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	2022-07-08	Q3 2023
Av/Co plan	2022-07-11	Q3 2023

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 by nagios probe org.nagiosexchange.Portal-WebCheck: https://argo-mon.egi.eu/nagios/cgi-bin/status.cgi?host=dirac.egi.eu

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 google spreadsheet. We will report here a summary of the assessment.

Risks analysis

Risk id	Risk description	Affected components	Established measures	Risk level	Expected duration of downtime / time for recovery	Comment

1	Service unavailable / loss of data due to hardware failure	All the service components	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	Medium	1 working day after the hardware failure recovery	the measures already in place are considered satisfactory and risk level is acceptable
2	Service unavailable / loss of data due to software failure	All the service components	Data protection with daily backups (6 months retention) of the entire database and on-the-fly backup of the binary logs.	Medium	1 working day	the measures already in place are considered satisfactory and risk level is acceptable
3	service unavailable / loss of data due to human error	All the service components	Data protection with daily backups (6 months retention) of the entire database and on-the-fly backup of the binary logs.	Medium	1 working day	the measures already in place are considered satisfactory and risk level is acceptable
4	service unavailable for network failure (Network outage with causes external of the site)	All the service components	Geographically distributed redundant Configuration Service. Redundant failover Request Management Service.	Low	1 hour after the network recovery	the measures already in place are considered satisfactory and risk level is acceptable
5	Unavailability of key technical and support staff (holidays period, sickness, ...)	Resources management. User support. Security infrastructure components	Automation of synchronization with BDII, VOMS, GocDB information indices. Automated resource monitoring service	Low	1 or more working days	the measures already in place are considered satisfactory and risk level is acceptable
6	Major disruption in the data centre. Fire, flood or electric failure for example	All the service components	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.	Medium	several weeks	the measures already in place are considered satisfactory and risk level is acceptable
7	Major security incident. The system is compromised by external attackers and needs to be reinstalled and restored.	All the service components	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.	Low	1 or more working days	the measures already in place are considered satisfactory and risk level is acceptable
8	(D)DOS attack. The service is unavailable because of a coordinated DDOS.	All the service components	Limited service queries queues avoiding dangerous overloading of the service components. Automatic service restart after going down due to an overload.	Low	1 hour	the measures already in place are considered satisfactory and risk level is acceptable
9	Resource Centres unavailability	The RCs used by the VOs	Regular update of site administrators contact information. Once the risk occurs, WMS admins will contact the site administrators to solve the unavailability.	High	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. It is possible that impact and likelihood have been overestimated, but over the next year they will keep track of future incidents to confirm or modify these values.

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:
  - **Maximum tolerable period of disruption (MTPoD)** (the maximum amount of time that a service can be unavailable or undelivered after an event that causes disruption to operations, before its stakeholders perceive unacceptable consequences): **2 days**
  - **Recovery time objective (RTO)** (the acceptable amount of time to restore the service in order to avoid unacceptable consequences associated with a break in continuity (this has to be less than MTPoD)): **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.

- The dedicated GGUS Support Unit will be used to report any incident or service request.
- 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.