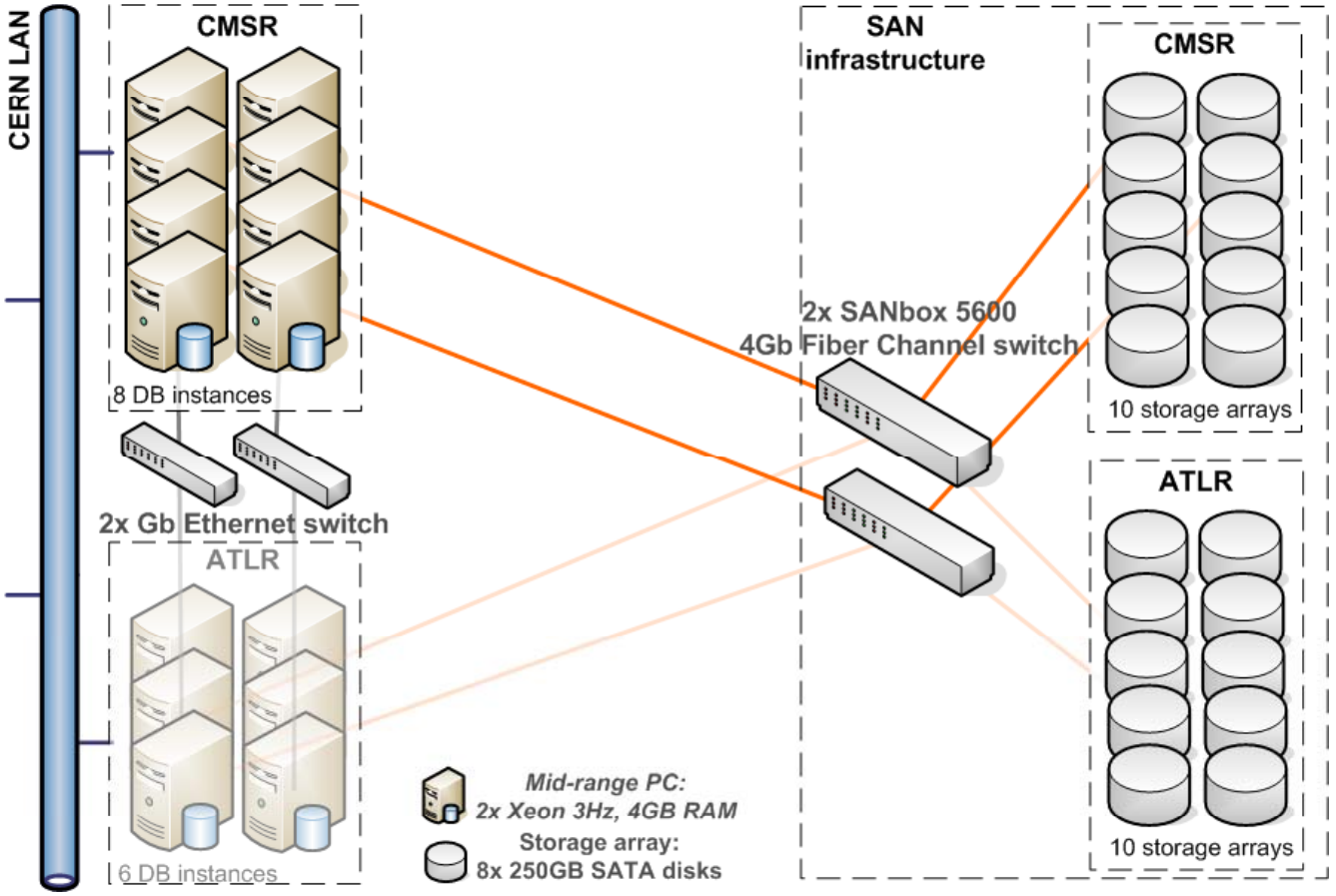


- 18 RACs up to 8 nodes running for production and integration services (LHC experiments, Grid, non-LHC)
  - 110 mid-range servers and 110 disk arrays (~1100 disks)
  - Or: 220 CPUs, 440GB of RAM, 300 TB of raw disk space(!)
- Recently connected to the 10 Tier1 sites for synchronized databases (3D project)
  - Sharing policies and procedures
- Team of 6 DBAs + Maria Girone service coordination and link to experiments
- 24x7 best effort service for production RACs
- Maintenance without downtime within RAC features
  - 99.99% production services availability (Oct2007)
    - Problems with SAM (max 3.5 hours unavailable)
  - 93.46% production servers availability (Oct2007)
    - Patch deployment, broken hardware

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

Architecture

- 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

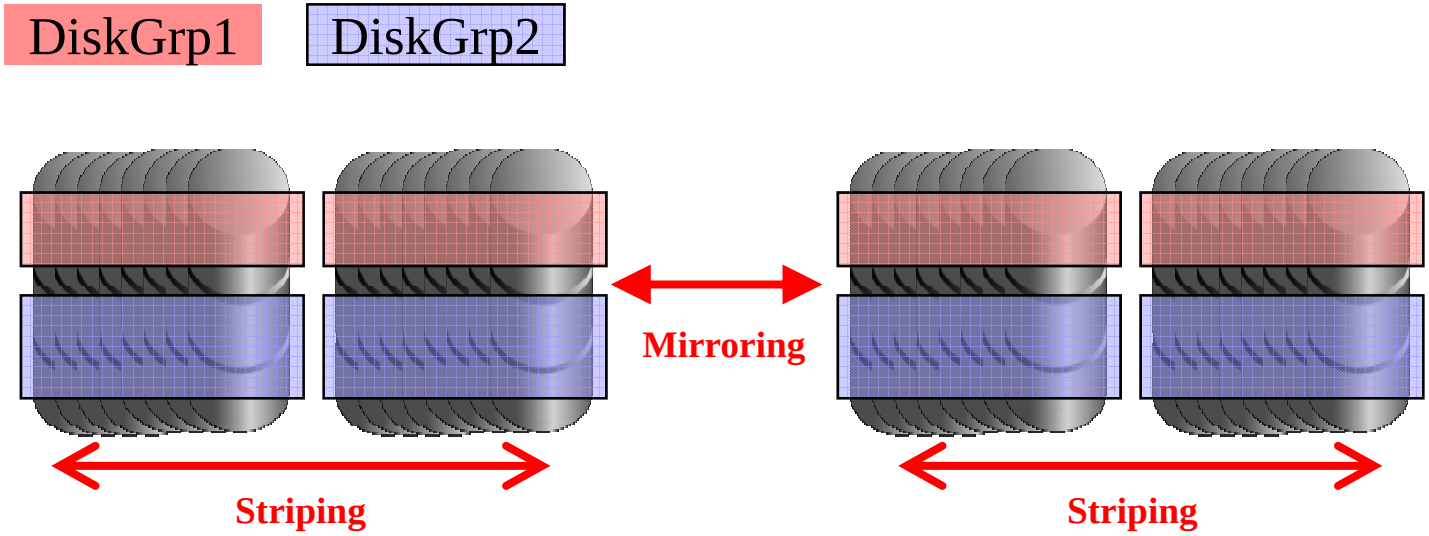
Architecture

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

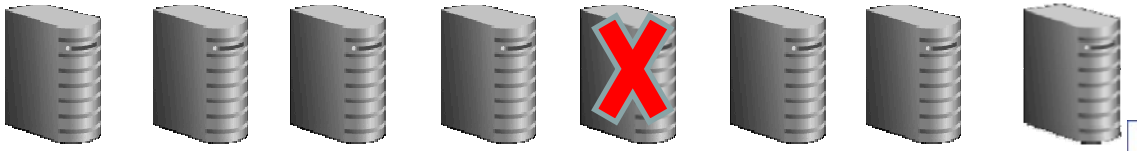
Architecture (services)

- 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

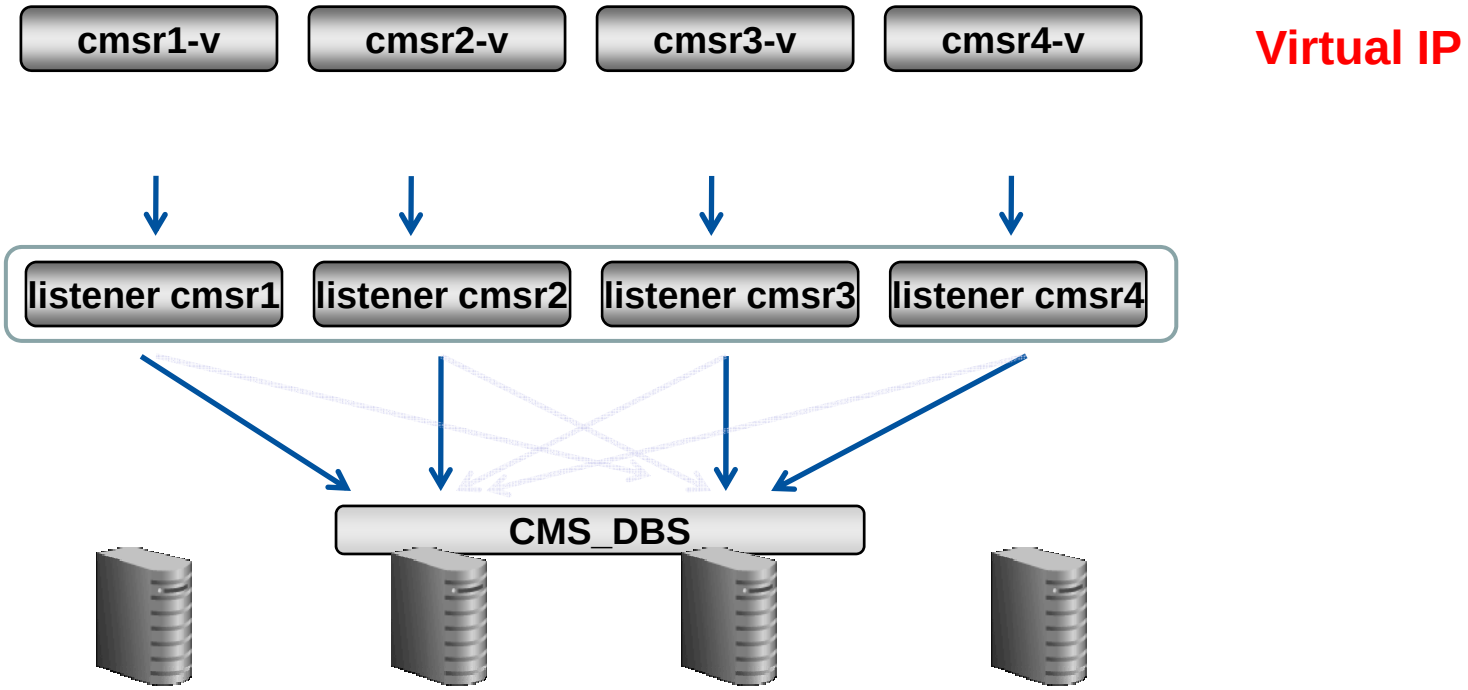


CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

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
```

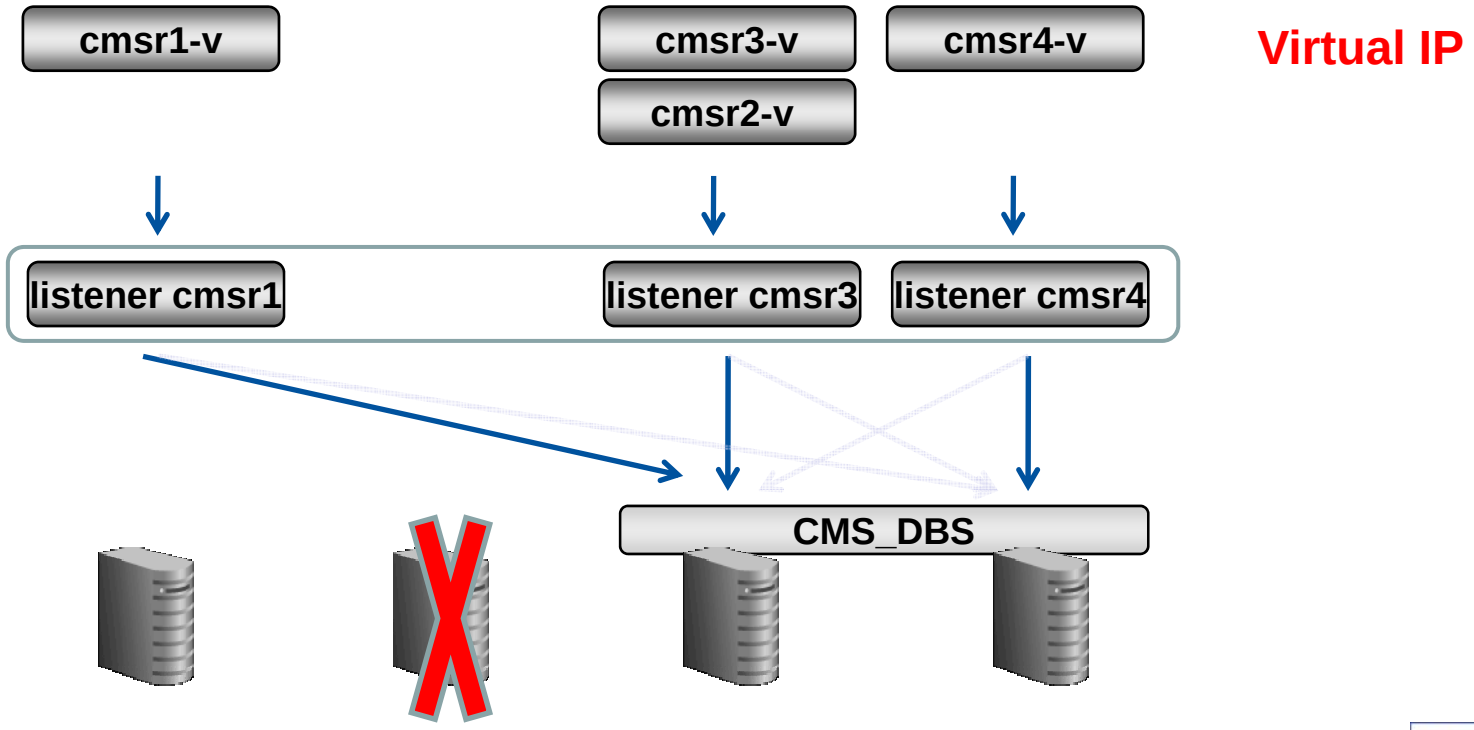


CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

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

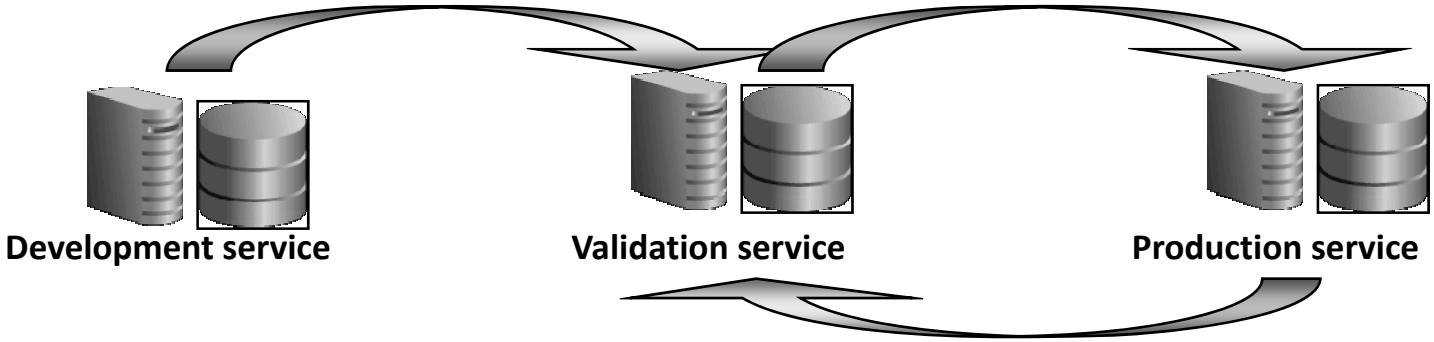
```
$ sqlplus cms_dbs@cms_dbs
```



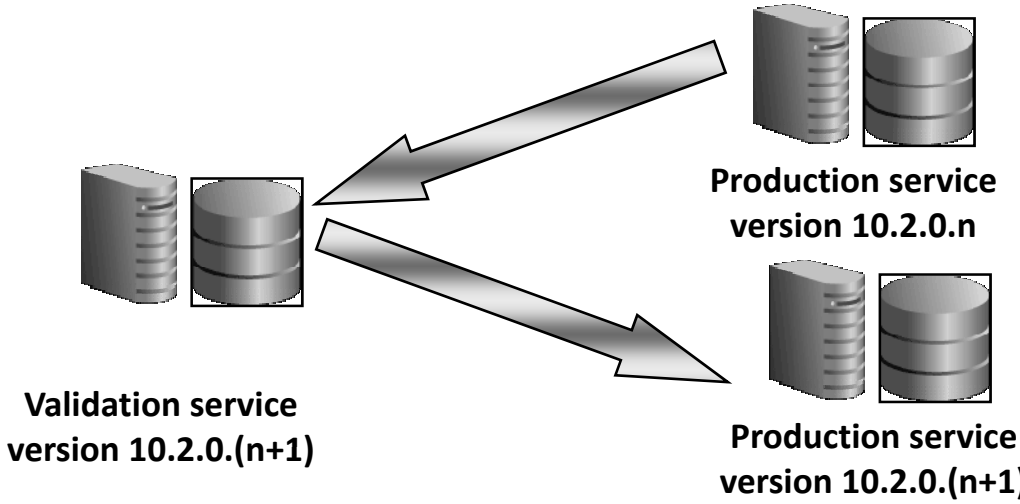
CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Apps and Database release cycle

- Applications' release cycle

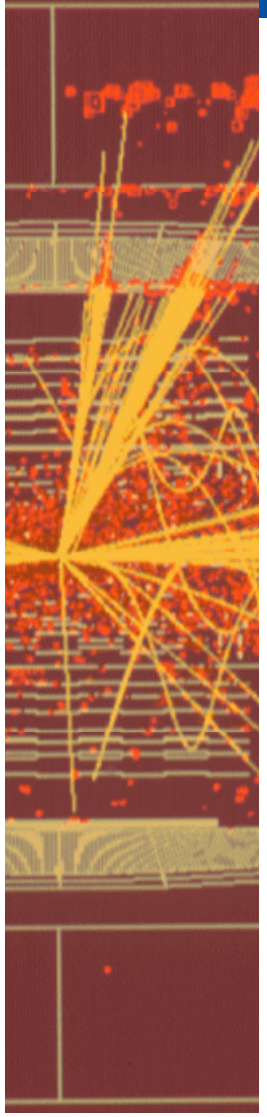


- Database software release cycle



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

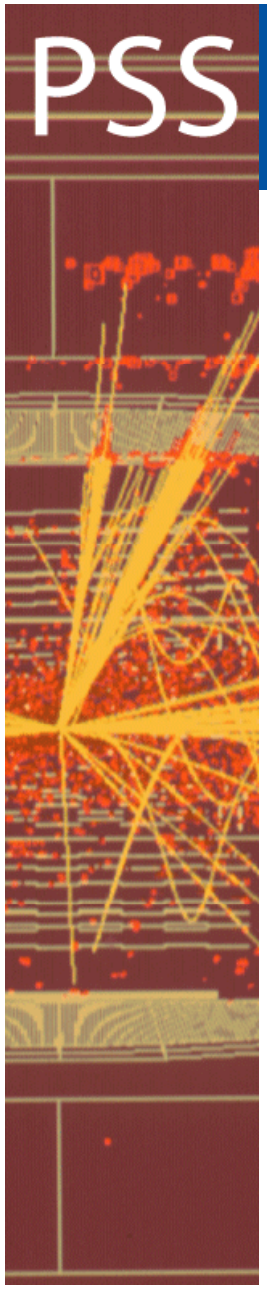
Administrative work



- Install and configure Oracle RAC databases
  - Expertise from several groups
- Apps developers support (remedy line)
- Account creation, application deployment
- On call rota 24x7
- Oracle and Linux patches deployment and test
  - /etc/nospma to avoid automatic OS updates
- Application optimization in cooperation with developers and experiment DBAs
  - Putting in place a policy for DB access for experiment DBAs
- Active monitoring

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

Monitoring

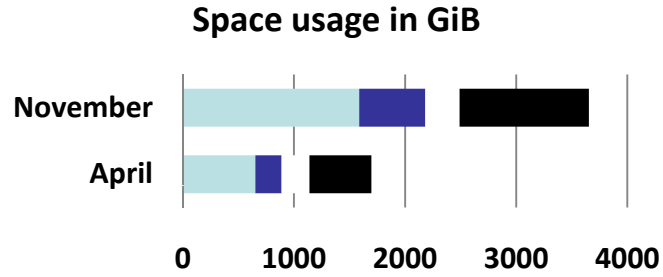


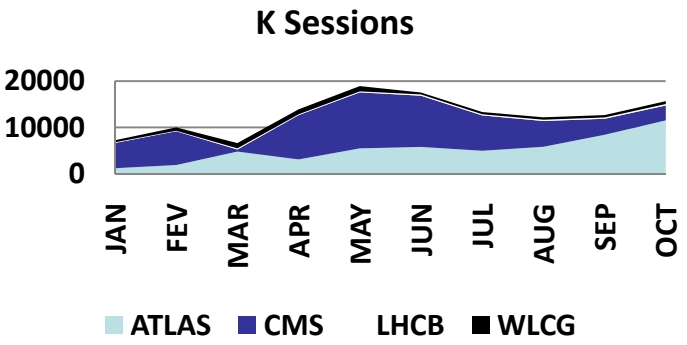
- 24x7 reactive monitoring
  - Lemon Alarms, Operators, SysAdmins
    - HW failures, OS problems, High load
  - Host and, instance and service availability
    - home grown monitoring
- Active monitoring
  - Oracle Enterprise Manager
    - Execution plans, resource usage per service
  - 3D monitoring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

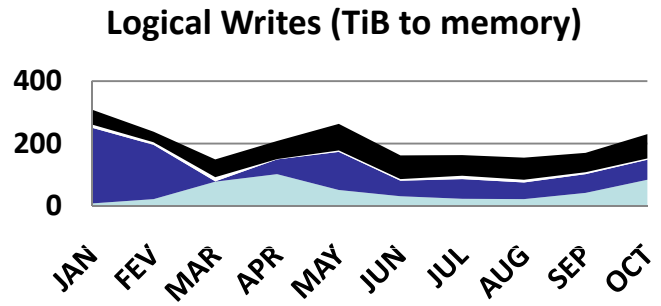
CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

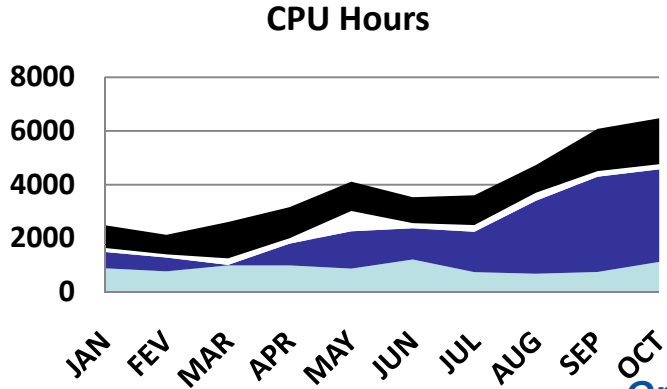
Current and Future usage

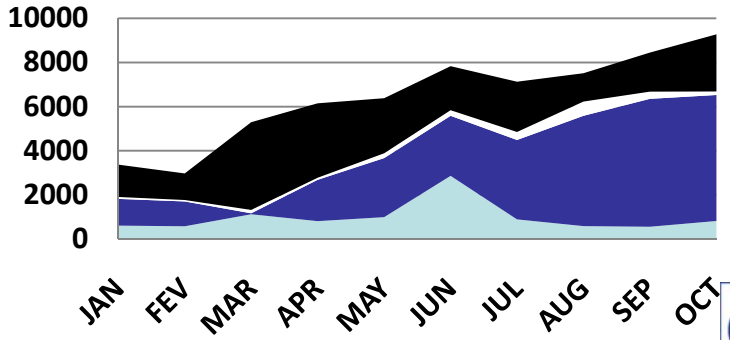
- Approached by the LHC experiments for service provision for the online database, including Online to Offline databases streaming











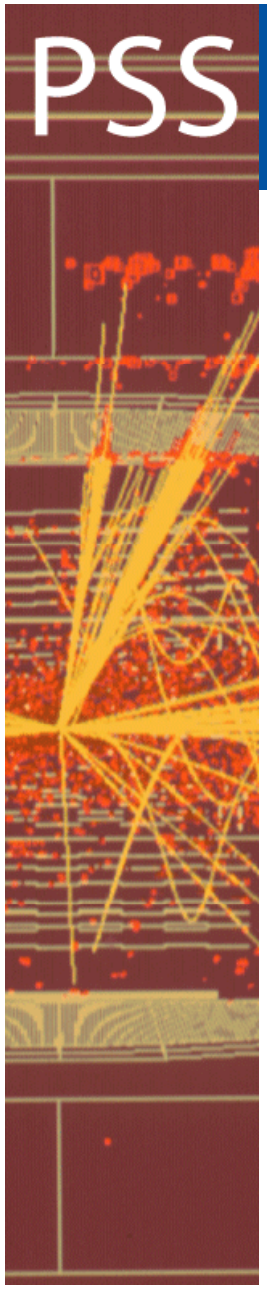
CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Main concerns (based on experience)

- Human errors
  - on administrative tasks (SAM problem on tidying up old data)
    - test better the procedures, not always easy task
- Human errors
  - By developers (SAM dropped tables)
    - Restrict access to production accounts to developers
- Logical corruption / Oracle SW bugs
  - VOMS/LFC data inserted in wrong VOs
    - Better testing on pilot environments before deployment of new patches in production

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

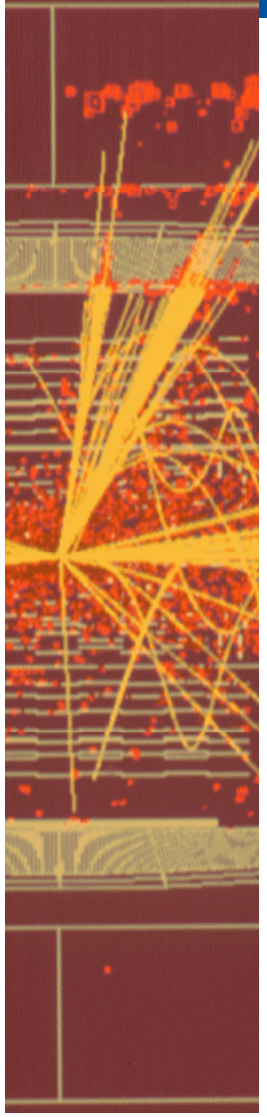
Main concerns (based on experience)



- Oracle software Security
  - Quarterly security patches released by Oracle
    - Firewalls, no default configuration, restrict access to essential
    - Do not publish connection data on web
- Increasing amount of stored data
  - Tapes slow as 5 years ago, backups take longer
    - Move to backup on disks
    - prune old redundant data/summarizing

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

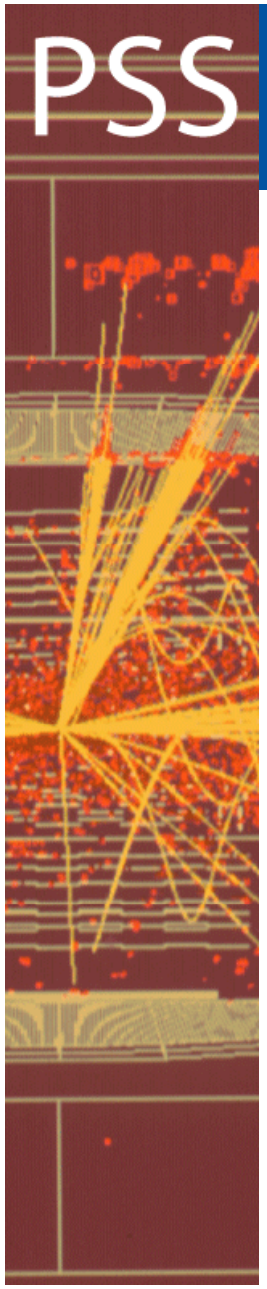
Future improvements



- Quad-core servers
  - Smaller RACs, 1 quad-core better than 4 dual CPU single core
- 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

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

Summary



- Clustering of redundant hardware
- Eliminate single points of failure
- Applications mapped to ‘oracle services’ to better allocate resources and avoid starvation
- Validation and production release cycles
- Active monitoring by 6 DBAs, reports
- Very good contact with experiment coordinators
- Average service availability 99.99% (oct 07)
- On-call 24x7 best effort service does not fulfill requirements from experiments
- Need cooperation of the application developers to hide glitches

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