OPENFABRICS ALLIANCE

12th ANNUAL WORKSHOP 2016

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)**
- **Industrial-strength fabric management (2015)**

- **Robust telco-grade hardware platforms – FPGA**
- **Vertical technology integration**
- **Commitment to open source, open standards, interoperability**

INFINICORTEX & INFINICLOUD

INFINICORTEX

A global network for beneficially aggregating data sources, HPC, storage and analysis

"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

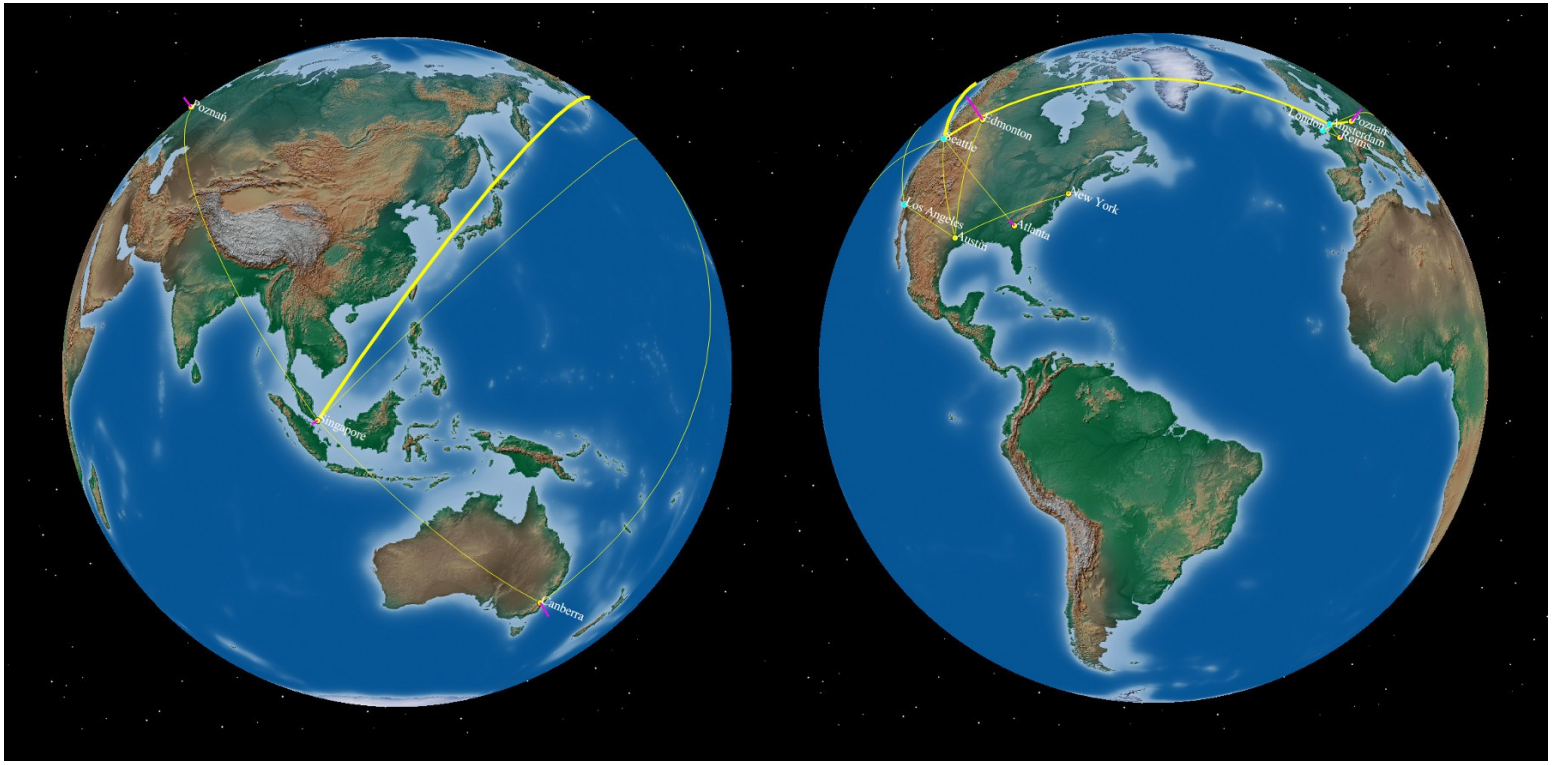
Chairman: **A*STAR Computational Resource Centre**
Director: **National SuperComputing Centre (NSCC)**

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**

A*STAR Computational Resource Centre



INFINICORTEX

A global network for beneficially aggregating data sources, HPC, storage and analysis

Team Singapore:

Team USA:

Team Europe:

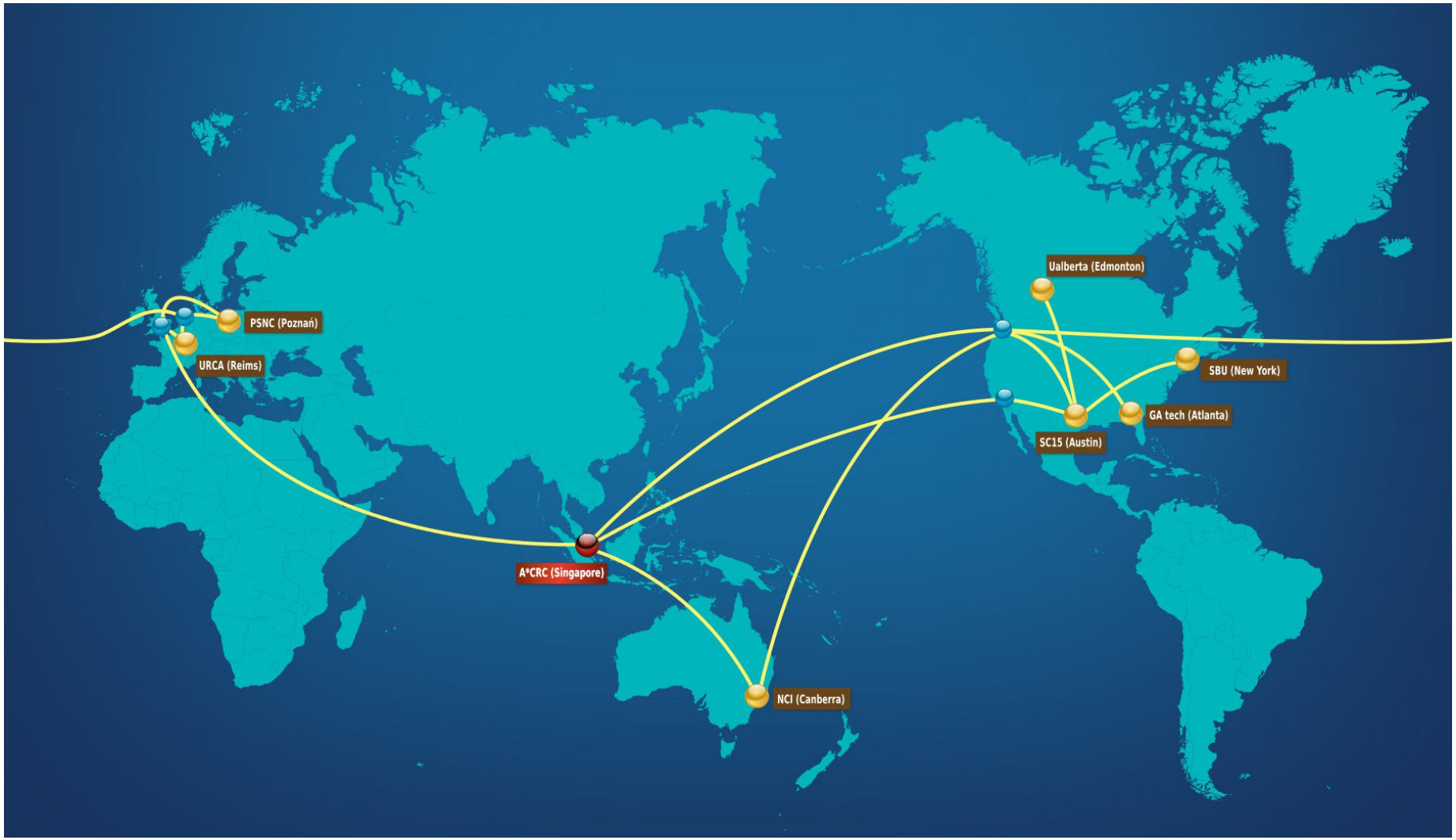
Team Canada:

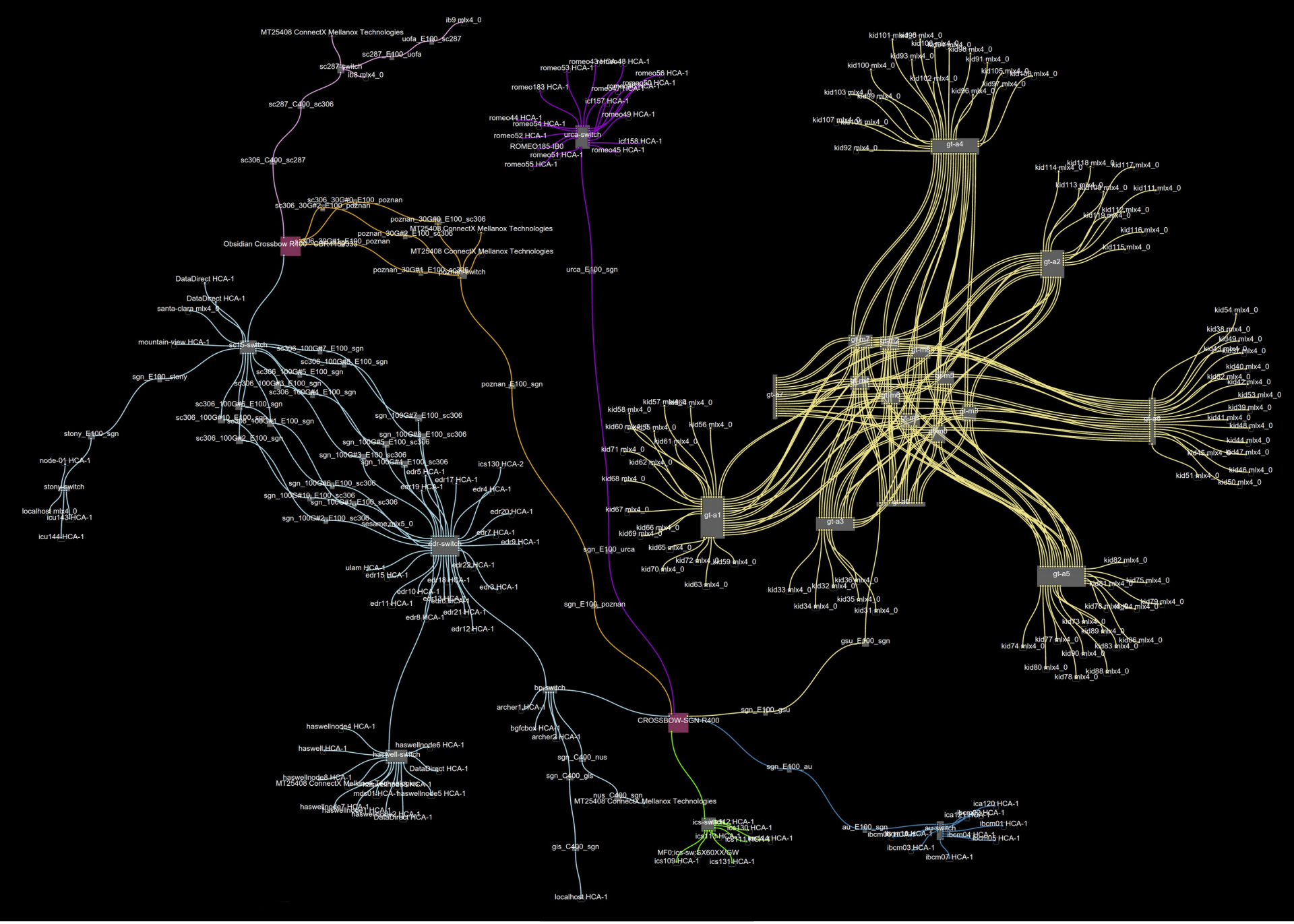
Team Australia:

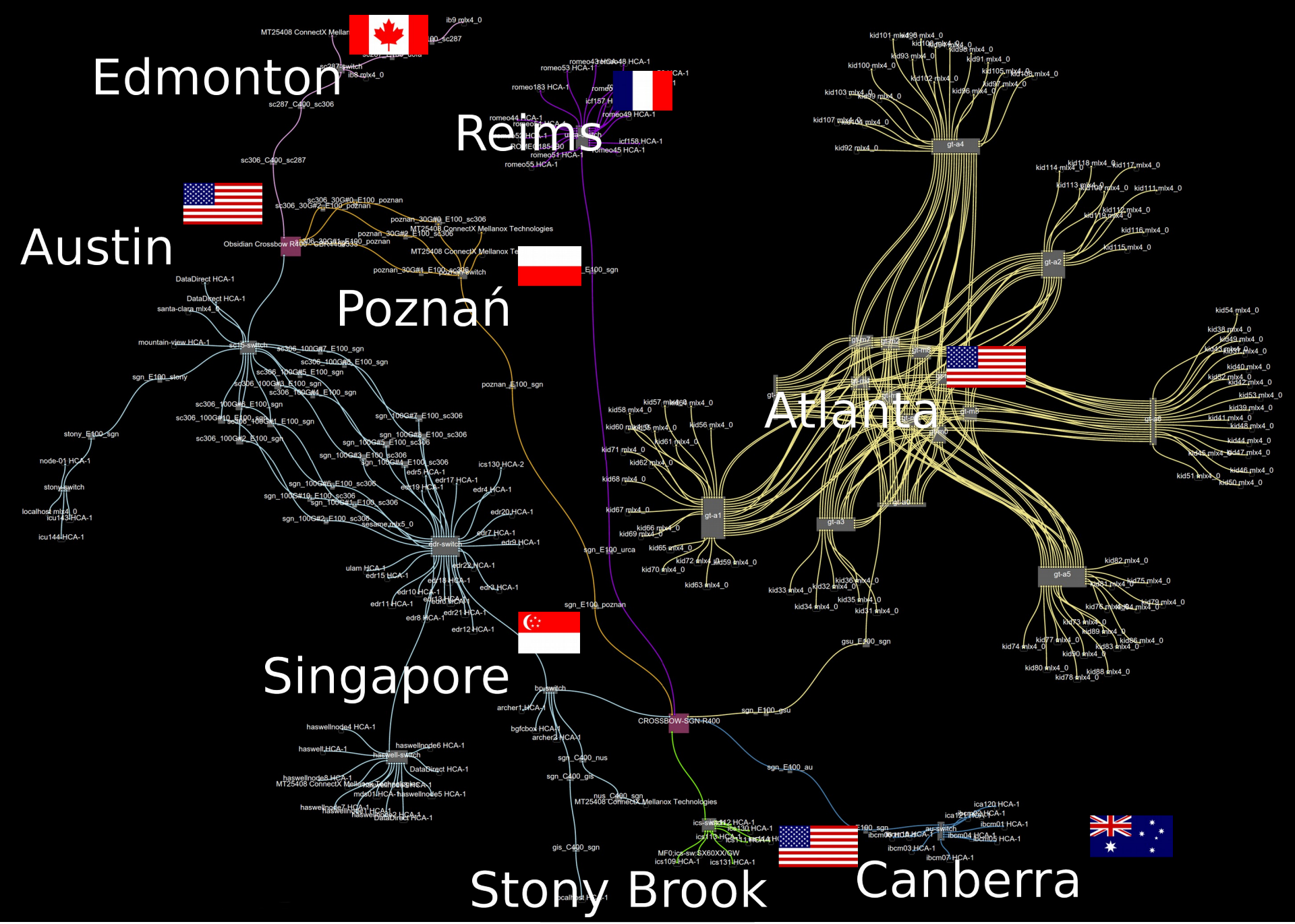
- **Bandwidth efficient global storage migration (asynchronous)**
- **Low latency metro area storage mirroring (synchronous)**
- **Cluster aggregation (MPI)**
- **Direct connect to streaming data sources (sequencers, physics...)**
- **Project remote high fidelity interactive visualisation**
- **Globally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

INFINICORTEX

A global network for beneficially aggregating data sources, HPC, storage and analysis

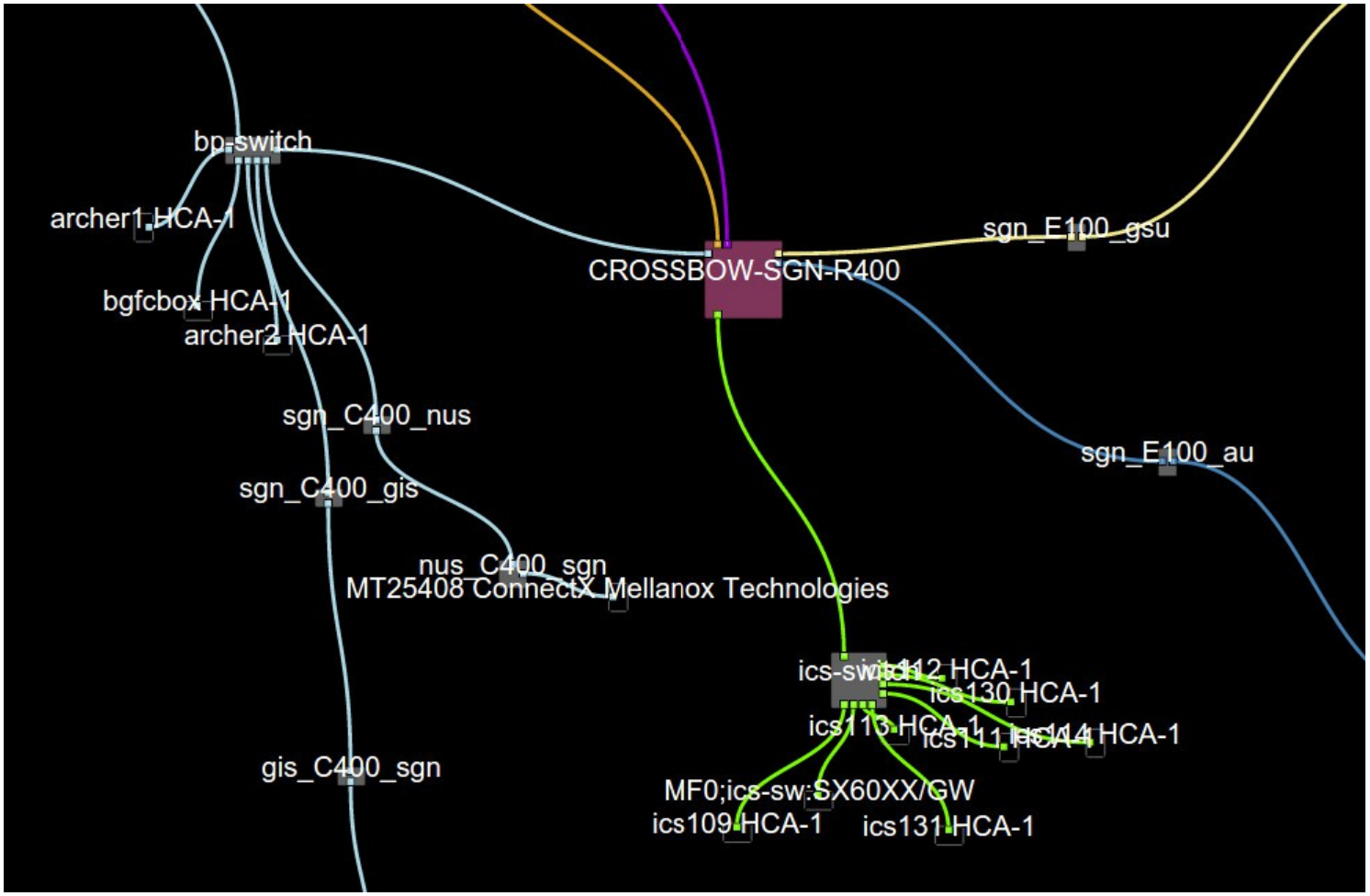



ib9 mlx4_0
MT25408 ConnectX Mellanox Technologies
uofa_E100_sc287
sc287_E100_uofa
sc287-switch
ib8 mlx4_0
sc287_C400_sc306
sc306_C400_sc287
romeo43 HCA-1
romeo48 HCA-1
romeo53 HCA-1
romeo56 HCA-1
romeo50 HCA-1
romeo183 HCA-1
icf157 HCA-1
romeo49 HCA-1
romeo44 HCA-1
romeo54 HCA-1
urca-switch
romeo52 HCA-1
icf158 HCA-1
ROMEO185-IB0
romeo45 HCA-1
romeo51 HCA-1
romeo55 HCA-1
kid101 mlx4_0
kid98 mlx4_0
kid106 mlx4_0
kid104 mlx4_0
kid93 mlx4_0
kid91 mlx4_0
kid100 mlx4_0
kid105 mlx4_0
kid102 mlx4_0
kid108 mlx4_0
kid97 mlx4_0
kid103 mlx4_0
kid99 mlx4_0
kid96 mlx4_0
kid107 mlx4_0
kid101 mlx4_0
kid92 mlx4_0
gt-a4
kid114 mlx4_0
kid118 mlx4_0
kid117 mlx4_0
kid113 mlx4_0
kid109 mlx4_0
kid111 mlx4_0
kid112 mlx4_0
kid110 mlx4_0
kid116 mlx4_0
kid115 mlx4_0
gt-a2
sc306_30G#0_E100_poznan
sc306_30G#2_E100_poznan
poznan_30G#0_E100_sc306
MT25408 ConnectX Mellanox Technologies
poznan_30G#2_E100_sc306
Obsidian Crossbow R400 - CBR4150533
sc306_30G#1_E100_poznan
MT25408 ConnectX Mellanox Technologies
poznan_30G#1_E100_sc306
poznan-switch
urca_E100_sgn
DataDirect HCA-1
DataDirect HCA-1
santa-clara mlx4_0
mountain-view HCA-1
sc15-switch
sc306_100G#7_E100_sgn
sc306_100G#6_E100_sgn
sc306_100G#5_E100_sgn
sgn_E100_stony
sc306_100G#3_E100_sgn
sc306_100G#4_E100_sgn
sc306_100G#8_E100_sgn
sc306_100G#10_E100_sgn
sc306_100G#1_E100_sgn
sgn_100G#7_E100_sc306
sc306_100G#2_E100_sgn
sgn_100G#8_E100_sc306
sgn_100G#5_E100_sc306
sgn_100G#3_E100_sc306
sgn_100G#4_E100_sc306
poznan_E100_sgn
stony_E100_sgn
node-01 HCA-1
stony-switch
localhost mlx4_0
icu143 HCA-1
icu144 HCA-1
ics130_HCA-2
edr5 HCA-1
edr17 HCA-1
edr19 HCA-1
edr4 HCA-1
sgn_100G#6_E100_sc306
sgn_100G#10_E100_sc306
sgn_100G#1_E100_sc306
edr20 HCA-1
sgn_100G#2_E100_sc306
sesame mlx5_0
edr7 HCA-1
edr9 HCA-1
edr-switch
edr22 HCA-1
ulam HCA-1
edr15 HCA-1
edr16 HCA-1
edr10 HCA-1
edr3 HCA-1
edr13 HCA-1
edr11 HCA-1
edr21 HCA-1
edr8 HCA-1
edr12 HCA-1
kid57 mlx4_0
kid58 mlx4_0
kid60 mlx4_0
kid55 mlx4_0
kid56 mlx4_0
kid61 mlx4_0
kid71 mlx4_0
kid62 mlx4_0
kid68 mlx4_0
kid67 mlx4_0
kid66 mlx4_0
kid69 mlx4_0
kid65 mlx4_0
kid72 mlx4_0
kid59 mlx4_0
kid70 mlx4_0
kid63 mlx4_0
gt-a1
gt-a3
kid54 mlx4_0
kid38 mlx4_0
kid49 mlx4_0
kid40 mlx4_0
kid52 mlx4_0
kid42 mlx4_0
kid53 mlx4_0
kid39 mlx4_0
kid41 mlx4_0
kid48 mlx4_0
kid44 mlx4_0
kid45 mlx4_0
kid47 mlx4_0
kid46 mlx4_0
kid51 mlx4_0
kid50 mlx4_0
gt-a6
kid82 mlx4_0
kid84 mlx4_0
kid75 mlx4_0
kid81 mlx4_0
kid79 mlx4_0
kid76 mlx4_0
kid94 mlx4_0
kid73 mlx4_0
kid89 mlx4_0
kid77 mlx4_0
kid86 mlx4_0
kid74 mlx4_0
kid83 mlx4_0
kid90 mlx4_0
kid80 mlx4_0
kid88 mlx4_0
kid78 mlx4_0
gt-a5
sgn_E100_urca
kid33 mlx4_0
kid32 mlx4_0
kid36 mlx4_0
kid34 mlx4_0
kid35 mlx4_0
kid31 mlx4_0
sgn_E100_poznan
gsu_E100_sgn
bp-switch
archer1 HCA-1
bgfcbox HCA-1
archer2 HCA-1
CROSSBOW-SGN-R400
sgn_E100_gsu
haswellnode4 HCA-1
haswellnode6 HCA-1
haswell HCA-1
haswell-switch
DataDirect HCA-1
haswellnode8 HCA-1
MT25408 ConnectX Mellanox Technologies
mds01-HCA-haswellnode5 HCA-1
haswellnode7 HCA-1
haswellnode5 HCA-1
DataDirect HCA-1
sgn_C400_nus
sgn_C400_gis
nus_C400_sgn
MT25408 ConnectX Mellanox Technologies
gis_C400_sgn
localhost HCA-1
sgn_E100_au
au_E100_sgn
ica120 HCA-1
ica121 HCA-1
ibcm01 HCA-1
au-switch
ibcm04 HCA-1
ibcm05 HCA-1
ibcm03 HCA-1
ibcm07 HCA-1
ics-sw-x312 HCA-1
ics130 HCA-1
ics110 HCA-1
ics111 HCA-1
MF0;ics-sw:SX60XX/GW
ics109 HCA-1
ics131 HCA-1


Edmonton
Austin
Rems
Poznań
Atlanta
Singapore
Stony Brook
Canberra
MT25408 ConnectX Mellanox Technologies
Obsidian Crossbow R400
CROSSBOW-SGN-R400
sc15-switch
edr-switch
gt-a1
gt-a2
gt-a3
gt-a4
gt-a5
bp-switch
haswell-switch
au-switch
stony-switch
sesame.mlx5_0
sc287_C400_sc306
sc306_C400_sc287
poznan_E100_sgn
sgn_E100_urca
sgn_E100_poznan
sgn_E100_gsu
sgn_E100_au
gsu_E100_sgn
sgn_C400_nus
sgn_C400_gis
gis_C400_sgn
nus_C400_sgn
sgn_E100_stony
stony_E100_sgn

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)**

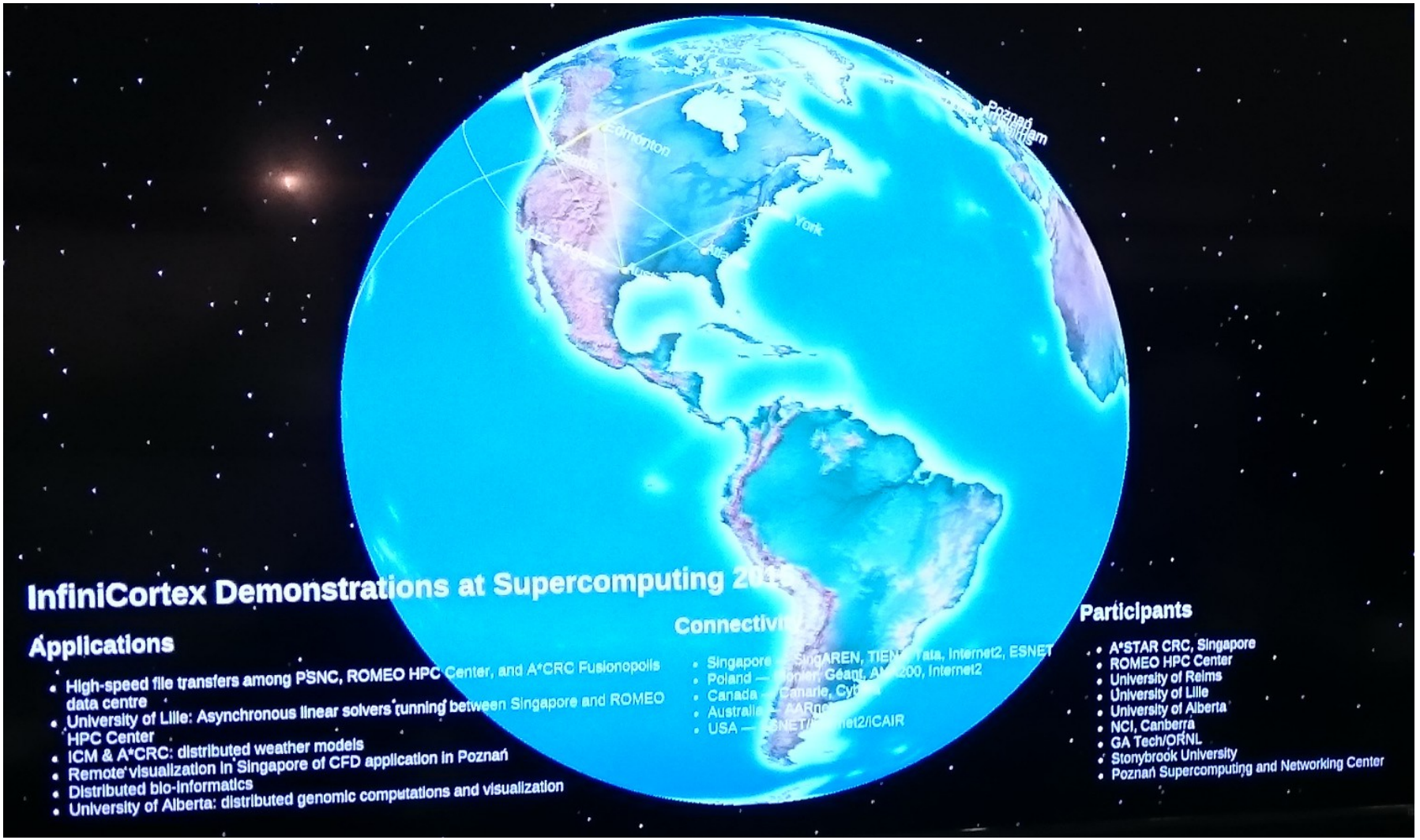
Security Functions:

- **Secure Global Range (Longbow E100)**

INFINICORTEX APPLICATIONS

INFINICORTEX APPLICATIONS

Enabling broad InfiniBand adoption through expanded capabilities



INFINICORTEX APPLICATIONS

InfiniCloud

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

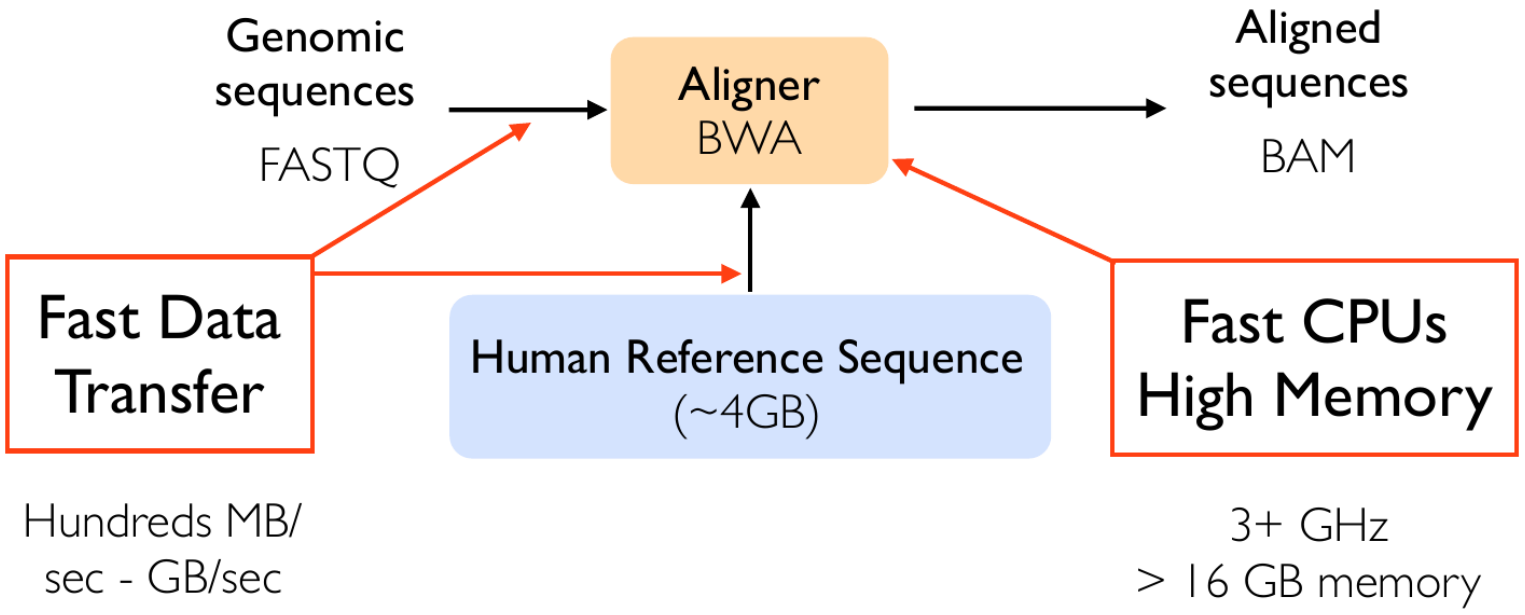
InfiniCloud

Computational Resource Centre
A*STAR

NCI
NATIONAL COMPUTATIONAL INFRASTRUCTURE

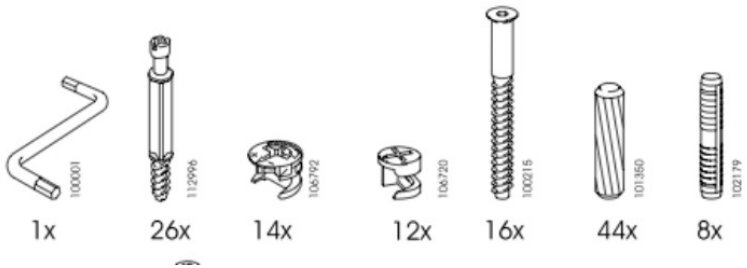
What components do we need?

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



What components do we need?

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

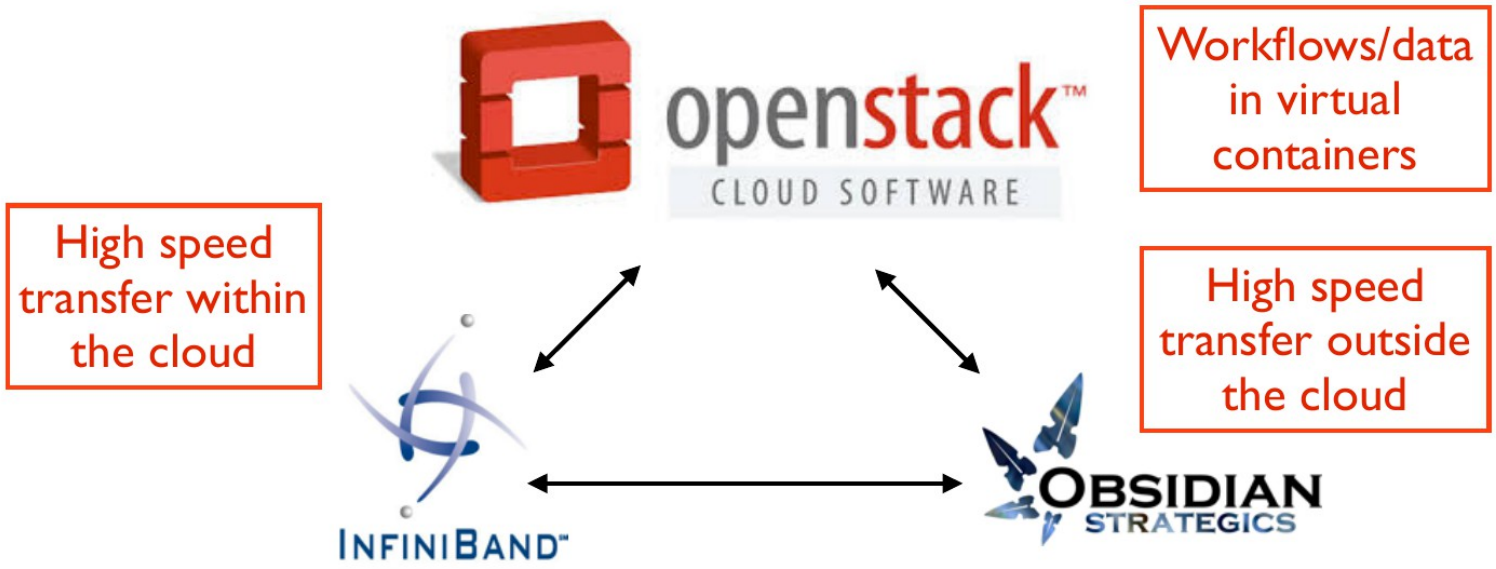


There are many parts (and different versions) in an analytical pipeline

Fitting them together properly can be challenging

INFINICORTEX APPLICATIONS

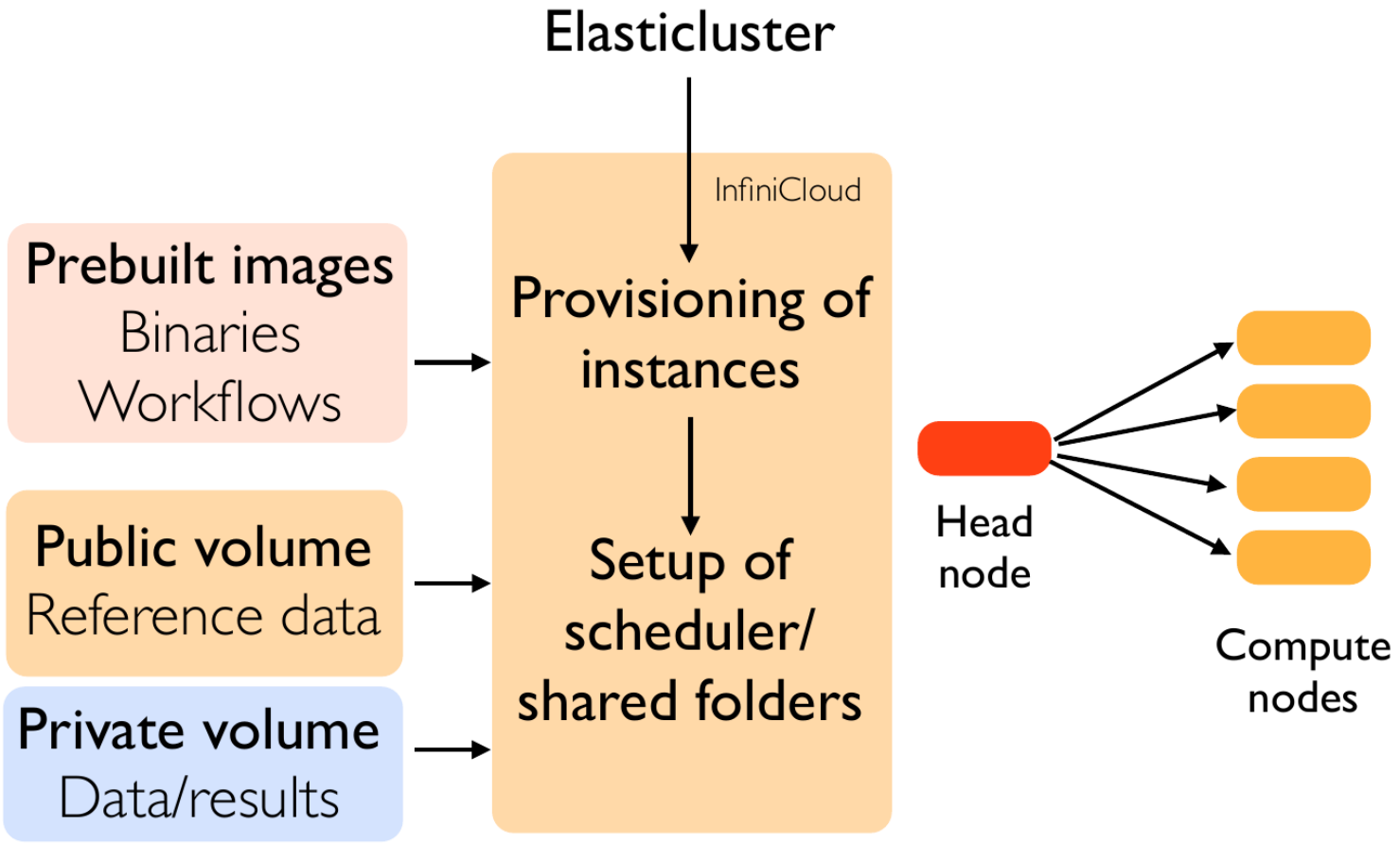
InfiniCloud: a flexible high performance cloud computing platform



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

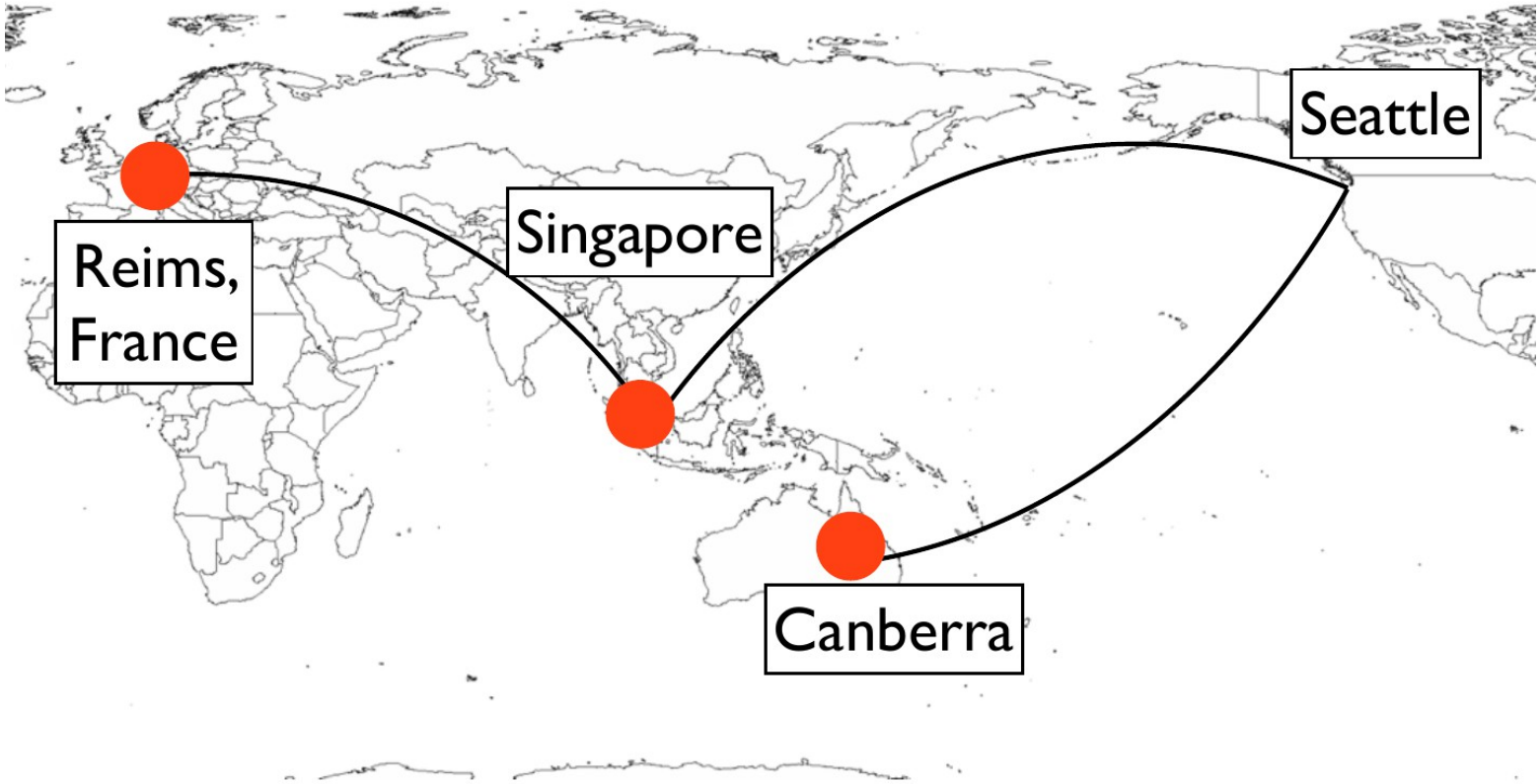
INFINICORTEX APPLICATIONS

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



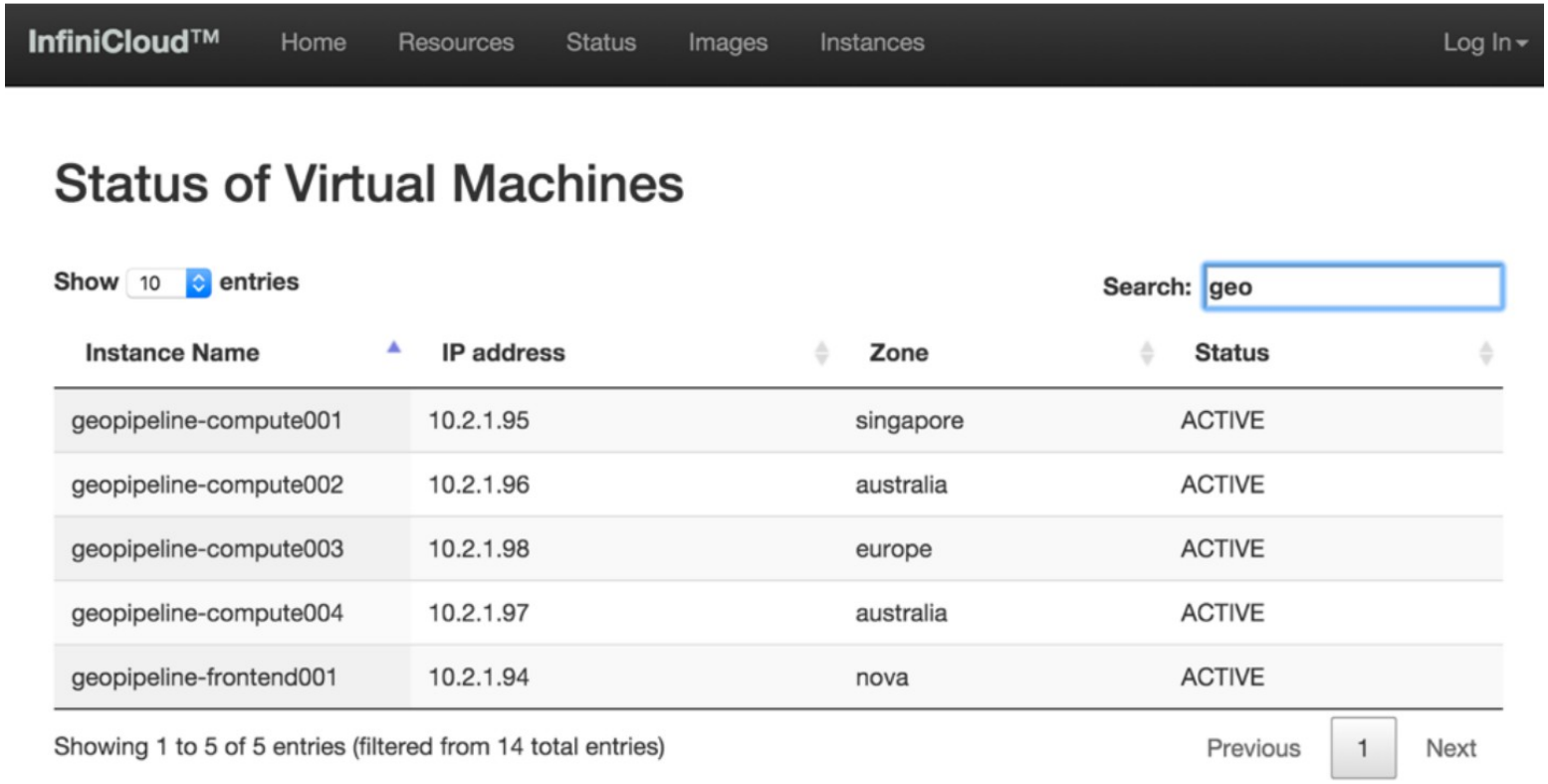
INFINICORTEX APPLICATIONS

A geo-distributed virtual cluster connected by long range InfiniBand

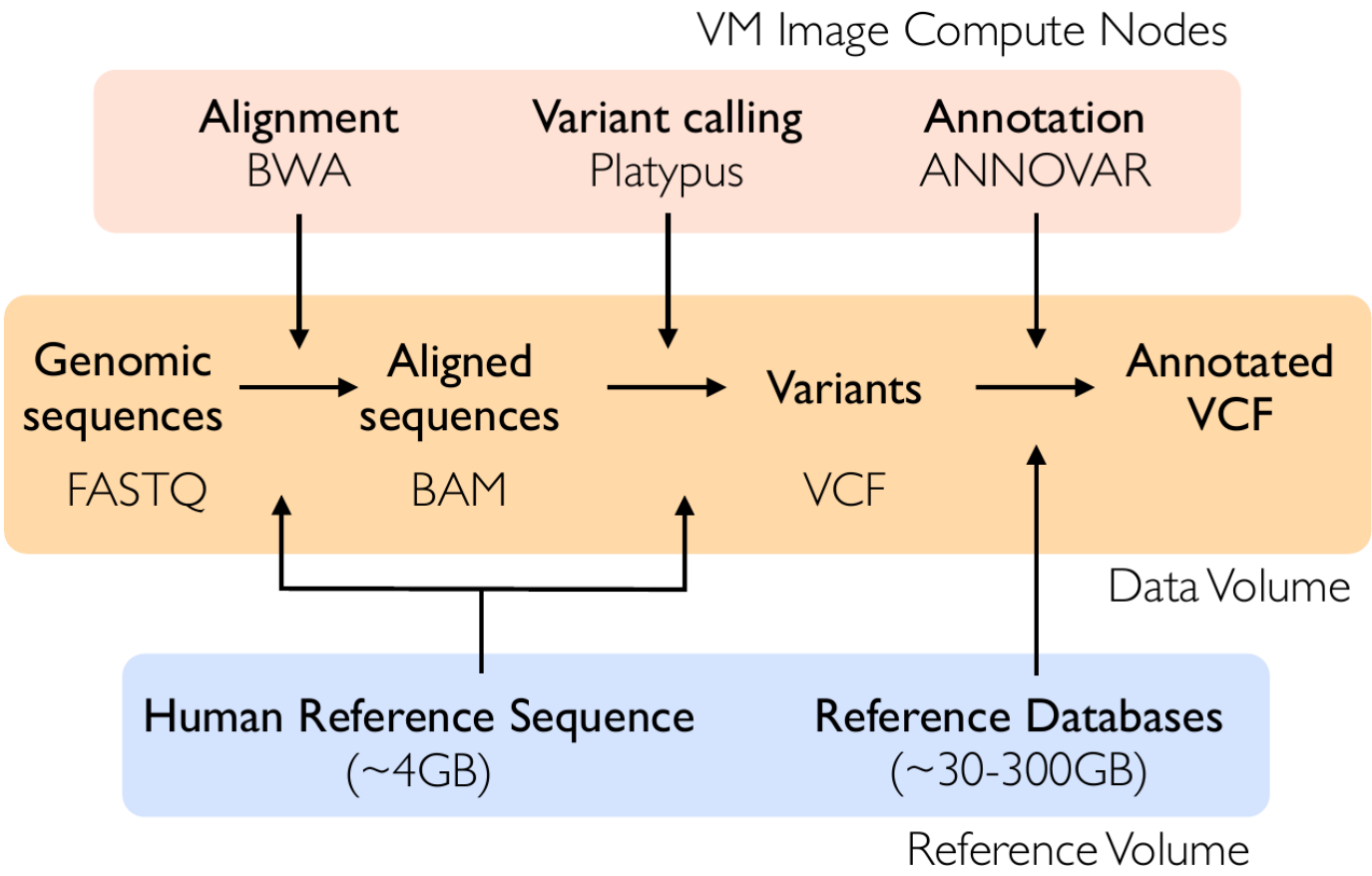


INFINICORTEX APPLICATIONS

Setup of geo-distributed virtual cluster



Demo pipeline for variant calling

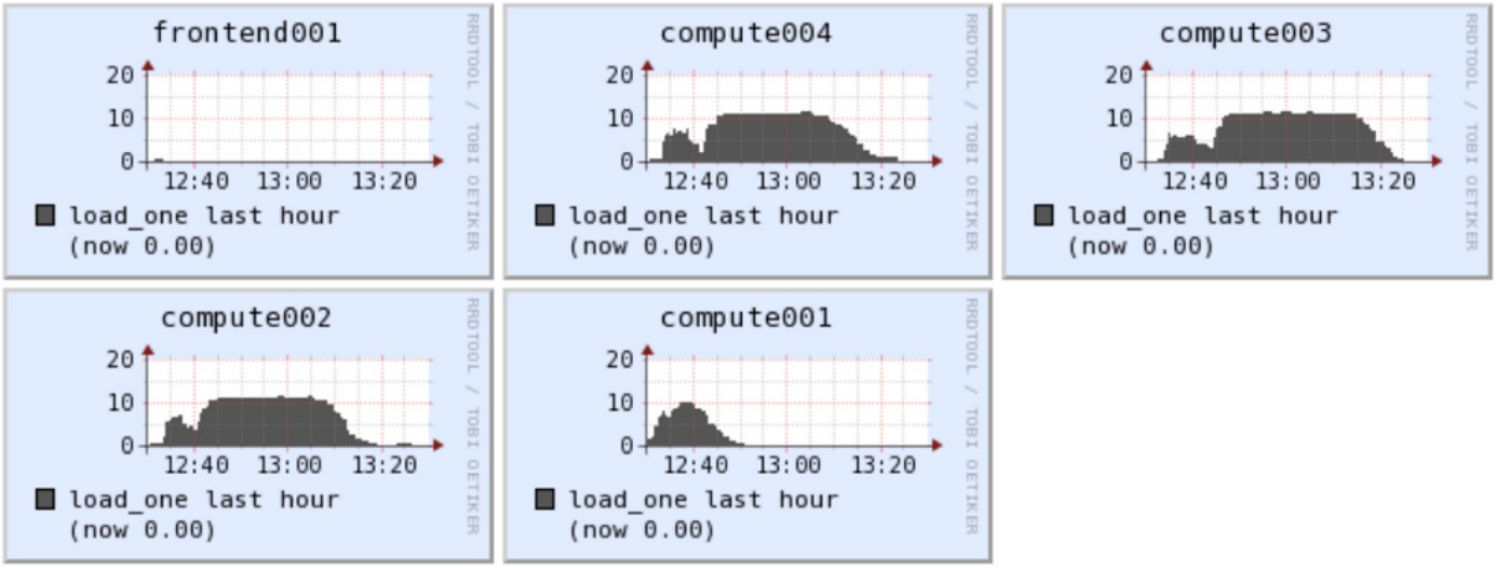


INFINICORTEX APPLICATIONS

Geo-distributed pipeline for identification of mutations in cancer samples

Australia

France



Australia

Singapore

INFINICORTEX APPLICATIONS

Acknowledgements

Computational Resource Centre

Institute of Molecular and Cell Biology

INFINICORTEX APPLICATIONS

Enabling broad InfiniBand adoption through expanded capabilities

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

BGFC

Bowman Global Fabric Controller

Middleware like InfiniCloud requires a fast, transparent, secure, scalable, segmentable, highly stable and manageable fabric.

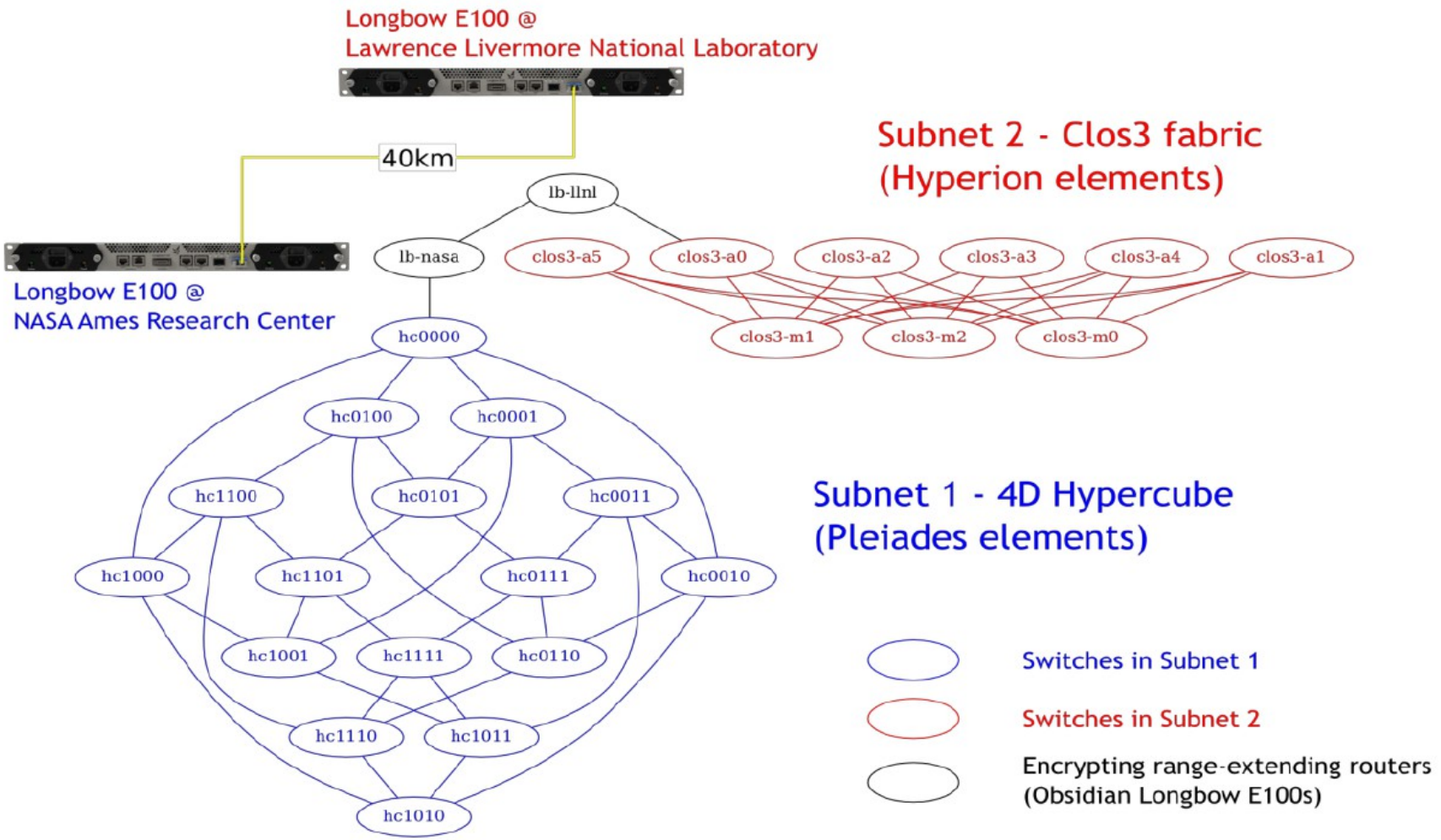
Such a fabric needs a controller that did not exist before BGFC.

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



BGFC

Bowman Global Fabric Controller

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.

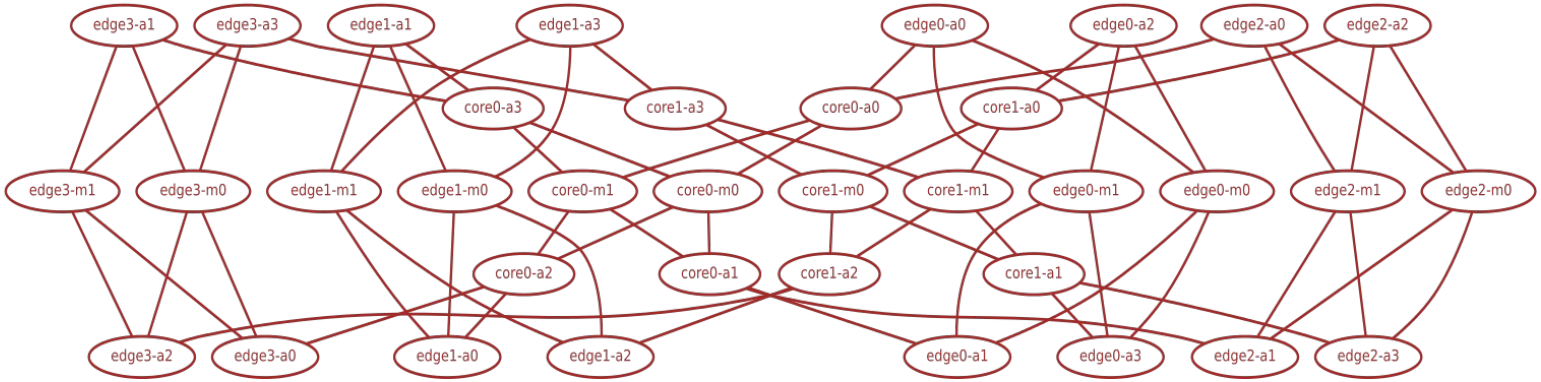
Built for a much grander mission than simple subnet management, BGFC is aimed at complex multi-domain fabrics supporting international traffic while preserving sovereign administrative domains.

BGFC

Bowman Global Fabric Controller

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



BGFC

Bowman Global Fabric Controller

Python topology prescription examples:

Simple, direct from templates ...

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

Customised from existing templates ...

```
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.**

BGFC

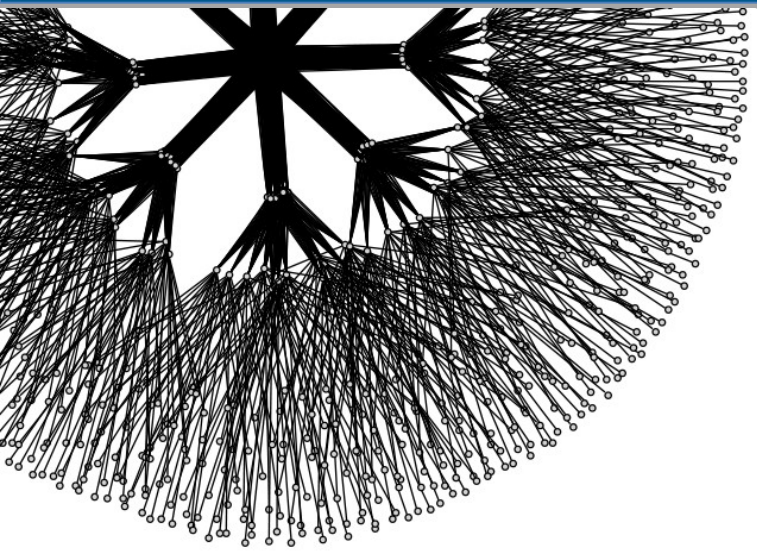
Bowman Global Fabric Controller

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.

BGFC

Bowman Global Fabric Controller



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.

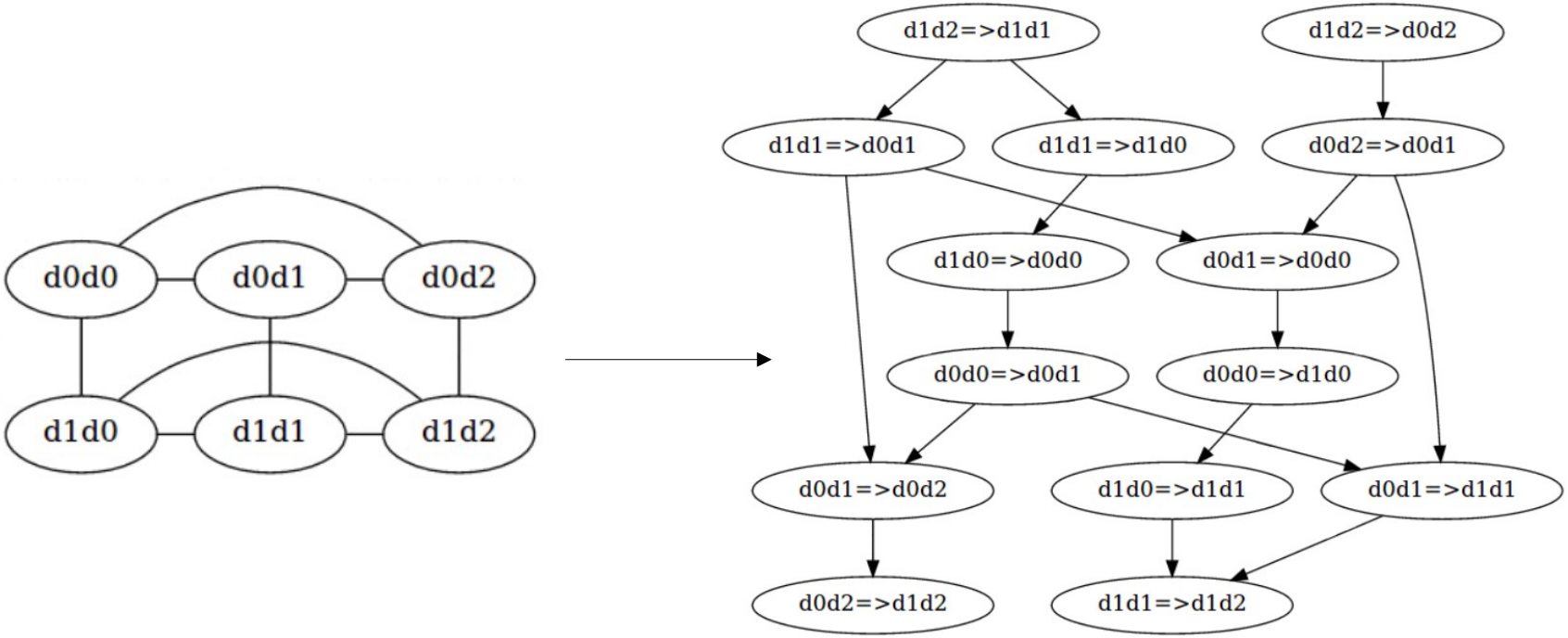
So, solve subgraph isomorphism and then get routing, rather than try to solve routing directly!

BGFC: **624 lines of Python**
OpenSM: **17,200 lines of C**

BGFC

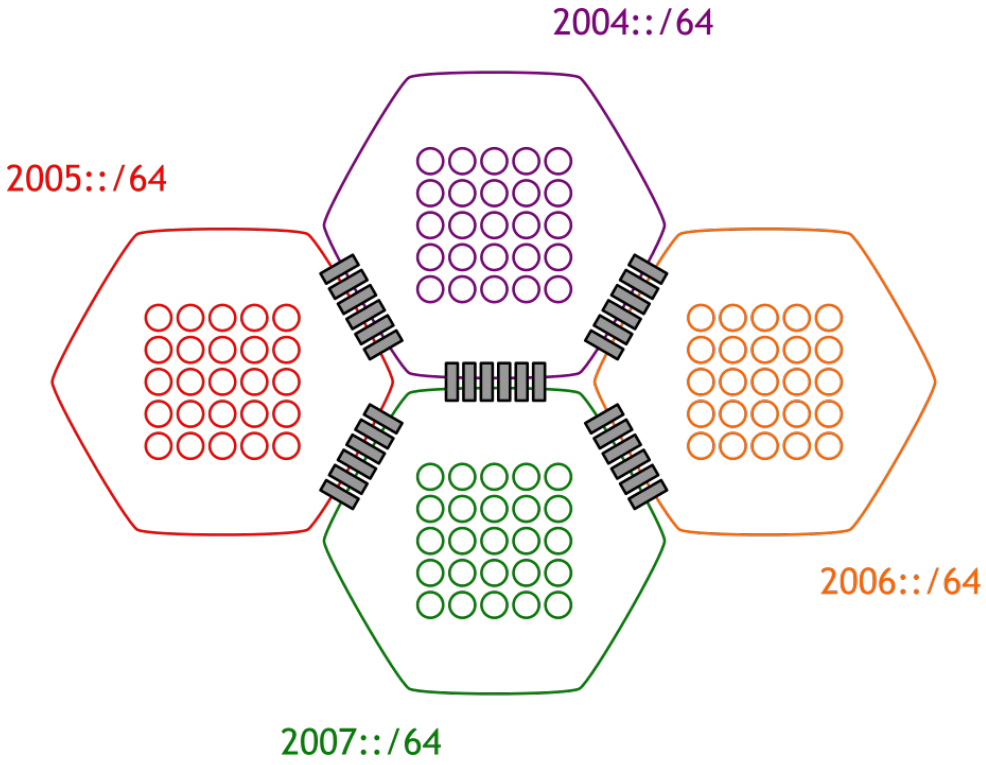
Bowman Global Fabric Controller

Example; a torus(2,3) and its acyclic flow group graph:



BGFC

Bowman Global Fabric Controller



Multi-subnet Native InfiniBand Routing

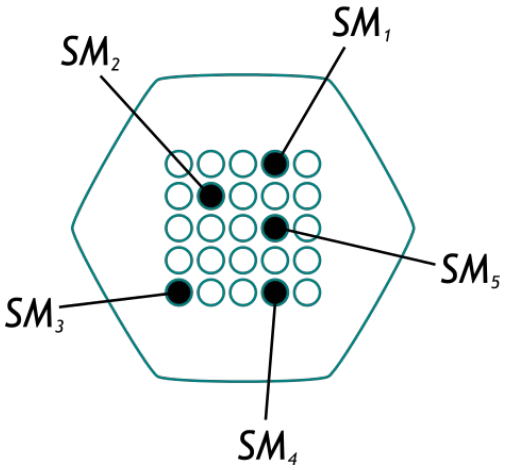
Administrative demarcation
Fault isolation
Performance at scale
Complex topologies
Inter-site separation

BGFC

Bowman Global Fabric Controller

BGFC

Bowman Global Fabric Controller



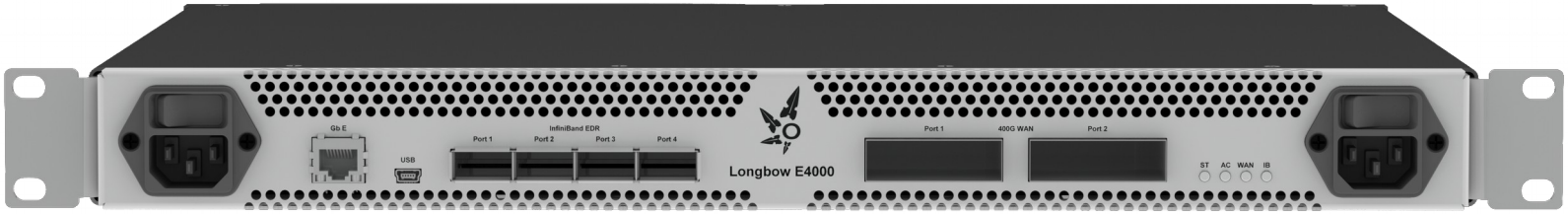
N-way Clustered Active SMs

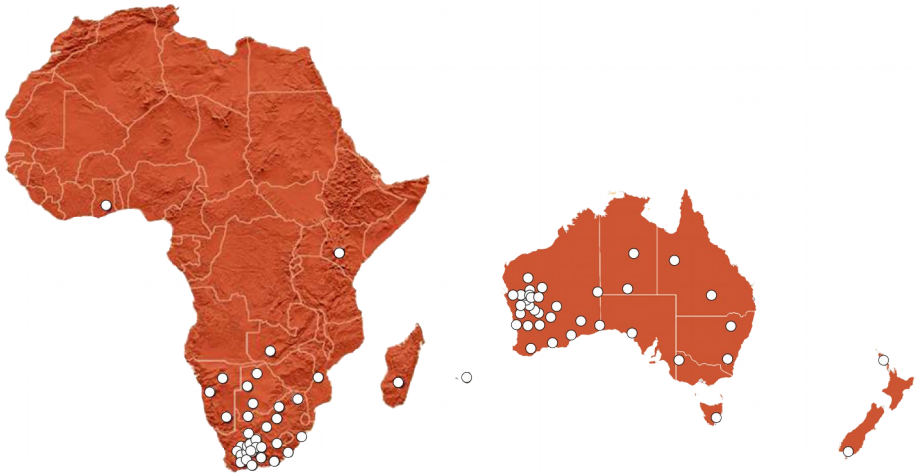
Parallel host-based subnet managers
High performance at scale
Lockless WODB architecture
Decentralised control
Extreme fault tolerance
C++11 implementation

FUTURES…

FUTURES

100 and 400Gbits/s Thresholds





OPENFABRICS ALLIANCE

12th ANNUAL WORKSHOP 2016

THANK YOU

David Southwell, CVO

Obsidian Strategics Inc.

OBSIDIAN STRATEGICS