D-SA1.4 – VERCE platform: third operation and monitoring report

20/11/2015

Project acronym:	VERCE
Project n°:	283543
Funding Scheme:	Combination of CP & CSA
Call Identifier:	FP7-INFRASTRUCTURES-2011-2
WP:	WP5/SA1, Management and operation of the research platform
Filename:	D-SA1.4.pdf
Authors[1]*:*	M. Carpenè, A. Gemünd, M. Lanati, G. Moguilny, V. Muraleedharan, D. Weissenbach.
Location:	http://www.verce.eu/Repository/Deliverables/RP4
Type of document:	Deliverable
Dissemination level:	Public
Status:	Final
Due date of delivery:	30/09/2014 – revised on 20/11/2015
Reviewers:	Thomas Garth, Alessandro Spinuso
Keywords:	iRODS, Inca, OTRS, Cloud, VM, gUse, Unicore, X.509, SHA2.

History	*Author*		*Date*	*Comments (and sections)*
1	G. Moguilny	(IPGP)	05/09	Initial skeleton
2	A. Gemünd	(SCAI)	17/09	Backup/restore of production VMs (section 2)
3	M. Carpenè	(CINECA)	18/09	Unicore details in the Inca section (4)
4	V. Muraleedharan	(IPGP)	23/09	iRODS update (section 1)
5	M. Lanati	(LRZ)	24/09	Completed the Inca section (4)
6	V. Muraleedharan	(IPGP)	26/09	End user support (section 5)
7	A. Gemünd	(SCAI)	29/09	Cloud evaluation (section 3)
8	G. Moguilny	(IPGP)	29/09	Summary and conclusions
9	D. Weissenbach	(IPGP)	02/10	Updated lists of resources (Appendices)
10	Th. Garth	(ULIV)	02/10	English improvement
11	A. Gemünd A. Spinuso	(SCAI) (KNMI)	07/10	Internal review revision
12	A. Gemünd	(SCAI)	20/11	Additional review

[1] Alphabetical order

Copyright notice

Contents

List of Figures

List of Tables

Executive Summary

SA1 is responsible for leading the operations and deployment activities of the VERCE platform. The activities of SA1 are aimed at providing a smooth operation of the infrastructure and services, to facilitate the usage of the platform and to provide support for users in case of issues.

The main objectives of the work package are to:

- Operate and manage a Virtual Organisation in coordination with SA3.
- Provide tools and services for unified access to the VERCE research platform combining Data, Grid and HPC resources.
- Provide and manage the distributed research platform.
- Provide and manage a set of application-tailored workbenches and enactment gateways, in coordination with SA2, to support specific use-case scenarios.
- Provide and manage a development testbed on which SA2 and JRA2 develop, integrate and evaluate the next releases of the VERCE platform.

During these last six months, the VERCE data platform has been upgraded, adding new resources based on different storage systems. To provide redundancy for data access and mitigate availability issues, the possibility of a replication mechanism for iRODS was investigated.

Maintenance and scalability of the VERCE Science Gateway have been improved by splitting the hosting of the Science Gateway frontend from the hosting of the backend web services. Particular emphasis has been placed on the upgrade of the gUse services (to at least version 3.6.6), to enable compatability with X.509 certificates using the SHA2 algorithm, and the selection of Linux distributions compatible with software requirements. Four new VMs have been installed in a Cloud environment.

A Cloud application allowing users to discover and download seismic traces through a Web interface and to process them using ObsPy is currently being evaluated by seismologists.

The Inca monitoring system now integrates the CINECA (Unicore 7) and INGV (new GridFTP mechanisms) updates.

The User support portal based on the OTRS ticketing system has been integrated to the VERCE Science Gateway and the support procedures have been refined.

The updated list of computing resources and software components available to VERCE can be found in the appendices pages 10 and 11.

1 Updated data platform with iRODS

The VERCE data platform is managing the data and meta data catalogue for the VERCE scientific use cases, e.g. it is used for storing synthetic waveforms and meta data. Maintenance and support of this data platform is provided by SA1.

New storage nodes

As has been reported before, the VERCE data platform consists of federated iRODS servers from different resource providers. Development and testing of new or enhanced services is carried out on the iRODS resource hosted in the EDIM1 OpenNebula cloud. In this reporting period, work focused in particular on stability testing and the integration into the Science Gateway. Apart from that, the storage resources which are managed by iRODS have been extended during this reporting period (see table 1 for an overview). Additional storage space with a volume of 1.5 TB, backed by NFS, has been added to the UEDINZone. Storage from LRZ, which had previously been added to this zone, was removed and the data it contained was migrated to this new storage. Additionally, a Hadoop cluster with a capacity of 6.5 TB at UEDIN was integrated with iRODS. Integration issues with the GridFTP DSI module are currently under investigation.

Table 1 – iRODS storage on September 2014.

Location of the zone server	iRODS zone	Zone members	Volume	Resource type
IPGP	IPGP05Zone	IPGP, CINECA, SCAI, INGV, UEDIN, ISTerre	200 GB	Linux
SCAI	verce-scai	SCAI, IPGP, UEDIN	8 TB	Linux
CINECA	CINECA	CINECA, IPGP, INGV	5 TB	Linux
INGV	INGV1211	INGV, IPGP, CINECA	20 GB	Linux
UEDIN	UEDINZone	UEDIN, IPGP, SCAI	1.5 TB 6.5 TB	Linux HDFS

Replication policies

To prevent a potential disruption of service for end-users in case one of the iRODS servers of the federation is offline (e.g. for maintenance), high-priority data should be replicated to different servers of the federation. This mechanism should preferably be transparent for users. Unfortunately, iRODS does not provide replication of files between different zones by default. VERCE is consulting EUDAT for a solution to this issue.

As an early test of the possibility to integrate all sites under same iRODS zone, LRZ resources have been integrated under the “UEDINZone” zone during RP3. However, this reduced the performance significantly since the VERCE HPC use case is generating a large number of relatively small files and the iRODS metadata catalogue has to be contacted for every operation. Therefore SA1 has separated the LRZ resources from the UEDINZone again, and will instead add them as a new zone to the platform.

Based on feedback from the last project review, the focus of the work has been shifted from the Data Intensive use case to completing the HPC use case including the misfit calculation. A task force which includes SA1 members was formed during a meeting devoted to the Data Intensive use case in July to provide support for the finalization of this case. Upcoming work in maintaining the data platform for SA1 includes integration of resources chosen by the task force to implement the use case, assisting SA3 to cache data from archives into iRODS, and creation of services to fetch data together with synthetics from the data platform.

2 Revision of the Science Gateway deployment

Motivation

In a joint discussion, SA1 and SA3 determined further requirements and possible improvements for the operation and hosting of the Science Gateway and its hosted auxiliary services (gUse). Among the defined points, the following were deemed particularly important:

- Upgrade of gUse services to at least release 3.6.6 to enable SHA2 compliance, as certificates recently issued by the IGTF member Certificate Authorities all use the SHA2 hashing algorithm.
- Maintain development and production instances of the Science Gateway and services to prevent disruptions for end-users during development and testing.
- Enable easy periodic backups of at least the production instances.
- Use the Linux distribution that is most compatible to the software requirements of gUse for the invocation of different Middleware. For example using specific Grid middleware from the gUse services requires installation of middleware clients on the gUse host, which is not always released for all distributions (e.g. EMI User Interface).
- (optional but desirable) Split hosting of Science Gateway (Liferay and deployed portlets) from gUse backend services hosting for better scalability and operations (e.g. Upgrades) as suggested by the gUse maintainers.

Due to resource restrictions in the prior hosting environment, it was decided that hosting of the Science Gateway and services would be migrated to the external Cloud cluster at SCAI. The Science Gateway and services were already hosted in Virtual Machines, so a comparatively simple migration would have been possible. However, due to the intention of splitting up the Science Gateway and gUse service hosting as well as the change of Linux distribution, it was preferred to deploy all components in freshly installed (higher capacity) VMs and migrate the content such as MySQL database content and files to the new instance subsequently.

Progress

Deployment of the Virtual Machines was completed without any difficulty. Attempting to install gUse in a split-deployment (separated frontend and backend) led to minor issues, that were solved with support from the SCI-BUS project. Migration of the data is being carried out by SA3 with support from SA1. Once the migration is fully completed, the backup strategy will be implemented.

Technical details

Four Virtual Machines were installed in the Cloud environment using Scientific Linux 6.5 for best compatibility with Grid middleware clients and the WS-PGRADE/gUse software stack:

- verce-portal: for hosting the Science Gateway production instance
- verce-guse: for hosting the gUse backend services (like DCI-Bridge)

- verce-portal-dev: for hosting the development instance of verce-portal
- verce-guse-dev: for hosting the development instance of verce-guse

The "-portal" VMs currently contain the gUse 3.6.6 frontend installation, consisting of Apache Tomcat 6.0.39, WSPGRADE (based on Liferay 6.1) and the corresponding frontend portlets. The "-guse" VMs contain the gUse 3.6.6 backend package consisting of Apache Tomcat 6.0.39 and the gUse services (DCI-Bridge, Repository, Storage, WFI, WFS, Stataggregator), as well as Grid middleware clients (EMI User Interface containing gLite, Globus and Unicore clients, VOMS clients etc.) and the EGI trustanchor.

The machines were deployed on an OpenNebula Cloud utilizing Xeon E5 v2 (Ivy Bridge) CPUs on the compute nodes. To supply sufficient resources for high load situations, the machines were configured to use 8 physical CPU cores (no over-commitment) and are able to scale up to 64GB of (not over-commited) memory.

For the backup, it is planned to use ad-hoc milestone snapshots of the full virtual machines as well as additional regular file and data backups. VM snapshots are copies of the complete state of a Virtual Machine, including its CPU state and RAM and disk content. To provide this feature (apart from additional advantages such as copy-on-write support), the QEMU Copy On Write (qcow2)[2] image format is used. The snapshots will be carried out ad-hoc after reaching important milestones of progression, e.g. when new features reach a production level quality. Additionally, regular dumps of the MySQL database and file-level backups are planned. Failover of the whole virtual machines in case of lower-level malfunctions (e.g. compute node outages) will be implemented using Cloud-layer tools (hooks to migrate the VMs).

3 Cloud evaluation

The evaluation of Cloud resources and the development of the Cloud application continued in this reporting period.

The proof-of-concept Cloud application allows users to discover and download seismic traces through a Web interface and to process them using ObsPy on Cloud resources. Data centers that provide a web service compliant with the web service standards of the International Federation of Seismograph Networks (FDSN) [3] are queried based on user input from the Web GUI. The User can then choose from the range of discovered traces the set he wants to download and process.

The processing itself is expressed as an ObsPy script that is entered by the user. With the help of a light-weight messaging service, the download and subsequent processing is distributed to "worker virtual machines" running on Cloud resources. To safeguard the execution, the processing is isolated using Docker sandboxes based on linux containers (LXC). The resulting MiniSeed data gets collectively archived and uploaded to Cloud storage using the libcloud abstraction layer (e.g. using S3 API) and offered for download to the end-user.

Development continued as planned in the last period and several user requests were addressed. Other improvements to the user interface composition and styling included the addition of the possibility to download an archive of the collected results, and syntax highlighting for the Obspy processing scripts. The application was presented on the EGI Community Forum 2014 and in the VERCE project review. It is publically accessible (see figure 1) and is currently being tested by additional users.

[2] The QCOW2 Image Format - https://people.gnome.org/~markmc/qcow-image-format.html
[3] http://www.fdsn.org/webservices/FDSN-WS-Specifications-1.1.pdf

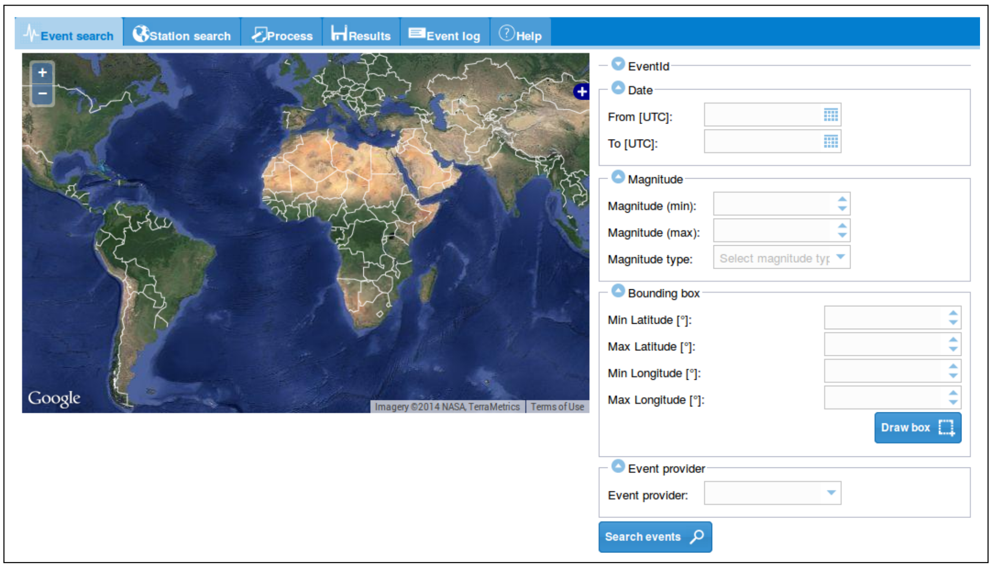

Figure 1 – `http://verce-fsp.scai.fraunhofer.de/`

4 Updated Monitoring with Inca

The deployment and integration of the Inca monitoring platform in the VERCE infrastructure have already been achieved during the last reporting period. The deliverables D-SA1.3 and D-SA1.3.1 give a detailed insight on how Inca works, its background, the resources monitored and the tests that are carried out.

During these months, Inca needed only minor interventions, mainly to address the updates of the infrastructure's components at INGV and CINECA.

The filesystem access mechanism underlying the GridFTP installation on the resource DEP_INGV-01 has changed. The GridFTP server running on this machine does not use standard local disk operations anymore, the administrators instead opted for an iRODS backend. The Inca reporter performing Data Access and Computing tests had to be adapted in order to face the different default path assigned to the remote (i.e., on the resource) Inca account executing the probe's commands.

CINECA installed the new version of the UNICOREX Server and UNICORE Front End on the production nodes. The UNICORE framework has been updated from UNICORE6 to UNICORE7, in particular the new UNICORE 7.0.3 version supports additional features, such as proxy certificates usage. The new release has been fully tested and UNICORE7 services have been added to Inca monitoring both on the HPC_CINECA-02 and HPC_CINECA-03 clusters (see the *Job Submission* section of the Inca consumer). Finally, the UNICORE BES (Basic Execution Services) interface has been configured in order to provide BES compliant access to both CINECA HPC clusters.

5 End-user Support

SA1 maintains a support system to provide a central contact point for users of the VERCE infrastructure and services. It is based on the open-source OTRS ticketing software and has been integrated to the VERCE Science Gateway (`http://portal.verce.eu/support`).

The ticketing system and support procedures have been extended to include members of different work packages besides SA1. The first level support and assignment of tickets to the corresponding experts based on the type of request is carried out by SA1. For example, if the issue is related to job submission in the gateway, the ticket is assigned to be solved by SA3 members. Supporteres can also proactively take ownership of tickets before they are assigned.

The following work packages are responsible for the different types of tickets

- Account - SA1
- Software Request - SA2
- Security Issue - SA1, SA3
- Service Availability - SA1, SA3
- Others - SA1

This ticketing system was introduced to the attendees of the webinar conducted in July together with the VERCE Science gateway. Up to the present, the usage of the support system has been low, but is expected to increase as the Science Gateway usage increases.

Conclusion and perspectives

The main achievements for this period are:

- the extension of the data platform based on iRODS with additional storage,
- a considerable improvement of the Science Gateway deployment by migration and distribution of the services,
- the inclusion of new kinds of tests to the the Inca monitoring system,
- the provision of a Cloud application allowing users to retrieve and process seismic traces from data centers using standard protocols,
- the integration of a frontend to the support system in the Science Gateway

Tasks for the upcoming period include:

- On the Data Platform:
  - investigate the interoperability issue of the iRODS GridFTP interface when used in combination with Hadoop,
  - simplification of the data replication mechanism for improved reliability.
- Maintenance and further improvement of the Science Gateway deployment.
- Implementation of a backup strategy for the Science Gateway and services.
- Deployment of upgraded software components, e.g. Dispel4py, SpecFEM3D at the VERCE sites.
- Support the transformation of the VERCE e-infrastructure framework into an operational state (e.g. Dispel4py workflows, Science Gateway).
- Support SA3 in integrating additional resources into the Gateway (e.g. resources using UNICORE middleware) for the execution of the simulation workflows.

Appendices

A Computing resources available

Table 2 – Hardware resources.

Designation	Resource Type	Short Name	Brief Description	Access Protocols	Accounting	Person in charge
HPC-CINECA-02	Computation	PLX GPU Linux Cluster	Model: IBM PLX (iDataPlex DX360M3).Architecture: Linux Infiniband Cluster, 274 IBM iDataPlex M3 Nodes, Processors: 2 six.cores Intel Westmere 2.40 GHz (548 processors, 3288 cores). GPU: 2 NVIDIA Tesla M2070/node for 264 nodes + 2 NVIDIA.Tesla M2070Q/node (for 10 nodes), RAM: 48 Gb/node.Internal Network: Infiniband with 4x QDR switches.Disk Space: 100 TB.Operating System: Red Hat RHEL 5.6, Peak Perf: 300 TFlop/s (142 TFlops sustained - Linpack benchmark)	ssh to login node	yes	Graziella
HPC-CINECA-03	Computation	BlueGene/Q	Architecture: 10 BGQ Frame with 2 MidPlanes each Front-end Nodes OS: Red-Hat EL 6.2 Compute Node Kernel: lightweight Linux-like kernel Processor Type: IBM PowerA2, 1.6 GHz Computing Nodes: 10.240 with 16 cores each Computing Cores: 163.840 RAM: 16GB / node; 1GB/core Internal Network: Network interface with 11 links ->5D Torus Disk Space: more than 2PB of scratch space Peak Performance: 2.1 PFlops	ssh to login node	no	Graziella
HPC-LRZ-01	Computation	SuperMIG/ SuperMUC	PRACE Tier 0 machine. SuperMUC (Summer 2012). Intel Xeon-Architecture, more than 110,000 cores. 3PFlop/s, 16cores/node with. 2GB/core memory and 40cores/node with 6.4 GB/core memory. 10 TB GPFS (aggregate BW 200 GB/s). linked via fully non-blocking. Infiniband network 1GB Ethernet to archive and backup	ssh to login nodes behind firewall, only accessible from IPs entered in firewall; then port 22 and globus ports range opened.	yes INCA, Dmon	Cerlane
DEP-INGV-01	Computation	selene	48 dual-proc AMD 6136 2.4GHz 8-core 64GB RAM (total 768 cores) 64 dual proc AMD 2378 2.4GHz 4-core 16GB RAM (total 512 cores) 36 TB scratch-grade storage	Account on request, then access via GSIssh (ports 22 and 2222). GRAM via Globus gatekeeper (port 2119).	no	Daniele
DEP-IPGP-06	Computation	S-CAPAD Dell cluster	96 CPU nodes,16 data intensive nodes. 4 GPGPU nodes. 1 SMP node, parallel storage : GPFS on 576 effective TB, a high-performance network with InfiniBand QDR switches.		yes	Geneviève
DEP-UEDIN-01	Computation	EDIM1	Data intensive cluster. 120 nodes, each with 6.24TB of storage. CPU atom dual core, 4GB ram	SSH to login node behind firewall only from IPs within Edin. Informatics or EPCC	yes Ganglia	Paul
DEP-ULIV-01	Data	Linux Storage	CPU Intel Xeon dual core 2.8 GHz, 2GB RAM, 2.8TB storage		no	Andreas
GRI-IPGP-03	Computation	EGI Cluster	SiteBDII: sitebdii1.ipgp.fr - ldap://sitebdii.ipgp.fr:2170/mds-vo-name=NSU01-PARISo=grid CREAM-CE: ce1.ipgp.fr - ce1.ipgp.fr:8443/cream-pbs-es DPM: dpm1.ipgp.fr SRM endpoint: srm://dpm1.ipgp.fr:8446/dpm/ipgp.fr/home/verce.eu/ GridFTP endpoint: gsiftp://dpm1.ipgp.fr:2811/dpm/ipgp.fr/home/verce.eu/ 136 CPU RAM: 1GB/CPU Storage: 8TB	gLite job/file operations, ssh to login on User Interface	yes	David
GRI-LRZ-02	Computation	Linux Cluster	> 2800 cores		yes (but only for internal purpose)	Cerlane
GRI-SCAI-01	Computation	EGI Cluster	32x Sun Fire X2200 M2 (Dual Opteron 2218, 2x2 cores), 8GB RAM, 250GB HDD 1 x Sun Fire X4600 M2 (Dual Opteron 8220, 4x2 cores), 78GB RAM, 2x140GB SAS HDD 8.5 TB NFS4 on ZFS shared filesystem (homes), 4.5 TB gLite DPM Storage Element (Grid) Connectivity: 2x Gbit Ethernet, Mellanox MT25204 10Gbps InfiniBand.	gLite, Unicore, Globus job/data access. GSISSH login to UI and dev-node (1, 2)	yes	André

Table 3 – Data and Software resources.

Designation	Resource Type	Short Name	Brief Description of Resource	Access Protocols	Installed Software	Accounting	Person in charge
GRI-EGI-01	Computation	EGI Virt. Org.	EGI Earth Science Research VO	gLite operations	gLite software stack	yes (EGI accounting)	André, David
GRI-EGI-02	Computation	EGI Virt. Org.	VERCE.EU VO	gLite operations	gLite software stack	yes (EGI accounting)	André
DAT-KNMI-01	Data	ORFEUS Data Center	Virtual European Broadband Seismograph Network,European Integrated Data Archive	Web services, interactive sessions	ArcLink		
DAT-IPGP-04	Data	COHER-D	IPGP Data Archive, NIED (Japan)	direct access from IPGP network		no	Geneviève

B Software components

Table 4 – Components (recommended by SA2) / resource

	TestBed									
	HPC-LRZ-01	HPC-CINECA-02	HPC-CINECA-03	DEP-UEDIN-01	DEP-ULIV-01	DEP-IPGP-06	DEP-INGV-01	GRI-SCAI-01	GRI-IPGP-03	GRI-LRZ-02
Codes										
SeisSol (1.0)	✓	✓	✓				◒	●		✓
SpecFEM3D (2.0)	✓	✓	✓			◒	●			✓
Scientific libraries										
Matpotlib (1.0.5)	✓	✓		●		●	✓	✓	✓	✓
ObsPy (0.9.2/0.8.3)	✓	✓		●		✓	●	✓	✓	✓
Metis (4.0.3 or 5.0.2)	✓	✓	●			●		✓	✓	✓
GMT (4.5.9)	✓	✓			●	✓	●	✓		✓
mpi4py (1.3.1)	✓	✓	✓	✓		◒	●	✓		✓
NetworkX (1.8.1)	✓	✓	✓	✓		✓	●	✓		✓
Environment / Tools										
iRODS (3.x)	✓	✓	✓	✓			✓	✓	✓	✓
iRODS-DSI (1.6)		✓	✓	✓			✓	●	✓	
GSI-SSH (GT 5.2.x)	✓	✓	✓				●	✓	●	✓
GridFTP (6.14)	✓	✓	✓	●			✓	✓	✓	✓
GRAM (5)							●	✓		✓
JSAGA (0.9.15)				●				✓	✓	
gLite CREAM (1.14.0)								✓	✓	
gLite WMS (3.4)								✓		

◒ Currently being installed ● Installed

✓ Available for VERCE users

An empty cell means not applicable or undecided as of yet