September 10 2006, Jamie Shiers

WLCG Service Availability Targets - CERN

Introduction

1. The WLCG Memorandum of Understanding (MoU) defines minimum levels of service that must be provided by the various sites that form the WLCG collaboration;
2. In this document we focus on the targets (maximum delay in responding to operational problems and average availability measured on an annual basis) that are currently defined for the Host Laboratory (i.e. CERN);
3. Taking the examples of "event reconstruction" and "distribution of data to Tier1 centres during accelerator operation", we identify the main WLCG and VO-specific services involved and discuss how the targets could be met together with possible resource implications.

Executive Summary

4. The services analysed in this document are characterised by strong dependence on both VO and IT provided services;
5. In the case of data export, a further tight coupling is introduced to storage services at the WLCG Tier1 sites;
6. Based on the experience gained through the Service Phases of SC3 and SC4, it is strongly felt that the current targets – both in terms of delay in response and average availability – cannot be met without on-call services, both overnight and weekends, by experts from the corresponding services;
7. These are required for *all* of the component services that are involved – listed below;
8. If this is not provided, downtimes overnight and extended downtimes during weekends can be expected on a regular basis;
9. A single interruption – e.g. from 02:00 to 10:00 (overnight) or worse, 14:00 Saturday to 10:00 Monday (typical examples from recent history) – can have a significant impact on the average availability (the target defined in the WLCG MoU being 99% during accelerator operation);
10. Whilst significant improvements can be expected with time, this is unlikely to be the case in the early years of LHC running;
11. Given the cross-site coupling in the case of data export, out-of-hours coverage, contact addresses and phone numbers plus agreed and tested procedures for problem resolution must be put in place;
12. Apart from a few specific cases (failover of network to backup path, reduction in overall batch capacity, …), the categories of service degradation (>20%, >50%) are not easily measurable or in some cases even definable;

MoU Targets

13. The MoU Targets for the accelerator centre are defined in the table below;
14. The document introduces the table with the following text:

The following parameters define the minimum levels of service. They will be reviewed by the operation boards of the WLCG Collaboration.

Service	***Maximum delay in responding to operational problems***			***Average availability[1] measured on an annual basis***	
	Service interruption	Degradation of the capacity of the service by more than 50%	Degradation of the capacity of the service by more than 20%	During accelerator operation	At all other times
Raw data recording	4 hours	6 hours	6 hours	99%	n/a
Event reconstruction or distribution of data to Tier-1 Centres during accelerator operation	6 hours	6 hours	12 hours	99%	n/a
Networking service to Tier-1 Centres during accelerator operation	6 hours	6 hours	12 hours	99%	n/a
All other Tier-0 services	12 hours	24 hours	48 hours	98%	98%
All other services[2] – prime service hours[3]	1 hour	1 hour	4 hours	98%	98%
All other services – outside prime service hours	12 hours	24 hours	48 hours	97%	97%

Event Reconstruction

15. It is assumed that event reconstruction is performed using the local batch system, i.e. LSF at CERN;

16. Other services involved include the conditions database service used by the experiment in question (an Oracle-based application for all except ALICE), the experiment-specific book-keeping system(s) (typically based on Oracle and/or MySQL), the LFC (either as a file catalog or as the basis of the CMS DLS), as well as CASTOR2;

17. In the recent ATLAS Tier0 exercise, DDM/LFC operations were decoupled leaving dependencies only on CASTOR, LSF and AFS;

18. In this exercise, AFS was the primary bottleneck and cause of job failures. This is being followed up (e.g. by the use of volume replication);

19. Overall LSF performed worse than in the previous test – leading to the suggestion that a dedicated instance for first pass processing might be needed;

20. CASTOR exceeded the goal of 1 week of stable operation but with a pool 2-times over-dimensioned and Atlas wasted time trying to understand its performance;

21. In summary, steps are being taken to ensure reliable services, although coupling to CASTOR, LSF and AFS (and presumably experiment-specific

[1] (time running)/(scheduled up-time)

[2] Services essential to the running of the Centre and to those who are using it.

[3] Prime service hours for the Host Laboratory: 08:00-18:00 in the time zone of the Host Laboratory, Monday-Friday, except public holidays and scheduled laboratory closures.

services) remains. All of these services are complex and problems typically require 'the expert' to be solved;

Distribution of Data to Tier1s

22. This activity is loosely coupled to the former, in that it requires the output of the reconstruction phase. It is, by definition, tightly coupled to the storage management services of the host laboratory (CASTOR + SRM, hence also Oracle and LSF), as well as the FTS (which also depends on Oracle), the experiment-specific framework that drives the FTS, as well as the corresponding storage management services at all of the Tier1 sites supporting a given VO;

23. Except in the case of failure or severe degradation of host laboratory services, problems with a single site can, in principle, be tolerated (provided that the site in question has the proven ability to rapidly catch up with a backlog, however caused (e.g. source/sink error, or both));

24. On the assumption that recovery from backlogs is demonstrated, expert coverage can probably be limited to ~12-16 hours per day. Although inter-site problems typically require dialog between experts on both sides, more than 2/3 of the data is sent to European sites, where the maximum time difference is 1 hour;

25. This does not mean that sites should not respond to site-local problems within a delay that is consistent with the MoU – both in terms of response time and in average availability – but allows for a more relaxed cross-site intervention procedure should it be proven that the service can indeed transparently and rapidly recover from any corresponding backlogs. It is thus essential that such recovery is demonstrated to all sites in the coming months;

Summary and Conclusions

26. The services listed in the WLCG MoU, namely Event Reconstruction and Data Distribution, are highly complex and rely on many component services that are typically complex in their own right;

27. The interaction between these components is highly non-trivial and problems typically require a high-degree of expertise and the skills of several teams;

28. The cross-site coupling introduced by Data Distribution to the Tier1s requires close collaboration between the corresponding teams at the different sites;

29. Experience with the WLCG Service Challenges, as well as the experiments' own Data Challenges, has shown that these high-level services can typically only run unattended for short periods of time, after which they either fail or rapidly decay;

30. We are currently a long way from the target of sustained, stable data export to all Tier1s at the rates needed for LHC operation;

31. 24x7 on-call services, manned by experts of the relevant IT and VO services, are required. The IT services involved are currently supported by CS, DES, FIO, GD and PSS groups. It is proposed that these groups arrange for stand-by ("piquet") coverage for the relevant services (networking services, AFS, LSF,

CASTOR, Oracle, LFC, FTS etc.), as provided for in the CERN Staff Rules and Regulations (chapter III, section 1);

32. In the case of data export to the Tier1 sites, corresponding on-call services are required at the Tier1s as well, together with inter-site contacts and escalation procedures;
33. We note that GGUS and COD currently provide a service during office hours (of the site in question) only, but should provide the primary problem reporting route during such periods. This requires that realistic VO-specific transfer tests are provide in the SAM (or equivalent) framework, together with the appropriate documentation and procedures;
34. The list of contacts and the procedures for handling out-of-hours problems will be elaborated by the WLCG Service Coordination team and presented to the Management Board for approval. These procedures will be constructed to facilitate their eventual adoption by standard operations teams, should extended cover ever be provided. We note that such a service may address problem determination, but will not, with the current structures, provide problem resolution.