WLCG Service Availability Requirements and Site Survey

The LCG MoU Service Targets

- These define the (high level) services that must be provided by the different Tiers
- They also define average availability targets and intervention / resolution times for downtime & degradation
- These differ from Tier to Tier (less stringent as N increases) but refer to the ‘compound services’, such as “acceptance of raw data from the Tier0 during accelerator operation”
- Thus they depend on the availability of specific components – managed storage, reliable file transfer service, database services, …
- An objective is to measure the services we deliver against the MoU targets
  - Data transfer rates
  - Service availability and time to resolve problems
  - Resources available at a site (as well as measured usages)
    - Resources specified in the MoU are cpu capacity (in KSi2K), tbyes of disk storage, tbytes of tape storage and nominal WAN data rates (Mbytes/sec)

CERN (Tier0) MoU Commitments

Service	Maximum delay in responding to operational problems			Average availability[1] on an annual basis	
	DOWN	Degradation > 50%	Degradation > 20%	BEAM ON	BEAM OFF
Raw data recording	4 hours	6 hours	6 hours	99%	n/a
Event reconstruction/data distribution (beam ON)	6 hours	6 hours	12 hours	99%	n/a
Networking service to Tier-1 Centres (beam ON)	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%

Tier 1 MoU committments

Service	Maximum delay in responding to operational problems			Average availability measured on an annual basis	
	Service stoppage	Degradation ... by more than 50%	Degradation ... by more than 20%	During accelerator operation	At all other times
Acceptance of data from the Tier-0 Centre during accelerator operation	12 hours	12 hours	24 hours	99%	n/a
Networking service to the Tier-0 Centre during accelerator operation	12 hours	24 hours	48 hours	98%	n/a
Data-intensive analysis services, including networking to Tier-0, Tier-1 Centres outside accelerator operation	24 hours	48 hours	48 hours	n/a	98%
All other services – prime service hours	2 hour	2 hour	4 hours	98%	98%
All other services – outside prime service hours	24 hours	48 hours	48 hours	97%	97%

Some of these imply weekend/overnight staff presence or at least availability.

Availability= time running/scheduled up-time

Prime time= 08:00-18:00 weekday in time zone of host laboratory

Tier-2 MoU Committments

Service	*Maximum delay in responding to operational problems*		*availability*
	Prime time	*Other periods*	
End-user analysis facility	**2 hours**	**72 hours**	**95%**
Other services	**12 hours**	**72 hours**	**95%**

No requirement for weekend/overnight staff presence

Nominal MoU pp running data rates CERN to Tier 1 per VO under LHC Design Luminosity

Centre	ALICE	ATLAS	CMS	LHCb	Rate into T1 (pp) MB/s
ASGC, Taipei	-	8%	10%	-	**100**
CNAF, Italy	7%	7%	13%	11%	**200**
PIC, Spain	-	5%	5%	6.5%	**100**
IN2P3, Lyon	9%	13%	10%	27%	**200**
GridKA, Germany	20%	10%	8%	10%	**200**
RAL, UK	-	7%	3%	15%	**150**
BNL, USA	-	22%	-	-	**200**
FNAL, USA	-	-	28%	-	**200**
TRIUMF, Canada	-	4%	-	-	**50**
NIKHEF/SARA, NL	3%	13%	-	23%	**150**
Nordic Data Grid Facility	6%	6%	-	-	**50**
Totals	-	-	-	-	**1,600**

These rates must be sustained to tape 24 hours a day, 100 days a year.

Extra capacity is required to cater for backlogs / peaks.

WLCG Internal Component Service Level Definitions and Availability Targets

Class	Description	Downtime	Reduced	Degraded	Availability
C	Critical	1 hour	1 hour	4 hours	99%
H	High	4 hours	6 hours	6 hours	99%
M	Medium	6 hours	6 hours	12 hours	99%
L	Low	12 hours	24 hours	48 hours	98%
U	Unmanaged	None	None	None	None

- Reduced defines the time between the start of the problem and the restoration of a reduced capacity service (i.e. >50%)
- Degraded defines the time between the start of the problem and the restoration of a degraded capacity service (i.e. >80%)
- Downtime defines the time between the start of a problem and restoration of service at minimal capacity (i.e. basic function but capacity < 50%)
- Availability defines the sum of the time that the service is down compared with the total time during the calendar period for the service. Site wide failures are not considered as part of the availability calculations. 99% means a service can be down up to 3.6 days a year in total. 98% means up to a week in total.
- None means the service is running unattended

Tier0 Services

Service	VOs	Class
SRM 2.1	All VOs	C
LFC global copy	LHCb	C
LFC local copy	ALICE, ATLAS	H
FTS	ALICE, ATLAS, LHCb, (CMS)	H
CE	All VOs	C
RB		C
Global BDII		C
Site BDII		H
Myproxy		C
VOMS		H→C
R-GMA		H

Required Tier1 Services

Service	VOs	Class
SRM 2.1	All VOs	H/M
LFC	ALICE, ATLAS	H/M
FTS	ALICE, ATLAS, LHCb, (CMS)	H/M
CE		H/M
Site BDII		H/M
R-GMA		H/M

Many also run e.g. an RB etc.

Required Tier2 Services

Service	VOs	Class
SRM 2.1	All VOs	H/M
LFC	ALICE	H/M
CE		H/M
Site BDII		H/M

To be checked with individual experiments/sites

Pre-workshop questionnaire 1

- For this session we sent 7 very basic questions to the Tier 1 sites and received replies so far from BNL, FZK, FNAL, SARA and TRIUMF.

1. Do you feel that you have a full understanding of the functionality and behaviour of the MoU services that you are required to provide?

All reply Yes and 3 with additions:

FZK: Yes but not fully implemented or tested under expected load. Also lack of T2 capacity gives us extra T1 roles not in the MoU.

FNAL: Yes but would like actual downtime to be measured, not that between checks.

SARA: Yes but would like clarification of which LCG components make up which MoU service.

Pre-workshop questionnaire 2

2. How does your site monitor the correct functioning of its LCG servers, both for the base system (machine+O/S), middleware and applications?

BNL: base system by Ganglia + Nagios, middleware by MonaLisa, OSG Gricat, LCG by SFT and gLite by SAM.

FZK: base system by Ganglia + Nagios. Will use Nagios for middleware. Would like list of service metrics/triggers

FNAL: NGOP + Remedy. Part of FNAL Computing Division 24 hour helpdesk infrastructure.

SARA: Argus for networking, Ganglia for cluster and dcache infrastructure.Own tools for data movement and also dcache tools, SFT and SAM. See later talk.

TRIUMF: base system by Ganglia and syslogd Logwatch. Dcache tools and SFT for grid. Looking at SAM.

Commonality: Ganglia, Nagios,SFT,SAM

CERN: Home built tools (LEMON,SURE)

Pre-workshop questionnaire 3

3. How does your site signal a failure in the correct functioning of its LCG servers, both for the system and applications?

BNL: both some automatic alarms and operators monitoring web interfaces to send critical alarms to on-call person. Developing a Nagios based system to report errors of LCG Grid servers.

FZK: Nagios sends emails. Externally visible web pages and system can send SMS to operator mobiles.

FNAL: NGOP agents run each 15 mins and page responsible

SARA: periodic check of tools. Argus sends emails for failures in services critical for data taking.

TRIUMF: automatic emails via Nagios for network, via raid monitor and smartd for disks. No paging experts yet.

Commonality: automatic emails or paging

CERN: 24 hour operator alarms with escalation procedures

Pre-workshop questionnaire 4

4. How is such a signal treated during

- normal working hours :
  - ALL: give immediate attention. FNAL pages responsible for critical services 24 by 7.
- weekday overnight :
  - BNL: best effort then first thing in morning.
  - FZK: voluntary but planning for 24 by 7
  - FNAL: 24 by 7.
  - SARA: No immediate attention.
  - TRIUMF: network problems are paged.
  - CERN: 24 by 7 sysadmin level cover but experts are voluntary (on-call service under discussion).

Pre-workshop questionnaire 4 (cont)

- weekend day :
  - BNL: best effort then first thing in morning.
  - FZK: voluntary but planning for 24 by 7
  - FNAL: 24 by 7.
  - SARA: No immediate attention.
  - TRIUMF: sporadic (keeping an eye on things)
  - CERN: 24 by 7 sysadmin level cover but experts are voluntary (on-call service under discussion).
- weekend overnight :
  - BNL: best effort then first thing in morning.
  - FZK: voluntary but planning for 24 by 7
  - FNAL: 24 by 7.
  - SARA: No immediate attention.
  - TRIUMF: No immediate attention
  - CERN: 24 by 7 sysadmin level cover but experts are voluntary (on-call service under discussion).
- Commonality: no overnight immediate response of experts

Pre-workshop questionnaire 5

5. What are the most common failures, both for the base system, middleware and applications?

BNL: System crashes under high load. SRM client failures blocking server connections leading to SRM outage. FTS failures to overwrite files during retry.

FZK: dcache gridftp servers and CE disappearing from info system (since fixed).

FNAL: poor middleware error handling. Bad user code on WNs and LCG modifying WN system parameters.

SARA: failing disks. Hanging file transfers killing dcache pools and gridftp doors. SRM timeouts, Oracle LFCs failing, CEs disappearing and hanging RBs.

TRIUMF: CERN-Triumf lightpath failures. Spurious errors in SFTs with insufficient information to find real error

Commonality: data movement/management

CERN: Some of everything but noticeably data movement/management

Pre-workshop questionnaire 6

6. How much manpower is dedicated to maintaining the monitoring?

BNL: 1 FTE maintaining infrastructure and 0.2 using it.

FZK: 1 FTE spread over 3-4 people.

FNAL: 1 FTE maintaining/improving the NGOP infrastructure and framework. Service responsibles must participate in monitoring.

SARA: 5 people involved part-time in different parts of the services

Triumf: 3 FTE are dedicated to Tier 1 operations which include monitoring and problem resolution.

Commonality: At least 1 FTE

CERN: 2 FTE and service responsibles write their agents

Pre-workshop questionnaire 7

7. How much manpower is dedicated to fixing failures?

BNL: Depends on production schedule.

FZK: No specific team – whole Tier 1 team participates in problem resolution/user support in their areas.

FNAL: Everyone in 10-person team participates and can call on other experts.

SARA: 5 people involved part-time in different parts of the services

Triumf: 3 FTE are dedicated to Tier 1 operations which include monitoring and problem resolution.

Commonality: Engineer level support

CERN: All experts contribute, 2 rotating engineer staff on weekday duty, probably 2 FTE total.