DOI: 10.14529/jsfi150302

InfiniCloud: Leveraging the Global InfiniCortex Fabric and OpenStack Cloud for Borderless High Performance Computing of Genomic Data

Kenneth Ban[1,2]***, Jakub Chrzeszczyk***[3]***, Andrew Howard***[3]***, Dongyang Li***[3]***, Tin Wee Tan***[2,4]

InfiniCloud is a geographically distributed, high performance InfiniBand HPC Cloud which aims to enable borderless processing of genomic data as the part of the InfiniCortex project. This paper provides a high-level technical overview of the architecture of InfiniCloud and how it can be used for high performance computation of genomic data in geographically distant sites by encapsulation of workflows/applications in Virtual Machines (VM) coupled with on-the-fly configuration of clusters and high speed transfer of data via long range InfiniBand.

Keywords: Genomics, Cloud-Computing, InfiniBand, Trans-continental, Virtualization, SR-IOV, OpenStack, HPC.

Introduction

The advent of big data has driven the need for flexible high performance computing platforms in order to analyze large amounts of data using user defined reproducible workflows, particularly in the emerging field of genomics and healthcare informatics. These workflows typically require a specific stack of applications with their operating system specific dependencies, which can be different for each pipeline and can frequently change over time as updates are released. In addition to the heterogenous nature of applications, such workflows demand high CPU performance paired with large memory capability as well as a high-performance interconnect for analysis of large genomic/healthcare datasets [9, 11].

In response to this growing need for high performance and flexible computing for analysis of large datasets [8], A*CRC and NCI teams collaborated to define a new cloud computing platform called InfiniCloud, which combines high performance cloud computing powered by OpenStack [6] with the high speed InfiniBand network architecture. This platform was optimized to provide high performance computing with minimal overhead within virtual instances, coupled with native InfiniBand protocol to provide high speed interconnect and transfer of data between the instances.

In cloud computing, resources are presented in a form of virtual machines (VMs). VMs are an abstraction layer which allows hardware resources of a physical system to be presented as number self-contained pools of virtual CPU cores, RAM, storage and network bandwidth that are used to run isolated operating system instances. These resources can be dedicated or shared, depending on performance requirements of the applications running in the cloud environment. The operating image can be created, customized, and versioned by users to ensure that the computing environment is reproducible and flexible. This is particularly important in the field of genomics, where processing pipelines are highly interconnected and can be dependent on a specific version of operating system, kernel, libraries and application binaries. This level of

[1]Institute of Molecular and Cell Biology, A*STAR, Singapore
[2]Dept. of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[3]National Computational Infrastructure - The Australian National University, Canberra, Australia
[4]A*STAR Computational Resource Centre (A*CRC), A*STAR, Singapore

flexibility is typically difficult to achieve on a traditional High Performance Computing cluster running multi-purpose system images.

Despite its advantages for reproducible and flexible computing, one major bottleneck in traditional cloud computing platforms is the inefficient and slow transfer of large datasets, commonly encountered in genomic analysis. To address this, we extended the InfiniCloud platform to address the need for efficient high speed transfers by leveraging on the long range Obsidian Longbow E100 InfiniBand extenders, enabling unprecedented high speed transfer of large datasets and VM images across trans-Pacific distances between two geographically distant InfiniCloud platforms in Singapore and Canberra. This capability enables borderless high performance cloud computing by high speed transfer of large datasets together with workflows/applications encapsulated in VMs. The workflows/applications in VMs can be parallelized in virtual instances by the on-the-fly setup of cluster compute nodes, thus opening the door for scaling up reproducible computing environments beyond any one single HPC cloud computing site.

We envision that the InfiniCloud platform combined with long range InfiniBand as part of a global fabric (InfiniCortex) [14] will enable seamless distributed high performance computing amongst geographically distant InfiniCloud nodes, breaking down barriers to meet the challenge of big data computing.

1. InfiniCloud Platform

The InfiniCloud platform was developed on the NCI and A*CRC hardware and is based on OpenStack cloud computing software stack with custom modifications.

1.1. Hardware Components

Currently, InfiniCloud consists of two sites: one located at the NCI (National Computational Infrastructure), in Canberra, Australia and the second at A*CRC, Singapore (fig. 1). The total count of compute cores available is 264, supporting 3TB of memory and a local storage capacity of 15TB (SSD and HDD). All instances are connected to the shared 56Gbit FDR IB fabric.

Figure 1. InfiniCloud sites (left: NCI, Canberra, right: A*CRC, Singapore)

1.1.1. Server Specifications

The overall design of each site is similar, utilizing a common InfiniBand interconnect. The server configurations are detailed in tab. 1 and tab. 2.

Table 1. NCI InfiniCloud configuration

Servers	10x Fujitsu PRIMERGY CX250
CPU	Intel Xeon E5-2650
Memory	256GB
Interconnect	Mellanox FDR
Local storage	1x Intel DCS3500 or 3x Intel DCS3500

Table 2. A*CRC InfiniCloud hardware configuration

Servers	6x SGI C1104-GP1
CPU	Xeon E5-2680
Memory	128GB
Interconnect	Mellanox FDR
Local storage	3x Intel DCS3500 or Micron M600

1.1.2. Local Area Network Components (each site)

The core of InfiniCloud is a global InfiniBand interconnect, which consists of a local FDR switch at each site to connect the local compute nodes, combined with an Obsidian Strategics Longbow E100 range extender connecting the AU and SG InfiniCloud network fabrics (tab. 3 and fig. 2).

Table 3. Network configuration

Switching	FDR IB
Range extender	Obsidian Strategics Longbow E100
Subnet manager	OpenSM (active:AU; standby: SG)

1.1.3. Global Area Network

To enable the global InfiniBand connection, the A*CRC and NCI teams worked closely with AARNet (AU), SingAREN (SG) and Pacific NorthWest GigaPop (PNWGP) in Seattle (USA) to secure a dedicated 10Gbit/s layer 2 link between Canberra and Singapore using spare fibre capacity. Due to the network topology connecting Australia (with the majority of the bandwidth provided to the more densely populated East Coast of Australia), the link was routed via the longer eastern path, crossing the Pacific Ocean twice through PNWGP in Seattle with an RTT of 305ms. In contrast, while exhibiting better delay characteristics more direct western path through Western Australia, Indian Ocean and Guam has limited capacity and is only capable of providing a 1Gbit/s connection (fig. 3).

1.2. InfiniCloud Installation and Configuration

All InfiniCloud systems run the following operating system, drivers and applications stack (tab. 4). Clusters consist of one dedicated management node, one dedicated controller node and

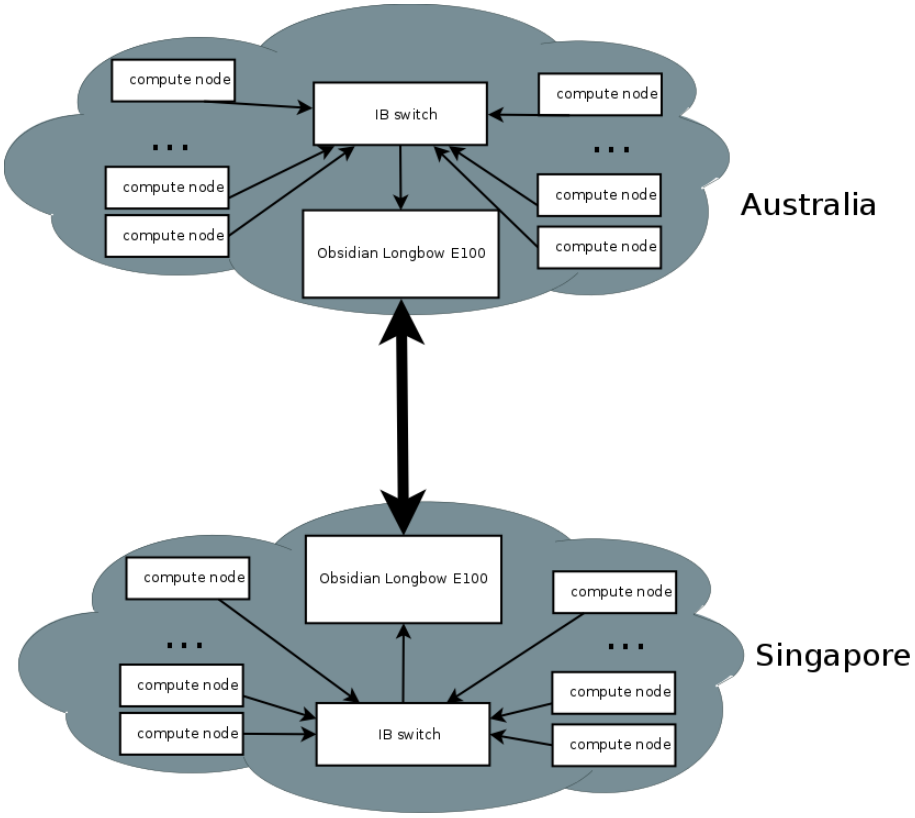

Figure 2. InfiniCloud Network Topology

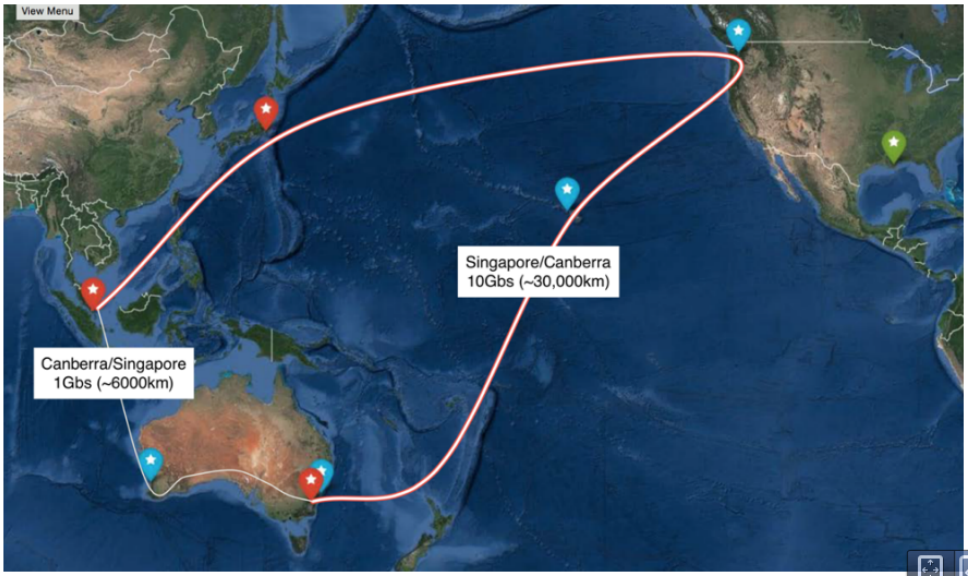

Figure 3. InfiniCloud wide-area networking

Table 4. InfiniCloud software stack

Operating System	CentOS 6.6 x86_64
InfiniBand drivers	Mellanox OFED 2.4
OpenStack version	Icehouse + InfiniCloud specific patches

a variable number of compute nodes (ranging from 4-8). All node classes are integrated to form a fully featured HPC Cloud.

The management node is used for bare metal provisioning and cluster-wide configuration. The controller node provides API, CLI and GUI access to the cloud and is responsible for managing all the core areas of cluster operation: identity management, scheduling, VM image storage, network management and providing an orchestration layer. Compute nodes provide CPU, RAM, storage and high performance SR-IOV networking [12] to the virtual instances. SR-IOV networking support is a requirement for enabling InfiniBand capability in virtual instances.

Building the InfiniCloud cluster required a high degree of customization in order to enable native InfiniBand capability in virtual instances, as well as to provide access to the global InfiniBand network connecting Australia and Singapore. Tab. 5 and fig. 4 show the list of these modifications: (i) A custom virtual interface module adds support for SR-IOV virtual function networking in the nova-compute component; (ii) an embedded switch module implements linking virtual functions to guests and enforces network access restrictions; and (iii) a custom DHCP server package adds InfiniBand support. On top of this, OpenStack out-of-tree patches were necessary in order to force the use of a single partition key, as required by the global InfiniBand fabric. After installing the additional modules and patches, compute nodes are configured to directly connect the HCA to the upstream network, bypassing the layer 2 agent traditionally present on OpenStack compute nodes, as this functionality is now provided by the embedded switch.

Table 5. InfiniCloud OpenStack Customizations

Neutron Server	enable SR-IOV and native IB capability
Neutron Networker	enable EoIPoIB support
Nova Compute	enable SR-IOV and native IB capability
Neutron Agent	enable SR-IOV and native IB capability
DHCP	enable IPoIB support
eswitchd	enable InfiniCortex global IB connectivity

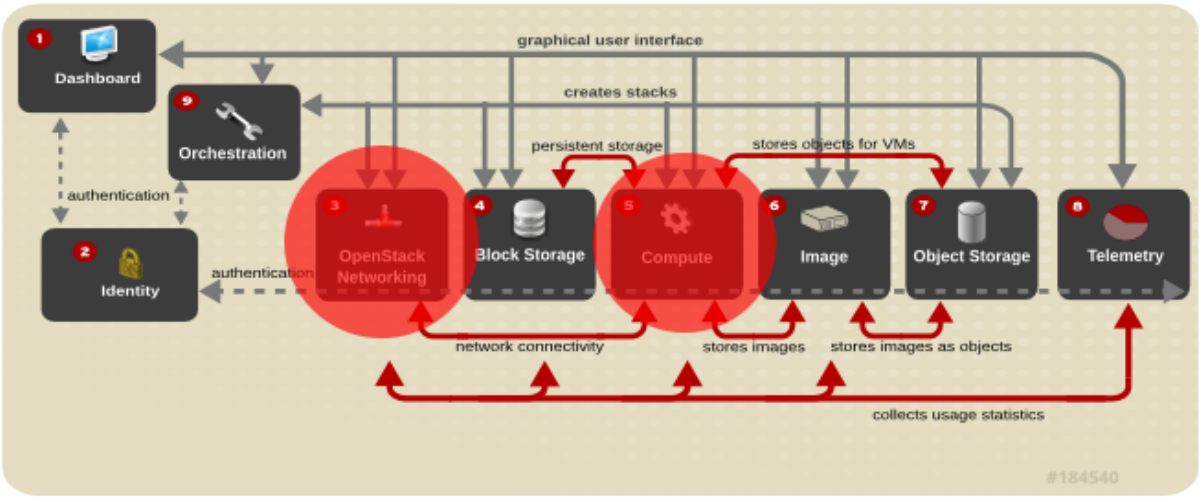

Figure 4. Overview of OpenStack components with customizations highlighted in red. Image adapted from access.redhat.com

2. InfiniCloud InfiniBand Capabilities

After cloud provisioning is complete and all the customizations required for global InfiniBand communications are in place, the system has the ability to provide virtual instances on demand, connected over InfiniBand with full ability to communicate with remote instances using RDMA over a trans-Pacific 10Gbit/s network.

2.1. High bandwidth capability — local connectivity

A high bandwidth capability within a cluster allows for the efficient transfer of data to compute nodes. Listing 1 demonstrates high bandwidth capability (∼6 GB/sec) between 2 virtual instances, close to the line rate on the FDR interconnect:

Listing 1. Local interconnect performance between a pair of virtual machines hosted in Singapore

```
---------------------------------------------------------------------------------
                    RDMA_Write BW Test
 Dual-port       : OFF          Device         : mlx4_0
 Number of qps   : 1            Transport type : IB
 Connection type : RC           Using SRQ      : OFF
 TX depth        : 128
 CQ Moderation   : 100
 Mtu             : 2048[B]
 Link type       : IB
 Max inline data : 0[B]
 rdma_cm QPs     : OFF
 Data ex. method : Ethernet
---------------------------------------------------------------------------------
 local address: LID 0x05 QPN 0x0a5e PSN 0x90c425 RKey 0xb8011700 VAddr (...)
 remote address: LID 0x1a QPN 0x0cac PSN 0x94503d RKey 0x7001182b VAddr (...)
---------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[MB/sec]    BW average[MB/sec]   MsgRate[Mpps]
 65536      5000             5984.52             5976.36            0.095622
---------------------------------------------------------------------------------
```

2.2. High bandwidth capability — global connectivity

Integral to the data transfer component is the use of the Obsidian Strategics dsync+ utility [1] which utilizes the RDMA (Remote Direct Memory Access) capabilities to provide long range InfiniBand RDMA transfers between InfiniBand-connected virtual instances. This high performance transfer capability uncouples the need for the data to be located close to the compute nodes, enabling the computing of data to scale beyond a single site.

As a proof-of-concept test of native InfiniBand transfers over long distances, we tested the processing of a large genomic dataset [7] that could be accelerated using large memory compute resources not readily available locally. Listing 2 shows the transfer of 381 GB of genomic data in under 9 minutes from NCI (Canberra, Australia) to A*CRC (Singapore) via the 10G link going through Seattle (∼30,000 km) using the dsync+ utility. In contrast, rsync transfer using TCP/IP over the same 10G link took 3 hours [10].

Listing 2. Global interconnect performance between a pair of virtual machines hosted in Singapore and Australia

```
[root@test01 ~]# dsync --direct-io --option Xfer::RDMA::Buffer-Size=5368709120 \
--option Xfer::RDMA::IO-Block-Size=10485760 \
192.168.200.144:/scratch/kuba/reference_dset/ /scratch/kuba-test/
Finished generating remote file list. 40 files, 3 directores, 381GB.
Finished checking local files. Need to get 40 files, 381GB.
Transfer xfer-ib-rdma network usage 3050B in 0s (10.0kB/s)
Transfer xfer-ib-rdma network usage 381GB in 8m19s (764MB/s)
Done. Transferred 381GB in 8m27s (752MB/s)
```

The remarkable performance observed with long range InfiniBand RDMA provides a significant improvement (∼20 fold) over standard TCP/IP protocols.

3. Using InfiniCloud for Parallelized Workflows in Genomic Analysis

The InfiniCloud platform provides a high performance cloud computing environment for flexible workflows, coupled with unprecedented high speed transfer of big data sets over large geographical distances. A key application that takes advantage of these high performance characteristics is the analysis of genomic sequences which has seen an exponential growth in demand with the advent of next generation sequencing technologies.

The rapid development of next generation of sequencing technologies has dramatically reduced the cost of sequencing genomes [13]. Previously, it took ∼USD $2.7 billion and 10 years to sequence one human genome, but currently the cost has dropped several orders of magnitude to ∼USD $1,000 per genome with the introduction of platforms such as the Illumina HiSeq X sequencer. This drop in cost coupled with the ability to sequence a complete human genome in a few days has driven the adoption of genomic sequencing in research labs as well as hospitals.

Although the cost and speed of sequencing has dramatically improved, the transfer and computation of the genomic data remains a bottleneck in translating that data into the insights needed for improving patient care [10]. Typically, sequencers are not colocated with the compute resources and require the transfer of data in a timely manner. For example, a single Illumina HiSeq X can sequence 32 whole human genomes a week, resulting in ∼6 TB of genomic data. Such volume of data would take over six and a half days to transfer on a dedicated 100Mbps TCP/IP network, assuming ideal conditions and 100 % efficiency. The same transfer could be completed in just under three hours, using long distance InfiniBand [10]. This high performance data transfer capability of native InfiniBand transport would provide the scalability to cope with the growth of genomic data, given the increasing adoption of genomic sequencing in clinical and research labs.

In addition, the computational analysis of genomic data for clinical use requires enforcement of reproducibility standards in addition to the data provenance and security guarantees needed to comply with ethical and legal privacy issues. A computational platform for clinical genomics needs to meet the following challenges:

- High speed data transfers from sequencing data stores to the computational platform
- Reproducible and well documented workflows that can be run on different hardware platforms

- Easy provisioning of compute clusters for processing genomic data from multiple samples using parallel workflows
- High CPU and network performance for rapid analysis of large datasets
- Mechanisms for data provenance and security (e.g. using ephemeral containers) for computation at remote sites

3.1. Provisioning of instances and on-the-fly setup of cluster compute nodes for parallel workflows

To address these challenges, we implemented a software stack on top of the InfiniCloud platform that leverages the use of VM instances or containers to encapsulate workflows, together with automated provisioning of virtual instances and the setup of virtual compute clusters for parallelized workflows.

We adapted ElastiCluster [2] for use on InfiniCloud to enable easy provisioning of instances and setup of virtual clusters for parallel workflows (fig. 5). In our configuration, ElastiCluster was used to provision instances and set up a virtual cluster consisting of a frontend node and a user-defined number of compute nodes. To enable cluster computing for parallel workflows, we configured ElastiCluster to install and setup the SGE job scheduler [5], Ganglia monitoring tools [3], and the IPython notebook interface [4]. The package versions used are listed in tab. 6.

Table 6. Cluster computing stack

IPython	Notebook shell (BASH/Python/R); IPython parallel engine	2.4.1
SGE	Grid engine job scheduler	6.2u5
Ganglia	Cluster monitoring (CPU/Memory/Network)	3.1.7

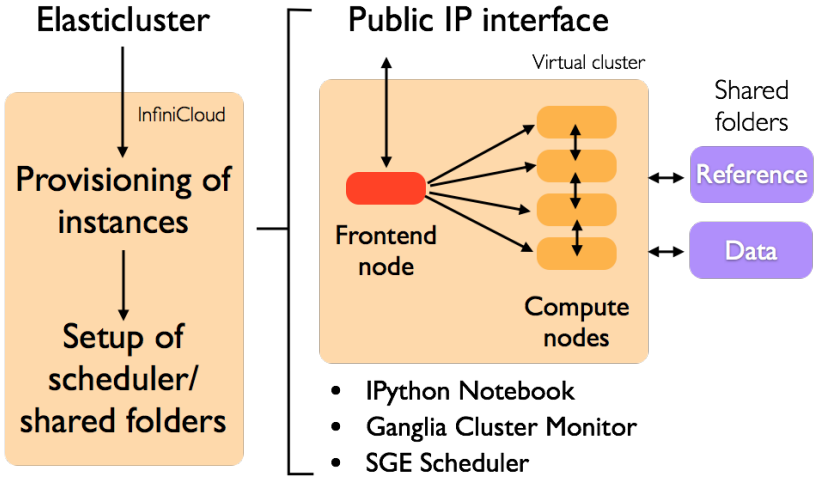

Figure 5. On-the-fly provisioning and setup of virtual clusters for parallelized workflows

In the final configuration, the setup provides SSH access, a web interface for cluster monitoring using Ganglia (fig. 6), and a versatile IPython Web Notebook interface for BASH/Python/R scripting (fig. 7).

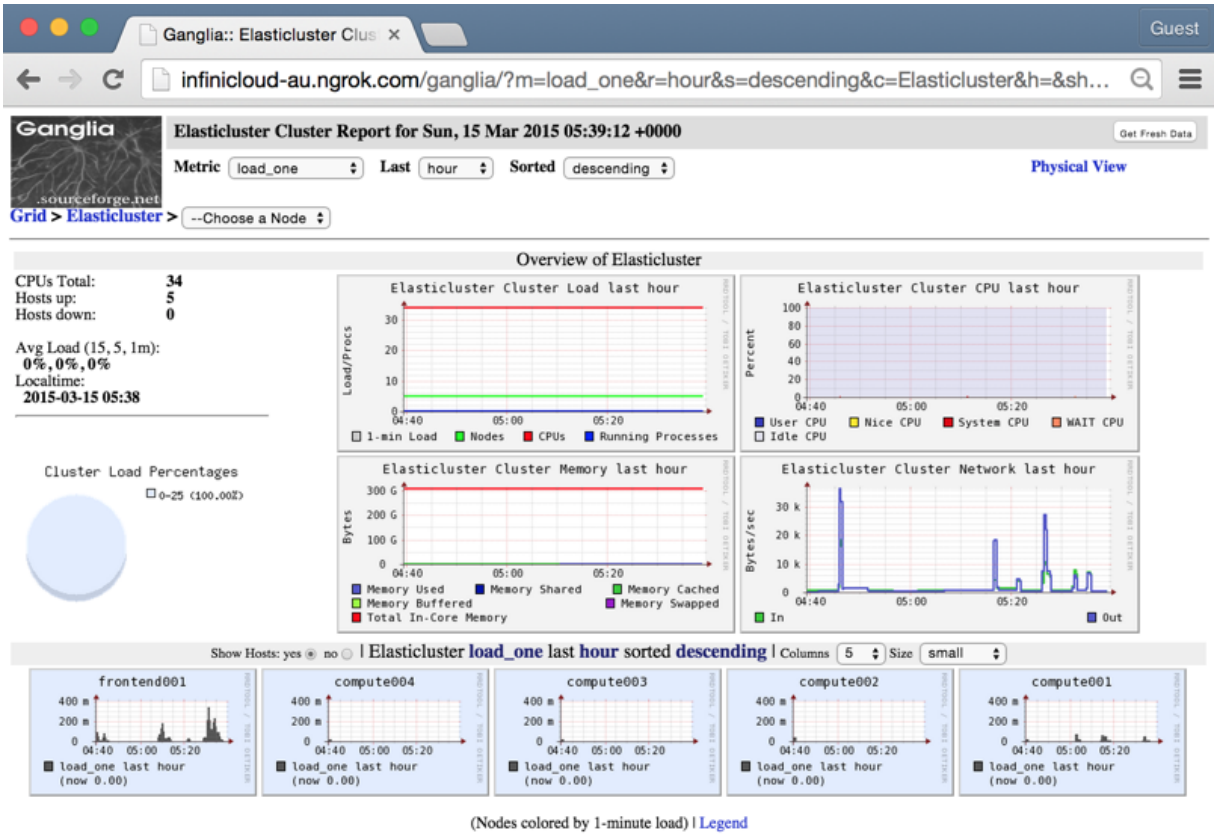

Figure 6. Ganglia cluster monitoring

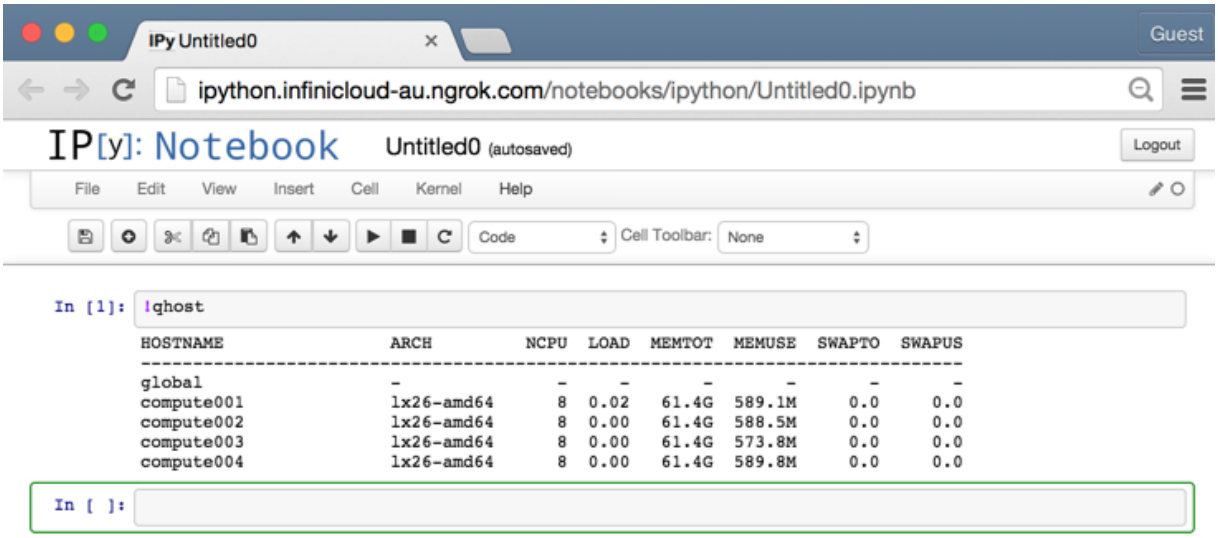

Figure 7. IPython notebook

3.2. Implementation of variant calling genome analysis pipeline

Next, we demonstrate how the on-the-fly provisioning and setup of virtual machines can be used to parallelize a genomic analysis workflow. We chose a clinically relevant workflow, called variant calling, that takes genomic sequences from cancer samples and detects mutations in genes that could be used to determine the prognosis of a patient, or to identify potential chemotherapy drugs that could be used for treatment. As each cancer sample can be analyzed separately, the workflow is amenable to simple asynchronous parallelization without any inter-process communication.

The implementation of genomic workflows typically involves several processing steps that are run using different applications that may vary in complexity of the setup dependencies. The ability to install and run them in a virtual instance allows the different applications to be set up to interoperate properly, then replicated for parallel workflows.

In this workflow, genomic sequences are processed in a pipeline through a series of steps using different applications to identify and annotate mutations (fig. 8). We use the pipeline application to orchestrate the steps in processing and to distribute the processing to the compute nodes using the SGE job scheduler:

1. Genomic sequences from each cancer sample are processed with an aligner — the application that compares the sequences to a human reference genome sequence and identifies the position and alignment of each sequence from the cancer samples.
2. The files from each cancer sample are processed by a variant caller program, which compares the aligned sequences to the human reference genome sequence to identify variations (substitutions, insertions, deletions) in the cancer samples.
3. The variant files from each cancer sample are annotated. A specialized application compares each variation to multiple databases to identify what the potential effects of each mutation have on regions in the genome.

The applications are pre-installed in the VM images together with their dependencies to enable portability between InfiniCloud platforms. The reference datasets required by the aligner, variant caller, and annotation tool are located in a data volume that can be mirrored between InfiniCloud platforms. The genomic dataset is isolated in a separate volume which also stores the results of the analysis (fig. 8); this isolation provides the flexibility for maintaining data provenance and security.

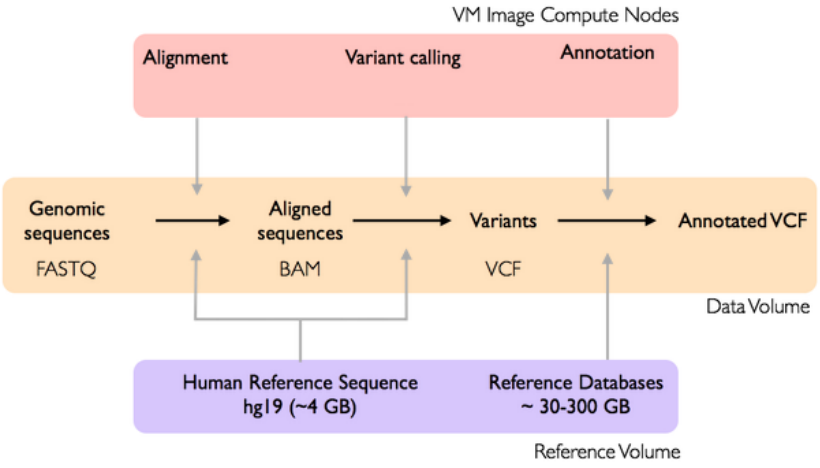

Figure 8. Workflow for variant calling of genomic data from cancer samples

3.3. Demonstration of genome analysis workflow for remote cloud computing

We demonstrate the computation of genomic sequences from multiple cancer samples on the InfiniCloud platform in Canberra, Australia from Singapore by remote provisioning of instances, setup of the cluster, and mounting of reference/data volumes (fig. 9).

- The VM images are mirrored from Singapore to Australia so that both sites have the same application/workflow backends for genomic analysis
- The common reference volume is automatically mirrored from Singapore to Australia and attached to the frontend and compute nodes
- The data volume is synchronized according to a user-defined workflow and attached to the frontend and compute nodes
- Genomic data is transferred from Singapore to Australia for computation
- Results are transferred back from Australia to Singapore
- The data volume on remote site can be deleted in cases where genomic data cannot be stored offsite for data provenance and security reasons

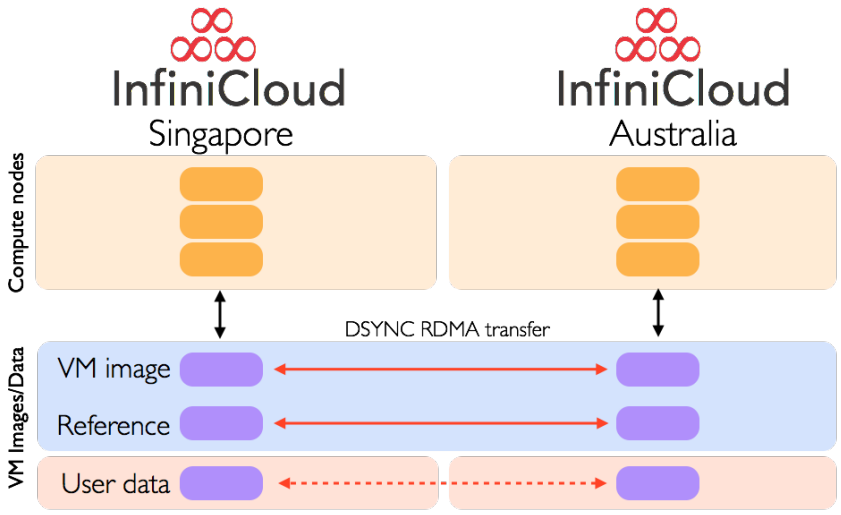

Figure 9. Borderless HPC cloud computing of genomic data across different sites for scalability

For the analysis, the genomic data is first transferred from Singapore into the data volume (Australia) using the dsync+ utility. Here, we achieve a transfer of ∼233 GB of data in 5.5 minutes (∼696 MB/sec) from Singapore to Australia via Seattle (∼30,000 km). The data from multiple cancer samples is then analyzed with the variant calling pipeline. Fig. 10 and fig. 11 show the CPU and network resource utilization during the pipeline run.

As an illustration of the output from the variant calling pipeline, fig. 12 shows mutations detected in a tumour suppressor gene (TP53) in a cancer sample which generally signifies a bad prognosis.

In summary, the high speed data transfer between InfiniCloud platforms can be used to allow scaling beyond a single site to speed up the parallel processing of data in cases where analysis is time-sensitive and/or constrained by local resources. Furthermore, the encapsulation of data and workflows within virtual machines provides one approach to maintain data provenance at the site of origin while harnessing the high performance computational resources at remote sites.

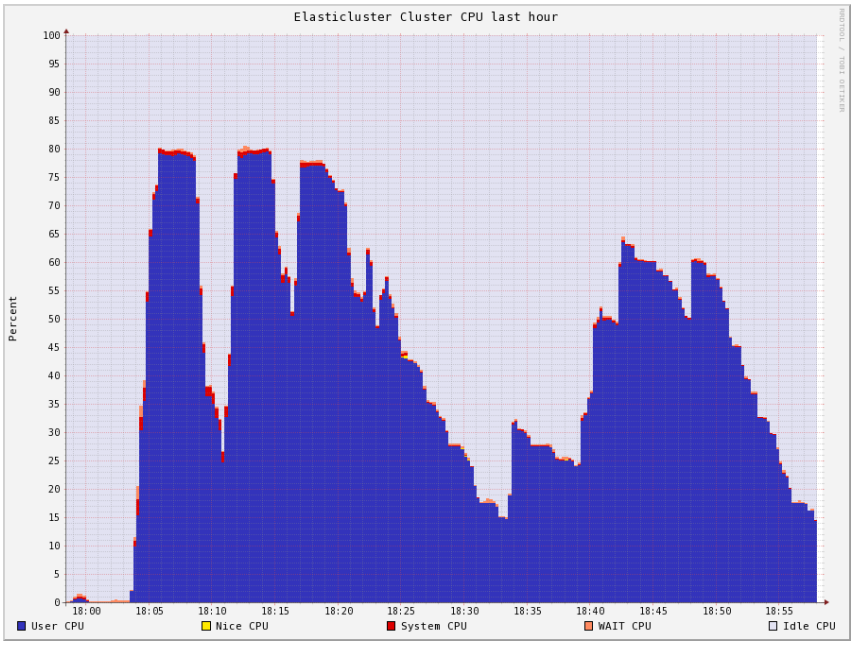

Figure 10. Aggregate CPU load

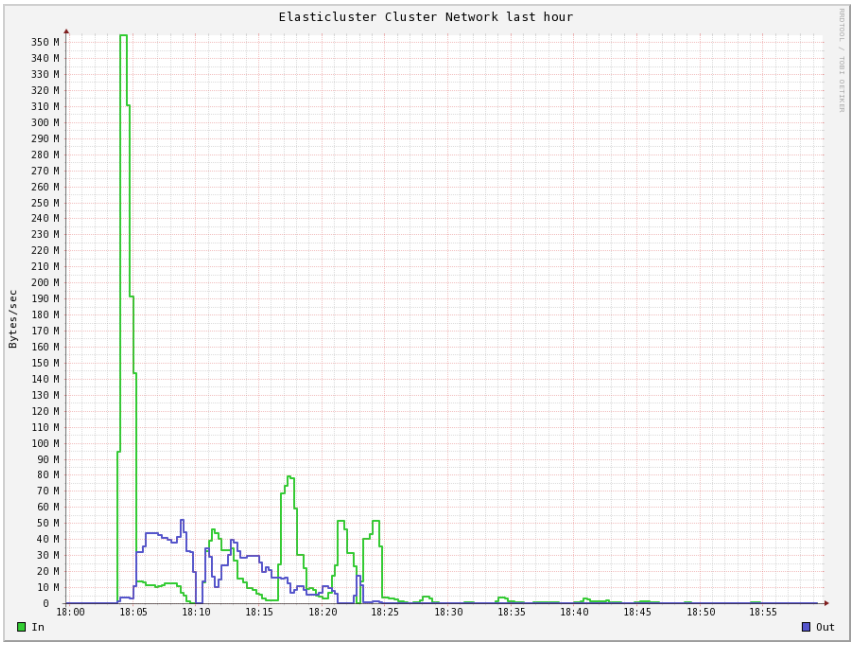

Figure 11. Aggregate network utilization

Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGen	ExonicFunc.refGene	snp138
chr17	7574025	7574025	C	-	exonic	TP53	frameshift deletion	.
chr17	7577531	7577531	G	-	exonic	TP53	frameshift deletion	.
chr17	7579472	7579472	G	C	exonic	TP53	nonsynonymous SNV	rs1042522
chr17	7579644	7579659	CCCCAGCCCTCCAGGT	-	intronic	TP53	.	rs146534833
chr17	7579801	7579801	G	C	UTR5	TP53	.	rs1642785

Figure 12. Mutations detected in the DNA sample

Conclusions

We present a new cloud computing platform called InfiniCloud, which combines high performance cloud computing powered by OpenStack [6] with the high speed/low latency of an InfiniBand network architecture. This platform delivers high performance computing with minimal overhead within virtual instances, coupled with native InfiniBand protocol for high speed interconnect transfer of data between the instances.

In addition, the InfiniCloud platform incorporates long range InfiniBand extension and enables unprecedented high speed transfers of large datasets such as genomic data and VM images across global distances. This capability enables borderless high performance cloud computing that integrates high speed transfer of large datasets together with workflows/applications encapsulated in VMs. This encapsulation allows easy parallelization of virtual instances and on-the-fly instantiation of cluster compute nodes using ElastiCluster.

We envision that the InfiniCloud platform combined with long range InfiniBand as part of the InfiniCortex global InfiniBand fabric [14], will enable seamless distributed cloud-based high performance computing amongst geographically distant InfiniCloud nodes, breaking down borders and illuminating the path to exascale computing to meet the challenge of supporting current and future big data computing needs.

This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI), which is supported by the Australian Government.

*This work was supported by the A*STAR Computational Resource Centre through the use of its high performance computing facilities.*

References

1. dsync+. `http://www.obsidianresearch.com/products/software/dsync+/index.html`.
2. ElastiCluster. `https://github.com/gc3-uzh-ch/elasticluster`.
3. Ganglia monitoring system. `http://ganglia.sourceforge.net/`.
4. IPython interactive computing. `http://ipython.org/`.
5. Open grid scheduler. `http://gridscheduler.sourceforge.net/`.
6. OpenStack cloud computing platform. `http://www.openstack.org`.
7. Georges A, Li Q, Lian J, O'Meally D, Deakin J, Wang Z, Zhang P, Fujita M, Patel HR, Holleley CE, Zhou Y, Zhang X, Matsubara K, Waters P, Graves JA, Sarre SD, and Zhang G. High-coverage sequencing and annotated assembly of the genome of the Australian dragon lizard Pogona vitticeps. *Gigascience*, 4(45), Sep 2015. DOI: 10.1186/s13742-015-0085-2.
8. Jakub Chrzeszczyk, Muhammad Atif, Joseph Antony, Dongyang Li, Matthew Sanderson, and Allan Williams. Perspectives on implementation of a high performance scientific cloud backed by a 56G high speed interconnect. HPC Advisory Council Event, Singa-

pore, http://www.hpcadvisorycouncil.com/events/2014/singapore-workshop/preso/12_ANU.pdf, November 2014.

9. Marius Hillenbrand, Viktor Mauch, Jan Stoess, Konrad Miller, and Frank Bellosa. Virtual InfiniBand clusters for HPC clouds. In *Proceedings of the 2nd International Workshop on Cloud Computing Platforms*, page 9. ACM, 2012. DOI: 10.1145/2168697.2168706.
10. Andrew Howard. NCI InfiniCloud: Expanding clouds with high speed InfiniBand interconnects. http://www-lk.apan.net/meetings/Fukuoka2015/Sessions/22/NCI_Howard_InfiniCloud_APAN_Fukuoka_TAN.pdf.
11. Wei Huang, Jiuxing Liu, Bulent Abali, and Dhabaleswar K Panda. A case for high performance computing with virtual machines. In *Proceedings of the 20th annual international conference on Supercomputing*, pages 125–134. ACM, 2006. DOI: 10.1145/1183401.1183421.
12. Jithin Jose, Mingzhe Li, Xiaoyi Lu, Krishna Chaitanya Kandalla, Mark Daniel Arnold, and Dhabaleswar K Panda. SR-IOV support for virtualization on infiniband clusters: Early experience. In *Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on*, pages 385–392. IEEE, 2013. DOI: 10.1109/ccgrid.2013.76.
13. E. R. Mardis. A decade's perspective on DNA sequencing technology. *Nature*, 470(7333):198–203, Feb 2011. DOI: 10.1038/nature09796.
14. Tin Wee Tan, Dominic S.H. Chien, Yuefan Deng, Seng Lim, Sing-Wu Liou, Jonathan Low, Marek Michalewicz, Gabriel Noaje, Yves Poppe, and Geok Lian Tan. InfiniCortex: A path to reach Exascale concurrent supercomputing across the globe utilising trans-continental InfiniBand and Galaxy of Supercomputers. Submitted to Supercomputing Frontiers 2015 conference proceedings, Singapore.

Received July 7, 2015.