SuMegha Cloud Kit: Create Your Own Private Scientific Cloud

Vineeth S Arackal, Arunachalam B, Kalasagar B, Sumit Kumar, Mangala N, Sarat Chandra Babu, Prahlada Rao, Sukeshini
Centre for Development of Advanced Computing, C-DAC Knowledge Park,
No.1 Old Madras Road, Byappanahalli, Bangalore-560038, INDIA
`vineeth,barunachalam,kalasagarb,ksumit,mangala,sarat,prahladab,sukeshini)@cdac.in`

Abstract— In this work, the authors discuss the implementation details of SuMegha Scientific Cloud Lab Kit, which enables users to set up their own private scientific cloud. The scientific cloud offers infrastructure, platform, and/or software services for the scientific community working on problems of modeling and simulation. It enables the creation of virtual clusters (High Performance Computing as a Service). The lab kit is very useful to the researchers and the student community, who want to set up a cloud, but lack the expertise to do so. It not only helps to effectively utilize the idle computers in the organization, but also provides MPI or Hadoop based clusters on demand to its users to solve compute-data intensive problems. The fact that the Cloud can be set up in a single desktop can greatly aid in adoption of Cloud Computing, since almost anyone can now establish a cloud in their premises. SuMegha is designed in a modular fashion using a judicious mix of open source and in-house developed components, such as Nimbus cloud middleware for virtual cluster creation (contextualization), Xen hypervisor and SuMegha Portal for enabling users to request the virtual machines/ clusters and other Services. Storage as a service is provided by the in-house developed CloudVault solution integrated into the lab kit. Supplemented with Software as a Service such as Seasonal Forecast Model (SFM), Next Generation Sequencing (NGS), etc., SuMegha Cloud Lab kit offers a comprehensive environment for scientific computing. We believe that this lab kit is very useful to the vast number of academic institutions, to easily set up the required Cloud Lab.

1.   Introduction

Cloud computing [1][2] provides on-demand access to shared pool of configurable computational resources like servers, storage, applications, networks and software that can be easily provisioned as and when required. Various cloud service models such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) exist today. Cloud computing has become very popular and is gaining acceptance in various domains such as e-governance, health, banking, scientific computing and big data. However, establishment of cloud environment, developing novel applications and migrating any existing applications to the cloud are the major challenges for cloud adoption.

Scientific computing [3] deals with solving large-scale scientific problems in domains such as material science, astrophysics and mechanical engineering,   by utilizing mathematical models and simulations. Running large simulations needs a significant amount of computing resources, often requiring the use of supercomputers, clusters or grids; which may not be available for the researchers and scientists. Scientific clouds can meet these huge computation and storage requirements of scientific community through infrastructure and storage services. Amazon [4] Elastic MapReduce is an example that gives processing capability for vast amounts of data stored over Amazon S3 service. SuMegha is a scientific cloud [5][16], developed by Centre for Development of Advanced Computing (C-DAC), offering High Performance Computing (HPC) as a service and Storage as a Service.
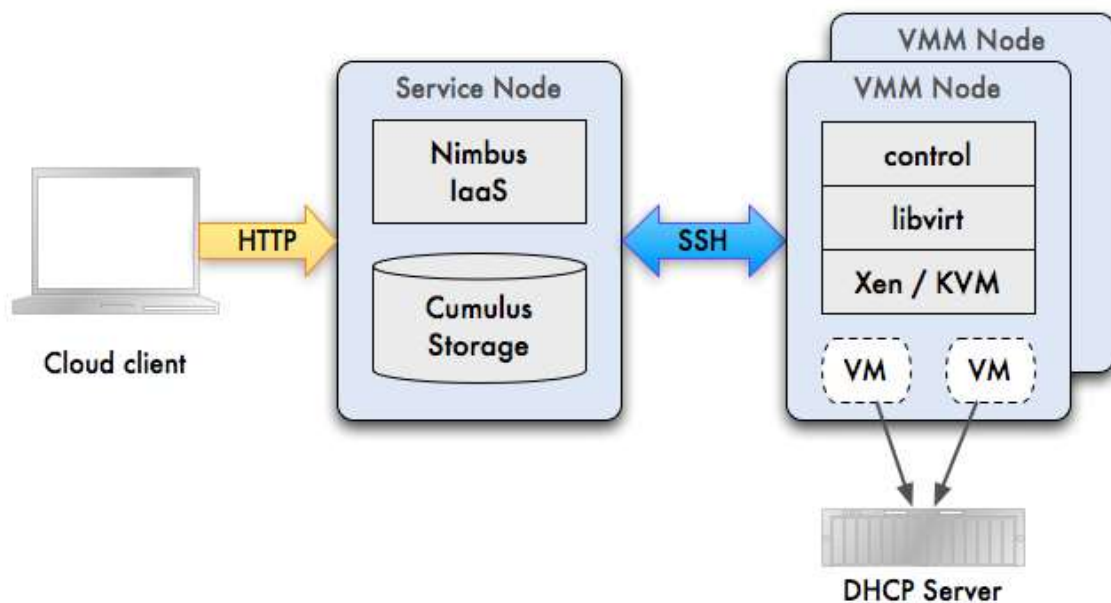

Figure 1:   Nimbus SuMegha Cloud Components

SuMegha Cloud Lab Kit is a sophisticated package that helps the users to set up their own private cloud in a short time frame. It enables quick installation of cloud components through automated installation scripts at each of the cloud layer like virtualization, cloud operating system, management, deploying user interfaces and running cloud services. The lab kit also provides the enablement of the complex parallel environments required for development and deployment of scientific applications.

This lab kit is specifically developed for entry level engineers to get better understanding of basics, necessary functions and features of scientific cloud computing, before working on Real-time environment. To set up Cloud laboratory for experimental as well as practical usage, this lab kit will be valuable for educational and training institutes. It will be useful for research organizations to identify issues and vulnerabilities of important aspects in cloud computing such as monitoring, security, resource usage, billing, etc. Figure 1 shows some of the components of SuMegha.

As a lab kit, it will be useful for the research scholars and college students to do research activity quickly in the areas of cloud computing with various service models (IaaS, PaaS and Saas) and need not worry about underlying technical components of the cloud. Along with the lab kit user benefits as mentioned in the above section, it is also having interactive installation & configuration of cloud stack. This lab kit is also having the option to set up and access the cloud vault cloud storage up to 5 GB.

The dissertations of this paper work are mentioned as follows: section 2 describes about the different related works on cloud lab kit and scientific cloud computing. Section 3 describes the need for the SuMegha lab kit. Section 4 explains about the architecture of this lab kit and describes various components. Section 5 explains various features and screen shots of the lab kit, section 6 and section 8 provides information about the various types of virtual machines and scientific cloud portal and section 8 concludes along with providing brief description about the future works.

2. Related Work

Scientific cloud or HPC as a Service enables greater systems flexibility and eliminates the need for exclusive hardware infrastructure per applications and would help researchers cope with exploring volumes of data that need to be analyzed to get better outcomes. Stratus Lab [8] is developing a complete, open-source cloud distribution that enables grid & non-grid resource centers to provide and to exploit an Infrastructure as a Service cloud. It generally improves the grid infrastructure with virtualization and cloud technologies. It is particularly focused on improving the usage of distributed computing infrastructures such as the European Grid Infrastructure (EGI).

Cumulus [9] is an open source community to build a scientific cloud for any data center. It is a storage cloud system that adapts available storage functionalities to provide improved upload/download interfaces compatible with S3. Open Cirrus [10] test bed is a collection of federated datacenters to build systems and services for open source research. It is defined to support research ideas into the design, provisioning, and managing the services at a global, multi-data center based scale. GridGain[11] is java based open source middleware for real time big data processing and analytics that scale up from one server to hundreds of physical machines.

The main objectives of the scientific cloud infrastructure facility are: (a) Easy to access the HPC resources for academic researchers & research communities (b) On demand provisioning of challenging parallel environments by providing virtual clusters (c) Enablement of scientific applications in the areas of climate modelling and bioinformatics. Subsequent sections provide details about the CDAC scientific cloud and how services offered by it are accessed.

3. Need for Sumegha Lab kit

SuMegha lab kit is required for educational organizations and training institutes for conducting various training programs such as MPI, Open MP and Hadoop[12] modules virtually without actually installing the modules on physical machines. This is possible by using the Golden Images of SuMegha cloud lab kit. For the purpose of conducting various experiments on scientific clouds, this lab kit will be used as test bed by domain experts and cloud application specialists. It is also needed for the research organizations who want to continue their research on Cloud computing areas such as monitoring, security, resource usage, billing.

4. Architecture

The architecture of SuMegha introduces the concept of two types of nodes, similar to client-server architecture: Virtual Machine Manager (VMM) Node and the Service Node (SN), as shown in Figure 2.

SuMegha Cloud Lab kit can have multi-node installation. There shall be only one service node, which acts as a controller. VMM Nodes can be installed on multiple machines and configured with service node. However, it is also possible to install all the components of SuMegha in a single desktop.
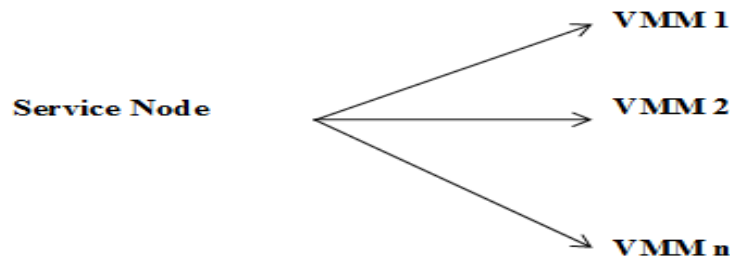

Figure 2: Basic Architecture of SuMegha

SuMegha Cloud lab kit consists of the following components as mentioned in Figure 3:
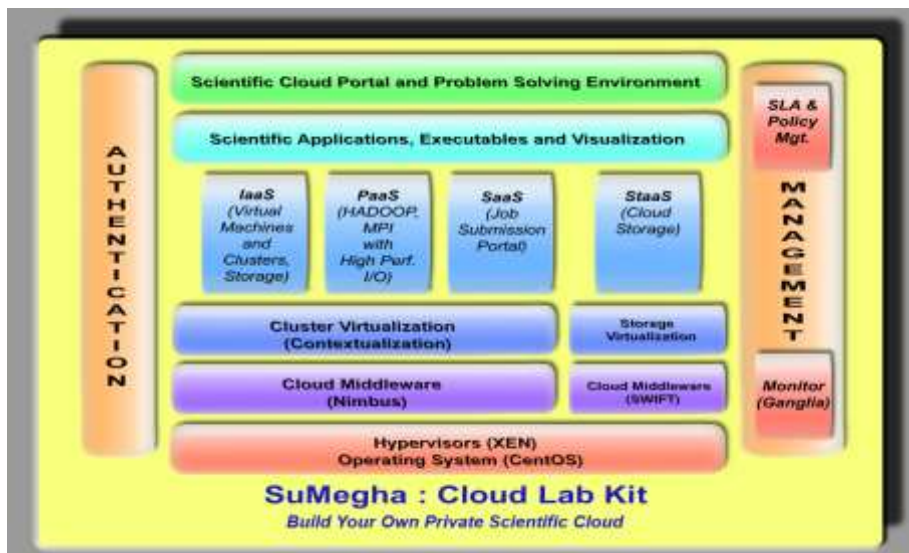

Figure 3: Architecture of SuMegha Lab kit

4.1 Hypervisor:

XEN[13], which is a Hypervisor using a microkernel design, providing services that allow more than one Operating Systems to execute on the same computer hardware concurrently, is used in SuMegha. Hypervisor basically virtualizes the physical resources to be shared among various operating systems.

4.2 Cloud middleware

Cloud middleware is an abstraction layer that shields system complexity and ensures effective communication mechanism between the cloud computing components. Nimbus, which contains the required software to manage cloud, is used as a cloud middleware in SuMegha. Nimbus[6,7] middleware consists of an integrated set of tools that deliver the power and versatility of infrastructure clouds to scientific users. Nimbus is preferred middleware because of the inbuilt support for virtual clusters. Being based on Linux operating system it will help in larger and faster adoption of the cloud.

4.3 Cluster virtualization:

Nimbus contextualization scripts are used to create virtual clusters using virtual machines. Contextualization will establish a connection between nodes of virtual cluster and provide each node the required information about other nodes of the cluster.

4.4 Storage virtualization:

OpenStack[14] object storage component called Swift is used for storage virtualization. Physical storage will get virtualized to provide isolated storage for every virtual machine or cluster.

4.5 IaaS:

Lab kit provides Infrastructure as a service [15] which refers to services that abstract user from the details of infrastructures like hardware resources, location, data partitioning, scaling, security, backup etc.

4.6 PaaS:

Platform as a service [18] offers a development environment to application developers. Application developers can develop and execute their applications on a cloud platform without the cost and complexity of buying and using the underlying hardware & software layers.

4.7 SaaS:

Software as a service (SaaS) model provides users access to application software and databases. Lab kit manages the infrastructure and platforms that run the applications. SaaS is more often referred to as on-demand software.

4.8 StaaS:

Storage as a Service[18] provides on-demand virtual storage to the cloud users. Each user/VM gets their own virtual storage which is isolated from others.

4.9 Scientific cloud portal:

Scientific cloud portal is a web interface which comes along with lab kit. Cloud portal allows users to create, delete and manage Virtual Machines and clusters from web interface.

4.10 Authentication:

Lab kit provides SSH Authentication for the services. Only authenticated users can access the resources/services.

PoS(ISGC 2016)024

5. Features

The cloud kit provides many features to the users to launch applications on the cloud by creating HPC virtual clusters using browser and managing the created virtual resources. Some of the salient features are as follows:

a. **Automated deployment software** for installation and configuration of private cloud on the recommended hardware to set up cloud lab with help from network administrator. This package contains virtualization software's, cloud middleware, and libraries to support virtualization

b. **Golden images:** Virtual machine images with CentOS are provided to create virtual machines

c. **Parallel programming environment:** Users can create MPI clusters which support scientific computing applications with MPI libraries

d. **Hadoop data intensive applications:** Users can create Hadoop clusters to run data intensive applications on Hadoop framework

e. **Creation of Virtual Machines/Clusters:** Users can create Virtual Machines/Clusters with single click by providing usage time and size of the image (CPUs and RAM).

f. **CDAC in house developed tools:** Along with lab kit, we are also providing the web interface cloud portal, job submission portal to MPI clusters and Problem Solving Environments (PSEs)

g. **Problem solving environments like SFM and NGS:** We are providing the problem solving environments such as SFM and NGS.

6. Types of virtual machines

User need to choose the image type, size and duration for which the instances / clusters are required. Cumulus image repository maintains the images uploaded by administrator, which contain preconfigured operating system software. The images we support are of raw type. After selection of image type, user needs to select the size (no. of vCPUs and RAM). The various size descriptions available are given in Table 1. The list of images will be displayed in the select box available in Instances page of cloud portal. In case of cluster creation user needs to mention number of nodes.

Table 1. Types of Virtual Machines

Images sizes	Description
Small	1 vCPU & 1GB RAM
Medium	2vCPU & 2GB RAM
Large	4vCPU & 4GB RAM

7. Distribution & installation

Distribution: SuMegha Cloud Lab Kit distribution is through DVD and its contents include the following:

- Installer scripts to set up private cloud i.e. deployment of Infrastructure as a Service – The users must give appropriate inputs when prompted, during the installation.

- Hypervisor

- Cloud Middleware

- Cloud portal

- Golden images with CentOS, OpenMP, MPI & HADOOP environments

- Preloaded HPC applications

- Manuals – Installation Manual and User Manual

Installation: This section deals with installation procedure for both VMM Node and SN in order to set up private cloud. It lists the pre-requisites for the installations, and pre and post configurations for launching portal through which the virtual machines or virtual clusters can be created and managed.

Table 2. Virtual Machine Manager Node

Operating System	CentOS Version > 6.2
CPU	One or more 64-bit x86 CPU(s), 1.5 GHz or above, 2 GHz or faster multi-core CPU recommended
RAM	Minimum 4 GB
Disk Space	Minimum 60 GB; Minimum 2GB for /boot partition
Network	Internet Connectivity

It is assumed that the administrator who performs the installation possesses the knowledge of basic UNIX system administration. He should be aware of machine details like Internet Protocol (IP) Address, NetMask, Gateway Address, Domain Name Server (DNS), Interface name (for creating bridge), etc. As SuMegha Cloud Lab Kit has been extensively tested on CentOS Version 6.4, it may work with higher versions of CentOS as well.

Table 3. Service Node

Operating System	CentOS Version > 6.2
RAM	4 GB
Disk Space	Minimum 100 GB
Network	Internet Connectivity
Software	Oracle JAVA 1.7, Python (2.6 – 3.0)

The requirements to set up the VMM node and SN as part of SuMegha Cloud Lab Kit installation are mentioned in Tables 2 and 3.

Just executing one script, the installation of SuMegha starts, first the user is required choose the option to install VMM node installation and then installation of Service Node. The first component of SuMegha Cloud Lab kit, the installation of Xen starts. The user is prompted whether to create a default Network Bridge "xenbr0" or not. A network bridge is a network device which connects more than one network segments Ensure that the pre-requisites are met.

Figure 4: Xenbr0 installation steps

After the network bridge is configured as shown in Figure 4, the installation of Xen, its relative packages and libvirt begins. Similarly all the packages will be installed and suitable network will be configured with minimal user intervention. The user can start using the creation of instances from scientific cloud portal as mentioned in the below section.

8.   Scientific Cloud Portal

Scientific Cloud Portal provides the facility to create Infrastructure as a Service by generating virtual machines and virtual clusters for development of cloud based applications. It also manages the VM resources for saving the images and destroying the VM and virtual cluster as per user convenience. Figure 5 provides the details of the Virtual Resources information in this portal.

Figure 5: Scientific cloud Portal – Running instances and VM Creation

The figure 5 also explains the virtual machines and virtual cluster creation from the browser information of repository images which are stored in the cloud using Cumulus package.

Figure 6 provides the details of the available information of repository images which are stored in the cloud using Cumulus package.

Figure 6: Scientific cloud Portal – Repository Images

The lab kit has several other packages such as job submission portal, storage management and PSE for specific applications. The following is the list of tools which will be used for the application development on cloud.

8.1 Cloud Job Submission Portal

This portal provides a web interface for submitting sequential and parallel jobs, i.e. HPC specific applications, to cloud computing environment which is offered as IaasS. It also supports monitoring the job status and output file management. This portal is specifically designed for the HPC users of the cloud to avail the virtual clusters and run their jobs without getting into the difficulties of typing commands on command line.

8.2 Scientific Cloud -PSE-SFM

PSE for SFM [15, 20] is a web-interface to interact with the virtual clusters provided by scientific cloud to run the SFM application for weather prediction.

11

This portal basically focuses on submission of SFM related inputs to the SFM application hosted on the virtual clusters, which processes the inputs. User can monitor the status of the job and also can view the outputs/ error of the submitted jobs. Web based file explorer is provided to view the files and the directories inside the user's home. The user can visualize the images based on his/ her requirements by suitable commands.

8.3   Scientific Cloud - PSE for NGS

Next Generation Sequencing (NGS) [16, 19] is used to analyze and process the data produced as a result of genome sequencing. Generally, the datasets produced are huge and require huge computation power and other resources. NGS data is for analyzing and processing which needs a work-flow where the results of one step need to be pipelined to the next step for further processing. Online NGS Tool is a web-based pipeline for genome sequence analysis, which is hosted on cloud and is provided to the users via internet.

Online NGS Tool works on MPI-enabled virtual clusters to provide maximum computation using parallel approaches provides storage for huge sequenced files and comes integrated with tools for pre-processing, mapping/aligning, and manipulating sequenced datasets.

9.   Challenges faced

The authors faced several challenges during installation, configuration, launching web applications from the VM and usage of SuMegha. During installation, while configuring Network Bridge, we need to provide the relevant interface name and it should exactly match the currently available interface; otherwise the entire system will hang.

It is necessary to provide the DHCP (Dynamic Host Configuration Protocol) server with the correct set of free IP (Internet Protocol) addresses. It is the DHCP server which will allocate IPs for the virtual machines and virtual clusters. If the correct set of IP addresses is not given, the virtual machines will not be accessible. Moreover, sufficient care should be taken to ensure that the free IPs are accessible over the local network; otherwise, the virtual machines will be created with IP addresses, but will be accessible only by virt-manager (virtual machine manager) commands.

Depending on requirement of the application, the virtual cluster or virtual machine requires the enhancement of either disk space or shared file system. Configuring the extra space was not a simple task.

For the Virtual cluster creation there is a need to assign a domain name for the Server node. We also faced problems while creating master and slave images for MPI cluster creation. During cluster creation, nimbus will create virtual machines. It will then execute contextualization scripts. During contextualization, master and slave will hang if any of the following is not done properly: Giving proper mpich location, doing proper nfs configuration, resolving of service node hostname (domain name) in network, stopping iptables and mounting of file system between master and slave. Just before saving virtual images, the existing ".ssh" directory should be removed, since otherwise the contextualization will not happen properly.

We analyzed and fixed these challenges by making necessary modifications in the scripting and terminating unwanted processes.

10. Conclusion And Future Work

Cloud Computing is becoming all pervasive. Many scientists predict that Cloud Computing may replace the traditional 'own and use' paradigm of computing. The SuMegha Cloud Lab Kit for setting up a private cloud has been discussed in this paper. It presents the need for cloud lab kit, and its features, architecture and components. SuMegha supports development and execution of very simple sequential applications, as well as complicated OpenMP/MPI/Hadoop applications. The authors feel that, such an environment can accelerate the adoption of Cloud computing, since it provides an easy-to-use platform for application development in the cloud. SuMegha is designed to support a wide variety of open source components, so that it will be useful to achieve different objectives of the users. We plan to make this lab kit available to an even larger audience by providing support for OpenStack and more hypervisors including KVM in the next version.

References

[1] Aminatul Solehah Idris, et al. *The Readiness Of Cloud Computing: A Case Study In Politeknik Sultan Salahuddin Abdul Aziz Shah, Shah Alam*- Paper published at International Conference on Computational Science and Technology – 2014 (ICCST'14) Pages: 1 – 5

[2] Yanuarizki Amanatulla, et al. *Toward Cloud Computing Reference Architecture: Cloud Service Management Perspective* - Paper published at International Conference on ICT for Smart Society (ICISS), 2013 Pages: 1 – 4

[3] Vineeth Simon Arackal, et al. SciInterface: *A Web-Based Job Submission Mechanism for Scientific Cloud Computing* Published at Second IEEE International Conference on Cloud Computing in Emerging Markets (IEEE CCEM 2013), October 2013, Bangalore

[4] Amazon Web services : *High Performance Computing* http://aws.amazon.com/hpc/

[5] Shivay Veer Sharma, et al. *Accessing E-Infrastructures using CDAC Scientific Cloud (CSC) Services*, Published at IEEE Cloud Computing for Emerging Markets,16-19th Oct 2013

[6] Chaker El Amrani, et al. *A Comparative Study of Cloud Computing middleware* - Paper published at 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing Pages: 690 – 693

[7] The Nimbus Toolkit: www.nimbusproject.org

[8] Struts Lab Open Source Cloud computing solution http://stratuslab.eu/

[9] John Bresnahan, et al. *Cumulus: An Open Source Storage Cloud for Science* http://www.nimbusproject.org/files/bresnahan_sciencecloud2011.pdf

[10] Roy Campbell, et al. *Open CirrusTM Cloud Computing Testbed: Federated Data Centers for Open*

[11] Java based open source middleware for real time big data processing and analytics, http://www.gridgain.com/

[12] Hadoop *Open-source software framework for distributed storage and distributed processing of very large data sets* https://en.wikipedia.org/wiki/Apache_Hadoop

[13] Citrix Xenserver, http://www.citrix.com/English/ps2/products/product.asp?contentID=683148

[14] Openstack swift documentation http://docs.openstack.org/developer/swift/

[15] Deepanshu S, et al. *Problem Solving Environment for Seasonal Forecast Model on CDAC Scientific Cloud* Second International Conference on Advances in Cloud Computing (ACC 2013), September 2013, Bangalore

[16] Aman A, et al. *Cloud NGS Tool: An online pipeline for NGS datasets* Second International Conference on Advances in Cloud Computing (ACC 2013), September 2013,Bangalore

[17] Payal Saluja, et al. *CDAC Scientific Cloud: On demand provisioning of HPC resources for Scientific Applications*- Paper published and presented at 18th International conference of parallel and distributed processing techniques and applications(PDPTA 2012) 6th-19th July 2012 at Las Vegas, Nevada, USA

[18] Services Models IaaS, PaaS and StaaS https://en.m.wikipedia.org/wiki/Cloud_computing#Cloud_Provider_Interface

[19] C.L.Biji, et al. *NGS read data compression using parallel computing algorithm* - Paper published at International Conference on Bioinformatics and Biomedicine (BIBM), 2015 Pages: 1456 – 1460

[20] Ramesh Naidu Laveti, et al. *Seasonal Forecast Modeling application on GARUDA Grid Infrastructure* at the International Symposium for Grids and Clouds (ISGC-2012), Academia Sinica, Taipei, 26th Feb - 2nd March, 2012.

PoS(ISGC 2016)024