An OpenNebula-based cloud computing environment for bioinformatics

Peter van Heusden, Long Yi and Alan Christoffels
South African National Bioinformatics Institute
University of the Western Cape, Private Bag X17, Bellville, 7535
Tel: +27 21 959 2356, Fax: +27 21 959 2512
Email: pvh@sanbi.ac.za, long@sanbi.ac.za and alan@sanbi.ac.za
Corresponding author: Peter van Heusden, pvh@sanbi.ac.za

***Abstract*—Bioinformatics is the discipline of solving problems in biology and medicine using computational resources (databases, analysis programs, etc). The field is characterised by rapid growth in data to be analysed and rapid evolution of analysis methods. The South African National Bioinformatics Institute (SANBI) provides post-graduate training in bioinformatics and hosts approximately 30 students. A cloud computing environment provides the opportunity to present SANBI students and researchers with virtual machines customised for different projects and users. In 2011 we started investigating cloud computing solutions and OpenNebula was chosen as a private cloud solution due to its low cost and its ability to leverage SANBI's Dell blade server infrastructure effectively. SANBI now hosts an OpenNebula environment spread over three blade servers with a total of 36 cores and 224 GB of RAM. The SANBI OpenNebula cloud hosts 15 virtual machines with workloads ranging from web server to high performance computing nodes. Typical virtual machines are based on an Ubuntu Linux image that integrates with our central configuration management server and allows for the creation of a new server in less than 30 minutes.**

***Index Terms*—cloud computing, OpenNebula, bioinformatics, private cloud, open source**

I. Introduction

The South African National Bioinformatics Institute (SANBI) was founded in 1997 to provide a site for research and postgraduate training in bioinformatics. Research necessarily involves the construction of a large number of different computing environments to meet the needs of the faculty and students at the Institute. In the past the purchase of computing resources for each project led to fragmentation of computing resources coupled with sub-optimal use of the computing needs of research groups such that there were periods of high intensity alternating with idle periods as results are examined and publications prepared.

Since 2009 SANBI has been deploying virtual machines, using a combination of the Xen and KVM hypervisors, to meet the needs of its research groups. These virtual machines were typically deployed ad-hoc and their life cycle (creation, running and termination) was manually managed. As the number of virtual machines in use increased it became apparent that the virtual machine infrastructure was high maintainance and fragile, and in 2011 SANBI initiated a collaboration with the University of the Western Cape Computer Science department to investigate cloud computing solutions. Cloud computing involves a group of machines configured in such a way that an end-user can request any number of virtual machines (VMs), and the cloud will spawn these VMs somewhere on the physical machines that it owns. The end-user is insulated from the details of the physical hardware on which their VM resides. Sempolinski et al [1] suggest, "[t]his kind of setup is ideal for applications where a specific hardware configuration is needed or users only occasionally need the high compute capacity."

After evaluating two open source cloud computing solutions, OpenStack[2] and OpenNebula[3], OpenNebula was chosen to manage SANBI's "private cloud". OpenNebula is deployed on three Dell M710HD blade servers, each having 12 CPU cores (from dual Intel E5649 CPUs) and between 64GB and 96GB of RAM. The cloud management software is installed on a virtual machine hosted on a Dell R710 rack server. The cloud is in an early stage of production and is hosting 15 VMs.

II. Cloud Architecture

The OpenNebula cloud management system consists of a central management node that runs the cloud controller (**oned**) and a number of cloud nodes that each run a supported hypervisor. OpenNebula interfaces to VM hypervisors using either **libvirt**, a Linux library that provides an abstract VM management interface, or the Amazon EC2 interface. At SANBI cloud nodes run CentOS 6.2 with the KVM hypervisor. No OpenNebula-specific software needs to be installed on the cloud nodes. Cloud nodes store their VM images on a SAN, accessed via iSCSI over a 10 gigabit Ethernet network.

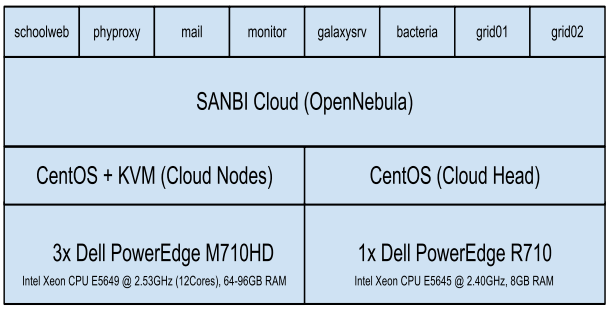

Fig. 1. The SANBI cloud infrastructure, virtual to physical layers.

The cloud controller (or head node) also supervises a collection of VM images, essentially operating system images that are instantiated to create VM instances. A set of command line tools and text configuration files control the operations of the OpenNebula cloud controller, and allow the

cloud operator to create new VM instances by combining a network specification, a resource specification (disk space, CPU and RAM allocation) and an image name. Commands also exist to gather an inventory of running VM instances, available operating system images, and virtual networks.

When a VM is created, OpenNebula creates the VM image on the cloud controller and then transfers it, typically using **ssh**, to an appropriate cloud node. The initial transfer of a new VM image can be somewhat time consuming since the deployment model does not support thin provisioning of virtual disk. Once the VM is instantiated it can at a later stage be migrated to a different cloud node, a useful feature if the physical machine supporting the cloud node requires maintenance.

SANBI virtual machine images have been customised to integrate with out central software provisioning system. On boot the newly created VMs register with our Puppet server and apply any configuration that is defined there. We have also automated the steps of VM provision in the form of a Python script that calls OpenNebula commands, with the result that a "standard" Ubuntu 12.04 VM can be created on the SANBI cloud very rapidly.

III. Applications

The SANBI cloud needs to support a combination of high performance computing, infrastructure services (mail, authorisation and file servers) and web-based application servers. The blade servers that support the cloud environment are running in a Dell PowerEdge M1000e enclosure and disk is provided by an Equallogic PS6010 Storage Area Network, a configuration that ensures high availability of the underlying physical server infrastructure. As a result we have virtualised most of our mission critical server infrastructure such as the Internet firewall, the Kerberos authentication server and the LDAP directory server.

The SANBI cloud also hosts a number of nodes for our compute grid. The compute grid is managed by a Sun Grid Engine scheduler and makes use of a combination of physical and virtual compute nodes. Bioinformatics computing problems tend to be CPU-bound and it has been shown [4] that the KVM hypervisor allows performance nearly identical to physical hardware for such workloads. Locating compute grid nodes in the cloud allows for dynamic provisioning of compute grid resources as the Institute's computing needs shift.

In addition to infrastructure servers and the high performance computing nodes, the SANBI cloud provides an environment for web-based application servers. Research results from the work of SANBI students and faculty are typically presented to the world through custom built web applications. Using VMs allows each research group to be given their own working environment where they can install and customise software for web servers, applications and databases as they see fit. In other words, VMs in the SANBI cloud enable personalised computing.

IV. Conclusion

Our research in cloud computing has led to the implementation of a OpenNebula-based compute cloud that is in use to provide a production environment for bioinformatics. Cloud computing hosting at the Institute has allowed for the advantages of a cloud environment, that is scalable, personalised computing, without the network overhead involved in using public computing clouds (such as Amazon EC2). An open source solution was chosen for reasons of cost and tighter integration with the existing open source software in use at SANBI.

As the volume of available biological data grows at a faster than exponential rate, we anticipate increasing our use of cloud computing technology to enable rapid deployment of new computing infrastructure. Towards this end we are investigating using cloud interface APIs to enable cloud based tools that can address both the SANBI private cloud and public cloud services (such as Amazon EC2 and possibly future South African scientific computing clouds) as demand for computing capacity scales up. We are also collaborating with the SAGrid project to investigate convergence between cloud computing and grid computing when it comes to distributed access to computing and data storage resources. Finally we continue to investigate OpenStack as an alternative cloud computing platform.

References

[1] P. Sempolinski and D. Thain, "A comparison and critique of eucalyptus, OpenNebula and nimbus," in *Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on*. Ieee, 2010, pp. 417–426.

[2] Rackspace Cloud Computing, "OpenStack open source cloud computing software," http://openstack.org/, 2012. [Online]. Available: http://openstack.org/

[3] OpenNebula Project, "OpenNebula: the open source solution for data center virtualization," http://opennebula.org/about:about, 2012. [Online]. Available: http://opennebula.org/about:about

[4] A. Younge, R. Henschel, J. Brown, G. von Laszewski, J. Qiu, and G. Fox, "Analysis of virtualization technologies for high performance computing environments," in *Cloud Computing (CLOUD), 2011 IEEE International Conference on*, Jul. 2011, pp. 9 –16.

Peter van Heusden is manager of computer systems at SANBI. He received a BSc in Computer Science from the University of Cape Town in 1993 and is presently completing his Honours in Computer Science with Unisa. His research interests include scientific workflow systems and bioinformatics development.

Long Yi received a MSc (cum laude) in Computer Science from the University of the Western Cape in 2007. He is a system developer at SANBI. His research interests include cloud computing and virtualisation.

Alan Christoffels is the DST/NRF Research Chair in Bioinformatics and Public health genomics and director of SANBI.