EGI Workload Manager Availability and Continuity plan

Table of contents

Introduction

This page reports on the Availability and Continuity Plan for **EGI Workload Manager - DIRAC4EGI** and it is the result of the risks assessment conducted for this service: a series of risks and treats has been identified and analysed, along with the correspondent countermeasures currently in place. Whenever a countermeasure is not considered satisfactory for either avoiding or reducing the likelihood of the occurrence of a risk, or its impact, it is agreed with the service provider a new treatment for improving the availability and continuity of the service. The process is concluded with an availability and continuity test.

	Last	Next
Risks assessment	2024-12	Q4 2025
Av/Co plan	2024-12	Q4 2025

Previous plans are collected here: https://documents.egi.eu/document/3597

Availability requirements and performances

In the OLA it was agreed the following performances targets, on a monthly basis:

- Availability: 99%
- Reliability 99%

Other availability requirements:

- the service is accessible through either X509 certificate or OAuth2 IdP (upcoming with the new release)
- The service is accessible via CLI and webUI

The service availability is regularly tested with metrics generic.http.connect-ssl and generic.certificate.validity:

- Status on ARGO UI

The performances reports in terms of Availability and Reliability are produced by ARGO on an almost real time basis and they are also periodically collected into the Documentation Database.

Over the past years, the Workload Manager hadn't particular Av/Co issues highlighted by the performances that need to be further investigated.

Risks assessment and management

For more details, please look at the Workload Manager Risk assessment. We will report here a summary of the assessment.

Risks analysis

Title	Risk description	Affected components of the service	Established measures	Risk level	Expected duration of downtime/ time for recovering	Treatment - Protective /mitigation measures - recovery activities - controls
Workload Manager Risk assessment	Service unavailable / loss of data due to hardware failure	All the service components	Reactive countermeasure: Data protection with daily backups (6 months retention) of the entire database and on-the-fly backup of the binary logs. Regular snapshots of virtual machines hosting DIRAC4EGI services. MySQL Database restoring from daily backups; point in time recovery, if needed, from binary logs backups. VM servers restored from snapshots.	(3) Medium	1 working day after the hardware faiure recovery	the measures already in place are considered satisfactory and risk level is acceptable
Workload Manager Risk assessment	Service unavailable / loss of data due to software failure	All the service components	Reactive countermeasure: Data protection with daily backups (6 months retention) of the entire database and on-the-fly backup of the binary logs. MySQL Database restoring from daily backups; point in time recovery, if needed, from binary logs backups.	(4) Medium	1 working day	the measures already in place are considered satisfactory and risk level is acceptable
Workload Manager Risk assessment	service unavailable / loss of data due to human error	All the service components	Reactive countermeasure: Data protection with daily backups (6 months retention) of the entire database and on-the-fly backup of the binary logs. MySQL Database restoring from backups for affected components. Restoring Configuration data from backups.	(4) Medium	1 working day	the measures already in place are considered satisfactory and risk level is acceptable
Workload Manager Risk assessment	service unavailable for network failure (Network outage with causes external of the site)	All the service components	Preventive countermeasure: Geographically distributed redundant Configuration Service. Redundant failover Request Management Service. Failover mechanism for recovering job outputs.	(2) Low	1 hour after the network recovery	the measures already in place are considered satisfactory and risk level is acceptable
Workload Manager Risk assessment	Not enough people for maintaining and operating the service	Resources management. User support. Security infrastructure components	Preventive countermeasure: Automation of synchronization with BDII, VOMS, GocDB information indices. Automated resource monitoring service. Training multiple system administrators. Involving new participants to the service administration group.	(2) Low	1 or more working days	the measures already in place are considered satisfactory and risk level is acceptable
Workload Manager Risk assessment	Major disruption in the data centre.	All the service components	Reactive countermeasure: Daily backups (6 months retention) of the entire database and on-the-fly backup of the binary logs. Regular snapshots of virtual machines hosting DIRAC4EGI services. Reestablinshing services in a different hosting environment. Restoring databases from backups if still available. Partial restoring of the File Catalogs contents from the storage elements information.	(4) Medium	several weeks	the measures already in place are considered satisfactory and risk level is acceptable

Workload Manager Risk assessment	Major security incident. The system is compromised by external attackers and needs to be reinstalled and restored.	All the service components	Reactive countermeasure: Daily backups (6 months retention) of the entire database and on-the-fly backup of the binary logs. Regular snapshots of virtual machines hosting DIRAC4EGI services. Reinstalling service components with the configuration restored from backups. Changing security tokens (logins, passwords) for accessing the service servers and databases. Assume that the database service is not affected, otherwise restoring the databases from backups.	(2) Low	1 or more working days	the measures already in place are considered satisfactory and risk level is acceptable
Workload Manager Risk assessment	(D)DOS attack. The service is unavailable because of a coordinated DDOS.	All the service components	Preventive countermeasure: Limited service queries queues avoiding dangerous overloading of the service components. Automatic service restart after going down due to an overload. Automatic recovery after the end of the DOS attack.	(1) Low	1 hour	the measures already in place are considered satisfactory and risk level is acceptable
Workload Manager Risk assessment	Resource Centres unavailability	None	Reactive countermeasure: Regular update of site administrators contact information. Once the risk occurs, WMS admins will contact the site administrators to solve the unavailability.	(4) Medium	1 or more working days depending on the site administrators response time	the measures already in place are considered satisfactory and risk level is acceptable

Outcome

The risk number 9 (Resource Centres unavailability) depends on the RCs whose resources are used by the VOs and in case of occurrence the Workload Manager provider can only mitigate the impact, as explained in the table above. It was agreed to include also this risk even if not completely control of the provider because it is related to an incident that can occur. The likelihood of the risk was set to 2 considering the history of incidents associated to this risk.

The level of other risks is acceptable and the countermeasures already adopted are considered satisfactory.

Additional Information

- There aren't special procedures to invoke in case of risk occurrence but the general administrator guide and generic internal procedures
- the Availability targets don't change in case the plan is invoked.
- recovery requirements:
  - **Recovery time objective (RTO)** (from ISO 22301: period of time following an incident within which a product or service must be resumed, or activity must be resumed, or resources must be recovered): **2 days**
  - **Recovery point objective (RPO)** (the acceptable latency of data that will not be recovered): n/a
- approach for the return to normal working conditions as reported in the risk assessment.
- In case of power supply problems:
  - There is a redundant power supply line (two independent lines), so the risk of a complete power outage is very much unlikely. To cover those rare cases there is a diesel electric generator (full tank capacity ~3000 L) that will keep running the essential services. The duration of the generator is approximately 1 hour.
  - CC-IN2P3 has 2 separate computing rooms and the self-sufficiency is assured by a combination of a diesel generator + a room of charged batteries. This will activate only if the two redundant power lines supplying CC fail simultaneously (very unlikely).
  - If needed, a controlled shutdown of the Workload Manager will be implemented to avoid data losses
  - Mainly then network related services ore the ones that must be kept up&running during exceptional situations
- if the data centre is down for several days and cannot be re-activated in reasonable time, the other IN2P3 sites would step up to temporarily host the service (which does not need a huge amount of resources)
- The dedicated GGUS Support Unit will be used to report any incident or service request.
- In case of exceptional situations, user can be informed through the usual channels, i.e., downtime declaration, broadcasts.
- The providers can contact EGI Operations via ticket or email in case the continuity plan is invoked, or to discuss any change to it.

Availability and Continuity test

The proposed A/C test will focus on a recovery scenario: the service is supposed to have been disrupted and needs to be reinstalled from scratch. Typically this covers the risks 1,2, and 7. The last backup of the data will be used for restoring the service, verifying how much information will be lost, and the time spent will be measured.

Performing this test will be useful to spot any issue in the recovery procedures of the service.

Test details

More details available on https://documents.egi.eu/document/3597 **The recovery test was performed on April 2020 but it is still considered valid: there is no need to repeat it**.

Test case	Simulation	Recover time	Actions	Status
Service/Agent crash that can be caused by some transaction failure, software error or human error	Kill the component process	few seconds	The component is restarted automatically by the system monitoring facility	PASS
Host failure, for example due to a power cut	reboot dirac4.grid.cyfronet.pl server	few minutes	The host rebooting sequence contains an automatic restart of all configured DIRAC components by using supervisord.	PASS
Installed software corruption	reinstall DIRAC software stack from scratch	10 - 15 minutes	Manual intervention: running dirac-install installer tool; verify that all the components properly restart.	PASS
Configuration files loss or corruption, for example, due to a hard disk failure.	BackUps of the local configuration files in a database or on another server	few minutes	Replace the lost configuration with a backup copy.	PASS
DB corruption and/or crash	recover from dump	5 - 30 minutes	Manual intervention by the IN2P3-CC database service administrators	PASS

Test outcome

The test can be considered successful: the service can be restored in few time if hardware, software or database failures occur.

Revision History

Version	Authors	Date	Comments
	Alessandro Paolini	2019-01-10	first draft, discussing with the provider
	Alessandro Paolini	2019-08-27	adding other availability requirements, and additional information for the risk assessment
	Alessandro Paolini	2019-11-25	page updated with additional availability requirements, and additional information section. Waiting for the recovery test, hopefully to be done in January.
	Alessandro Paolini	2020-04-14	added details about the recovery test provided by the supplier. Plan finalised.
	Alessandro Paolini	2021-05-11, 2021-05-31	starting the yearly review. (https://ggus.eu/index.php?mode=ticket_info&ticket_id=151951). Minor changes, review completed.
v. 8	Alessandro Paolini	2022-07-11	yearly review; added the risk about RC unavailable; updated the MTPoD; no need to perform a new recovery test.
v.9	Alessandro Paolini, Catalin Condurache, Gino Marchetti	2023-03-03	during the last ISO20k certification audit, it was noticed that the duration of MTPoD and RTO was the same, while actually the RTO is expected to be shorter. We had a meeting with the supplier where we agreed to shorten the RTO from 2 days to 1 day, also in accordance to the outcome of the recovery test.
v.15	Alessandro Paolini, Catalin Condurache, Gino Marchetti	2023-08-14, 2023-09-05	yearly review; updated "av. req. and performance" section; removed RTO; MTPoD renamed as RTO; updated risk analysis table; updated the section "Additional Information" with some more details about continuity and recovery during exceptional situations.
v.18	Alessandro Paolini, Andrei Tsaregorodtsev	2024-12-18	yearly review; updated the link to the monitoring metrics; risk assessment moved to confluence and updated; updated the outcome section; no need for a new recovery test.