Deployment of an e-Infrastructure for Academic Research

Collins N. Udanor[1(✉)], Florence I. Akaneme[2], Stephen Aneke[1], Blessing O. Ogbuokiri[1], Assumpta O. Ezugwu[1], Chikaodili H. Ugwuishiwu[1], Carl E.A. Okezie[2], and Benjamin Ogwo[3]

[1] Department of Computer Science, University of Nigeria Nsukka, Nsukka, Nigeria
{collins.udanor,stepehen.aneke,blessing.ogbuokiri,
assumpta.ezugwu,chikodili.uwguishiwu}@unn.edu.ng
[2] Department of Plant Science and Biotechnology,
University of Nigeria Nsukka, Nsukka, Nigeria
{florence.akaneme,carl.okezie}@unn.edu.ng
[3] Department of Vocational Teacher Preparation,
State University New York, Oswego, NY 13126, USA
Benjamin.ogwo@oswego.edu

Abstract. One of the greatest problems researchers in Africa face, according a 2007 UNESCO report is a chronic lack of investment in facilities for research and teaching. This has both affected the quality and quantity of research output from institutions of higher learning, with the ripple effect of stagnating industrialization and R&D processes. This paper presents the design and implementation of an e-infrastructure, which is made up of cloud and grid computing clusters domiciled in University of Nigeria Nsukka (UNN). The project has objectives, such as to deploy an Identity Provider (IdP) based on Simple Access Markup Language (SAML) that uses robot certificates to authenticate users on the cloud and grid infrastructures, deploy a Web 2.0 based Science Gateway application server that enables researchers have access to simulation, and modeling applications in their research domains on the infrastructure. As well as implement a Virtualized cluster for big data analytics. Results from one of the applications developed and deployed on the infrastructure show over 60 % predication accuracy while participation database in the infrastructure has reached up to 350 users.

Keywords: Cloud computing · Science gateway · e-Infrastructure · IdP · Clusters · Robot certificates

1 Introduction

One of the greatest problems researchers in Africa face, according a 2007 UNESCO report is a chronic lack of investment in facilities for research and teaching [1]. According to Thomson Reuters National Science Indicators database, between 1999 and 2008 the quantity of papers published in the three highest publishing countries in Africa stands as follows; Egypt in the North, 30,000; Nigeria in the middle, 10,000; and South Africa in the South, 47,000. Yet, Netherlands alone produces over 27,000 every year. Researchers, who would want to publish papers with impact factors like Thomson

R. Glitho et al. (Eds.): AFRICOMM 2015, LNICST 171, pp. 57–65, 2016.
DOI: 10.1007/978-3-319-43696-8_6

Reuters, would usually travel to countries where the facilities exist. Research in the 21st century requires skills in the area of the 4Cs (critical thinking and problem solving, communication, collaboration, and creativity and innovation), all which are addressed by Grid and Cloud Computing. Most research tools today are soft tools. And most research works done today are based on the use of software simulation, modeling or analytic tools. High Performance Computing (HPC) comes with multi-cores of processing power, petabytes of storage and myriad of software tools that range from the modeling of bridges to protein synthesis, etc.

E-Infrastructures can be defined as networked tools, data and resources that support a community of researchers, broadly including all those who participate in and benefit from research [2]. According to [2], the term e-Infrastructure comprises very heterogeneous projects and institutions within the scientific community. E-Infrastructures include services as diverse as the physical supply of backbone connectivity, single- or multi-purpose grids, supercomputer infrastructure, data grids and repositories, tools for visualization, simulation, data management, storage, analysis and collection, tools for support in relation to methods or analysis, as well as remote access to research instruments and very large research facilities.

In 2011 the University of Nigeria Brain Gain Initiative (BGI) project successfully set up the first ever Grid Computing Infrastructure in Nigeria (The Lion Grid), under the funding of UNESCO and HP [3, 4]. The Grid computing infrastructure enables researchers to run jobs remotely from their PCs, monitor the job and receive results. Because of this success UNESCO granted the UNN BGI further sustainability funding from June till December, 2013 when the project ended. The BGI project organized several seminars and workshops within and outside the university, including one workshop at Federal University Ndufu-Alike (FUNIA) Ebonyi state, Nigeria in July 2013 which had over 150 academics in attendance, a presentation at the West and Central Africa Research and Education Network (WACREN) conference in 2013 at the National Universities Commission (NUC) Abuja Nigeria, eI4Africa International Thematic Workshop in March 2014 at the University of Lagos (UNILAG) [5], two workshops in UNN for academics and researchers in which staff of the Centre for Atmospheric Research Ayingba (CAR), a unit of the National Space Research and Development Agency (NASRDA) attended in July 2014, and a presentation at UNESCO headquarters, Paris in September, 2013.

Since the establishment of the High performance Computing (HPC) infrastructure (the Lion Grid), a number of research applications have been deployed and put to use by researchers, including the OpenFoam application for fluid dynamics used by some PhD research students in Mechanical Engineering, UNN, Plantisc, a Plant Tissue Culture micro propagation simulation software, which achieved over 60 % predication accuracy [6] developed locally by the UNN BGI team, etc. Active research on the project continued even after the end of the BGI project. The team has continued to work hard on sustaining the project by deploying more applications for researchers and extending the infrastructure to include the science gateway cloud and big data cluster components. The Lion Grid was established to meet the challenges faced by researchers as well as limit brain drain. In view of recent changes in technologies it has become expedient to upgrade the infrastructure.

In mid-2012, a new approach to grid computing was introduced by the eI4Africa FP7 project funded by the European Commission (DG CONNECT). The aim of the eI4Africa

project was to boost the Research, Technological Development and Innovation (RTDI) potential of African e-Infrastructures and to support policy dialogues and EuroAfrican cooperation in the framework of the joint Africa-EU Strategic Partnership [7]. The eI4Africa project developed the grid Science Gateway (SGW) for Africa [8], a cloud based infrastructure for research application repository. This new infrastructure brings the grid closer to individuals as against the Virtual Organization method that involved the use of Digital Certificates that were difficult to obtain. The Science Gateway uses robot certificates issued by Identity providers (IdPs), which can now be located at individual institutions, rather than Certificate Authorities (CAs).

2 The Proposed Infrastructure

It is against this background that we proposed to develop an institutional based Cloud computing infrastructure that has both an IdP and a Grid Science Gateway, a repository for simulation, modeling and analytic applications. Simulation helps a scientist test his ideas before practice, detect problems in workflows and enables one increase business productivity as well as minimize the number of experimental trials in the laboratory, where applicable. This infrastructure could be adopted by ngREN as a service provider (SP) in the recently commissioned ngREN network of interconnected universities in Nigeria, which till now has no services running on it. This project will bring e-infrastructures such as the grid and SGW with a repository of simulation and analytic software tools like R, GATES, Octave, OpenFoam, Clustalw, etc. closer to the researchers in Nigeria by limiting network latency and improve response time. It can also be used by other NRENs within and beyond the WACREN region.

The processes of implementing this project includes the following tasks:

1. Upgrade the current infrastructure for the Lion Grid by adding a Cloud computing component made up of additional high performance servers with Redundant Arrays of Independent Disks (RAIDs) and virtual machines. The Cloud will offer Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) based on multi-tenant approach.
2. Add Solar panels to the existing power backup infrastructure currently made up of a 5KVA inverter and 8 nos of 400 AHr deep cycle batteries. The addition of the solar unit will ensure additional 18–20 h backup, in addition to that of the inverter, generator and public electricity supplies. If this is achieved, there will be a guaranty of 99.9 % server availability (uptime) comparable to most other Grid and Cloud infrastructures and ISPs based in the US and Europe.
3. Develop a Cloud Computing portal with IdP authentication interface that authenticates users of the Grid Science Gateway, and grant them access to a repository of research applications for simulation, modeling, forecasting, analytics, etc.
4. Develop sample simulation applications for researchers as well as port existing ones not readily available to the proposed Cloud infrastructure, in collaboration with our colleagues in the eI4Africa project.
5. Conduct training for support staff and advocacy to the university community.

3 Design and Implementation

The Architecture of the proposed Cloud Infrastructure for Lion Grid UNN is shown in Fig. 1. The implementation strategies are in two phases, systems deployment & configuration and software development.

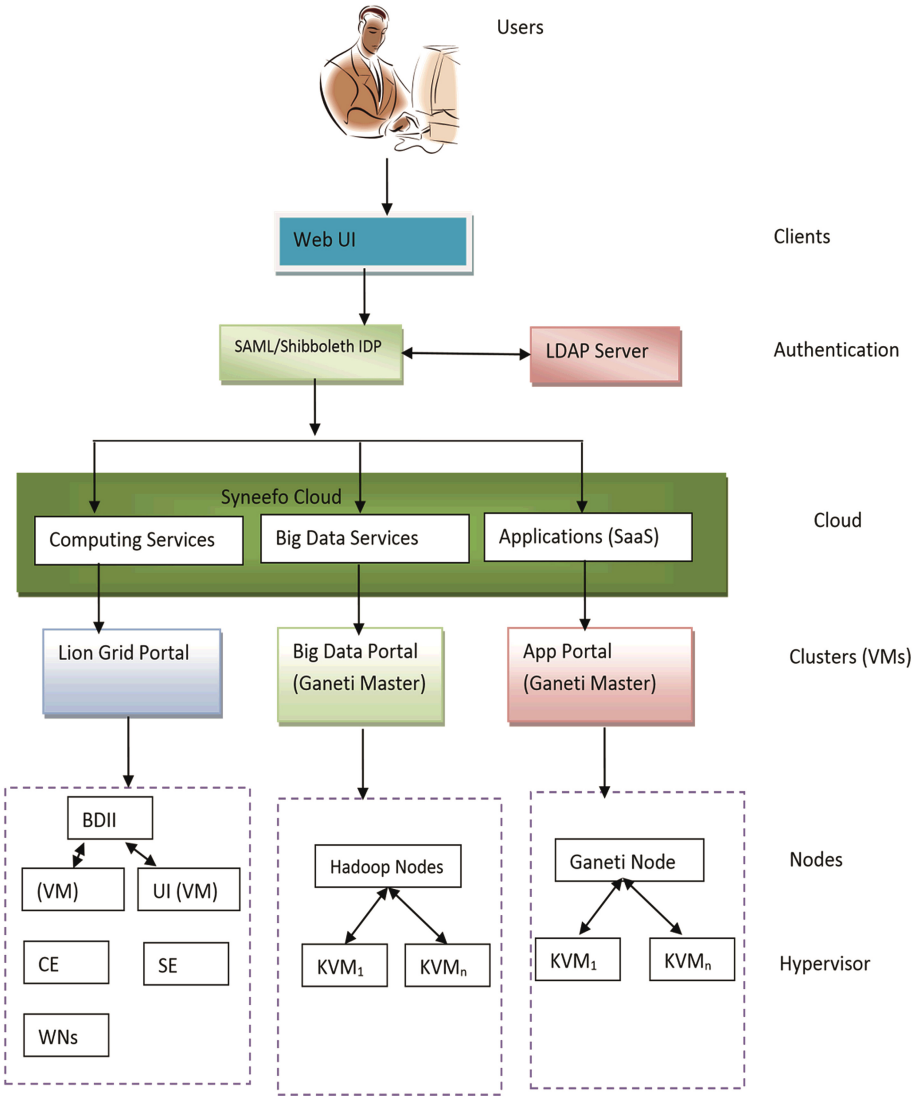

Fig. 1. The architecture of the proposed cloud infrastructure for lion grid UNN

3.1 System Deployment and Configuration

While work is completed on the Grid computing services, which is already configured and running, work is still in progress on the other components of the infrastructure. Due to the challenges of low computing resources, we have built a number of virtual machines using KVM, a kernel virtualization hypervisor in Linux platforms. The VMs are used to build computing clusters, such as the Hadoop cluster. The VMs will also increase the number of available machines to users for computing and running user applications and storages. While Scientific Linux, SL 5.5-64bit is installed on the physical machines for the Grid middleware (gLite 3.2), the Debian 6.5, 64bit distribution of Linux is installed on all the virtual machines. The clusters will be managed by the Ganeti Manager. Ganeti is a cluster management system developed by Google [9]. Each of the machines have a dedicated public IP address, since we have already acquired a/24 bundle of IP-v4 addresses, enough for all our machines. Table 1 summarizes the machine configurations.

Table 1. Machine configuration details

S/N	Machine SPEC	Host name	VMs
1.	IDP HP PROLIANT ML 110 G6 SERVER, 2.8GHZ, core i3, 2BG RAM, 250 GB HDD, VT-x	idp.grid.unn.edu.ng	Nil
2.	LDAP SERVER ML 110 G6 SERVER, 2.8GHZ, core i3, 2BG RAM, 250 GB HDD, VT-x	ldap.grid.unn.edu.ng	1. VM (APP)
3.	BDII HP PROLIANT GL360, G5 SERVER, 2.4GHZ XEON, 4 GB RAM, 1 TB HDD	bdii.grid.unn.edu.ng	1. (UI) Grid.unn.edu.ng 2. VM2
4.	WN HP PROLIANT GL360, G5 SERVER, 2.4GHZ XEON, 4 GB RAM, 1 TB HDD	wn01.grid.unn.edu.ng	1. VM1(Hadoop) 2. VM2(Hadoop)
5.	CE HP Workstations	ce.grid.unn.edu.ng	1. VM1(Hadoop)
6.	SE HP Workstations	se1.grid.unn.edu.ng	

One physical server is dedicated for use as the IdP machine, another for the SGW, and another for storage using RAID-5. At present we have two HP Proliant 360 GL 6 servers with 8 CPU cores each, 1 TB of storage, and 4 GB RAM each, as well as two HP Proliant ML 110 G5 servers with 2 CPU cores each, 250 GB storage and 2 GB RAM each, in addition to two HP Z-workstations.

Installation and configuration of the services like the IdP, SGW will be done remotely by using repositories from cloud clusters like Ansible playbooks. The Simple Access Markup Language (SAML) and Shibboleth will be the principal tools to be used in configuring the IdP and SGW. Figure 2 shows a screen shot of the Lion Grid portal.

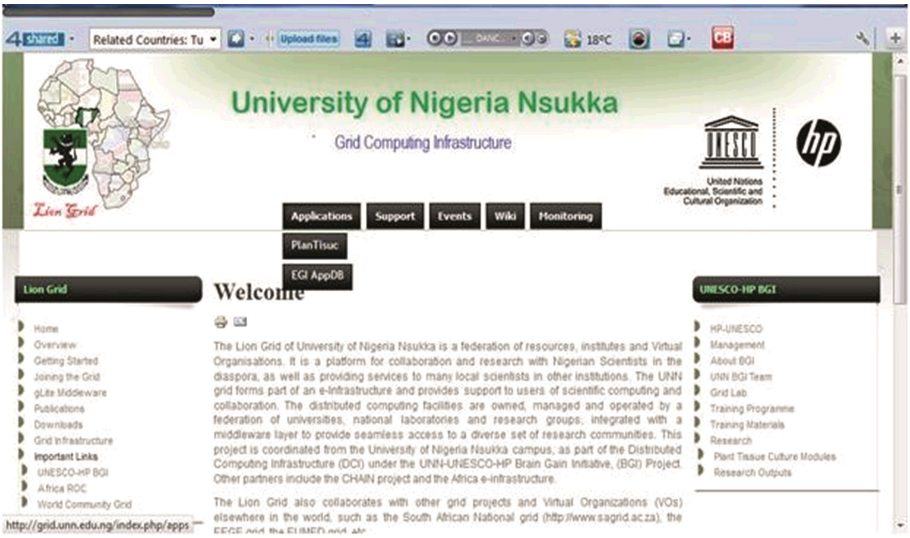

Fig. 2. Lion grid UNN portal

Figure 3 shows the micro-propagation of plant tissue culture experiment laboratory, while Figs. 4 and 5 show the input and results screen shots of the plant tissue culture prediction software (Plantisc), respectively.

Fig. 3. Tissue culture experiments in the lab

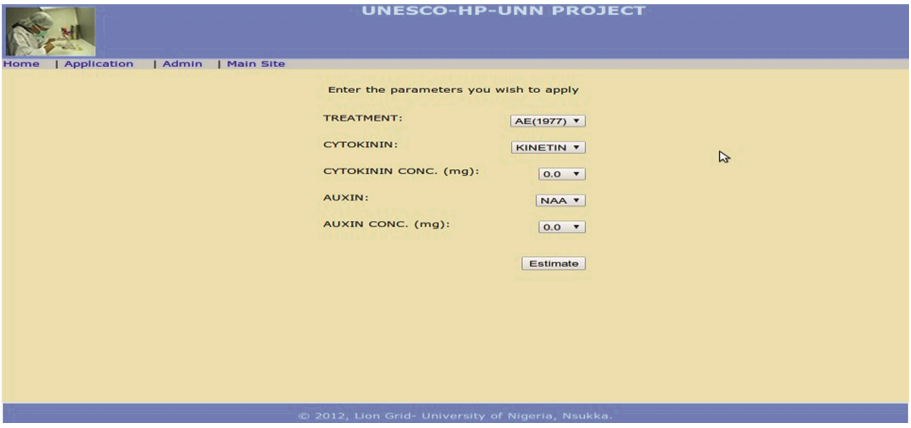

Fig. 4. Plantisc software showing auxin combinations input

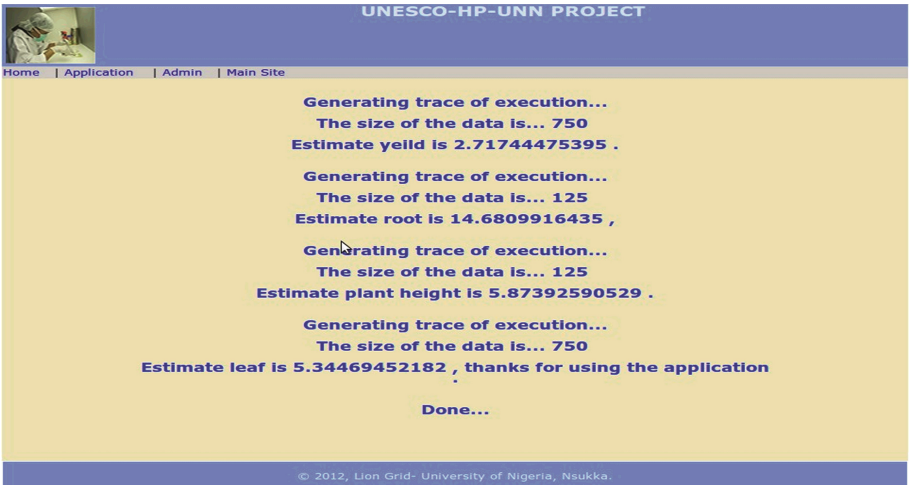

Fig. 5. The result page of the Plantisc application

3.2 Big Data Analytic Application Deployment

Big data can be processed faster and more efficiently than it would be in the more conventional supercomputer architecture in enterprise settings that relies on parallel file system where computation and data are connected to high speed networks. This can be achieved by creating a cluster of distributed computer nodes on which Apache Hadoop tool is installed. Hadoop is an open source framework written in Java for distributed storage and distributed processing on very large datasets on distributed clusters [10]. Hadoop is useful for pre-processing data to identify macro trends or find nuggets of information, such as out-of-range values. It enables businesses to unlock potential value from new data using inexpensive commodity servers. Organizations primarily use

Hadoop as a precursor to advanced forms of analytics. Hadoop can scale down a large data into smaller pieces distributed among computers in a cluster to be processed simultaneously using Hadoop's MapReduce tool. The core of Apache Hadoop consists of a storage part, Hadoop Distributed File System (HDFS) and a processing part, the MapReduce. MapReduce has been used by Google over the years for managing user data.

3.3 Application Deployment on the SGW

In collaboration with the eI4Africa project, we shall deploy some already existing applications such as R for statistical computing and graphics, GNU Octave for numerical computation, Clustalw for multiple alignment of nucleic acid and protein sequences, WRF for weather research and forecasting, Hadoop for big data analytic and prediction, etc. on the proposed e-infrastructure.

3.4 Application Development

Having set up the physical hardware systems and configured the various services on them, the next phase is to develop a number of applications. These will include:

The Cloud Portal: The portal is a single entry port into the infrastructure. It allows users create their accounts and sign in into the system for authenticated by the IdP. It also provides links to applications in the Science Gateway that users can run, among other features.

Streaming Apps: Work is in progress on developing applications that will be able to search social networks like Twitter, Facebook, etc. to extract high volume unstructured data. The application will analyze the data and plot various visualizations that will be used to gain insights into specific areas of user interests. Users will be able to use this app to do different types of searches and show different relationships. The Twitter streaming app has already been developed using Python programming language, and is currently undergoing testing.

3.5 Training

As reported earlier, a number of trainings have been conducted, yet more are scheduled, which will include: Training of researchers on the use of the cloud infrastructure and demonstration of the applications. This will also include how to use our application to extract data from social networks and analyze them.

Training of developers: We shall identify young promising programmers within the university and invite them to be trained on how to develop applications for the SGW. This will ensure sustainability of the project.

4 Conclusion

The impact of e-infrastructures on academic research is huge. According to a survey carried out by [2], more than 85 % of e-Infrastructure users classify e-Infrastructure as important or very important to their work. Most would also see their research work or programmes impaired if the e-Infrastructure did not exist. E-infrastructure opens doors for collaboration, innovation and communication among researchers in virtual communities. The impact of e-infrastructure also brings about the integration or separation of e-Infrastructures at national and disciplinary levels, different organizational and business models, considerations of research communities' needs and practices in the services provided by e-Infrastructures. The research community is in dire need of the innovative tools promised by the e-infrastructure. Over and again we have seen the enthusiasm and eagerness with which the workshops are received and we are certain that this infrastructure is not only timely but a must-have for all institutions that promote credible and cutting-edge research in this 21st century.

Acknowledgment. We acknowledge the support from the United Nations Education Scientific and Cultural Organization (UNESCO), for providing the funds for this research project and Hewlett Packard (HP) for providing the equipment for the deployment of the grid infrastructure. And also University of Nigeria for providing the enabling ground and other forms of support.

References

1. Adams, J., King, C., Hook, D.: Global Research Report Africa. Thomson Reuters publishers, Philadelphia (2010)
2. The Role of e-Infrastructures in the Creation of Global Virtual Research Communities. European Communities (2010)
3. Lion Grid. http://grid.unn.edu.ng
4. Brain Gain Initiative. http://www.unesco.org/new/en/education/themes/strengthening-educationsystems/higher-education/reform-and-innovation/brain-gain-initiative/
5. The 3rd eI4Africa Thematic Workshop. http://ei4africa.eu/wp-content/plugins/alcyonis-eventagenda//files/Training_experience_in_Nigeria_with_the_NgREN_Identity_Provider.pdf
6. Akaneme, F.I., Udanor, C.N., Nwachukwu, J., Ugwuoke, C., Okezie, C.E.A., Ogwo, B.: A grid enabled application for the simulation of plant tissue culture experiments. Int. J. Adv. Comput. Sci. Inf. Technol. **3**(3), 227–242 (2014)
7. ei4Africa. http://ei4africa.eu/
8. Africa Grid Science Gateway. http://sgw-africa.grid.org
9. Apache Hadoop. http://en.wikipedia.org/wiki/Apache_Hadoop
10. Ganeti. http://docs.ganeti.org/ganeti/2.13/html/