WLCG Service & Operations: Issues & Concerns

Jamie.Shiers@cern.ch

~ ~ ~

WLCG ←→ Oracle Operations review meeting, April 2009

Databases - Plenary

- The Online systems use a lot of data-bases:
  - Run database, Configuration DB, Conditions DB, DB for logs, for logbooks, histogram DB, inventory DB, …
  - Not to forget: the archiving of data collected from PVSS (used by all detector control systems)
- All experiments run Oracle RAC infrastructures, some use in addition MySQL, object data-base for ATLAS Configuration (OKS)
- Administration of Oracle DBs is largely outsourced to our good friends in the CERN IT/DM group
- Exchange of conditions between offline and online uses Oracle streaming (like replication to Tier1s)

Databases - Track Summary

- Web Service
  - Standard tools or home-grown
  - API and Web pages
- Distribution, Oracle Streams
- Proxies and caches, Frontier
- Client-Server
- Oracle "by design", SQLite
- Conditions databases

What Needs to be Improved?

- We still see “emergency” interventions that could be “avoided” – or at least foreseen and scheduled at a more convenient time
  - Often DB applications where e.g. tables grow and grow until performance hits a wall → emergency cleanup (that goes wrong?); indices “lost”; bad manipulations; ...
- We still see scheduled interventions that are not sufficiently well planned – that run well into “overtime” or have to be “redone”
  - e.g. several Oracle upgrades end last year overran – we should be able to schedule such an upgrade by now (after 25+ years!)
- Or those that are not well planned or discussed with service providers / users and have a big negative impact on ongoing production
  - e.g. some network reconfigurations and other interventions – particularly on more complex links, e.g. to US; debugging and follow-up not always satisfactory
- There are numerous concrete examples of the above concerning many sites: they are covered in the weekly reports to the MB and are systematically followed up

Much more serious are chronic (weeks, months) problems that have affected a number of Tier1 sites – more later…

Major Service Incidents

- Quite a few such incidents are “DB-related” in the sense that they concern services with a DB backend
  - The execution of a “not quite tested” procedure on ATLAS online led – partly due to the Xmas shutdown – to a break in replication of ATLAS conditions from online out to Tier1s of over 1 month (online-offline was restored much earlier)
  - Various Oracle problems over many weeks affected numerous services (CASTOR, SRM, FTS, LFC, ATLAS conditions) at ASGC → need for ~1FTE of suitably qualified personnel at WLCG Tier1 sites, particularly those running CASTOR; recommendations to follow CERN/3D DB configuration & perform a clean Oracle+CASTOR install; communication issues
  - Various problems affecting CASTOR+SRM services at RAL over prolonged period, including “Oracle bugs” strongly reminiscent of those seen at CERN with earlier Oracle version: very similar (but not identical) problems seen recently at CERN & ASGC (not CNAF...)
- Plus not infrequent power + cooling problems **[+ weather!]**
  - Can take out an entire site – main concern is controlled recovery (and communication)

At the November 2008 WLCG workshops a recommendation was made that each WLCG Tier1 site should have at least 1 FTE of DBA effort.

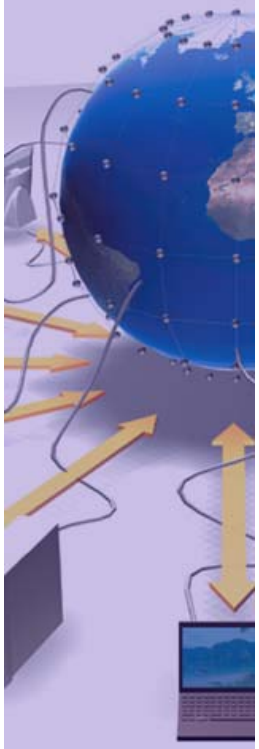
This effort (preferably spread over multiple people) should proactively monitor the databases behind the WLCG services at that site: CASTOR/dCache, LFC/FTS, conditions and other relevant applications.

The skills required include the ability to backup and recover, tune and debug the database and associated applications.

At least one WLCG Tier1 does not have this effort available today.

ALICE critical services

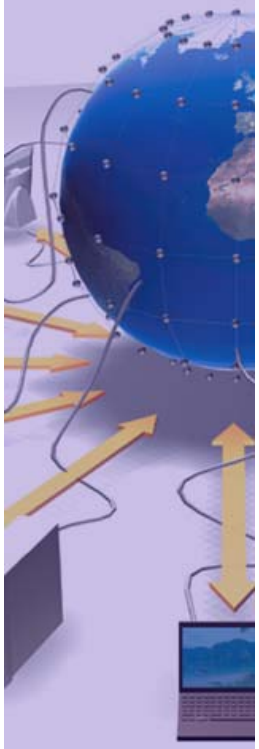
Critical services	Rank	Comment
AliEN	10	ALICE computing framework
Site VO boxes	10	Site becomes unusable if down
CASTOR and xrootd at Tier-0	10	Stops 1st pass reco (24 hours buffer)
Mass storage at Tier-1	5	Downtime does not prevent data access
File Transfer Service at Tier-0	7	Stops 2nd pass reco
gLite workload management	5	Redundant
PROOF at Tier-0 CERN Analysis Facility	5	User analysis stops



Rank 10: critical, max downtime 2 hours
Rank 7: serious disruption, max downtime 5 hours
Rank 5: reduced efficiency, max downtime 12 hours

CERN IT Department
CH-1211 Genève 23
Switzerland

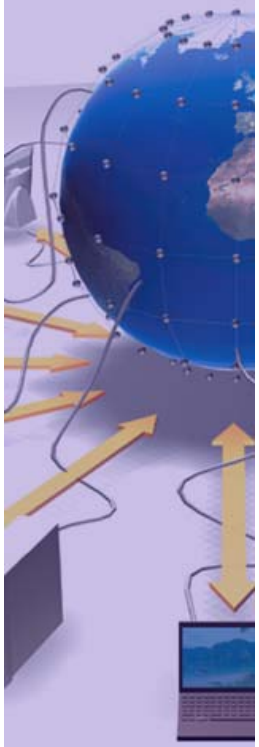
ATLAS criticality



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

CERN IT Department
CH-1211 Genève 23
Switzerland

ATLAS critical services



Rank	Services at Tier-0
Very high	Oracle (online), DDM central catalogues, Tier-0 LFC
High	Cavern→T0 transfers, online-offline DB connectivity, CASTOR internal data movement, Tier-0 CE, Oracle (offline), Tier-0/1 FTS, VOMS, Tier-1 LFC, Dashboard, Panda/Bamboo, DDM site services/VO boxes
Moderate	3D streaming, gLite WMS, Tier-0/1/2 SRM/SE, Tier-1/2 CE, CAF, CVS, Subversion, AFS, build system, Tag Collector

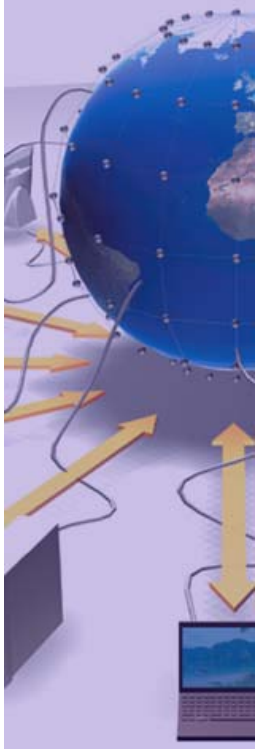
Rank	Services at Tier-1
High	LFC, FTS, Oracle
Moderate	3D streaming, SRM/SE, CE

Rank	Services at Tier-2
Moderate	SRM/SE, CE

Rank	Services elsewhere
High	AMI database

CERN IT Department
CH-1211 Genève 23
Switzerland

CMS criticality



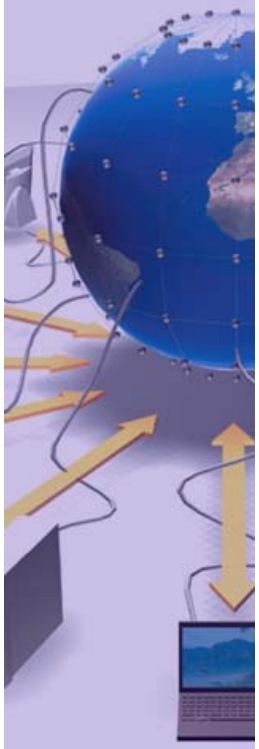
CMS gives a special meaning to all rank values

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 for 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	

Rank 10: 24x7 on call
Rank 8,9: expert call-out

CERN IT Department
CH-1211 Genève 23
Switzerland

CMS critical services



Rank	Services
10	Oracle, CERN SRM-CASTOR, DBS, Batch, Kerberos, Cavern-T0 transfer+processing
9	CERN FTS, PhEDEx, FroNTier launchpad, AFS, CAF
8	gLite WMS, VOMS, Myproxy, BDII, WAN, Non-T0 prod tools
7	APT servers, build machines, Tag collector, testbed machines, CMS web server, Twiki
6	SAM, Dashboard, PhEDEx monitoring, Lemon
5	WebTools, e-mail, Hypernews, Savannah, CVS server
4	Linux repository, phone conferencing, valgrind machines
3	Benchmarking machines, Indico

CERN IT Department
CH-1211 Genève 23
Switzerland

LHCb critical services

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	

Rank	Services
10	Tier-0 CASTOR, AFS, CERN VO boxes (DIRAC3 central services), Tier-0 LFC master, Oracle at CERN
7	VOMS, CERN FTS, Tier-0 ConditionDB, LHCb bookkeeping service, Oracle Streams, SAM
5	Tier-0/1 CE and batch, Tier-0 gLite WMS, Tier-1 ConditionDB
3	Tier-1 SE, Tier-0/1 Replica LFC, Dashboard, Tier-1 FTS, Tier-1 gLite WMS
1	Tier-1 VO boxes

CERN IT Department
CH-1211 Genève 23
Switzerland

Service Impact

- At least for ATLAS and CMS, database services and services that depend on them are ranked amongst the most critical of all!

➢ **We must remember our motto "by design" in ensuring that not only are service interruptions and degradations avoided where possible but also that any incidents are rapidly resolved.**

10: Very high	interruption of these services affects online data-taking operations or stops any offline operations
7: High	interruption of these services perturbs seriously offline computing operations

Concrete Actions

1. Review on a regular (3-6 monthly?) basis open Oracle "Service Requests" that are significant risk factors for the WLCG service (Tier0+Tier1s+Oracle)
   - The first such meeting is being setup, will hopefully take place prior to CHEP 2009
2. Perform "technology-oriented" reviews of the main storage solutions (CASTOR, dCache) focussing on service and operational issues
   - Follow-on to Jan/Feb workshops in these areas; again report at pre-CHEP WLCG Collaboration Workshop
3. Perform Site Reviews – initially Tier0 and Tier1 sites – focussing again and service and operational issues.
   - Will take some time to cover all sites; proposal is for review panel to include members of the site to be reviewed who will participate also in the review before and after their site

The Goal

- The goal is that – by end 2009 – the weekly WLCG operations / service report is quasi-automatically generated 3 weeks out of 4 with no major service incidents – just a (tabular?) summary of the KPIs

➢ **We are currently very far from this target with (typically) multiple service incidents that are either:**
  - New in a given week;
  - Still being investigating or resolved several to many weeks later

- By definition, such incidents are characterized by severe (or total) loss of service or even a complete site (or even Cloud in the case of ATLAS)

Examples and Current High-Priority Issues

- Atlas, Streams and LogMiner crash
- CMS Frontier, change notification and Streams incompatibility
- Castor, BigID issue
- Castor, ORA-600 crash
- Castor, crosstalk and wrong SQL executed
- Oracle clients on SLC5, connectivity problem
- Restore performance for VLDB

Summary of Issues

Issue	Services Involved	Service Request	Impact
Logminer	ATLAS Streams	SR 7239707.994	PVSS, conditions service interruptions (2/3 days); Replication to T1s severly affected
Logminer	Streams		Proposed 'workaround' not acceptable -> data loss!
Streams/ Change notification	CMS Frontier	SR 7280176.993	Incompatibility with Streams
BigID	CASTOR	SR 7299845.994	Service interruption
"Crosstalk"	CASTOR + ??	SR 7103432.994	Logical data corruption
ORA-600	CASTOR	SR 7398656.993	Service interruption
	Clients on SL5		Clients can't connect!

Status of Discussions with Oracle

- After some initial contact and e-mail discussions, a first preparatory meeting is scheduled for April 6th at Oracle's offices in Geneva

- There is already a draft agenda for this meeting
  - Introductions & Welcome
  - WLCG Service & Operations – Issues and Concerns
  - WLCG Distributed Database Operations: Summary of Current & Recent Operational Problems
  - Oracle Support – Reporting and Escalating Problems for Distributed Applications (Customers)
  - Discussion on Frequency, Attendance, Location, …
  - Wrap-up & Actions
    - http://indico.cern.ch/conferenceDisplay.py?confId=53482

➢ **A report on this meeting will be given at the Distributed DB Operations meeting to be held in PIC later in April**
  - Agenda: http://indico.cern.ch/conferenceOtherViews.py?view=standard&confId=54037

Outlook

- Depending on the outcome of the discussions with Oracle, we would expect to hold such “WLCG Oracle Operational” (W) reviews roughly quarterly
- RAL have already offered to host such a meeting
- The proposed initial scope is to:
  - Follow-up on outstanding service problems;
  - Be made aware of critical bugs in production releases, together with associated patches and / or work-arounds.

➢ **Target: reliable & available services “by design”**