InfiniCortex - From Proof-of-concept to Production

Gabriel Noaje[1], *Alan Davis*[1], *Jonathan Low*[1], *Lim Seng*[1], *Tan Geok Lian*[1], *Łukasz P. Orłowski*[1,2,3], *Dominic Chien*[1], *Liou Sing-Wu*[1], *Tan Tin Wee*[1,4,5], *Yves Poppe*[1], *Ban Hon Kim Kenneth*[6], *Andrew Howard*[7], *David Southwell*[8], *Jason Gunthorpe*[8], *Marek T. Michalewicz*[9,1,2]

The global effort to build ever more powerful supercomputers is faced with the challenge of ramping up High Performance Computing systems to ExaScale capabilities and, at the same time, keeping the electrical power consumption for a system of that scale at less than 20 MW level. One possible solution, bypassing this local energy limit, is to use distributed supercomputers to alleviate intense power requirements at any single location. The other critical challenge faced by the global computer industry and international scientific collaborations is the requirement of streaming colossal amounts of time-critical data. Examples abound: i) transfer of astrophysical data collected by the Square Kilometre Array to the international partners, ii) streaming of large facilities experimental data through the Pacific Research Platform collaboration of DoE, ESnet and other partners in the US and elsewhere, iii) the Superficilities vision expressed by DoE, iv) new architecture for CERN LHC data processing pipeline focussing on more powerful processing facilities connected by higher throughput connectivity.

The InfiniCortex project led by A*STAR Computational Resource Centre demonstrates a worldwide InfiniBand fabric circumnavigating the globe and bringing together, as one concurrent globally distributer HPC system, several supercomputing facilities spanned across four continents (Asia, Australia, Europe and North America). Using global scale InfiniBand connections, with bandwidth utilisation approaching 98% link capacity, we have established a new architectural approach which might lead to the next generation supercomputing systems capable of solving the most complex problems through the aggregation and parallelisation of many globally distributed supercomputers into a single hive-mind of enormous scale.

Keywords: InfiniCortex, InfiniBand, global supercomputer connectivity, superfacilities, InfiniCloud, HPC Cloud, supercomputer networking, HPC workflows, ADIOS.

Introduction

This article is a final report of *InfiniCortex I* project led by A*STAR Computational Resource Centre in Singapore. We document an implementation of a worldwide InfiniBand fabric bringing together several supercomputing facilities spanned across the globe to create a galaxy of supercomputers [12]. *InfiniCortex I* project represents a huge collaboration effort of several agencies and universities in Singapore (A*STAR, NSCC, NUS, NTU, SingAREN) together with more than 20 international partners around the globe.

After successfully demonstrating the first 100Gbps transcontinental InfiniBand connection connecting Singapore and United States of America at the annual Supercomputing Conference 2014, in New Orleans, USA [10], the award winning InfiniCortex project [1] grew rapidly demonstrating every year novel capabilities.

[1] A*STAR Computational Resource Centre (A*CRC), 1 Fusionopolis Way, #17-01 Connexis, 138632 Singapore
[2] Institute for Advanced Computational Science, Stony Brook University, New York, USA
[3] Department of Applied Mathematics and Statistics, Stony Brook University, New York, USA
[4] Singapore National Supercomputing Centre (NSCC), Singapore
[5] Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[6] National University of Singapore, Singapore
[7] The Australian National University, Canberra, Australia
[8] Obsidian Strategics Inc, Canada
[9] Interdisciplinary Centre for Mathematical and Computational Modelling (ICM), University of Warsaw, Poland

Over the last few years several unprecedented elements have been showcased:

- largest ever spanning InfiniBand network – a ring-around-the-world with most of the segments at 100Gbps and few at 10-30Gbps;
- eight InfiniBand subnets created using InfiniBand routers and demonstrated InfiniBand routing using BGFC (Bowman Global Fabric Controllers [5]);
- InfiniCloud: the first ever true high throughput global span HPC cloud allowing instances provisioning across four continents: Asia, Australia, North America and Europe [7–9].

During the last three years, InfiniCortex and numerous applications utilising this concept and infrastructure, have been successfully demonstrated at several international events: Supercomputing Frontiers 2015 and 2016 in Singapore; ISC15 and ISC16 in Frankfurt, Germany; TNC15 in Porto, Portugal and TNC16 in Prague, Czech Republic and finally at SC14 (New Orleans) and SC16 (Austin), USA.

Hence we have furnished a sufficient proof of concept demonstrations exhibiting the effectiveness of the proposed solutions. Several elements are already being implemented in Singapore and elsewhere as production solutions enabling higher bandwidth and security.

In the next section we will describe in some detail the third stage on our *InfiniCortex I* project which took place during the Supercomputing 2016 conference in Salt Lake City, USA. In Sub-Section 1.1 we will provide a list of all our collaborators in this project, followed in Sub-Section 1.2 with a description of a global scale network infrastructure, and, in Section 1.3, details of a number of application demonstrations prepared with our partners from the Oak Ridge National Laboratory, Fermilab, Stony Brook University, George Washington University, USA; University of Reims Champagne-Ardenne, France; Pozna? Supercomputing and Network Centre and Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Poland. In Section 2 we online our plans for InfiniCortex 2 phase of our project, and finally Section 3 contains conclusions of this report.

1. InfiniCortex Demonstrations at SuperComputing 2016

The *International Conference for High Performance Computing, Networking, Storage and Analysis - SC16*, the 28th annual international conference of high performance computing, networking, storage and analysis, celebrated the contributions of researchers and scientists – from those just starting their careers to those whose contributions have made lasting impacts. The conference drew more than 11,100 registered attendees and featured a technical program spanning six days. The exhibit hall featured 349 exhibitors from industry, academia and research organizations from around the world.

During the conference, Salt Lake City also became the hub for the world's fastest computer network: SCinet, SC16's custom-built network which delivered 3.15 terabits per second in bandwidth. The network featured 56 miles of fiber deployed throughout the convention center and 32 million dollars in loaned equipment. InfiniCortex is build on top of the SCinet network with support from the SCinet team and in collaboration with various netowrking orgnizations.

1.1. Partners

The following partners were involved in the InfiniCortex demonstrations at SC16:

- A*STAR Computational Resource Centre* - Singapore
- Oak Ridge National Laboratory (ORNL) - USA

- Fermilab - USA
- Stony Brook University (SBU)* - USA
- George Washington University (GWU)* - USA
- University of Reims Champagne-Ardenne (URCA)* - France
- Poznań Supercomputing and Network Centre (PSNC)* - Poland
- Interdisciplinary Centre for Mathematical and Computational Modelling (ICM)* - Poland

Locations marked with an asterisk denote locations where a Longbow InfiniBand range extenders were installed for the SC16 demo.

1.2. Network Infrastructure

The InfiniBand network ran on top of the dark-fibre network infrastructure prepared by A*CRC Network team in collaboration with various networking organisations (SingAREN, TEIN, GEANT, PIONEER, RENATER, Internet2, SCinet). A total of five E100 Longbows were used to connect the SC16 show floor to A*CRC in Singapore providing a 50Gbps InfiniBand link.

A global WAN InfiniBand network has been setup with 4 distinct subnets:

- National Supercomputer Centre, Singapore
- Interdisciplinary Centre for Mathematical and Computational Modelling (ICM), Poland
- A*CRC, Singapore + URCA, France + George Washington University, USA
- Stony Brook University, USA

using Obsidian's BGFC InfiniBand subnet manager [5] capable of InfiniBand routing between subnets.

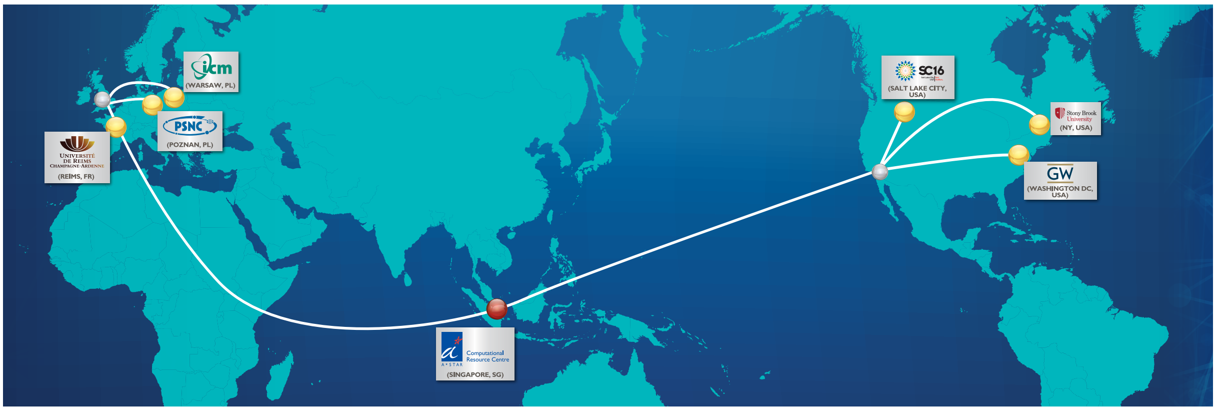
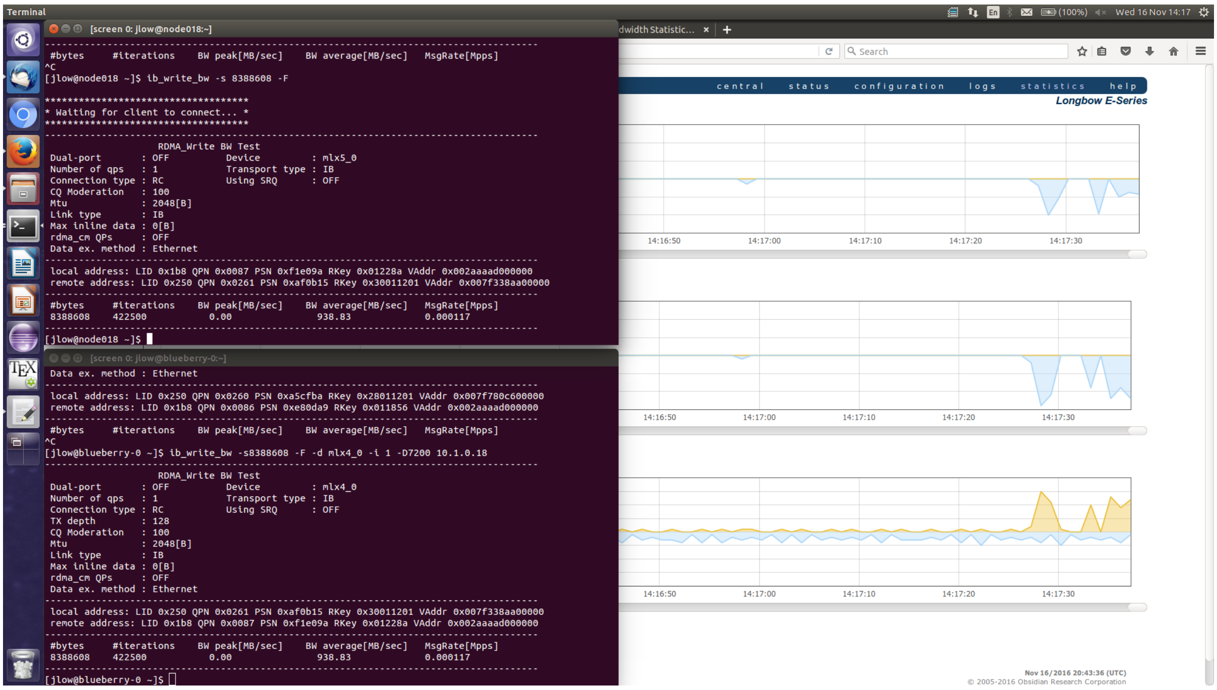


Figure 1. SC16 InfiniCortex global coverage map

1.2.1. Performance metrics

InfiniBand

- Each Longbow E100 link is capable of 10Gbps of InfiniBand traffic. All 5 usable Longbows were stress-tested for bandwidth. All five Longbows were pumping 10Gbps simultaneously - giving 50Gbps raw bandwidth (40Gbps usable data bandwidth due to 2 in every 10 bits for encoding).
- The bandwidth test of a single Longbow shows 938.83 MB/s between a server at SC16 and a remote server at A*CRC in Singapore.



```
---------------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[MB/sec]    BW average[MB/sec]   MsgRate[Mpps]
^C
[jlow@node018 ~]$ ib_write_bw -s 8388608 -F

************************************
* Waiting for client to connect... *
************************************
---------------------------------------------------------------------------------------
                    RDMA_Write BW Test
 Dual-port       : OFF          Device         : mlx5_0
 Number of qps   : 1            Transport type : IB
 Connection type : RC           Using SRQ      : OFF
 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 0x1b8 QPN 0x0087 PSN 0xf1e09a RKey 0x01228a VAddr 0x002aaaad000000
 remote address: LID 0x250 QPN 0x0261 PSN 0xaf0b15 RKey 0x30011201 VAddr 0x007f338aa00000
---------------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[MB/sec]    BW average[MB/sec]   MsgRate[Mpps]
 8388608    422500           0.00               938.83              0.000117
---------------------------------------------------------------------------------------
[jlow@node018 ~]$
```

```
 Data ex. method : Ethernet
---------------------------------------------------------------------------------------
 local address: LID 0x250 QPN 0x0260 PSN 0xa5cfba RKey 0x28011201 VAddr 0x007f780c600000
 remote address: LID 0x1b8 QPN 0x0086 PSN 0xe80da9 RKey 0x011856 VAddr 0x002aaaad000000
---------------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[MB/sec]    BW average[MB/sec]   MsgRate[Mpps]
^C
[jlow@blueberry-0 ~]$ ib_write_bw -s8388608 -F -d mlx4_0 -i 1 -D7200 10.1.0.18
---------------------------------------------------------------------------------------
                    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 0x250 QPN 0x0261 PSN 0xaf0b15 RKey 0x30011201 VAddr 0x007f338aa00000
 remote address: LID 0x1b8 QPN 0x0087 PSN 0xf1e09a RKey 0x01228a VAddr 0x002aaaad000000
---------------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[MB/sec]    BW average[MB/sec]   MsgRate[Mpps]
 8388608    422500           0.00               938.83              0.000117
---------------------------------------------------------------------------------------
[jlow@blueberry-0 ~]$
```

Figure 2. Global RDMA test between Singapore and Salt Lake City using Longbows E100

- An aggregated total bandwidth of 75 - 80 Gbps was utilised with link sharing with Tokyo-Tech University

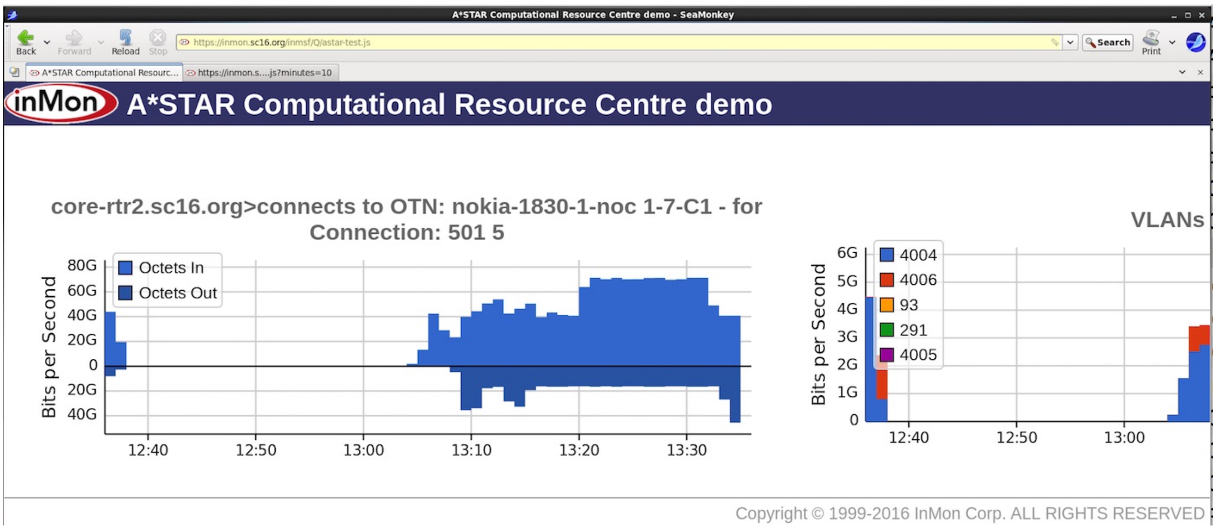

A*STAR Computational Resource Centre demo - SeaMonkey
Back
Forward
Reload
Stop
https://inmon.sc16.org/inmsf/Q/astar-test.js
Search
Print
A*STAR Computational Resourc...
https://inmon.s....js?minutes=10
inMon
A*STAR Computational Resource Centre demo
core-rtr2.sc16.org>connects to OTN: nokia-1830-1-noc 1-7-C1 - for Connection: 501 5
Bits per Second
80G
60G
40G
20G
0
20G
40G
Octets In
Octets Out
12:40
12:50
13:00
13:10
13:20
13:30
VLANs
6G
5G
4G
3G
2G
1G
0
4004
4006
93
291
4005
12:40
12:50
13:00
Copyright © 1999-2016 InMon Corp. ALL RIGHTS RESERVED

Figure 3. Screen-shot of the bandwidth utilisation during SC16 demos showing 75-80 Gbps data transfer between Tokyo University of Technology and A*CRC booth at the show floor

- The whole SC16 exhibit utilised just over 800 GBps bandwidth.
- A dsync+ test was attempted to transfer a large dataset from storage in A*CRC to SC16. The initial transfer was 6MB/s due to heavy packet loss on the link - despite no recorded packet loss during the bandwidth stress tests. There was no time left to diagnose the issue.

100G Ethernet

Andrew Howard from NSCC, Canberra, Australia conducted the following additional tests:

- iperf3 network test showed 16-23Gbps per UDP stream, 17.2Gbps for TCP.

- Lim Seng from A*CRC did further Ethernet bandwidth tests and was able to add additional bandwidth to the network.

1.3. Demonstrations

During SC16 A*CRC and several partners demonstrated five applications that were using the long range InfiniBand connectivity. The following sections will briefly describe the details and achievements of each applications.

1.3.1. Demonstrations with Oak Ridge National Laboratory (ORNL) / Fermilab / Stony Brook University (SBU)

The demonstration with ORNL and SBU called *Remote Fusion Experiment Data Analysis Through Wide-Area Network* consisted in remote data processing capability of large and high-throughput science experiment through cross-Pacific wide area networks and showed how one can manage science workflow executions remotely by using ORNL ADIOS data management system and FNAL mdtmFTP data transfer system. In this demonstration our partners presented a fusion data processing workflow, called Gas Puff Imaging (GPI) analysis, to detect and trace blob movements during fusion experiment. GPI data streams were being sent from Singapore to Fermilab for near-real time analysis, while ADIOS was managing analysis workflows and mdtmFTP transports stream data.

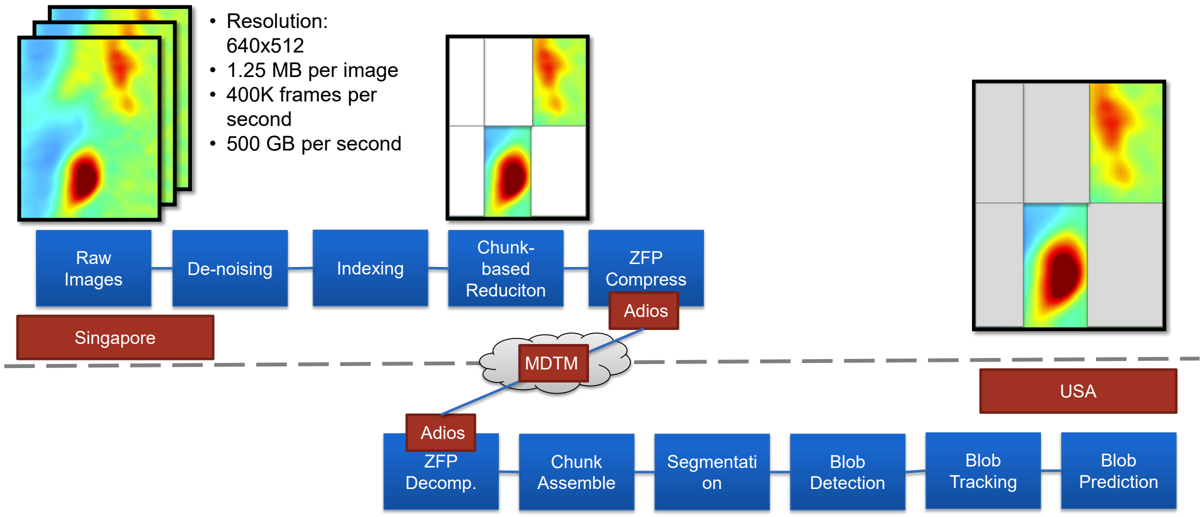


Figure 4. NSTX GPI Workflow With ADIOS and MDTM

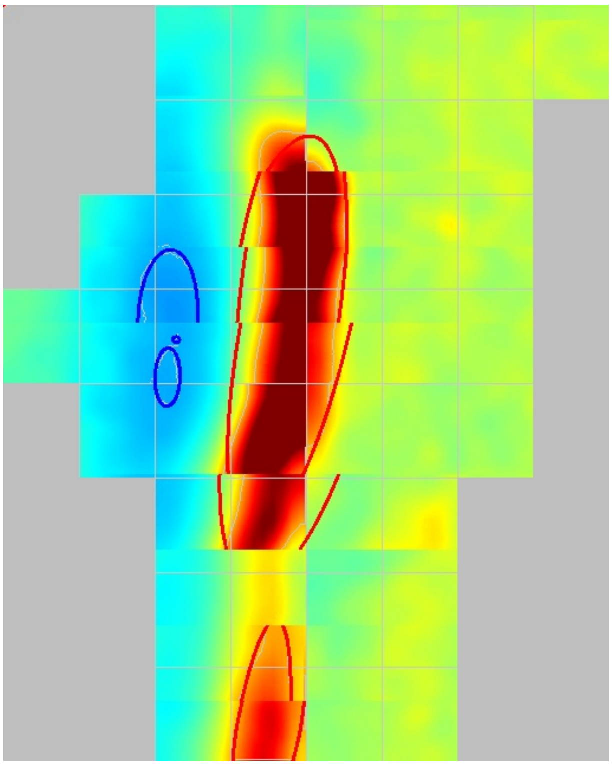


Figure 5. ORNL demonstration screenshot

Figure 5 shows a screenshot of the live demo running in SC16. The demonstration was susscesfully relying on the mdtmFTP data transfer system developed in Fermilab.

Accomplishments and problems encountered:

We encountered severals problems with this demo especially because the Stony Brook servers were available only a few days before SC16. Most of the tests have been done between Singapore and Fermilab, however even in this scenario a lot of problems arose from the fact this was not a dedicated L2 circuit and several firewalls were blocking the communication on each side.

Ultimately the problems have been solved. ORNL team have had plans to show this demo once again in 2017. The demo is part of a bigger collaboration between ORNL and Japan in the ITER project.

1.3.2. Demonstrations with George Washington University (GWU)

A Preliminary Study of Executing Parallel Applications over a Long-Range-IB network was showcased using mpiBLAST - a freely available, open-source, parallel implementation of NCBI BLAST. The mpiblast was run on multiple nodes on a cluster comprising in 4 nodes at GWU and 3 nodes at A*CRC and communicating was done via mpi/LHIB.

The experiment consisted in the following steps:

- A protein reference database (524603 protein entries, size of 153MB) was prepared and distributed to all of the compute nodes of the cluster (on both USA and Singapore).
- The database was fragmented into 64 smaller fragments by running the program: mpiformatdb.
- A subset of the protein sequences (from the reference database) was used for the protein blast search (blastp). The total number of sequences used is 7516.

- The total execution time of the blastp (protein blast search) was measured against the different number of compute nodes used. Two set of measured were conducted by using case1: compute nodes on USA only and case2: compute nodes on both USA and Singapore.
- The mpiblast command is as follows:

  ```
  mpirun -np 9 -mca btl openib,self -hostfile hostfile mpiblast -copy-via=cp
  /-concurrent=8 -use-parallel-write -query-segment-size=1000 -p blastp
  /-d AR.faa -i testInput.faa -o testOutput.txt
  ```

  where:
  –np changing from 9, 17, 33 (for 1 node, 2 nodes and 4 nodes)
  –concurrent changing from 8, 16, 32 (for 1 node, 2 nodes and 4 nodes)
  –copy-via specify to use system cp command to copy fragment database files onto nodes.
  –mca btl specifies to use openib as the communication protocol
  –hostfile specifies the host machines being used, for example:
  10.1.1.30 (node at USA)
  10.1.1.20 (node at Singapore)
  –AR.faa is the protein reference database
  –testInput.faa is the protein blast input sequence file
  –segment-size specifies the job size (no. of sequence send from master mpi process to the workers mpi processes to work; i.e. for controlling task granularity)

Preliminary results of the tests are shown in figures 6 and 7.

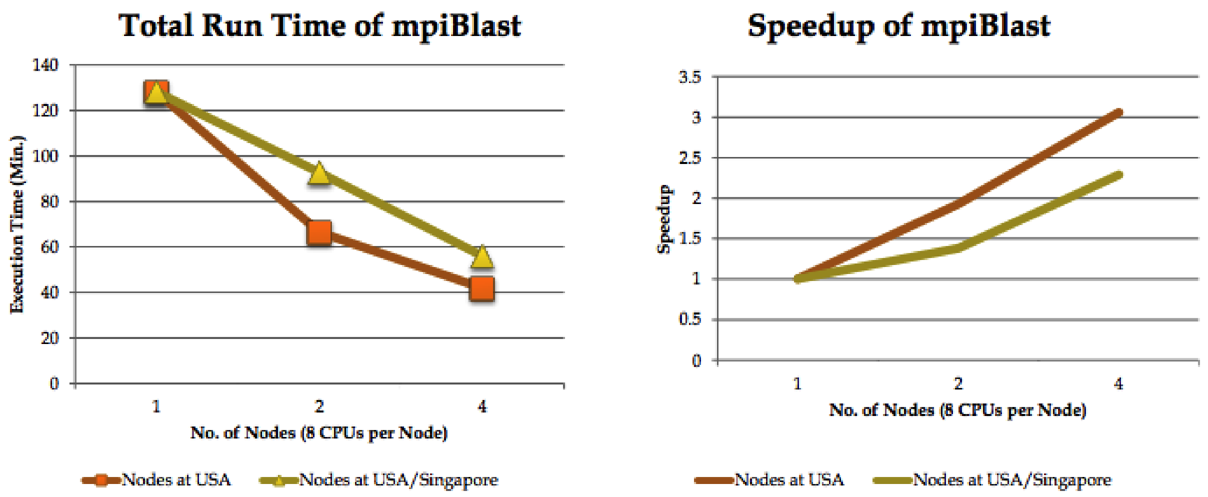


Figure 6. Input sequences: 7516

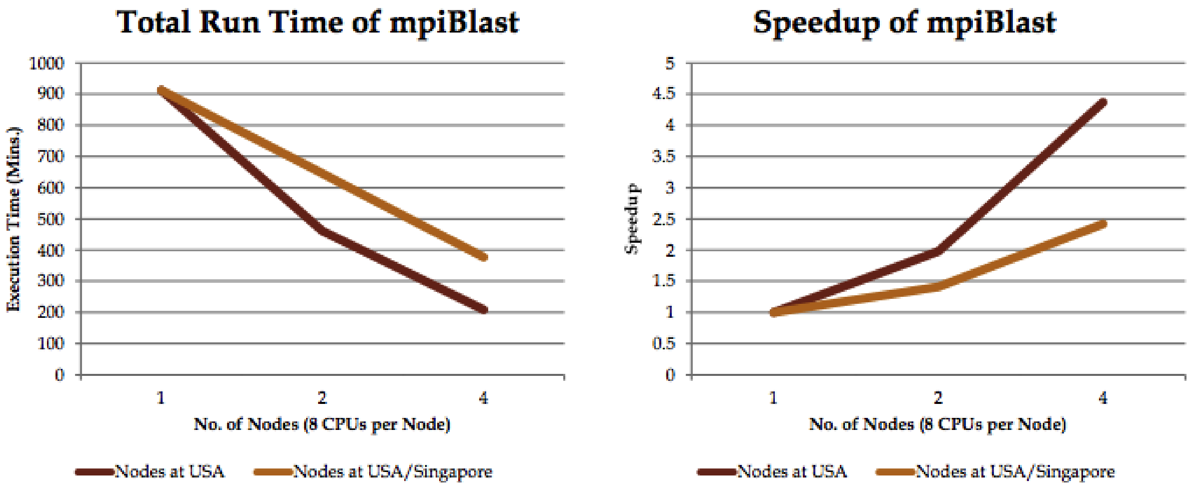


Figure 7. Input sequences: 48844

The following conclusions were drawn after analizing the results:

- Up to 32 individual tasks were running on different nodes. In figures 6 and 7, the computational scalability is shown with different problem sizes using different numbers of computing nodes, and also reveal the difference of localising the process on one side versus distributed process across InfiniCortex infrastructure.
- Super linear scalability is observed for the processes running only on the US nodes. Super linear scalability is unusual and there is not guarantee if cluster size is expended. The main reason behind this super linear scalability is unclear yet, but it may be because the overhead of the initial data distribution is smaller amongst the nodes connected within the same Infiniband switch.
- Basically, a linear scalability for the task distributed across InfiniCotex is observed, and it is the ideal case for a parallel computational process. So it should be considered as a successful demonstration.
- Many large scale scientific and engineering computational tasks can be divided into many small sub-tasks using data partitioning strategy, and then computed in parallel using MPI (Message Passing Interface) protocol to distribute tasks and data on different computer nodes. However, the efficiency of the rapid data exchange amongst the sub-task is very sensitive to the network latency, and thus certain computational tasks are inherently not scalable on the InfiniCortex infrastructure.
- To hide the inevitable latency due to the distance with a large data transfer, we have successfully demonstrated a number of workflows since SC14 (i.e. pipelining different stages of a task on the systems in different locations), but this is the first time we demonstrate solving a single computational task on two HPC clusters across continents using MPI. This specific application was succesfully run because it is inherently embarrassingly parallel, no data exchange required among the sub-tasks.
- Despite the success of this demonstration, we have encountered several difficulties:
  - OpenMPI was unable to build on the cluster on US side, and the issue was eventually resolved by the A*CRC software team. It was mainly due to the version of OpenMPI being too new for the building scripts that were used.

- We observed a big overhead (up to 2 minutes) for the initialisation of MPI if the job are distributed across InfiniCortex. It was confirmed that this overhead does not affect the accuracy of the computation, but the cause is not clear yet.
- Because the servers in US and SG were using different types of processors and additional time was necessary for tuning the data set between the two clusters. Such heterogeneity made difficult a fair comparison.

1.3.3. Demonstrations with University of Reims Champagne-Ardenne (URCA)

rCUDA (`http://www.rcuda.net`) is a development of the Parallel Architectures Group from Universitat Politecnica de Valencia (Spain). rCUDA enables the concurrent remote usage of CUDA-enabled devices in a transparent way. Thus, the source code of applications does not need to be modified in order to use remote GPUs but rCUDA takes care of all the necessary details. Furthermore, the overhead introduced by using a remote GPU is very small.

ROMEO HPC Center succesfully tested and run rCUDA inside their 260 GPUs cluster for the past year. The purpose of the SC16 demonstration was to test rCUDA over InfiniCortex and analyse the behaviour of the framework over extremely long distances. For the demo purpose the rCUDA server was installed in Singapore on a machine that didn't have any GPUs attached. The rCUDA client was installed on 18 servers in Reims each having 2 K20x. A standard matrix-matrix multiplication example from the CUDA SDK was then run on the Singapore machine which transparently sent all CUDA calls to the servers in Reims where the computation was actually taking place on the GPUs and then it was collecting the final result.

The rCUDA developers from Universitat Politecnica de Valencia developed a small graphical interface that was showing the performance of one single node compared to the performance of running the same code over several nodes. This was clearly showing the performance and benefits of running the rCUDA framework. The main advantage is that all the GPUs in the cluster are exposed as a big pool of resources for each node.

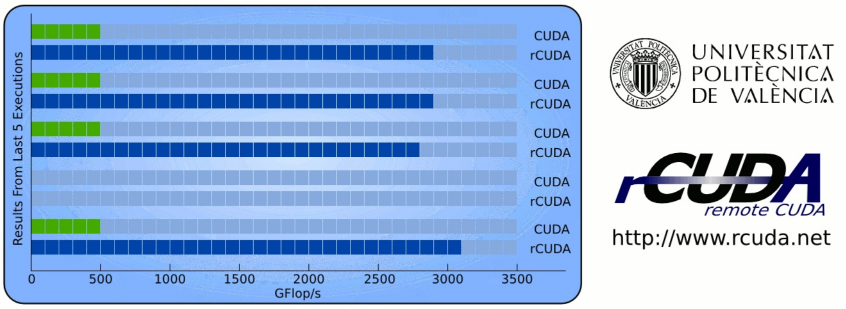


Figure 8. rCUDA performance running on 18 GPUs as compared to standard CUDA running on 1 GPU

The demo was eventuallly configured to run on IPoIB. For reasons that were not determined before the beginning of the conference the IB version of rCUDA was always freezing in an initialization stage. The presence of the Obsidian R400 router and BGFC and different OFED stacks were initially thought to be the problem. However even after the router was removed from the configuration and the OFED stacks synchronized the rCUDA was not able to work in native IB mode although other tests like ib_pingpong were succesful. rCUDA developers suggested that further tests are run after SC16 in order to determine why their framework is not able to function properly in a native long range IB setup.

1.3.4. Demonstrations with Poznań Supercomputing and Network Centre (PSNC)

Vitrall (`http://apps.man.poznan.pl/trac/vitrall-test`) is a distributed web based visualization system. It is under development at the Applications Department of the Poznań Supercomputing and Networking Center.

A typical scenario of Vitrall usage is a real-time visualization of a complex 3D content using remote servers equipped with modern multi GPU solutions, like nVidia Tesla. Following frames are compressed as JPEG pictures and sent over HTTP protocol to clients. Still, they may be displayed on an attached screen or projector. Users may watch the same content from different points of view simultaneously. Information about users' input is sent back to the Vitrall Visualization Server using WebSocket protocol – part of HTML5 specification.

Rendering process was distributed among two locations: Poznań and Singapore - every second frame will be rendered in Singapore, and every other frame in Poznań and then accessed by client through Singapore. At the exhibition floor in SC a regular web browser was used to connect to the Vitrall instances and and visitors were able to interact with the presented 3D scene providing a smooth animation. Web browser uses WebSocket to connect with Vitrall server controller instance and after a new rendering session is established, client starts to send input messages. Controller instance interprets those messages and continuously applies changes to the authoritative state of presented 3D scene. That state is then incrementally replicated to both rendering instances - only those need access to a GPU device. One such instance runs locally with the controller instance (in Singapore), and the other runs in Poznań. Controller instance decides which frame to render where, sends rendering requests to rendering instances and notifies the client where following frames will be available. The client then requests those frame using HTTP in the way that frame rendered in Poznań are accessed through Singapore.

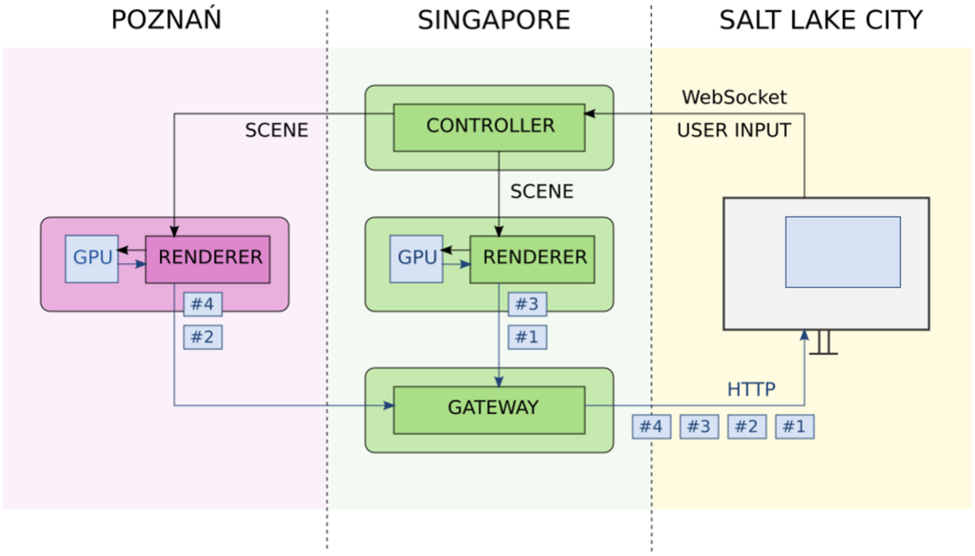


Figure 9. Vitrall distributed web based visualization system

Accomplishments and problems encountered:
The demo was succesfully run between PSNC and A*CRC without any major problems. Figure 10 shows a screenshot of the demo running in SC16. The performance was quite good as the interactivity with the scene was almost seameless. The only problem is that all servers involved in the demo requier quite powerfull GPUs.

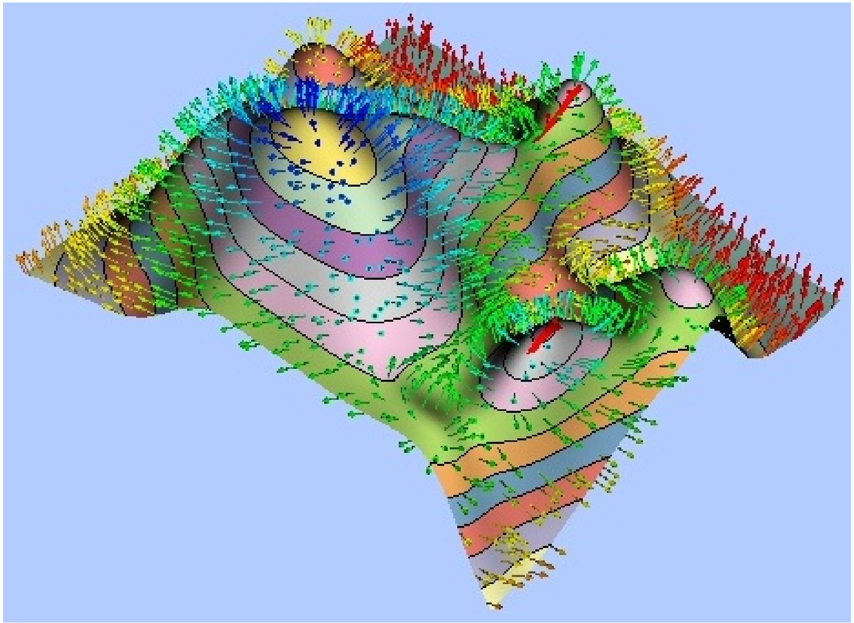


Figure 10. PSNC demonstration screenshot

1.3.5. Demonstrations with Interdisciplinary Centre for Mathematical and Computational Modelling (ICM)

The concept of demo was to show a basic proof-of-concept solution for globally remote visualization, where the simulation (or datasets), the visualization pipeline and the user are globally dispersed and connected only by a global network like InfiniCortex. A scientific example chosen for this demo was the numerical weather forecasting. The model itself (run by the Interdisciplinary Centre for Mathematical and Computational Modelling at the University of Warsaw (ICM)) is an iterative process of predicting an atmosphere state condition dynamics based on initial weather. In each step a significant number of data is created being a multivariate dataset on a three dimensional non-uniform grid over a modelled area. Observations of the evolution of a running simulation are possible by visualization of the consecutive iterations and provide the insight to the simulation. For the purpose of this a demo a dynamics of cloud coverage over central Europe was chosen. From the implementation perspective the demo consists of three layers: a simulation layer (or data layer), a visualization layer and end-user layer. The three layers were globally spatially disjoint and combined by a global interconnect. The simulation layer was physically located in Singapore (ACRC) and the running simulation was mimicked by incremental creation of new time step data files in the shared filesystem (each new file represents a simulation dump of a single iteration). The filesystem based on BeeGFS was remotely shared via InfiniCortex network and used by the visualization layer to access datasets. The visualization layer was physically located in Poland (ICM, Warsaw) and based on a dedicated visualization server running both the remote visualization middleware and the visualization software. VisNow (`http://visnow.icm.edu.pl`) was used as the visualization platform for implementation of the whole visualization pipeline. Visualization was created to show the orography of central Europe (static baseline layer) and the semi-transparent representation of cloud coverage was animated looping over the available iterations. A dedicated data access module in VisNow was monitoring the remote filesystem for presence of new time steps. Incremental dataset diffs were read in remotely via a shared filesystem. The remote visualization middleware was based on NICE Desktop

Cloud Visualization (DCV) platform, providing a remote desktop with dedicated data streaming for 3D OpenGL graphics and hardware compression. The end-user layer, being the interactive visualization, was physically located in the USA (SC16, Salt Lake City) and was running on a DCV thin client. On the DCV client-server path both InfiniCortex and Internet connections were tested.

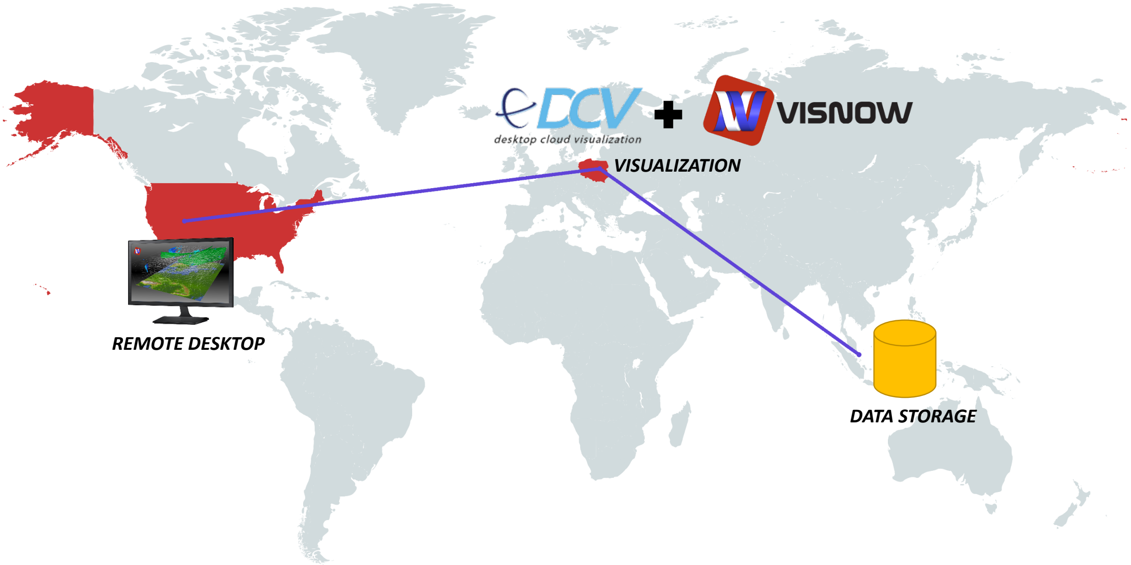


Figure 11. Remote visualization workflow

Accomplishments and problems encountered:

The proposed demo was successfully configured and run on a basic dataset of numerical weather forecast. As a proof-of-concept solution the demo showed the possible application of a globally connected supercomputer based on InfiniCortex network. At the same time novel knowledge was gathered on technical bottlenecks of the proposed visualization ecosystem and several concepts of improvement solutions were defined.

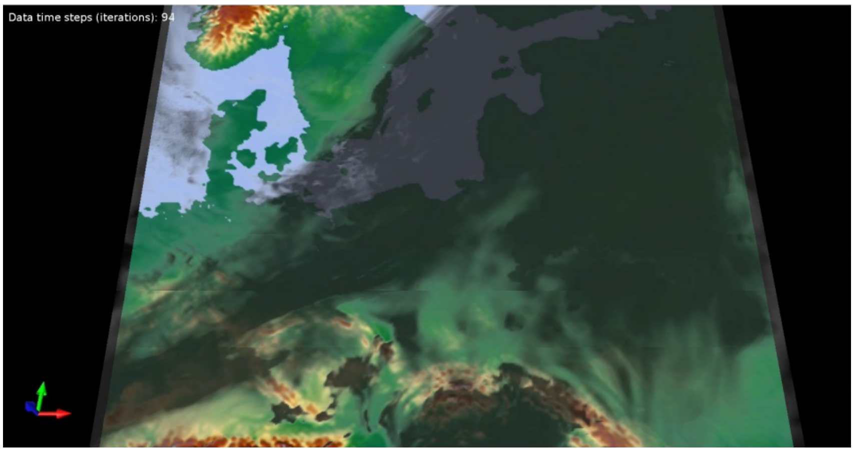


Figure 12. ICM demonstration screenshot

Some problems were encountered fdue to the fact that a BeeGFS parallel file system was being mounted over the InfiniCortex. This configuration was not tested and fine tunned previously leading to access time outs to the file system.

2. InfiniCortex Phase 2

Although over the last three years the InfiniCortex project ran more as a proof-of-concept to demonstrate several unprecedented features, currently it is entering phase "InfiniCortex 2" where many of the features demonstrated start being integrated in production systems creating symbiotic collaborations and developing new projects [11].

At the beginning of the project the international connectivity was obtained through the goodwill of several international carriers. Since 2015 Singapore has its own permanent links with US and Europe dedicated to research. The National Supercomputing Centre of Singapore (NSCC) currently funds the enhancement of the connectivity and through SingAREN MoUs have been signed for co-funding of permanent international links (100 Gbps towards US with Internet2 and 10Gbps towards Europe with TEIN*CC). These links will open the world and benefit the entire research community in Singapore, thus creating a symbiotic relationship between several local entities. The link to US set to open new oppotunities such as the participation in the Global Research Platform, an international extension of the Pacific Research Platform [3]. Similar international connectivity are currently being negotiated with Australia and Japan.

In the next years, NSCC and A*CRC are set to work on implementing a nation-wide InfiniBand fabric to interconnect several academic and industrial sites in Singapore which will provide high throughput, low latency direct connection to the supercomputing facilities in Fusionopolis. The first steps have already been made with the launch of the NSCC who has its main stakeholders campuses (NUS, NTU, GIS) connected through InfiniBand directly to the main supercomputer. This infrastructure provides researchers in remote campuses an unparalleled fingertip access to HPC resources. This initiative to allow a wide access to the HPC resources will continue in the future under NRF funding for National Research Infrastructure.

Genome Institute of Singapore (GIS) has the largest sequencing facility in South-east Asia at their facilities at the Genome Building, Biopolis. GIS relies on in-house storage as well as storage in Matrix Building Biopolis (100m away) and on storage and high performance compute facilities 2km away at the A*STAR Computational Resource Centre, Fusionopolis. Going forward, GIS will be processing up to thousands of exomes on a regular basis, processing capacity needs to be ramped up to cope with this demand of several Terabytes of data per day emerging from their sequencing labs. This means that in-house generated sequence data must be safely stored for data regulation compliance reasons, as well as transferred and stored at remote location pending computational processes such as the quality control step, read mapping step (high memory), variant call steps (embarrassingly parallel) and annotation steps involving different types of software with different hardware requirements. To avoid a data-bottleneck, we have constructed on top of standard TCP/IP network of A*STAR's next generation ExaNet a 500Gbps Infinera CloudExpress 1 Ethernet link plus a Mellanox MetroX and Obsidian Longbow InfiniBand interconnections between Biopolis and Fusionopolis. The 500 Gbps link runs over a dedicated dark fibre and is the fastest point-to-point link in Asia, and the fastest known link in the world dedicated for Next Generation Sequencing (NGS) analytics. By 2017 the whole bandwidth capacity between Biopolis and Fusionopolis with exceed 1.2 Tbps. In 2016, a 1 Terabyte RAM node was installed in GIS Biopolis linked by InfiniBand to the new NSCC supercomputer with 10 PByte storage (up to 500Gbytes/sec I/O using DDN's state-of-the-art Infinite Memory Engine, accessing a dedicate genome analytic queue on the 1PFLOPS NSCC supercomputer at Fusionopolis. This will become one the world fastest and biggest distributed genome processing system and will ensure that genome analytics can be scaled up to thousands of genomes to be processed routinely per month for biomedical research which will be a current practice in the framework of projects like Genome Asia 100k.

3. Conclusions

InfiniCortex project started by A*CRC under leadership of Dr Marek Michalewicz in 2014 has gained a lot of interes both in Singapore and abroad. The A*CRC group which was responsible for the InfiniCortex project was recognized through several awards during the past few years, the most prestigious one being the *Innovative Project Gold Award* from the Ministry of Trade and Industry of Singapore in 2015 [1]. InfiniBand connectivity is still regarded as a high-end HPC oriented interconnect however features such as high-bandwidth and security which are demonstrated advantages over the classic TCP/IP are now recommending it for production environments outside the walls of a datacentre.

InfiniCortex project could serve as a very useful prototypical infrastructure for a number of Big Scientific Data projects currently being developed: i) distribution of data to the international partners from the Square Kilometre Array [4], ii) streaming of large facilities experimental data through the Pacific Research Platform collaboration of DoE, ESnet and other partners in the US and elsewhere [3], iii) the Superficilities vision expressed by DoE [6], and iv) new architecture for CERN LHC data processing pipeline focussing on more powerful processing facilities connected by higher throughput connectivity [13] .

Data connectivity between key HPC centres and countries has been defined as one of the priority areas of newly established EuroHPC programme. *"HPC is developing to cope with the constant increase in data volumes and flows. A recent report projects that annual global IP traffic will reach 2.3 zettabytes by 2020 – or 504 billion DVDs per year."* [2]

The authors are firmly convinced that *InfiniCortex I* project provides very well defined and tested path to realising some of the goals of EuroHPC connectivity agenda.

4. Acknowledgements

Authors would like to acknowledge all the partners around the world that supported and worked hard to make the InfiniCortex project a success.

USA:

ORNL: Scott Klasky, Jong Youl Choi, Matthew Wolf, Dave Pugmire, Manish Parashar

Fermilab: Wenji Wu, Jason Wang, Liang Zhang

Stony Brook University: Robert Harrison, Firat Coskun, Behzad Barzideh, Michael White, Yuefan Deng, Tahsin Kurc

George Washington University

Adam Kai Leung Wong, Jaroslav Flidr, Donald DuRousseau, Andrew Gallo

France: University of Reims Champagne-Ardenne

Arnaud RENARD, Michaël Krajecki

Spain: Universitat Politècnica de València

Carlos Reaño, Federico Silla

Poland:
Poznań Supercomputing and Network Centre (PSNC): Artur Binczewski, Tomasz Szewczyk, Tomasz Piontek, Piotr Śniegowski
Interdisciplinary Centre for Mathematical and Computational Modelling (ICM), University of Warsaw: Marek Michalewicz, Bartosz Borucki, Konrad Bierzuński, Jaroslaw Skomial
Commercial Partners
SingAREN: John Kan, Francis Lee, Lawrence Wong
Obsidian: David Southwell, Jason Gunthorpe, Bill Halina
DDN: Atul Vidwansa, Susan Presley
Huawei: Chenxingying Nick, Robin Shi Bin

References

1. A*CRC team received five awards in 2015:. Singapore Ministry of Trade and Industry Awards 2015: Innovative Project Gold Award 2015; A*STAR Awards 2015: STAR Innovation Award 2015; FutureGov Singapore Award: Technology Leadership Award; CIO Asia 100: Honouree 2015; Singapore Public Service Most Innovative Project: Merit Award 2015.

2. EuroHPC. https://ec.europa.eu/digital-single-market/en/news/eu-ministers-commit-digitising-europe-high-performance-computing-power. Last accessed: May 10, 2017.

3. Pacific research platform. http://prp.ucsd.edu/. Last accessed: May 10, 2017.

4. Square kilometer array. http://skatelescope.org/signal-processing/. Last accessed: May 10, 2017.

5. Obsidian introduces the Bowman Global Fabric Controller. http://www.obsidianresearch.com/archives/all/2015/Bowman-Global-Fabric-Controller.html, November 2015. Last accessed: May 10, 2017.

6. Kate Antypas. Superfacility: How new workflows in the DOE Office of Science are influencing storage system requirements. http://storageconference.us/2016/Slides/KatieAntypas.pdf, May 2016.

7. Kenneth Hon Kim Ban, Jakub Chrzeszczyk, Andrew Howard, Dongyang Li, and Tin Wee Tan. Infinicloud: Leveraging the global infinicortex fabric and openstack cloud for borderless high performance computing of genomic data. *Supercomputing Frontiers and Innovations*, 2(3):14–27, 2015. DOI:10.14529/jsfi150302

8. Jakub Chrzeszczyk, Andrew Howard, Andrzej Chrzeszczyk, Ben Swift, Peter Davis, Jonathan Low, Tin Wee Tan, and Kenneth Ban. Infinicloud 2.0: distributing high performance computing across continents. *Supercomputing Frontiers and Innovations*, 3(2):54–71, 2016. DOI:10.14529/jsfi160204

9. Jonathan Low, Jakub Chrzeszczyk, Andrew Howard, and Andrzej Chrzeszczyk. Performance assessment of infiniband hpc cloud instances on intel haswell and intel sandy bridge architectures. *Supercomputing Frontiers and Innovations*, 2(3):28–40, 2015. DOI:10.14529/jsfi150303

10. Marek Michalewicz, David Southwell, Tin Wee Tan, Yves Poppe, Scott Klasky, Yuefan Deng, Matthew Wolf, Manish Parashar, Tahsin Kurc, C.S. Choong-Seock Chang, Satoshi Matsuoka, Shin'ichi Muira, Jakub Chrzęszczyk, and Andrew Howard. InfiniCortex: concurrent supercomputing across the globe utilising transcontinental InfiniBand and Galaxy of Supercomputers. *Supercomputing 2014: The International Conference for High Performance Computing, Networking, Storage and Analysis, At New Orleans, LA, USA*, November 2014.

11. Marek T Michalewicz, Tan Geok Lian, Lim Seng, Jonathan Low, David Southwell, Jason Gunthorpe, Gabriel Noaje, Dominic Chien, Yves Poppe, Jakub Chrzęszczyk, et al. Infinicortex: present and future invited paper. In *Proceedings of the ACM International Conference on Computing Frontiers*, 267–273. ACM, 2016. DOI:10.1145/2903150.2912887

12. Łukasz Orłowski, Yuefan Deng, and Marek Michalewicz. Galaxies of supercomputers and their underlying interconnect topologies hierarchies. In *International Supercomputer Conference, Leipzig, Germany*, 2014.

13. Gianfranco Sciacca. Big Data Science Accessing High-End HPC. https://www.digitalinfrastructures.eu/sites/default/files/LHConCRAY-DI4R2016-v2.pdf, October 2016.