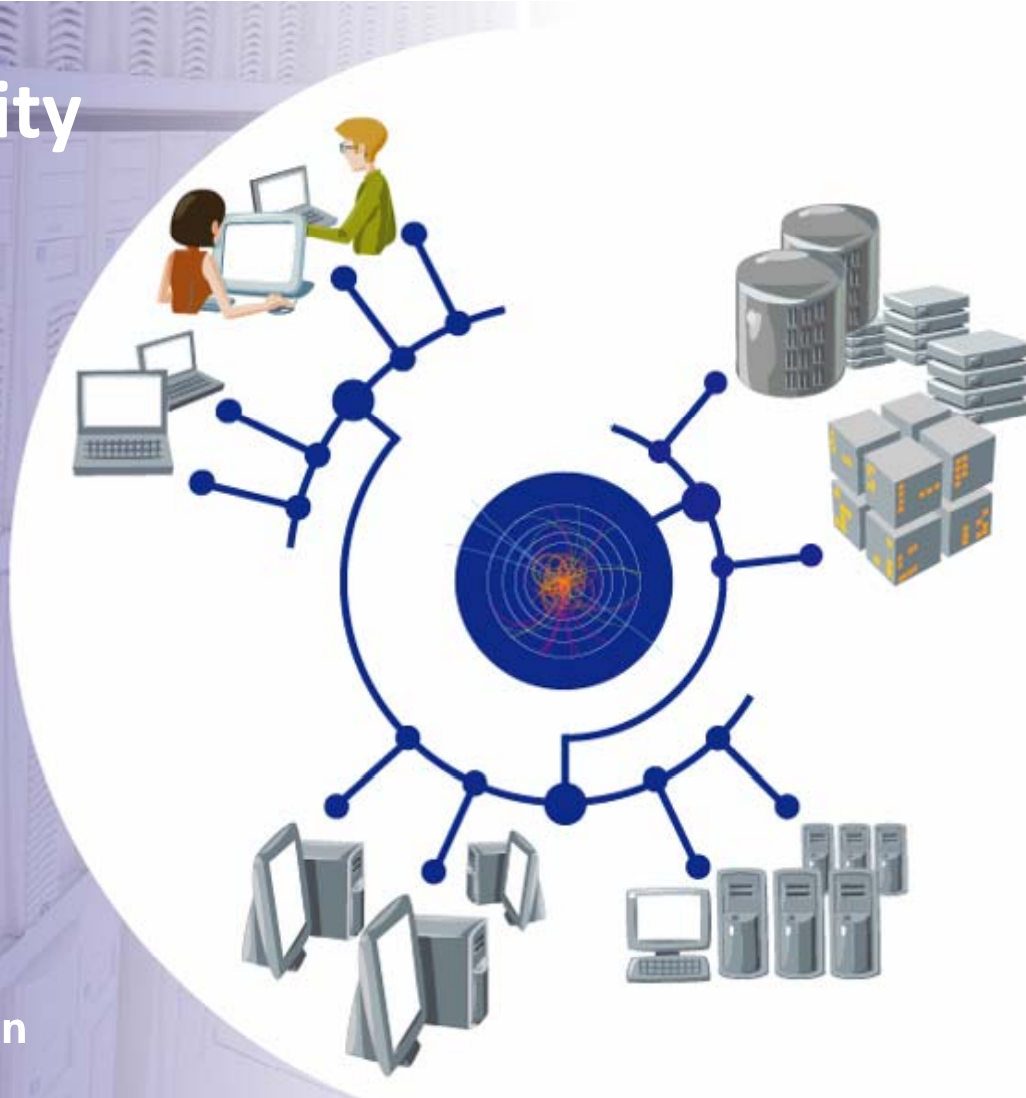




# WLCG – Worldwide LHC Computing Grid

## Service Reliability

Workshop Introduction,  
November 26<sup>th</sup> 2007



Jamie Shiers  
WLCG Service Coordination





# Agenda

- The goal of this workshop is to understand how to build, deploy and operate robust and reliable services
- The driving force is the list of critical services as viewed by the four LHC experiments – plus also WLCG MoU requirements
- These will be presented next, followed by techniques and experience with wide-scale service delivery
- **💣 Please note – there are constraints! Presenting a request here is not a guarantee that it can be met!**
- There is no workshop summary *per se*, but there will be reports to next week's WLCG Overview Board (OB) and Grid Deployment Board (GDB)
- It is also foreseen to extend this discussion to WLCG Tier2 sites, possibly at the [April 2008 Collaboration Workshop](#)



## Pros & Cons – Managed Services

☺ **Predictable service level and interventions; fewer interventions, lower stress level and more productivity, good match of expectations with reality, steady and measurable improvements in service quality, more time to work on the physics, more and better science, ...**

☹ **Stress, anger, frustration, burn-out, numerous unpredictable interventions, including additional corrective interventions, unpredictable service level, loss of service, less time to work on physics, less and worse science, loss and / or corruption of data, ...**



**This workshop is about the 1<sup>st</sup> column**



# CMS Critical Services ([wiki](#))

Rank	Definition	Max. Downtime	Comments
11	CMS Stops Operating	0.5 hours	Not covered yet
10	CMS stops transferring data from 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		
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		
... more ...			



# ATLAS Critical Services (PDF)

Tier	Service	Criticality	Consequences of service interruption
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 ...			



# LHCb Critical Services ([CCRC08 wiki](#))

Service	Criticality
CERN VO boxes	10=critical=0.5h max downtime
CERN LFC service	10
VOMS proxy service	10
TO 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 ...	



## 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





## ATLAS Critical Services ([PDF](#))

Tier	Service	Criticality	Consequences of service interruption
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 ...

CHEP 2007



## CMS Critical Services ([wiki](#))

Rank	Definition	Max. Downtime	Comments
11	CMS Stops Operating	0.5 hours	Not covered yet
10	CMS stops transferring data from 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		
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		

... more ...

CHEP 2007



## ALICE critical services list

- WLCG WMS (hybrid mode OK)
  - LCGRB
  - 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

CHEP 2007



## 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 ...

CHEP 2007





# Some First Observations

- Largely speaking, requirements on services are more stringent for Tier0 than for Tier1s than for Tier2s...
  - Some lower priority services also at Tier0...
- Maximum downtimes of 30' can only be met by robust services, extensive automation and carefully managed services
  - Humans cannot intervene on these timescales if anything beyond restart of daemons / reboot needed (automate...)
- 💣 **Interventions out of working hours are currently "best effort" - there is (so far) no agreement regarding on-call services (CERN)**
- Small number of discrepancies (1?):
  - ATLAS streaming to Tier1s classified as "Moderate" - backlog can be cleared when back online, whereas LHCb classify this as "Serious" - max 8 hours interruption
  - Also, ATLAS AMI database is hosted (exclusively?) at LPSC Grenoble and is rated as "high" (discussions re: IN2P3/CERN)
- Now need to work through all services and understand if "standards" are being followed and if necessary monitoring and alarms are setup...
- Do we have measurable criteria by which to judge all of these services? Do we have the tools? (Again < CCRC'08...)



# When to apply updates / upgrades?

- An issue that we have still not concluded on is when to apply needed updates / upgrades
- I assume that we agree that major changes, machine room configurations etc are done **outside** the period of LHC operation
  - And carefully planned / scheduled / tested...
- But priority bug / security fixes are a fact of life!

## Options:

1. Schedule during machine stop / technical development
2. Schedule when necessary - sufficient buffering / redundancy must be built in so no loss of data occurs in short downtimes and active processing of the data will **definitely** occur even with beam off
3. Are there any others?

# Robust Services


- Services deployed at CERN with a view to robustness:
  - h/w, m/w, operational procedures, alarms, redundancy (power, network, middle-tier, DB b/e etc.)
- This was done using a ‘service dashboard’ & checklist at the time of SC3 & re-visited recently
  - Extensive documentation on robustness to specific failure modes – highlights where to improve ([FTS2.0](#))
- Some degree of ‘rot’ – needs to be followed regularly
- Some middleware improvements still required...
- Sharing of experience / training would be valuable

# Main Techniques

## ➤ Understanding of implications of service downtime / degradation

- Database clusters – *not* a non-stop solution; requires significant (but understood) work on behalf of application developer & close cooperation with DBAs
- Load-balanced middle Tier – well proven; simple(!)
- H/A Linux as a stop-gap (VOM(R)S); limitations
- Follow-up: workshop at CERN in November following recent re-analysis with Tier1s and m/w developers
- DB experts & developers will share knowledge

# Running the Services

- Daily operations meeting as a central point for following service problems & interventions
- Excellent progress in integrating grid services into standard operations
- Consistent follow up – monitoring, logging, alarming – efficient problem dispatching
-  **Still some holes – not everyone is convinced of the necessity of this...**
- Despite what experience tells us...

# Scheduled Interventions

- Still the reason for most downtime – security patches, upgrades (fixes) etc.
  - Often several interventions at main sites / week
- Impact can be reduced by performing some interventions in parallel (where this makes sense)
- ☺ **An increasing number of interventions can already be done with zero user-visible downtime**
- In particular true for LFC; FTS has some features to minimize impact; down-time of ½ day per year to introduce new version (schema changes – forward planning reduces this)
- CASTOR interventions (a few per year) also ½ day downtimes
- Done by VO; during shutdown / technical stop?
  - Significant pressure to look at any data also during these periods – is zero user-visible downtime possible for storage services?



# Unscheduled Interventions

- ☹ By far the worst – power & cooling!
  - These have to be addressed by sites directly
- Beyond that: relatively few major downtimes (I am not talking about on-going reliability – this has to be addressed too!)
  - LFC: short-term panic last summer (ATLAS) – problem with alarms – solved by escalation in a few hours (PK / expert call-out over night)
  - FTS: service degradations – solved by restart of daemons (power cycle would also have worked!)
  - CASTOR: ‘stuck’ processes/daemons, still some improvements in monitoring needed; some well known problems have required new versions – need to test extensively to minimize risk of ‘late surprises’
  - DB services: again some stuck clients / services – rapidly resolved by expert intervention
- 💣 Single points of failure – and complexity – are the enemies!

# Other Problems

- Still see far too many ‘file system full’ & ‘system overload’ type problems
- This is being actively addressed by the monitoring working groups
  - “You can’t manage what you don’t measure”
- Another problem that has affected many services – and is independent of DB technology – is ‘DB house-keeping’
  - Largely table defragmentation or pruning...
  - Tom Kyte: *“It’s a team effort...”*
  - In some cases, even a need for “DB 101” ...
- This (to me) underlines the need for a ‘WLCG Service’ view, following Ian Foster’s vision:

# Ian Foster's Grid Checklist

3. A non-trivial level of service is achieved:

*"A Grid allows its constituent resources to be used in a coordinated fashion to deliver various qualities of service, relating for example to response time, throughput, availability, and security, and/or co-allocation of multiple resource types to meet complex user demands, so that the utility of the combined system is significantly greater than that of the sum of its parts."*

# On-Call Services

- ‘PK’ working group<sup>1</sup> [identified](#) need for on-call services – at least for initial years of LHC running – in the following areas:
  - CASTOR (and related services, such as SRM);
  - Database Services for Physics (also Streams replication?);
  - Grid DM (FTS, LFC etc.)
- These are clearly not the only services required for e.g. data acquisition / first-pass processing, but:
  - It has been shown over a period of years that on-call experts (at CERN) can readily solve problems;
  - On-call teams (still under discussion) appear viable in these areas (and are needed!)
- What is (urgently) needed at T1/T2 sites?
  - Batch, storage services? File transfer and conditions support?

# Other Services

- There are clearly other services that are essential to the experiments' production
  - AFS, LSF, ..., phones, web services, ...,
- However, it is not obvious (see PK WG report) that an on-call rota for these services:
  - Could realistically be staffed;
  - Is actually needed
- Relatively stable services with infrequent expert call-out
- This reflects not only the service maturity but also the care taken in setting up the service itself
- ☎ Named experiment contact decides when intervention is needed & calls console operators who have list of experts

# Cross-Site Services

- This is an area that is still not resolved
- Excellent exposé of the issues involved at WLCG Collaboration Workshop in Victoria
- But not presented due to time!
- Will follow-up at [WLCG Service Reliability](#) workshop at CERN in November (26+)
- Emphasizes the need for consistent (available) logging and good communication between teams
- “UTC” – the Time of the Grid!



# Guess-timates

- Network cut between pit & B513
  - ~1 per decade, fixed in ~4 hours (redundant)
- Oracle cluster-ware “crash”
  - ~1 per year (per RAC?) – recovery < 1 hour
- Logical data corruption – database level
  - ~1 per decade, painful recovery (consistency checks)
  - \* Scripts run directly against the (LFC) DB – much higher
- Data corruption – file level
  - Being addressed – otherwise a certainty!
- Power & cooling
  - Will we get to (<) ~1 per site per year? Soon?
- Critical service interruption – 1 per year per VO?
  - Most likely higher in 2008...

# Specific Actions

- ✓ **Need ‘Critical Service’ list from all 4 experiments**
  - Large degree of commonality is expected
  - Target November workshop to present these lists **(or better first results...)**
- Need to work through existing check-list and ensure all issues addressed
  - Use a ‘Service Dashboard’ just as for core ‘WLCG’ services
- Propose: all such services followed up on daily / weekly basis using standard meetings and procedures
  - This includes all the things we have come to know and love: Intervention plans; announcements & post-mortems
- This is basically an extension and formalisation of what is done now
- A formal program to follow up on this work is recommended
  - e.g. specific targets & deadlines; actions on named individuals...
- ¿ Establish & document some “best practices” (requirements?) for new developments → future projects ?

# In a Nutshell...

Services	
ALL	<a href="#">WLCG</a> / “ <a href="#">Grid</a> ” standards
KEY PRODUCTION SERVICES	+ Expert call-out by operator
CASTOR/Physics DBs/Grid Data Management	+ 24 x 7 on-call

# Summary – We Know How to Do It

- Well-proven technologies & procedures can have a significant impact on service reliability and even permit transparent interventions
- We have established a well-tested checklist for setting up and running such services
- These services must then be run together with – and in the same manner as – the ‘IT (GRID) ones’
- These techniques are both applicable and available to other sites (T1, T2, ...)
- ✓ Follow-up: WLCG Service Reliability w/s: Nov 26+
- **Report back to next OB & GDB (Dec 3 & 5)**

**BACKUP**



# Ticklist for new service



- User support procedures (GGUS)
  - Troubleshooting guides + FAQs
  - User guides
- Operations Team Training
  - Site admins
  - CIC personnel
  - GGUS personnel
- Monitoring
  - Service status reporting
  - Performance data
- Accounting
  - Usage data
- Service Parameters
  - Scope - Global/Local/Regional
  - SLAs
  - Impact of service outage
  - Security implications
- Contact Info
  - Developers
  - Support Contact
  - Escalation procedure to developers
- Interoperation
  - ???
- First level support procedures
  - How to start/stop/restart service
  - How to check it's up
  - Which logs are useful to send to CIC/Developers
    - and where they are
- SFT Tests
  - Client validation
  - Server validation
  - Procedure to analyse these
    - error messages and likely causes
- Tools for CIC to spot problems
  - GUIS monitor validation rules (e.g. only one "global" component)
  - Definition of normal behaviour
    - Metrics
- CIC Dashboard
  - Alarms
- Deployment Info
  - RPM list
  - Configuration details (for yaim)
  - Security audit