



GridPP

UK Computing for Particle Physics

GridPP Project Management Board

Tier-1 Disaster Management and Business Continuity

Document identifier :	GridPP-PMB-149-TierOneDisasterManagement.doc
Date:	25/11/2010
Version:	1.1
Document status:	Final
Author	PMB

Disaster Management

The Tier-1 disaster management system consists of a set of general procedures that formalise the way serious situations are handled and set of specific procedures and contingency plans for a few obvious likely scenarios.

The current hybrid processes evolved upon the experience gained from developing two previous systems. The first of these consisted of a single monolithic document that tried to address both our strategic response to a major incident and specific contact names and escalations for very specific failures. As the document grew in size, the gaps in knowledge grew more rapidly than the areas of certainty. Eventually it became clear that it was unlikely that a maintainable plan could be constructed in a reasonable time by this process. The second system consisted of a risk analysis and a set of brief, contingency plans (with contact names and escalation actions and times) addressing areas of high risk. This had the benefit of constraining the size of the task, but led to the re-documentation of standard operational procedures and discrepancies between procedures in different contingency plans.

The current Disaster Management System grew from consultations with a staff on site who had responsibility for site contingency plans. It also took into account the views of the experiments as to what constituted a disaster. An informal review of the historical operating experience at the Tier-1 was also carried out. One common theme from this work was that generated contingency plans (both ours and other people's) tended to be operational and start suddenly (eg a fire or power failure) but the ones actually experienced were slow moving (often running for many months) and hard to recognise to start with (eg firmware problems, growing concern about fire risk presented by certain equipment, etc). The strategy evolved therefore to create a Disaster Management System which:

- Handles all potential disasters in a similar way.
- Identifies common features and trigger levels to allow us to spot events before they blossom into disaster
- Modifies existing processes as little as possible, doesn't document activity which is routinely handled through normal processes
- Builds specific contingency plans that add to the general response in specific circumstances but copes well with incidents for which no contingency plans exist.
- Triggers early, triggers often, to respond ahead of curve and to make use of the system routinely. This:
 - Stops the system decaying.
 - Gives operational and project management benefits.

Existing Systems

There were a wide range of processes already in place:

- The site has a set of contingency plans and a process for handling site related incidents (such as a fire). Likewise CICT, the department responsible for the Tier-1's machine room infrastructure and site networking, has its own set of contingency plans. Therefore the Tier-1 plans do not need to directly address site related infrastructure issues but do need to interact with the teams dealing with them.
- The Tier-1 operates a daytime Production Team who have a set of standard response procedures and contact points ensuring that the operational status of the service is rapidly disseminated. Therefore the Tier-1 contingency plans do not need to address what to do when a service exception occurs nor who to notify. Routine production operations should handle this smoothly.
- The Tier-1 already operates a large out-of-hours callout system. This provides a 24x7x365 emergency contact point, a clear point of initial responsibility for all operational incidents, procedural documentation (including offsite copies) and escalation procedures. There are 5 staff on-call at any given time with access to remote system management tools. Therefore the Tier-1 contingency plans do not need to address how to trigger a response to out of hours incidents nor how to maintain routine procedural documentation.

- In a service the size of the tier-1, hardware failure is a routine occurrence. Resilience to this kind of problem and system recovery from catastrophic hardware failure is a normal part of the Tier-1's operation. Fabric Management systems allow rapid system imaging and an automated tape management system with off-site fire-safe ensures that critical backups are likely to be retained. Therefore the Tier-1 contingency plans do not need to address how to rebuild individual systems.

Overview of Disaster Management System

As a matter of routine operation, many incidents occur which can have a major short-term impact on the service. It is not intended that disaster management should immediately interfere with the normal “production” response to major but relatively short-term events. For example, loss of power in the machine can have a major impact on the service but provided recovery is proceeding along normal lines then the existing effective command, control and communications production infrastructure should be allowed to proceed unhampered. Only where the normal production response is not proceeding along expected avenues and timescales will the disaster response plan be triggered.

By disaster is meant any incident that has a significant long term impact on:

- Safety
- Service Commitments
- Reputation
- Finance

The response to events escalates in stages:

Stage	Managed by	Description
1	Disaster Controller	Initial informal assessment of situation. Don't interfere when not necessary. Response entirely managed via standard operational channels and escalations. If appropriate set deadline for remedial work to succeed. Continue to monitor. Take immediate control in an emergency.
2	Disaster Assessment Team	Disaster possible. Disaster Controller takes executive authority to authorise emergency action as necessary. First formal assessment. Oversight by Disaster Assessment Team. Reallocation of internal team resources to meet threat. Notification mainly by normal operational channels.
3	Disaster Management Team	Disaster increasingly likely. Broaden oversight to Disaster Management Team. Deploy all resources available. Advise operational contacts that contingency plans may need to be activated. Advise Communications team of situation. Escalate management chain advisories.
4	Disaster Management Team	Disaster. Work to minimise impact. Contingency plans activated. Publicity and communications in operation.

Membership of the teams is as follows:

Team	Membership
(Duty) Disaster Controller	Defaults to the Tier-1 Manager, but a documented decision tree exists that allows a Disaster Controller to be identified. An optimal Disaster Controller is appointed after the first Disaster Assessment meeting.
Disaster Assessment Team	Tier-1 Manager, Group Leader, Tier-1 Production Manager, Disaster Controller (if not one of previous), relevant special expert(s) if appropriate
Disaster Management Team	Disaster Assessment Team plus Director e-Science RAL, Services Division Head e-Science RAL, GRIDPP Project Leader, GRIDPP Production Manager, GRIDPP UB Chair

At each meeting of the Disaster Team an assessment of the situation is carried out and compared against criteria for escalation. An estimate of likelihood and projected time to escalation is also made. At each escalation well-defined actions and communications are described in the General Incident Response Plan and these may be supplemented by specific additional contingency plans.

Contingency Plans

A risk analysis was carried out to identify some of the more obvious threats (see Appendix 1 – Risk Analysis and Contingency Plans) for each risk a contingency plan was prepared. These document, trigger levels for escalation, actions at each level and existing mitigations. Given the existing production/operations processes described above and the detailed general response plan many of these contingency plans tend to be very brief documenting trigger levels and special contacts and actions specific to the incident.

Operational Experience

The Tier-1 has operated the current disaster management system for over 18 months now. In that time ten incidents have been entered into the system, two reaching level 3 and one (subsequently forked into two separate tracks) reaching level 4. The full list of incidents handled is provided in Appendix 2 – Catalogue of Incidents Managed. For each of these a set of meeting reports exists and, for operational incidents, a WLCG Service Incident report is also available.

We have found that the General Management process has been extremely useful particularly when coupled with a culture of “trigger early, trigger often”. The contingency plans have been less useful in practice, although the process of constructing them has been helpful. The well-defined trigger levels between stages are very useful, helping to concentrate minds and give a sense of urgency to the response. However, the detailed planned actions and communications have been of less value. For events such as a network break that is outside our control, these special actions tend to be obvious management escalations. For events such as a major operational break the contingency plan is more detailed, but experience has shown that real in real life the situation is rarely as simple as that planned for. Careful analysis and discussion is usually needed before critical actions are carried out.

As was recognised when the system was first set up, the incidents have been a mix of about 50% relatively short term operational problems and the remaining 50% very long term issues. With the longer running issues in particular the challenge seems to be to shut down the special track and hand back responsibility and actions to existing established bodies and entities once there is no longer any real likelihood of a disaster occurring. Another issue has been that on occasion the Disaster Team have been ahead of other existing structures and have occasionally usurped roles where it might have been better to push for action elsewhere.

Overall we consider the system to have been a considerable success, pro-actively managing incidents that have required it while successfully differentiating those from what are actually normal operational problems.

Case Study – Swine Flu

On the Friday 11th June 2009 it was reported that WHO had declared a global flu pandemic in response to the outbreak of swine flu. The Tier-1 recognised this as a potential threat to its future operations and, on the 19th June, triggered its disaster management response. The Disaster Assessment Team met on 25th June (see Appendix 3 – Swine Flu Initial Assessment Report) and concluded that there was a real likelihood that swine flu might cause sufficient staff loss to severely impact (or even halt) the service. Work on the a contingency plan was completed the same day (see Appendix 4 – Example Contingency Plan.) Because of the pre-existing framework and procedures the task of producing the contingency plan was a matter of only a few hours.

Over the next two months (despite work pressure from the recent machine room migration and subsequent operational problems) the team continued to develop remote working infrastructure in response to this risk and even held a work-from-home day. By early September the team concluded that preparations were sufficient and as the estimated likelihood of escalation was falling, regular review meetings were suspended in favour of tracking by the established STFC monitoring meeting. The process was officially stood down once the site-monitoring meeting ceased to operate in March 2010. Although (fortunately) the preparations put in place were

never needed, much of the remote management infrastructure remains useful and was used by many staff during the snowy days of January 2010.

Business Continuity

National and International Context

The Tier-1 forms an essential part of the UK infrastructure for GRIDPP, providing national services that the Tier-2 sites depend on. Most of these services are potential candidates for hot, warm or cold standby at other sites – either among the UK Tier-2 sites or located at other Tier-1 sites. The current status of critical services is given in the table below:

National top level BDII	The Tier-1 runs a top level BDII service for the whole of the UK. Although some of the larger UK sites chose to run their own top level BDIIs these are being phased out in favour of an international failover (NL-T1 in the case of the UK). The Tier-1 is also experimenting with locating top-level BDII servers in the Amazon EC2 service.
Caching services	The Tier-1 and UK Tier-2s use a number of caching proxy services hosted at the Tier-1, such as FroNTier for ATLAS, Squid for CMS and the LHCB LFC. Such caches are a hot topic within wLCG, and their use is likely to expand both in terms of VO uptake and a diversification of applications. Failover to other sites is suitable in some cases, and such configuration is already in place for the production caches at the Tier-1.
myproxy	There is no alternative service at present, although configuration of other MyProxy services (such as during the ATLAS/R89 migration) is straightforward and is limited to adding support for UK WMSs. There is a standby service, and there are plans to relocate this to R27.
ATLAS LFC service	Critical for UK ATLAS activities, and the VO model places custodial responsibility for all UK ATLAS replicas in this service at the Tier-1. ATLAS are planning to move to international hot standby. It is planned to have at RAL a standby service in R27, and the Fabric and Oracle layers have been reengineered prior to deployment of this architecture.
FTS	No alternative service at present. In an emergency the channel configuration could be set up on another international FTS elsewhere in the world. Planned to have RAL standby service in R27
Workload Management System (WMS)	The RAL Workload Management System (WMS) is one of several within the UK and alternatives exist internationally also

The Tier-1 CASTOR service is also an essential staging point for datasets destined for the Tier-2s. It is not practical to replicate the whole CASTOR service, but tests were carried out in 2009 to use an alternative Tier-1 as the main feed for data replication. A number of projects are underway to test caching strategies for file replication and these too are likely to allow the Tier-2 sites to decrease their dependence on the Tier-1 for data access.

Experiment Response to Tier-1 Incidents

Each LHC experiment has its own specific requirements as to how it should handle extended downtimes at one of its Tier-1s. In the case of CMS these plans were well described at a recent WLCG workshop where the VO had a different response for short (< 2-3 days), medium (1-2 weeks) and long (>2 weeks). This consisted of shifting production workloads to other available sites and where appropriate moving the flow of custodial

data. What was notable was that for longer outages, CMS would not by default choose to move missing datasets back to the "lost" site.

RAL Standby Service

The best solution for nationally critical services is to ensure sufficient national or global resilience within the project so that the loss of the Tier-1 has negligible impact. However the Tier-1 itself needs to plan how to handle major incidents such as a fire in the main machine rooms.

Work is underway to construct a small, low cost, standby Tier-1 infrastructure in the old R27 machine rooms in the ATLAS centre (500M from the new R89 machine room) some testing and procurement has already been carried out and work is expected to be completed during 2011. The strategy is to provide a tiered approach to service restoration and rebuild based mainly on dual-purpose equipment and services.

- For trivial multi-hosted services, site a few hosts in ATLAS Centre. Service stays up even during an incident or planned operational event such as power maintenance.
- For some critical database services (eg the CASTOR databases) maintain a synchronised warm spare copy. Can be brought online if needed, but operationally useful anyway (eg to allow maintenance on production servers).
- Maintain copies of instancing and (as it develops) virtualisation infrastructure. Allows us to rebuild any Tier-1 system in ATLAS Centre if we have the hardware.
- Maintain our main testbed infrastructure in ATLAS Centre. Provides an initial source of hardware to feed the standby instancing service.
- Then make use in situ of any other hardware on the RAL site that can be made available by emergency loan. In the past we have successfully borrowed Tier-2 disk servers, overlaying our own network address space across the RAL site backbone.
- Eventually an emergency procurement of hardware could deliver more capacity into ATLAS Centre or elsewhere on site.

Because much of this infrastructure must exist in any case as part of the normal resilience required, the additional cost has actually been very low, mainly consisting of some additional networking and SAN equipment. Although this infrastructure would not protect against a major site network failure it does provide options to protect against smaller incidents. If the service was located off the RAL site then problems with the network could be avoided, but at the cost of increasing complexity for operation as well as probably losing some of the dual use functions.

Appendix 1 – Risk Analysis and Contingency Plans

Reference	Description	Example	Affects	Data lost	Impact	Likelihood	Duration
DRT0	General response for all incidents	Supplement or use on own	All Vos	Any	Any	Any	Any
DRT1	Tier-1 damaged	Fire in machine room	All Vos	Yes	Unavailable	Very unlikely	Very long
DRT2	Loss or corruption of CASTOR metadata catalogue	Catalogue lost - no backups	All VO's	Yes	Unavailable	Unlikely	Long
DRT3	Loss of significant amount of user data	Data accidentally deleted - 1 in 50 files identified as corrupted	VO	Yes	Degraded	Very unlikely	Long
DRT4	Site incident causes Tier-1 to be temp inoperable	Electricity supply failure	All Vos	No	Unavailable	Very likely	Short
DRT5	Site incident causes Tier-1 to be inoperable	Transformer fire	All Vos	No	Unavailable	Very unlikely	Very long
DRT6	Break in connection to SJ5 network	Fire in site network machine room	All Vos	No	Unavailable	Very unlikely	Long
DRT7	Break in connectivity to OPN	Digger cuts fibre	All Vos	No	Degraded	Unlikely	Long
DRT8	Disruption to site network infrastructure	Router keeps crashing	All Vos	No	Degraded	Unlikely	Long
DRT9	Extended DoS attack		All Vos	No	Degraded	Very unlikely	Long
DRT10	Late delivery of equipment or acceptance failure	Supplier goes out of business	All Vos	No	Degraded	Likely	Very long
DRT11	Systematic failure of a batch of equipment	Backplanes burn out	All Vos	No	Degraded	Likely	Very long
DRT12	Failure of component service	CE fails	VO	No	Degraded	Very likely	Short
DRT13	Staff level falls below level required for operation	Ban on recruitment	All VO's	No	Degraded	Unlikely	Very long
DRT14	Major Security Incident	Multiple root compromise	All VO's	No	Unavailable	Likely	Long
DRT15	Disease	Swine flu	All VO's	No	Degraded	Very unlikely	Long

Appendix 2 – Catalogue of Incidents Managed

Reference #	Description	Date Entered	Elapsed days	# reviews	Worst level	Comments
1	Power Failure	24/3/09	2	3	2	Nearly routine operation, but restart delay triggered response
2	Home filesystem file loss	22/4/09	1	0	1	Turned out to be a minor problem. Better safe than sorry.
3	Swine Flu	19/6/09	270	8	2	No operational impact but a lot of preparatory work done. This work provides long term benefit in terms of remote operations capability.
4	Cooling failure	11/08/09	7	4	3	Lack of understanding of new cooling system led to decision to delay restart until problem clearly understood. Substantial changes to cooling system were implemented over subsequent 12 months.
5	Water leakage into robot	17/08/09	210	4	2	Condensation from upstairs chillers caused small volume of water to dribble into robot. No data loss but remedial work on office cooling system was tracked for 6 months
6	Disks fail acceptance	3/08/09	230	3	2	Disk server delivery failed acceptance tests. Problem due to hardware interoperation issue. Only 3 reviews as activity spun off to procurement team who held regular series of reviews.
7	Dirty UPS power feed caused instability on core ORACLE database hardware	5/10/09	390	16	4	Interaction between UPS supply and load led to 3KHz wave on current supply. All core database services destabilised and had to be migrated to alternative hardware. Acute phase lasted 7 days but cause properly understood and satisfactory solution identified.
8	Data loss from CASTOR	22/10/09	1	2	4	Hardware problems during recovery operation for incident 7 led to an out of date version of a database being made live. Ten days data lost on LHCB instance.
9	Disks fail acceptance	6/5/10	180	2	1	Another disk server procurement acceptance failure. Eventually traced to RAID card firmware problem and cards swapped out.
10	Dust in machine room	7/6/10	70	5	2	Cladding from under floor cooling pipes eroded by cooling airflow. Dust may have presented a health hazard (did not), high level might lead to closure of robot service.

Appendix 3 – Swine Flu Initial Assessment Report

Title	H1N1 Pandemic				
Description	Swine flu pandemic may cause substantial staff absence.				
Incident Date/Time	19/6/09	Incident #	3	Current Stage	2
Meeting Date/Time	25/6/09 10:00	Relevant Contingency Plan	DRT15(under construction)	Next Formal Review	2 July
Elapsed Time		Likelihood of escalating	50%	Time to escalation	8-12 weeks
Current Team Membership	Name		Role		
	Andrew Sansum		Duty Disaster Controller		
	Dave Corney		Group Leader		
	Gareth Smith		Production Manager		
Confirmed Disaster Controller	David Corney				
GRIDPP Incident Report URL	N/A				
Summary of Situation	Recognition of growing risk from H1N1				
Conclusions	Urgent need to identify and address priority areas for mitigation. Need to complete contingency plan DRT15.				
Perceived Risks	Self evident see- contingency plan. Service will degrade or become inoperable as staff numbers fall (in team or services we depend on).				
Proposed Additional Team Members	Responsible person within ESC for H1N1 response (Neil Geddes?)				
Planned Actions					
ID	Owner	Required by	Action		
1	DC	26/06/09	Book next meeting		
2	DC	29/06/09	Identify department level member		
3	DC	Before next meeting	Understand existing STFC plans		
Planned Communications					
ID	Owner	Required by	Action		
4	RAS	26/06/09	Distribute initial assessment according to general response plan.		
5	DC	26/06/09	Follow up with team comments on our response.		
6	RAS	29/06/09	Train David to be disaster controller		

Appendix 4 – Example Contingency Plan

		Disaster Reference: DRT15
1	Type:Final	
2	Area Tier-1	
2	Description A public health issue severely reduces staffing level to critical levels	
3	Further	details
The spread of a contagious or infectious disease such as Influenza may cause illness within the team or team members' dependents. Reducing staff numbers to such low levels the service can no longer be operated. Site may close, suppliers may be affected or external services the Tier-1 depends on may degrade or fail.		
4	Services/areas	affected
All services		
5	Severity/Impact (priority if multiple events)	
<p>a) <i>Scope</i>: All VOs</p> <p>b) <i>Data lost</i>: None</p> <p>b) <i>Impact</i>: Degraded or Unavailable</p> <p>c) <i>Likelihood</i>: Very Unlikely</p> <p>d) <i>Time to resolve</i>: Long</p>		
6	Response coordination	
Normal disaster management process		
7	Response Plan	
This is supplementary to our General Disaster Response plan		
	Stage	
	1	Criteria WHO alert level pandemic level 5. Reports Unexpected absence of two or more members of the team with illness (but no wider scale disease).
		Actions <ul style="list-style-type: none"> • Review STFC advice • Review government advice (if problem is national) • Review WHO advice • Review local newspaper reports
		Communications Standard communications: <ul style="list-style-type: none"> • Obtain intelligence from staff off sick or carrying out caring role.
	2	Criteria WHO has an active situation classified as pandemic level 6. Illness observed within UK that has the potential to spread in extent with an impact likely to cause operational difficulties (without WHO pandemic warning).
		Actions <ul style="list-style-type: none"> • Ensure remote operation capability is optimized. • Increase priority to documentation efforts • Increase priority to staffing resilience • Monitor situation

	<p><u>Communications</u></p> <ul style="list-style-type: none"> • Standard
3	<p><u>Criteria</u> Staff diagnosed to be sick. Occurrence of disease within local community staff have recent direct contact with. Closure of laboratory or advice to home work.</p>
	<p><u>Actions</u></p> <ul style="list-style-type: none"> • Consider team dispersal to work from home • Ban work related travel • Discourage attendance at large scale events.
	<p><u>Communications</u></p> <ul style="list-style-type: none"> • Standard
4	<p><u>Criteria</u> Faults occurring that cannot be fixed by remaining staff. Access to site lost to such an extent that risk to equipment may occur. Failure in external infrastructure we depend on.</p>
	<p><u>Actions</u></p> <ul style="list-style-type: none"> • Prioritise national and global core services. • Terminate service if hardware safety cannot be maintained
	<p><u>Communications</u></p> <ul style="list-style-type: none"> •
8	<p><u>Existing Mitigation</u></p> <p>On-call system provides means to operate service remotely. On-call documentation provides documents for many routine interventions On-call rota provides contact details for most team members Most of team have laptops Most of team have broadband or 3G modem.</p>
	<p><u>Future Mitigation</u></p> <p>Obtain Clarification on site policy and priorities Identify all critical external “services” we depend on (eg concern about site networking). Plan for entire team to work remotely. Ensure sufficient laptops, 3G etc Ensure sufficient capacity in remote access system (eg pptp will collapse, email failure) Ensure maximum access to irc Plan for remote conferencing Clarify decision making hierarchy Maximise coverage and documentation of IPMI and remote power management Site supportive of team dispersal Site provide means to encourage partial working while carrying out caring role at home. Hygiene policy within e-Science Review physical resources How to protect the data (mothball) Fall back contacts (when existing SOPs break down).</p>
9	<p>Approval Date: _____ Review date: _____</p>