NCI InfiniCloud

Expanding Clouds with High Speed InfiBand Interconnects

Andrew Howard
High Performance Computing and Cloud Team
andrew.howard@anu.edu.au
Thanks to my colleagues who did all the hard work

• Jakub Chrzeszczyk, NCI Cloud Team Architect
• Dr. Kenneth Hon Kim Ban, A*STAR
• Jonathan Low, A*STAR
• NCI - Australia’s National Computational Infrastructure

High Performance Computing
Clouds
Storage
Networks
• VirtualBox
  • Service development
• OpenStack
  • Nectar (Icehouse)
  • RHOS (Juno)
  • Tenjin (Icehouse)
  • InfiniCloud (Icehouse)
• Development and testing
• VMware
  • Production and Internal services

exploring Docker
• Can we use OpenStack to create “mini HPC” instances with InfiniBand to support workflows which perform better with local SSD storage?

• Can we use this as a provisioning system for bare metal only or can we also run Virtual Machines?

• What performance cost do we pay for InfiniBand within a VM?
HPC capabilities

- InfiniBand
- RDMA
- MPI
- IPoIB
- Lustre
  - Performs well for large sustained I/O
  - Performs poorly from small scattered I/O

Cloud convenience

- Rapid deployment, provisioning and Elastic capabilities
- NFS & Ceph
- SSD storage
- Secure
- High speed data access
NCI InfiniCloud: HPC InfiniBand performance in a Cloud

NCI InfiniCloud

OpenStack

Lustre File System

InfiniBand

Lustre File System
Can we make this work at intercontinental distances and demonstrate at SC14?
• NCI and A*Star collaboration
• Goals
  • Utilise trans-Pacific extended InfiniBand and shared SuperComputer resources to accelerate DNA analysis
  • Transfer large (~300Gb) genetic sequence data sets generated in Canberra from NCI to A*Star Singapore for analysis on the A*Star Aurora large memory system with the results (~1350Gb) visualised at the show floor in New Orleans (SC14) from a cluster at GaTech in the US
  • Utilise NCI High Performance InfiniBand Cloud HPC systems for visualisation of genetic data results produced by SG Aurora system
Illumina HiSeq 2500

GENETIC SEQUENCE: GATACGGAGTTTA……A 381GB

NCI AU

A*Star SG

Process with BWA on Aurora
Canberra/Singapore
1Gbs (~6000km)

Singapore/Canberra
10Gbs (~30,000km)
Crossbow

Longbow E-Series
Experiment 1: AU to SG large data transfer, process in SG on Aurora, return results data to AU
Experiment 2: SG to AU data transfer, process on InfiniCloud, view results at SC14
AU to SG data transfer speed

<table>
<thead>
<tr>
<th>Data transfer method</th>
<th>Average Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>rsync NCI to A*:Star R&amp;E path</td>
<td>4h 7m</td>
</tr>
<tr>
<td>rsync NCI to A*:Star 10Gbs IPoIB</td>
<td>3h</td>
</tr>
<tr>
<td>DSYNC 1Gbs IB path</td>
<td>1h 21m</td>
</tr>
<tr>
<td>DSYNC 10Gbs IB path (~26,000Km)</td>
<td>7m</td>
</tr>
</tbody>
</table>

298Gb data set
SG to AU data transfer speed

1,143Gb data set

- rsync A*Star to NCI R&E path: 12h 33m
- rsync A*Star to NCI 10Gbs IPoIB: 9h
- DSYNC 1Gbs IB path: 4h 5m
- DSYNC 10Gbs IB path (~26,000Km): 24m
SG to AU Observed data transfer rate at 10Gb using Obsidian DSYNC

<table>
<thead>
<tr>
<th>Time Unit</th>
<th>Data Transfer Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 second</td>
<td>900 MB</td>
</tr>
<tr>
<td>1 minute</td>
<td>54 GB</td>
</tr>
<tr>
<td>1 hour</td>
<td>3.24 TB</td>
</tr>
<tr>
<td>1 day</td>
<td>77 TB</td>
</tr>
<tr>
<td>1 week</td>
<td>539 TB</td>
</tr>
<tr>
<td>1 month</td>
<td>2.3 PB</td>
</tr>
<tr>
<td>Time Period</td>
<td>Bare Metal</td>
</tr>
<tr>
<td>----------------</td>
<td>-------------</td>
</tr>
<tr>
<td>1 second</td>
<td>900 MB</td>
</tr>
<tr>
<td>1 minute</td>
<td>54 GB</td>
</tr>
<tr>
<td>1 hour</td>
<td>3.24 TB</td>
</tr>
<tr>
<td>1 day</td>
<td>77 TB</td>
</tr>
<tr>
<td>1 week</td>
<td>539 TB</td>
</tr>
<tr>
<td>1 month</td>
<td>2.3 PB</td>
</tr>
</tbody>
</table>
Not a /dev/zero to /dev/null copy operation

All tests used actual data from an Illumina 2500 Genetic sequencer or an uncompressed HD (1920x1080) video file to replicate actual and not synthetic workflows

SSD were node attached for data storage
Suggested DSYNC parameter settings for 305ms delay

dsync \ 
   –option Xfer::RDMA::Buffer-Size=2146435072 \ 
   –option Xfer::RDMA::IO-Block-Size=10485760 \ 
   –option Xfer::RDMA::RDMA-Block-Size=524288
NCI InfiniCloud: HPC InfiniBand performance in a Cloud
• OpenStack Cloud supported by a 56Gbs FDR InfiniBand fabric
  • High Performance IB MPI
  • High Performance Lustre File System access
• Built using Mellanox Neutron modules
• Flexible top performance computational resources ‘on demand’
• HPC quality, Curated and Managed OS and application stack supported by an I/O architecture which includes local Solid State Disk and high performance large file storage using Lustre
- Nova to Neutron transition
  - nova-networks replaced with many private neutron-networks
- Public networks can connect tenants to the Internet
- Networks can remain purely private
- Virtual routers
- VMs can be multi-homed
- Provides users with flexible SDN network setup
Neutron supports many plugins including
- **Mellanox**
- **neutron-agents**

Mellanox plugins provide support for 56Gb Ethernet or Native InfiniBand operation

However specific hardware and version requirements (firmware etc) add complexity for provisioning
• Native IB installation of OpenStack requires the same degree of modification to core components as 56Gb Ethernet…… PLUS
• Replacing default DHCP with custom DHCP agent with IPoIB support
• OpenSM configuration mapping PKIs and Ethernet VLANs
• Neutron, hypervisor OS and nova-compute need to be configured
• https://wiki.openstack.org/wiki/Mellanox-Neutron-Icehouse-Redhat-InfiniBand
• External communication is exclusively Ethernet
• OpenSM is required to maintain fabric migrating to BGFC (feature request submitted for operation in Virtual mode)
• VMs hosted at NCI in Australia can communicate over RDMA with Physical and VMs hosted at A*Star in Singapore
• Obsidian DSYNC data transfers yield ~800MB/s within VM
• Ongoing collaboration between NCI, A*Star and NUS
• Now processing 1000Gb+ genomics datasets with ElasticCluster
• OpenSM stability
  » Pushing it in ways it wasn’t designed for
• Great data transfer numbers achieved with Obsidian DSYNC (even in a VM)
  • How can we utilise this technology for other large data transfers?
• IPoIB performance tuning
• Understand components we need to do further work on
• Observed ~10-15% performance “cost” for virtualisation of IB for large data transfers
• Further testing and characterisation of MPI performance and inter-VM performance has been completed by A*Star and published at Supercomputing Frontiers 2015
• “Let’s open this up to an actual bioinformatician who understands the limitations but is prepared to experiment”
• Dr. Kenneth Hon Kim Ban A*Star SG
- Provisioning
- ElasticCluster
- Data ingest
- Processing
- BWA
- Data export
• Dynamic provisioning of NFS servers for data access and multiple compute nodes into a BIO cluster
• 1:1 Core to VM CPU ratio
• Centos 6.6 guest
• *module* package manager (HPC compatibility)
• DSYNC for data movement between AU and SG
• Astronomy
• Existing radio and optical facilities
• Skymapper
• SKA Pathfinder and SKA

• Geographically sharing data
• Climate
• Environment
• Biology and Health
• Plant Phenomics
• Genetic Sequencing
Illumina HiSeq 2500
- 50-1500 Gb data set
- 1-6 days per run
- 1 or 2 Flow Cells
- 2 billion reads per Flow Cell

HiSeq X Ten
- 900-1800 Gb data set
- < 3 days per run
- 1 or 2 Flow Cells
- 3 billion reads per Flow Cell
- Equivalent to 25 * HiSeq 2500
What benefits can this provide?

- Personalised Medicine and Pharmaceuticals
- Significant cost reductions in Genome mapping (from $1B to around $1,000 per run) moves this into the pricing realm of other standard clinical diagnostic tests
- Better drug selections
- Better patient outcomes
• InfiniCortex: concurrent supercomputing across the globe utilising trans-continental InfiniBand and Galaxy of Supercomputers
• Galaxies of Supercomputers and their underlying interconnect topologies hierarchies
• Performance Assessment of InfiniBand HPC Cloud Instances on Intel Haswell and Intel Sandy Bridge Architectures
• InfiniCloud: Leveraging the Global InfiniCortex Fabric and OpenStack Cloud for Borderless High Performance Computing of Genomic Data and Beyond
On the path to global Exascale computing

• Lots of interesting questions for our community to research and answer:
  • How can we optimally support our legacy TCP/IP workloads in this environment?
  • Can InfiniBand scale?
  • Can we accelerate classes of science problems?
  • What are the performance characteristics of HPC systems with a high speed interconnect but a predictable transcontinental latency?
  • What are the optimal routing topologies? Can they be virtualised on a workflow basis for optimised communication?
  • How can we manage Authentication, Authorisation in a collaborative Multi National environment?
This is only one of many great stories of pioneering science built on the foundation of evolving and powerful high performance networks and supercomputers.

It only takes someone with an interest in trying something new, building trust and achieving something astounding to change the world.
This project made possible with the kind assistance of

A*Star
ANU John Curtin School of Medical Research
Obsidian Strategics
SingAREN
AARNet
Pacific Northwest GigaPOP
ESNet

and anyone else we’ve forgotten to thank