EXPERIENCES WITH LARGE-SCALE MULTI-SUBNET INFINIBAND FABRICS

David Southwell, CVO
Obsidian Strategics Inc.

[ April 7th, 2016 ]
TAKING INFINIBAND FURTHER
VISION

Enabling broad InfiniBand adoption through expanded capabilities

is about:

- Range extension – global reach using standard WANs (2005)
- In-line AES cryptography – encryption and authentication (2008)
- Native IB routing – multiple subnets, compound topologies (2013)

- Robust telco-grade hardware platforms – FPGA
- Vertical technology integration
- Commitment to open source, open standards, interoperability
INFINICORTEX & INFINICLOUD
“A Galaxy of Supercomputers” was initial motivation at A*STAR:

(Agency for Science Technology and Research - Singapore)

Marek Michalewicz, Ph.D

CEO: A*STAR Computational Resource Centre

Tin Wee TAN, Ph.D

Chairman: A*STAR Computational Resource Centre
Director: National SuperComputing Centre (NSCC)
INFINICORTEX
A global network for beneficially aggregating data sources, HPC, storage and analysis

- Concept developed by A*STAR CRC, Singapore
- An infrastructure for novel HPC workflows
- Entirely InfiniBand based
INFINICORTEX
A global network for beneficially aggregating data sources, HPC, storage and analysis

Team Singapore:

Team USA:

Team Europe:

Team Canada:

Team Australia:

- Bandwidth efficient global storage migration (asynchronous)
- Low latency metro area storage mirroring (synchronous)
- Cluster aggregation (MPI)
- Direct connect to streaming data sources (sequencers, physics...)
- Project remote high fidelity interactive visualisation
- Globally dispersed HPC stream processing (geo-pipelining)
INFINICORTEX
A global network for beneficially aggregating data sources, HPC, storage and analysis
INFINICORTEX
A global network for beneficially aggregating data sources, HPC, storage and analysis
INFINICORTEX
A global network for beneficially aggregating data sources, HPC, storage and analysis
INFINICORTEX
A global network for beneficially aggregating data sources, HPC, storage and analysis

Management Functions:
- Bowman B400

Interconnect Functions:
- Optical Processing (Rainbow C9)
- Metro Range (Longbow C100)
- Regional Range (Longbow C400)
- Routing (Crossbow R400-6)

Security Functions:
- Secure Global Range (Longbow E100)
INFINICORTEX APPLICATIONS
**INFINICORTEX APPLICATIONS**

Enabling broad InfiniBand adoption through expanded capabilities
High performance cloud computing platform for data-intensive workflows

Kenneth Ban
Dept of Biochemistry, NUS & IMCB, A*STAR

Jakub Chrzeszczyk
National Computational Infrastructure, ANU
INFINICORTEX APPLICATIONS

Breaking out of the traditional HPC mold

Specialized applications
High performance

Flexible
Virtualization overhead

InfiniCloud
What components do we need?

• **High CPU/memory and network performance** for rapid analysis of large datasets

  - **Fast Data Transfer**: Hundreds MB/sec - GB/sec
  - **Human Reference Sequence** (~4GB)
  - **Aligner BWA**: Genomic sequences (FASTQ)
  - **Aligned sequences** (BAM)
  - **Fast CPUs**
  - **High Memory**: 3+ GHz, > 16 GB memory
INFINICORTEX APPLICATIONS

What components do we need?

- **Reproducible** and well documented workflows that can be run on different hardware platforms

There are many parts (and different versions) in an analytical pipeline

Fitting them together properly can be challenging
INFINICORTEX APPLICATIONS

InfiniCloud: a flexible high performance cloud computing platform

- **Cloud** infrastructure for **flexible** computing
- High speed/low latency **Infiniband** interconnect
- **Long-range** Infiniband (global reach)
INFINICORTEX APPLICATIONS

On-the-fly provisioning of virtual clusters for distributed computing

Elasticluster

Prebuilt images
- Binaries
- Workflows

Public volume
- Reference data

Private volume
- Data/results

Provisioning of instances

Setup of scheduler/shared folders

InfiniCloud

Head node

Compute nodes
INFINICORTEX APPLICATIONS

A geo-distributed virtual cluster connected by long range InfiniBand
INFINICORTEX APPLICATIONS

Setup of geo-distributed virtual cluster

Status of Virtual Machines

<table>
<thead>
<tr>
<th>Instance Name</th>
<th>IP address</th>
<th>Zone</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>geopipeline-compute001</td>
<td>10.2.1.95</td>
<td>singapore</td>
<td>ACTIVE</td>
</tr>
<tr>
<td>geopipeline-compute002</td>
<td>10.2.1.96</td>
<td>australia</td>
<td>ACTIVE</td>
</tr>
<tr>
<td>geopipeline-compute003</td>
<td>10.2.1.98</td>
<td>europe</td>
<td>ACTIVE</td>
</tr>
<tr>
<td>geopipeline-compute004</td>
<td>10.2.1.97</td>
<td>australia</td>
<td>ACTIVE</td>
</tr>
<tr>
<td>geopipeline-frontend001</td>
<td>10.2.1.94</td>
<td>nova</td>
<td>ACTIVE</td>
</tr>
</tbody>
</table>
Demo pipeline for variant calling

- **Genomic sequences**
  - FASTQ

- **Aligned sequences**
  - BAM

- **Variants**
  - VCF

- **Annotated VCF**

**Data Volume**

- Human Reference Sequence (~4GB)
- Reference Databases (~30-300GB)

**Reference Volume**
INFINICORTEX APPLICATIONS

Geo-distributed pipeline for identification of mutations in cancer samples
INFINICORTEX APPLICATIONS

Acknowledgements
InfiniCloud is interesting for several reasons:

- An InfiniBand based HPC cloud overlay on InfiniCortex
- Virtualised HPC nodes through OpenStack containers (SRIOV)
- Spans multiple InfiniBand subnets
- All WAN links are hardware AES encrypted (bioinformatics...)
- All WAN link end points are hardware authenticated
- Implements high performance data flow pipeline across the globe!
- Fully automated across heterogeneous metal
- Sustained wire speed operation (simultaneous streaming and computation)
THE HEART OF THE FABRIC: BGFC
Middleware like InfiniCloud requires a fast, transparent, secure, scalable, segmentable, highly stable and manageable fabric.

Such a fabric needs a controller that did not exist before BGFC.

Matt Leininger (LLNL) and Bob Ciotti (NASA Ames) approached Obsidian in 2011 describing similar challenges with their LAN, CAN and WAN InfiniBand deployments.

Not seeing OpenSM as a viable platform upon which to build, Obsidian responded with a green field fabric controller architecture that would address immediate requirements and many more besides...
BGFC
Bowman Global Fabric Controller

Subnet 1 - 4D Hypercube (Pleiades elements)

Subnet 2 - Clos3 fabric (Hyperion elements)

Switches in Subnet 1

Switches in Subnet 2

Encrypting range-extending routers (Obsidian Longbow E100s)
Now 5 years into the adventure, Obsidian has completed phase I (LLNL and NASA) and phase II (A*STAR) of the original development program.

Phase III is imminent (TRL-9 testing at scale), and BGFC will be open-sourced thereafter.

Built for a much grander mission than simple subnet management, BGFC is aimed at complex multi-domain fabrics supporting international traffic while preserving sovereign administrative domains.
**Exact and Mathematically Perfect LFTs**

Graph theory-based subgraph isomorphism
Guaranteed deadlock-free routing
Reliably supports very large subnets
Deterministic QoR
Python scripts for topology descriptions
Python topology prescription examples:

Simple, direct from templates ...

```python
from bgfc.template import *
topology = {"my-topo": Hypercube(9)}
```

Customised from existing templates ...

```python
from bgfc.template import *
class MyNetwork(ClosTree5):
    def __init__(self):
        ClosTree5.__init__(self, core=(36, 18), cores=2, edge_conns=18//2)
    def disperse_edges(self):
        # [.. insert site-specific wiring function ..]
    topology = {"my-topo": MyNetwork()}
```

, or arbitrarily complex by deriving from the Topology class.
Precise topology definitions allow simple mathematical routing functions, but also provide a powerful means of detecting unintended deviations.

BGFC uses persistent topology and LFT solution databases to make it easy to ensure a fabric initialises the same way if required, increasing reproducability of fabric behaviour and drastically reducing initialisation time.
Core routing concept:

Optimal dead lock free IB routing is a NP problem with complexity related to the number of buffers, and no apparent easy shortcuts.

Subgraph isomorphism is a NP problem with complexity related to the number of switches, with well known shortcuts.

So, solve subgraph isomorphism and then get routing, rather than try to solve routing directly!

BGFC: 624 lines of Python
OpenSM: 17,200 lines of C
Example; a torus(2,3) and its acyclic flow group graph:
Multisubnet Native InfiniBand Routing

- Administrative demarcation
- Fault isolation
- Performance at scale
- Complex topologies
- Inter-site separation

BGFC
Bowman Global Fabric Controller
Why is it called an InfiniBand 'subnet'...

...if there is only one in your cluster?
N-way Clustered Active SMs

- Parallel host-based subnet managers
- High performance at scale
- Lockless WODB architecture
- Decentralised control
- Extreme fault tolerance
- C++11 implementation
FUTURES...
FUTURES
100 and 400Gbits/s Thresholds
THANK YOU

David Southwell, CVO

Obsidian Strategics Inc.