Environmental Science Federated Cloud Platform in the BSEC Region

H. Astsatryan[1], A. Hayrapetyan[1], W. Narsisian[1], V. Sahakyan[1], Yu. Shoukourian[1], A. Stanciu[2] and G. Neagu[2]

[1]Institute for Informatics and Automation Problems of the National Academy of Sciences of Armenia, Yerevan, Armenia
[2]National Institute for Research & Development in Informatics, Bucharest, Romania

Abstract— Various types of research infrastructures available in the region of Black Sea Economic Cooperation (BSEC) make it possible to handle large data sets and provide significant computational resources. The environmental science community has a vital role in the region by exploiting the available computational and storage resources using several digital models, special data infrastructures and tools. The main aim of the article is not only to present the extended federated cloud platform for target user communities, but also to introduce the environmental science potential and geoprocessing facilities that may benefit from the suggested platform.

Index Terms—BSEC, cloud, GIS, Federated, Environmental Science, OpenNebula, Ozone.

1 Introduction

In recent years, state-of-the-art research infrastructures have been deployed in the Black Sea region (BSEC) via a number of targeted initiatives funded by the European Commission [1-2] and national Governments. Various types of research infrastructures available in the region make it possible to handle large data sets and provide significant computational resources. However, the complexity of the infrastructures is often discouraging new and inexperienced users and impedes the use of such technologies in their application domains.

Federated cloud infrastructure, that is a service platform [3] based on OpenNebula open-source cloud management system [4], has been deployed in the BSEC region enabling user communities from participant countries (Armenia, Georgia, Moldova and Romania) to work in a virtualized environment by providing them with virtual machines, networks and storages. The resource providers within the BSEC cloud share the local resources (computational, storage and network) using the oZones component of OpenNebula, which allows for the centralized management of multiple instances of OpenNebula (zones), managing in turn potentially different administrative domains. This enables user communities to use the local or remote resources and makes the regional collaboration of user communities easier. The SunStone graphical user interface is also provided to the user communities to manage the resources and use them in a more optimized way.

The main objective of this article is to present the environmental science potential and geoprocessing facilities available in the extended BSEC federated platform. The remaining part of the paper is organized as follows: Section 2 introduces the related work developed within the European Grid Initiative (EGI) on setting up the EGI production platform technical architecture based on a federated cloud infrastructure. Section 2 is explains the extensions implemented in the BSEC federated cloud platform. In Section 4 the specificity of requirements and needs of the Environmental Science community is underlined and illustrated with some applications developed in support of this community in BSEC participant countries. Section 5 presents the results of experiments demonstrating the capacity of the proposed infrastructure to meet these specific requirements as well as the penalizing network impact on the efficiency of these applications. Some concluding remarks are provided in Section 6.

2 Related Work

In recent years there are several ongoing efforts towards creating new federated computing infrastructures based on the existing resource providers. For example the EGI.eu foundation as a part of EGI operates the grid and federated cloud based infrastructures [5].

The architecture of the proposed EGI platform is based on a core infrastructure on top of which are deployed several technologies such as the new middleware for high-throughput data analysis, and specific virtual research environments which could use either the cloud resources or the data analysis platform. The core infrastructure platform should include the main components that are needed to support the operation, such as the monitoring system, information system, service registry and accounting system.

The current grid middleware solutions envisaged to be integrated with the core infrastructure are EMI-3 HTC grid platform, IGE-2, UNICORE and QCG, and the cloud management toolkits currently used are Openstack, Opennebula and StratusLab.

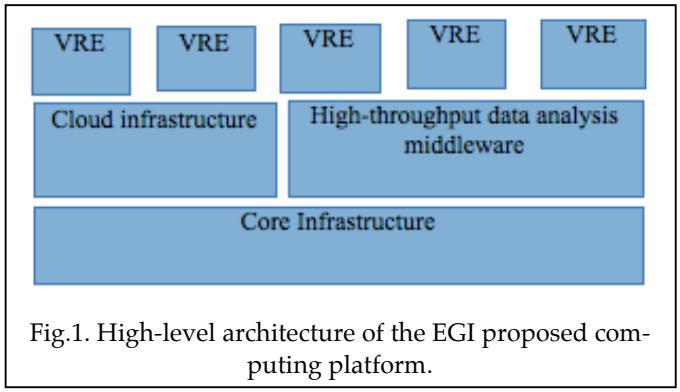

Fig.1. High-level architecture of the EGI proposed computing platform.

International Journal of Scientific & Engineering Research Volume 5, Issue 1, January-2014
ISSN 2229-5518

The cloud infrastructure which was built by EGI provides access to computing resources using both EC2 and OCCI query interfaces, cloud storage based on CDMI endpoints, and a Virtual Appliance Marketplace. In Fig.1 is depicted a high level architecture of the proposed EGI computing platform.

One important use case for the virtualized resources provided by the cloud infrastructure is of domain specific e-laboratories [6]. These are virtual appliances deployed on local or remote cloud resources that can be customized with specific configurations for dedicated tasks which are performed by the members of the scientific communities.

3 BSEC Extended Federated Cloud Platform

The BSEC infrastructure provides user communities with ability to build their own system images by installing and configuring their required software, which is often difficult for non expert users to take advantage of nowadays computational infrastructures or even to use the available advanced libraries, because this often requires to be familiar with middleware and tools, parallel programming techniques and packages. Thus besides of building a platform where user communities can work and collaborate, there is also a strong need to provide them with preconfigured system images, which are ready to use just after running them on virtual machines. After the analyzes of the needs of the target user communities in the region, several images of virtual machines have been created and in a few of them user friendly graphical web interfaces have been developed on the top of the tools that the users need for their research. This approach allows the users to have access to the resources both via standard command line tools and via graphical user interfaces in some cases. Besides the computational resources and services, the suggested platform consists of local data repositories that may be used for regional applications in an efficient way (see fig. 2). For example in the case of Armenia, the data repository is a spatial data infrastructure [7], which consists of geographic data, metadata and tools.

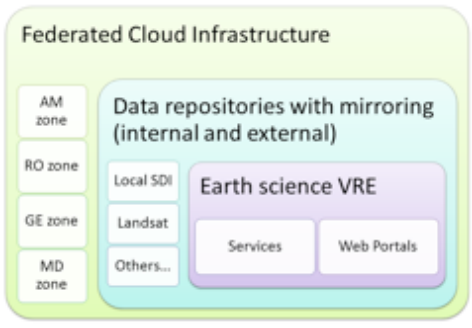

Fig.2. BSEC extended federated cloud platform

4 Environmental Science Applications in BSEC Region

Due to its intensive data processing and highly geographically distributed organizations, the multidisciplinary Environmental science community is uniquely positioned for the uptake and exploitation of the suggested infrastructure. According to the authors' experience in setting up and management of the SEE-GRID eInfrastructure for regional eScience (www.see-grid-sci.eu, www.egi.eu) this community is active and plays an active role in BSEC participant countries nowadays. The requirements and needs of this community are different in terms of computational facilities, network usage scenarios, bandwidth, location of input and output data, frequency and size of file transfers between the home site (where input data are obtained) and execution sites (where data is being processed), frequency and size of transfers of data between execution site and home site.

Many regional applications in this domain need a huge amount of computational resources, such as implementation of numerical weather prediction models with coarse resolution [8] in order to have hydrometeorological information of finer resolution and to analyze changes of hydrometeorological elements at the regional scale, or modeling multiple air quality issues [9], including tropospheric ozone, fine particles, toxics, acid deposition, and visibility degradation.

One of the key activities in the region is the management, processing, and sharing of geospatial data in the environmental domain. As a geospatial data, satellite images don't only have spatial dimensions, but are also characterized by radiometric, spectral and temporal dimensions. Satellite images are important sources of raw observational data about the Earth, but they typically need to be processed and analyzed to be transformed into useful information.

Several applications for parallel Geo-processing of Satellite Image Indices have been developed in the region that use parallel computing resources for complex calculations of indices based on remote sensing imageries. For example, in case of the Armenian application, a parallelization method has been used [10], which automatically chooses the target computational resources depending on the complexity of the operations and the amount of data. The algorithm optimizes the processing by selecting the best methodology (e.g., serial or parallel) and the number of cores to efficiently manipulate and distribute the data. The application uses the OGC Web Processing Service [11] standard for handling requests and responses from the web. Users can either upload their images or point to a remote image and start the calculation. Then the system performs the calculation and sends an email to the user with details about the results of the calculation. The developed instance for the environmental science community (see fig. 3) consists of packages for geoprocessing, digital models, libraries and tools that are being used by communities in the region.

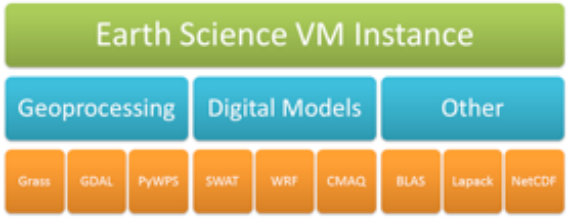

Fig.3. The structure of the VM image for environmental science community

International Journal of Scientific & Engineering Research Volume 5, Issue 1, January-2014
ISSN 2229-5518

5 Experiments

The network impact plays an important role on the efficiency of the environmental applications that are manipulating a large amounts of data. For example the multispectral satellite image processing, such as the calculations of indices of the satellite images from the raw observational data is a good example showing the complexity of handling and organizing the data movement and simulations. But in case of some indices the transfer of satellite images between zones in BSEC infrastructure may take more time than the calculations itself.

A series of experiments have been carried out in order to analyze the network impact on the above mentioned applications by taking into account the data transfer time.

Two identical virtual machines (2 CPU cores, 2048 MB of RAM) located in Armenian and Romanian zones of the BSEC federated infrastructure have been used for the experiments. Both of the virtual machines were instances of the mentioned VM image for the environmental science community. The connectivity capacity inside the zones was gigabit and between the zones was about the 50Mbit/s. Images with various sizes indicated in table 1 were downloaded from the Landsat public imagery [12] and stored in data storage of Armenian zone. The imagery represents the world's longest continuously acquired collection of space-based moderate-resolution land remote sensing data.

TABLE 1
LANDSAT IMAGES USING FOR THE EXPERIMENTS

	Rows	Columns	Cells	Size (in MB)
1	5770	6040	34850800	≈ 32
2	7697	8054	61991638	≈ 64
3	11545	12080	139463600	≈ 128
4	15200	16100	244720000	≈ 256

The calculation of the Normalized Difference Vegetation Index (NDVI) [13] is used for the experiments, because the index is widely used by user communities in the region. The scenario of the experiments includes the calculation of the satellite images based on parallel approach using two cores both on local and remote virtual machines. The analyzes show that the calculation times were nearly the same in both virtual machines but the time of data transfer from the data repository to the virtual machine in its local zone and the virtual machine in the remote zone was fairly different and in case of remote virtual machine it was also quite unstable, because it mostly depends on the network bandwidth, which is dedicated for different applications in different time periods. Thus the time of data transfer to remote zone through the global network is not only being affected by the amount of the data itself but also by the work of other services which use the network channel.

Because the calculation of NDVI requires two different bands of satellite images (Red and NIR Infrared), which were described in table 1, the actual data that was transferred during each experiment was twice bigger than the size of each image. The average time of data transfers and actual calculation process during the experiments were benchmarked (see fig. 4)

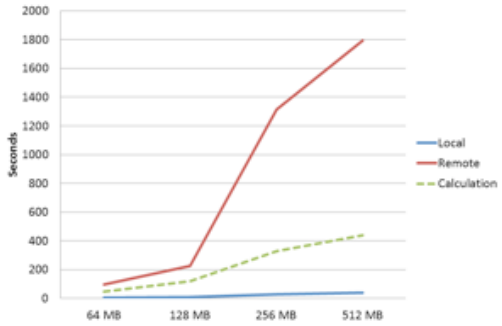

Fig. 4. Data transfer and calculation time

6 Conclusions

The suggested federated cloud platform offers computational and storage facilities for target user communities of the BSEC region. It simplifies the work of user communities by providing them with ready to use virtual machine images. The optimization mechanisms deployed on the platform, like, for example, parallel computation of some time consuming tasks, help to use the computational resources in a more efficient way. The platform also makes it easies for participating countries to share national computational and storage resources with each other.

The experiments showed that the platform works effectively for solving regional problems that require large amounts of computational resources. In such cases when the data manipulations required for solving the problems are not very large the platform works effectively even when the actual calculation happens not in the zone where the data is stored. But in some cases, when the data that need to be transferred from the storage to the remote zone for processing is large, the data movement can take a number of times longer than the calculation itself and thus reduce the efficiency of the entire process. Also in some cases different tasks are using the same data for the calculation, so if the calculation will take place in remote zone, the same data will need to be transferred through the global network several times.

So to make the regional collaboration and resource sharing more efficient in such cases, it was decided to deploy data mirroring mechanism in the platform to merge the data between the storages of different zones. This mechanism will solve the problems related to data transfer described above.

Acknowledgment

This work was supported by the Organization of the Black Sea Economic Cooperation through the project "Experimental Deployment of an Integrated Grid and Cloud Enabled Environment in BSEC Countries on the Base of g-Eclipse".

International Journal of Scientific & Engineering Research Volume 5, Issue 1, January-2014
ISSN 2229-5518

REFERENCES

[1] Integrated Sustainable Pan-European Infrastructure for Researchers in Europe, http://www.egi.eu/about/egi-inspire/
[2] High-Performance Computing Infrastructure for South East Europe's Research Communities, http:/www.hp-see.eu
[3] Hrachya Astsatryan, Andranik Hayrapetyan, Wahi Narsesian, Peter Bogatencov, Nicolai Iliuha, Ramaz Kvatadze, Nugzar Gaamtsemlidze, Vladimir Florian, Gabriel Neagu, Alexandru Stanciu, "Deployment of a Federated Cloud Infrastructure in the Black Sea Region," *Proceedings of the International Conference on Computer Science and Information Technologies (CSIT'2013)*, pp. 283-285, Yerevan, Armenia, September 23-27, 2013.
[4] OpenNebula Project (2008) http://www.opennebula.org/
[5] Michel Drescher, Tiziana Ferrari, David Wallom, "EGI position paper for a Federated Cloud," *EGI Towards Horizon 2020 Workshop*, Amsterdam, 2013.
[6] Gergely Sipos, Peter Solagna, Salvatore Pinto, Tiziana Ferrari, David Wallom, "EGI position paper for Data Services," *EGI Towards Horizon 2020 Workshop*, Amsterdam, 2013.
[7] H. Astsatryan, W. Narsisian, V. Ghazaryan, A. Saribekyan, Sh. Asmaryan, V. Muradyan, Y. Guigoz, G. Giuliani, N. Ray, "Toward to the Development of an Integrated Spatial Data Infrastructure in Armenia," *Proceedings of the ICT Innovations 2012 Conference*, ISSN 1857-7288, pp. 85-93, 12-15 of September 2012, Ohrid, Macedonia.
[8] H. Astsatryan, V. Sahakyan, Yu. Shoukourian, A. Shahnazaryan, Z. Petrosyan, R. Abrahamyan, H. Melkonyan, L. Vardanyan, "Sensitivity of WRF Data Assimilation for Heavy Rainfall event over the Territory of Armenia," *Proceedings of the 7th European Conference on Severe Storms (ECSS2013)*, extended abstract, #212, 3 - 7 June 2013, Helsinki, Finland.
[9] A. Mirzoyan, H. Saroyan, G. Minasyan, V. Sahakyan, Yu. Shoukourian, H. Astsatryan, "Environment for Access to the Inventory of Stationary Point Sources of Emissions of Air Pollutants in Armenia," *Proceedings of the International Conference on Computer Science and Information Technologies (CSIT'2013)*, pp. 453-455, Yerevan, Armenia, September 23-27, 2013.
[10] D. Petcu, D. Gorgan, F. Pop, D. Tudor, D. Zaharie, "Satellite Image Processing on a Grid-Based Platform," *International Journal of Computing*, Research Institute of Intelligent Computer Systems, Ternopil National Economic University. 2008, Vol. 7, Issue 2, pp. 51-58.
[11] Shashi Shekhar, Hui Xiong, "Web Processing Service," *Encyclopedia of GIS*, 2008, p 1269.
[12] Landsat public imagery, http://landsat.usgs.gov
[13] R.E. Crippen, "Calculating the vegetation index faster," *Remote Sensing of Environment*, 34, 71-73, 1990.