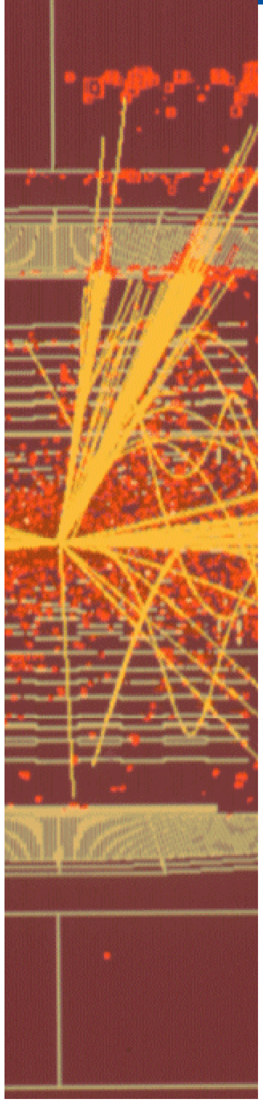


Database Services for Physics Plan for 2008

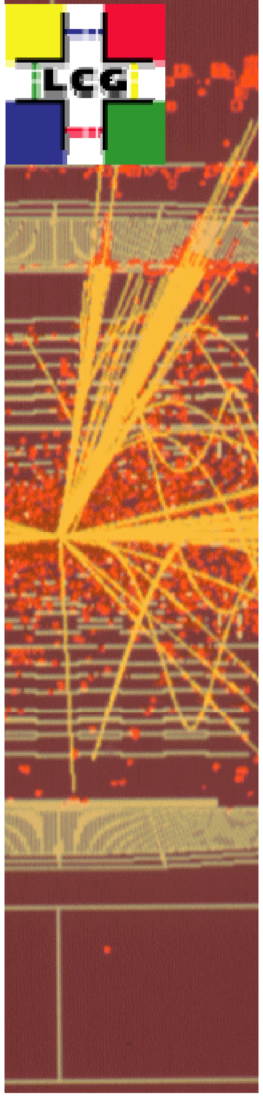
Maria Girone, CERN

WLCG Service Reliability Workshop

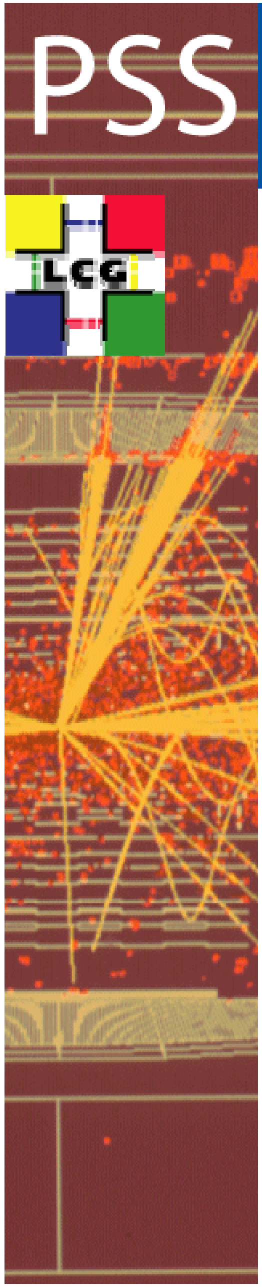
26-30th November 2007

Phydb.support@cern.ch

DB Services for Physics

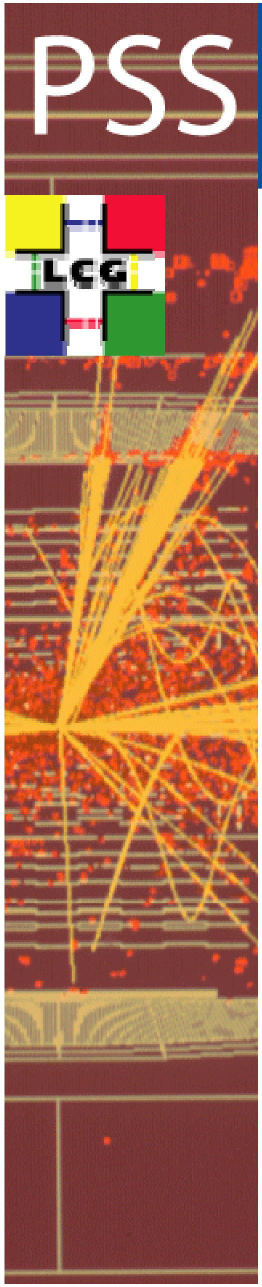


- Database services for physics are classified by the experiments among the highly critical services
  - 6 DBAs (5 in 2008) for 24x7 on "best effort"
  - Need to match service level expectations
    - E.g. 30 mins maximum down-time
- We have set up a database infrastructure for the WLCG
  - RAC as building-block architecture
    - Several 8-node clusters at Tier0
    - Typically 2-node clusters at Tier1
  - Homogeneous h/w and s/w configuration
  - Scalability and high availability achieved
  - ➢ Most of maintenance operations w/o down-time!
  - Backup and software update common policies
- At Tier0, three service levels are essential
  - Development, test and production levels



DB Services for Physics (2)

- The hardware is deployed at the IT computer centre; the production clusters are connected to the critical power (UPS and diesel)
  - Discussing now with FIO for the allocation of the new hardware. Current proposal is to accommodate half of the servers out of critical power
- Need to review proper use of roles
- Need to address security issues from shared-accounts and passwords distribution and the external access to databases
- Approached by the LHC experiments for service provision for the online databases

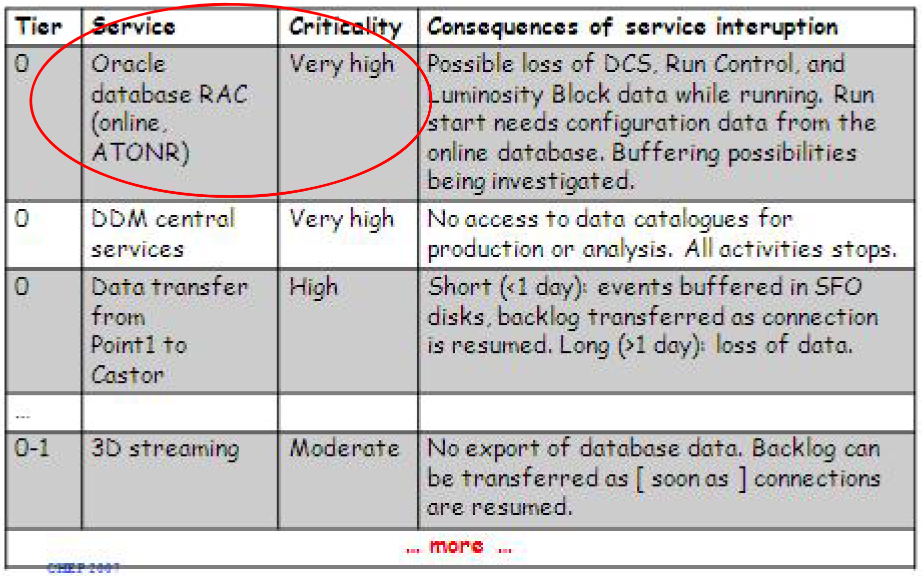


Streams Operations

- Oracle streams replication in production since April 2007 between
  - Tier0 and 10 Tier1 sites
    - ATLAS and LHCb
  - Online and offline at Tier0
    - ATLAS, CMS and LHCb

- Streams procedures included in the Oracle Tier0 physics database service team
  - Procedures review by Eva Dafonte Perez
  - 8x5 coverage
  - Optimized the redo log retention on downstream database to allow for sufficient resynchronization window without recall from tape (for 5 days)
  - Need to automate more the split-merge procedure when one site has to be dropped/re-synchronized
    - Progress in Oracle 11g but we need a stop-gap solution

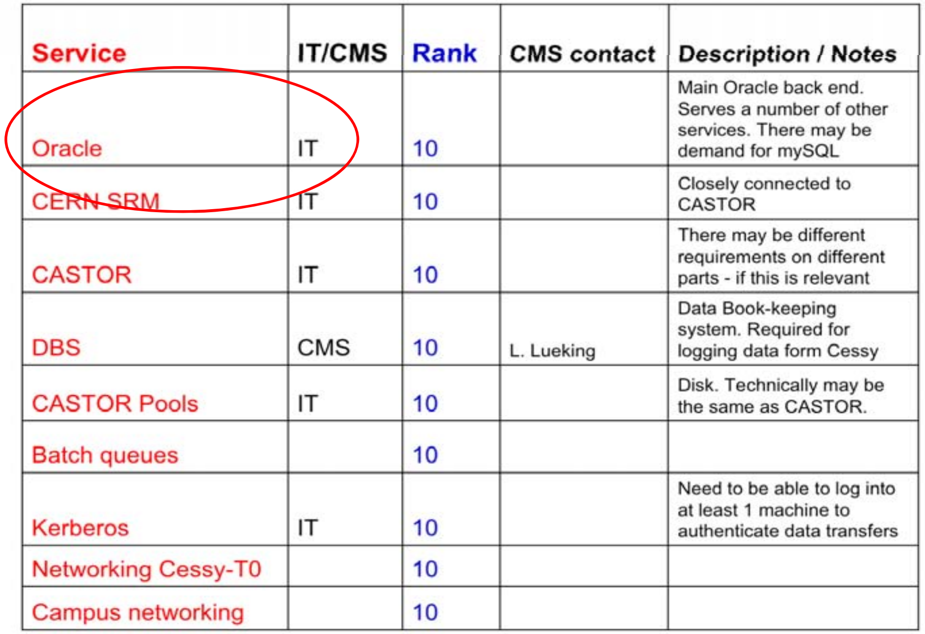
ATLAS Critical Services (PDF)



Tier	Service	Criticality	Consequences of service interuption
0	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	DDM central services	Very high	No access to data catalogues for production or analysis. All activities stops.
0	Data transfer from Point1 to Castor	High	Short (<1 day): events buffered in SFO disks, backlog transferred as connection is resumed. Long (>1 day): loss of data.
...			
0-1	3D streaming	Moderate	No export of database data. Backlog can be transferred as [soon as] connections are resumed.

... more ...

CMS Critical Services (wiki)

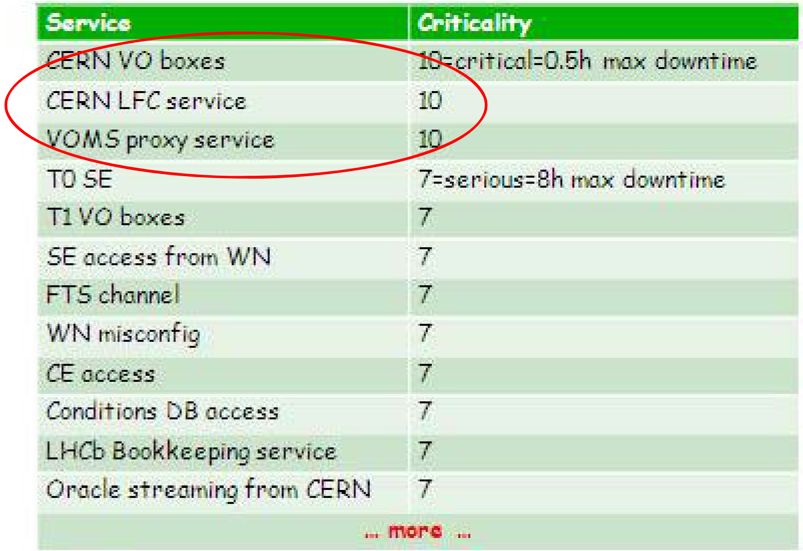


Service	IT/CMS	Rank	CMS contact	Description / Notes
Oracle	IT	10		Main Oracle back end. Serves a number of other services. There may be demand for mySQL
CERN SRM	IT	10		Closely connected to CASTOR
CASTOR	IT	10		There may be different requirements on different parts - if this is relevant
DBS	CMS	10	L. Lueking	Data Book-keeping system. Required for logging data form Cessy
CASTOR Pools	IT	10		Disk. Technically may be the same as CASTOR.
Batch queues		10		
Kerberos	IT	10		Need to be able to log into at least 1 machine to authenticate data transfers
Networking Cessy-T0		10		
Campus networking		10		

ALICE 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

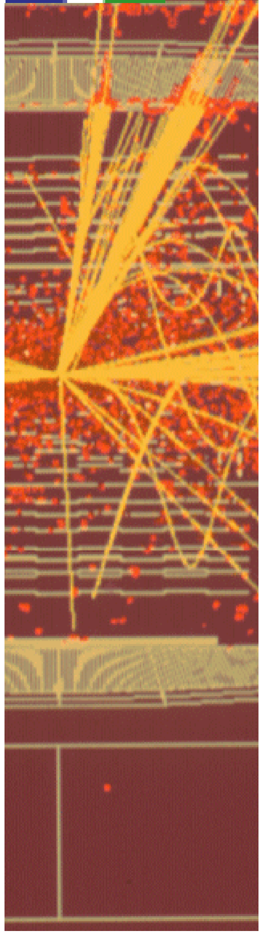
LHCb Critical Services (CCRC08 wiki)



Service	Criticality
CERN VO boxes	10=critical=0.5h max downtime
CERN LFC service	10
VOMS proxy service	10
T0 SE	7=serious=8h max downtime
T1 VO boxes	7
SE access from WN	7
FTS channel	7
WN misconfig	7
CE access	7
Conditions DB access	7
LHCb Bookkeeping service	7
Oracle streaming from CERN	7

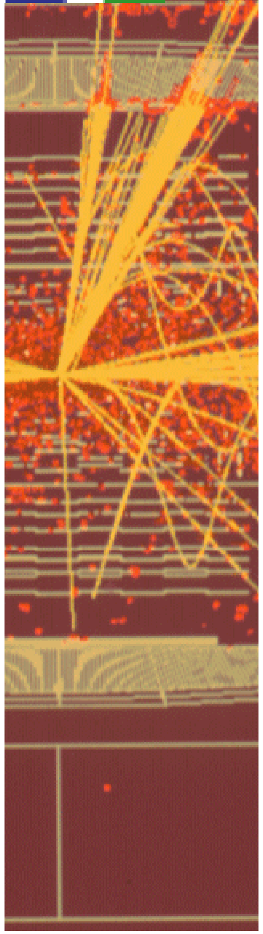
... more ...

Current Set-Up



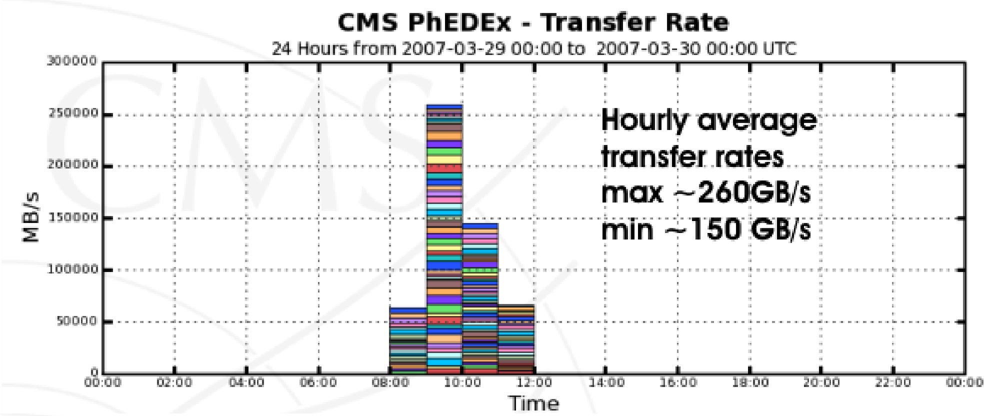
- RAC on commodity hardware - Full redundancy!
  - Linux RHES4 32bit as OS platform, Oracle ASM as volume Manager
  - Dual-CPU P4 Xeon @ 3GHz servers with 4GB of DDR2 400 memory each
  - SAN at low cost
    - FC Infortrend disk arrays, SATA disks, FC controller
    - FC QLogic switches SANBox (4Gbps)
    - Qlogic HBAs dual ported (4Gbps)
- Service Size
  - 110 mid-range servers and 110 disk arrays (~1100 disks)
  - In other words: 220 CPUs, 440GB of RAM, 80TB of effective disk space
- About 20 validation and production RACs, up to 8-node clusters

New Hardware Set-up



- New servers and disk arrays expected in Jan 2008
  - 34 dual-CPU quad-core Xeon processors servers, with 16GB of FB-DIMM memory
    - For memory and CPU intensive jobs a quad-core server performs as five-node RAC of our current set-up
  - 60 disk arrays (16 disks of 400GB each)
    - A total of 100 TB of effective space for the production services
- Would like to migrate the all our production RACs to the new hardware
  - Good for services that don't scale over multiple nodes
  - Will be on 64 bit, Oracle 10.2.0.4
  - Target date for deployment is March 2008 for CCRC'08. Need feedback from the experiments
- Will migrate our integration RACs to 64bit in January 2008
  - will give two months time for tests

Quad-core Performance Testing

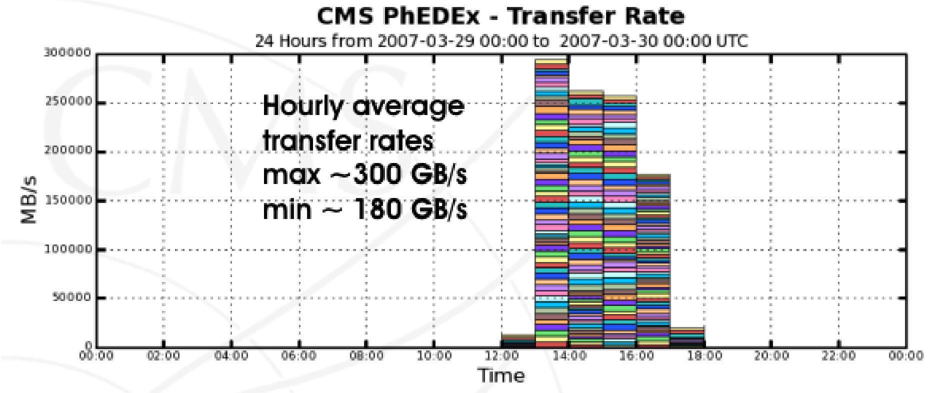


CMS PhEDEx - Transfer Rate

24 Hours from 2007-03-29 00:00 to 2007-03-30 00:00 UTC

Hourly average transfer rates max ~260GB/s min ~150 GB/s

6-node RAC



CMS PhEDEx - Transfer Rate

24 Hours from 2007-03-29 00:00 to 2007-03-30 00:00 UTC

Hourly average transfer rates max ~300 GB/s min ~ 180 GB/s

Quad-core server

A single quad-core server is able to handle PhEDEx-like workload (a transaction oriented application) even more efficiently then a 6-node RAC

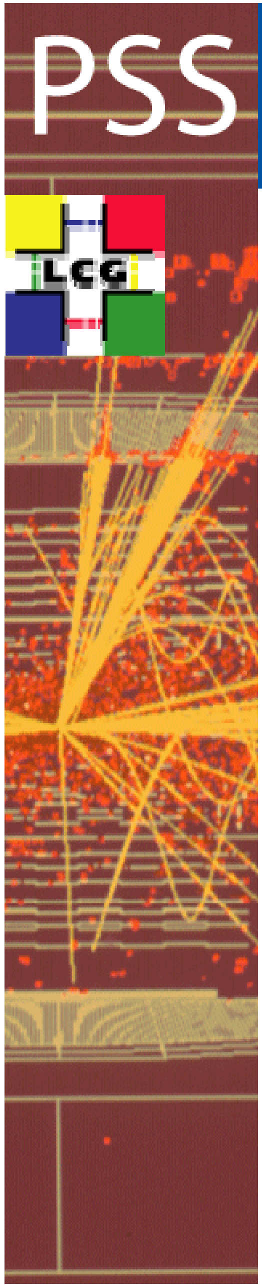
Hardware Allocation in 2008

- Production databases for LHC:
  - 3 or 4-node clusters built with quadcore CPU machines (24-32 cores per cluster)
  - 48-64 GB of RAM per cluster
  - Planning for >10k IOPS
  - TBs of mirrored space

- Integration and test systems:
  - Single core CPU hardware
  - Usually 2 nodes per cluster
  - Usually 24-32 disks

- 64bit version of Linux and Oracle software

- Migration tools have been prepared and tested to minimize the downtime of the production RACs
  - More details in Jacek Wojcieszuk's talk



Conclusions

- Database Services for physics at CERN run production and integration Oracle 10g services
  - **Designed to address the reliability, performance and scalability needs of WLCG user community**
    - *Application developers need to follow guidelines to profit from it*
    - *Approached by the LHC experiments for service provision for the online databases*
- Connected to the 10 Tier1 sites for synchronized databases since April 2007
  - Sharing policies. Need to discuss on procedures
- Would like to complete the migration of the productions RACs by March 2008
  - In time for CCRC'08
- Planning now the service growth for 2009-2010