DOI: 10.14529/jsfi200202

Long Distance Geographically Distributed InfiniBand Based Computing

***Karol Niedzielewski*[1], *Marcin Semeniuk*[1], *Jarosław Skomiał*[1], *Jerzy Proficz*[2], *Piotr Sumionka*[2], *Bartosz Pliszka*[2], *Marek Michalewicz*[1]**

Collaboration between multiple computing centres, referred as federated computing is becoming important pillar of High Performance Computing (HPC) and will be one of its key components in the future. To test technical possibilities of future collaboration using 100 Gb optic fiber link (Connection was 900 km in length with 9 ms RTT time) we prepared two scenarios of operation.

In the first one, Interdisciplinary Centre for Mathematical and Computational Modelling (ICM) in Warsaw and Centre of Informatics – Tricity Academic Supercomputer & networK (CI-TASK) in Gdańsk prepared a long distance geographically distributed computing cluster. System consisted of 14 nodes (10 nodes at ICM facility and 4 at TASK facility) connected using InfiniBand. Our tests demonstrate that it is possible to perform computationally intensive data analysis on systems of this class without substantial drop in performance for a certain type of workloads. Additionally, we show that it is feasible to use High Performance Parallex [1], high level abstraction libraries for distributed computing, to develop software for such geographically distributed computing resources and maintain desired efficiency.

In the second scenario, we prepared distributed simulation - postprocessing - visualization workflow using ADIOS2 [2] and two programming languages (C++ and python). In this test we prove capabilities of performing different parts of analysis in seperate sites.

Keywords: HPC, distributed computing and systems, InfiniBand, federated supercomputing, geographically distributed workflows, ADIOS, HPX, High Performance Parallex.

Introduction

Growth of computing capabilities in connection with big data manipulation and analysis gives us new tools for broadening knowledge and bringing new scientific breakthroughs. However new possibilities introduce new challenges. Great scale of stored information requires new approches to data storage and manipulation. Sometimes data movement between data centres requires a lot of time (measured in days or months) or is even impossible because of property rights (the data is owned by one entity and cannot be shared). Such cases occur more and more frequently and demand close cooperation between data centres. Collaboration between multiple computing centres is referred to as federated computing and will be one of the key components of High Performance Computing (HPC) in the future.

Between 2014–2016, A*STAR Computational Resource Centre (A*CRC) in Singapore engaged in exploration of long-range InfiniBand technology to build globally distributed concurrent computing system called InfiniCortex. These exploration led to integrating computing resources over four continents and six countries connected with RDMA enabling InfiniBand fabric [3–6].

The technology for long-haul global reach extended InfiniBand has been created by two companies: Obsidian Strategics, a Canadian company [7, 8], which apparently is not in operation anymore, and Bay Microsystems which was recently bought by Vcinity [9].

Mellanox Technologies built MetroX InfiniBand long-haul extenders [10], but they have limited range of about 40 km.

[1]Interdisciplinary Centre for Mathematical and Computational Modelling (ICM), University of Warsaw, Warsaw, Poland

[2]Centre of Informatics – Tricity Academic Supercomputer & networK (CI TASK), Gdańsk University of Technology, Gdańsk, Poland

The extended range InfiniBand has been initially used for remote storage, and for moving very large data (so-called "Large Data") between the sites. In 2007 it was reported that *"Obsidian Research Corporation's Longbow Campus products have enabled NASA to relocate 15 percent (1,536 processors) of its high-ranking SGI Altix-based Columbia Supercomputer to another facility and connect both locations without any performance degradation."* [11].

The early instances of long range InfiniBand connectivity were implemented at:

1. NASA: connecting Pleiades at NASA Ames Research Center and Hyperion elements at Lawrence Livermore National Laboratory with Obsidian Longbow extenders.
2. NASA: secure wire-speed storage synchronisation between NASA Ames Research Center in California and NASA Goddard Space Flight Center in Maryland.
3. Arizona State University testbed [12].
4. Swiss Supercomputer Center: *"We evaluated the Obsidian Longbow InfiniBand Range Extender with the overall goal to ensure continuous availability of GPFS through the complete CSCS relocation period by running one single GPFS file system over both sites. The geographical distance between the current and the future location is about 3 km, the measured distance of dark fiber is 10 km. The evaluation results for the range extender are encouraging and are in line with our expectations and requirements."* [13].
5. IT Centers at the Heidelberg to Mannheim Universities [14].

InfiniCortex built by A*CRC over three year period 2014–2016 was by far the largest and the most extensive, global scale InfiniBand distributed concurrent computing system ever built. A notable application created on the top of InfiniCortex was InfiniCloud – a globally distributed cloud infrastructure used to run cancer mutation calling pipeline over four continents [15, 16].

Recently an idea of Superfacilities was formulated within the US Department of Energy Labs. Superfacilities would encompass supercomputing resources with large scale data storage, large scale experimental facilities, mathematical methods, software and human expertise – and, of course, with all infrastructure elements connected with super-efficient network fabric [17–19].

In the words of Gregory Bell, with creation of Superfacilities *"Scientific progress will be completely unconstrained by the physical location of instruments, people, computational resources, or data."* [20].

It should be noted, that InfiniCortex created several years earlier was a precursor of a DoE defined Superfacility. European prototype, named Fenix Infrastructure is currently being built by five major supercomputing centres [21].

Based on the success and experiences of the global scale InfiniCortex infrastructure, Singapore implemented a country-wide STAR-N Singapore InfiniBand Fabric connecting Nanyang Technological University, Singapore National University and A*STAR into one 100 Gbps InfiniBand network. The fabric is based on the shorter range Mellanox MetroX extenders. It allows for easy access to ASPIRE Supercomputer based at the National Supercomputer Centre at the A*STAR central location through login nodes at remote locations, and efficient data transfer between the sites [22].

The main objectives of our project, reported here, were to:

- establish the first long-haul InfiniBand connection between two Polish HPC centres, which will serve as the first step towards federating all Polish HPC centres;
- test Vcinity 40 Gbps long-range InfiniBand technology over distance of 900 km;
- run High Performance ParalleX enabled application over long-haul distributed network;
- test ADIOS workflows over this distributed infrastructure.

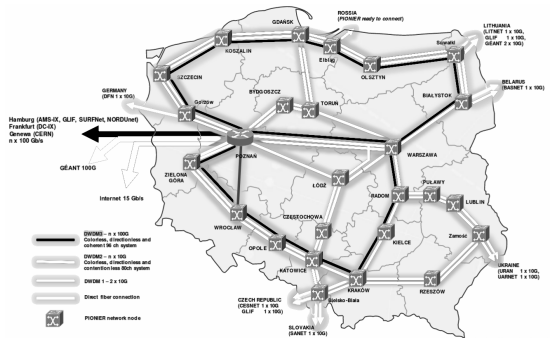
To test technical possibilities of future collaboration, ICM and TASK teams decided to test 40 Gb InfiniBand connection over optic fiber link in various scenarios.

The first step was to prepare a long distance geographically distributed computing cluster and to examine its data analysis capabilities. We demonstrate possibilities of using high level abstraction libraries for distributed computing (High Performance Parallex) to develop software for such clusters. What is more we show that some of the workflows (with low comunication requirements) can perform without drop in performance on such distributed clusters. More comprehensive tests involving MPI all-reduce algorithms on this distributed computing cluster were presented in a separate conference report [23]. The second test was focused on distributing different parts of data analysis workflow between seperate sites. Here we show that it is feasible to implement efficient distributed workflows using geographically distributed hardware configuration.

The paper is organized as follows. In Section 1 we describe our distributed computing cluster in two separate locations. It includes hardware, software and storage specification. In Section 2 we present the results of data analysis performed using geographically distributed computing cluster. Section 3 presents the capabilities of distributed simulation-postprocessing-visualization ADIOS workflow on the distributed infrastructure. The last section, Conclusions, contains a summary of the study and provides some hints to the future activities.

1. Testbed

For testing purposes ICM in collaboration with TASK, prepared a distributed computing cluster that consisted of the nodes located at ICM datacenter in Warsaw, and some nodes at TASK datacenter in Gdańsk. Facilities which are about 350 km apart were connected using Pionier academic network fiber optic link running in a round-about way over ∼900 km path (Fig. 1a and 1b).



(a) Map of Poland with Pionier network architecture and 100 gb links marked in black

(b) Zoom in on the map of Poland with Pionier network architecture. Link used in tests is marked in black

Figure 1. Pionier network architecture map (courtesy of Artur Binczewski, PSNC, Pozna)

1.1. Hardware

There were 10 compute nodes at ICM site. Each node was a dual socket HUAWEI RH-1288 v3 server with two Intel E5-2680 v3 CPUs, four 6 TB SATA drives and 128 GB DDR4 RAM. Each CPU has 12 cores and operate at 2.50 GHz clock frequency. At TASK facility there were four nodes. Each diskless node was HPE ProLiant XL230a Gen9 server with two Intel E5-2670 v3 processors and 128 GB DDR4 RAM. Each CPU has 12 cores operating at 2.3 GHz clock frequency.

InfiniBand interconnect in both clusters consisted of Mellanox SX6025 – InfiniBand switching system with 36 (FDR) 56 Gb/s ports and 4 Tb/s aggregate switching capacity. Each server was equipped with Mellanox FDR (56 Gb/s) Connect-X3 interface card used for the InfiniBand link and 1GE link for management. InfiniBand Extenders used in this tests were IBEX G40 – QDR InfiniBand RDMA based Extension Platform[3] and each was equipped with one QDR InfiniBand interface and one 40 GE port. IBEX G40 form factor is 1U rack unit and its power consumption is less than 140 W. Total buffer capacity allows extending InfiniBand connection up to 15,000 km.

The 100 GE circuit spanned between Warsaw and Gdańsk is routed via Balystok and was delivered by Pionier Polish National Research and Education Network in cooperation with Poznan Supercomputing and Networking Center. The ∼900 km long circuit introduces 9 ms RTT latency that is consistent with theoretical results calculated using (1):

$$ping = \frac{l}{V_{glass}} = \frac{1800\ km}{200\ \frac{km}{ms}} = 9\ ms, \tag{1}$$

where l – length of optic fiber connection, V_{glass} – velocity of light in glass.

Storage was located at ICM and shared using Network File System (NFS) technology, therefore nodes on TASK side had to download dataset before analysis.

1.2. Software

For testing purposes we decided to use Multidimensional Feature Selection algorithm implemented together with High Performance Parallex (HPX) [24]. We chose this application because of its very good parallel scaling on our Okeanos Cray XC40 supercomputer (each node equipped with 24 Intel Xeon E5-2690 v3 cpu cores). The scaling results are presented in Fig. 2. We can see that analysis of Madelon dataset exhibits almost perfect parallel scaling up to 64 nodes (1,536 cores), and then deteriorates due to the size of the problem being too small (starvation). Possible applications of Multidimensional Feature Selection exhaustive search include many domains of science such as genomics, economics, social sciences and others. Full details of this work can be found in [24].

For MPI connectivity openMPI v3.1.4 [25] was used. HPX was built using this openMPI library. MPI processes count was equal to the number of used computing cores (number of cores * number of nodes).

Tests on a distributed cluster were performed using Madelon dataset. Madelon [26] is a synthetic dataset with 2,000 objects and 500 variables that can be accessed from the UCI Machine Learning Repository [27] that was prepared in csv format. Data was located on ICM side, therefore nodes on TASK side had to download dataset before analysis. Jobs were invoked on ICM side, therefore latency of the excecution on TASK side was ∼5 ms caused by the connection latencies.

[3]The test InfiniBand Range Extenders were provided by Vcinity, Inc. and 2CRSI SA.

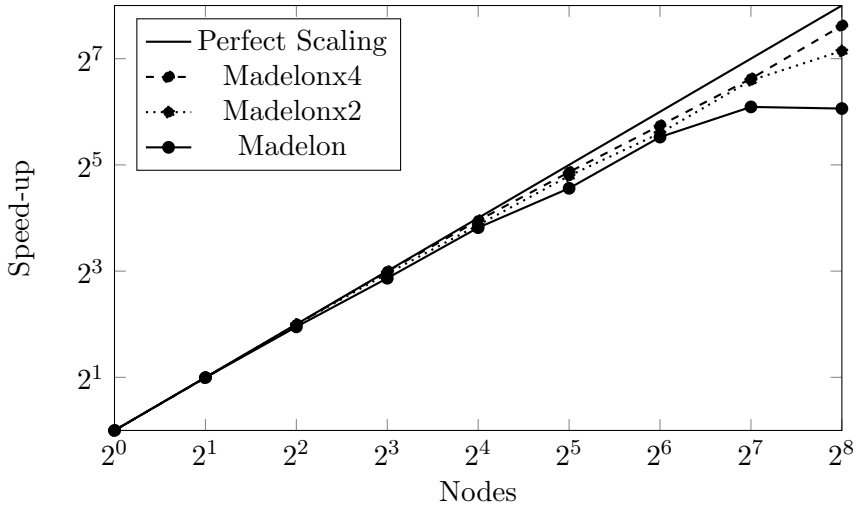


Figure 2. Measured speedup of 3-Dimensional analysis with 100 discretizations on different Madelon dataset sizes and with first algorithm implementation [24] (e.g. dataset that is twice as big has twice the number of variables. Additional variables are copies of variables from the original data set). Each node was equiped with 24 Intel Xeon E5-2690 v3 cores

2. Results

We tested the first implementation of MDFS [24] on groups of nodes of varying sizes and locations. Scenarios were prepared so that the amount of work was split evenly between sites (ICM and TASK) or was performed on nodes located only at one site.

We decided to perform 2-Dimensional analysis tests because it is the minimal size of the problem that fits well on up to 4 nodes. The measured time of the analysis performed on different configurations of nodes is presented in Fig. 3 and in Tab. 1. Speed-up of analysis is seen in Fig. 4 and the Tab. 2.

The results are presented using following coding (2):

$$G[Number\ of\ nodes\ on\ TASK\ side\ (Gdansk)]W[Number\ of\ nodes\ on\ ICM\ side\ (Warsaw)]$$
$$Example:\ G2W3\ -\ 2\ nodes\ on\ TASK\ sideand\ 3\ nodes\ on\ ICM\ side. \quad (2)$$

Table 1. Measured time of Madelon dataset analysis at ICM-TASK

Nodes configutation	Number of nodes	Measured time [s]
G1W0	1	66.4
G0W1	1	60.9
G2W0	2	33.4
G1W1	2	33.5
G0W2	2	30.8
G2W2	4	16.9
G0W4	4	16.0

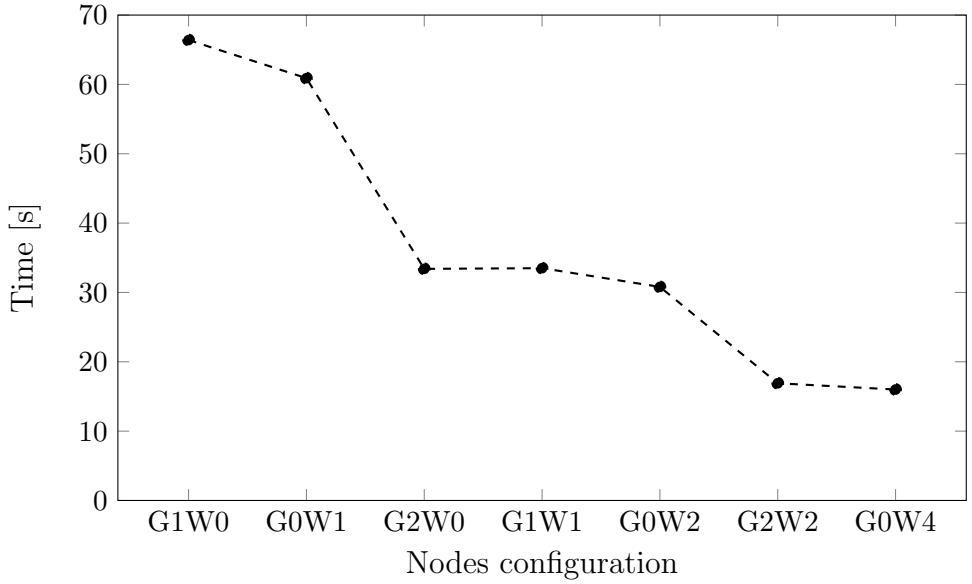


Figure 3. Measured time of madelon dataset analysis at ICM-TASK

Table 2. Measured speedup of madelon dataset analysis at ICM-TASK

Nodes configuration	Number of nodes	Measured speedup
G1W0	1	0.9
G0W1	1	1
G2W0	2	1.8
G1W1	2	1.8
G0W2	2	1.9
G2W2	4	3.6
G0W4	4	3.8

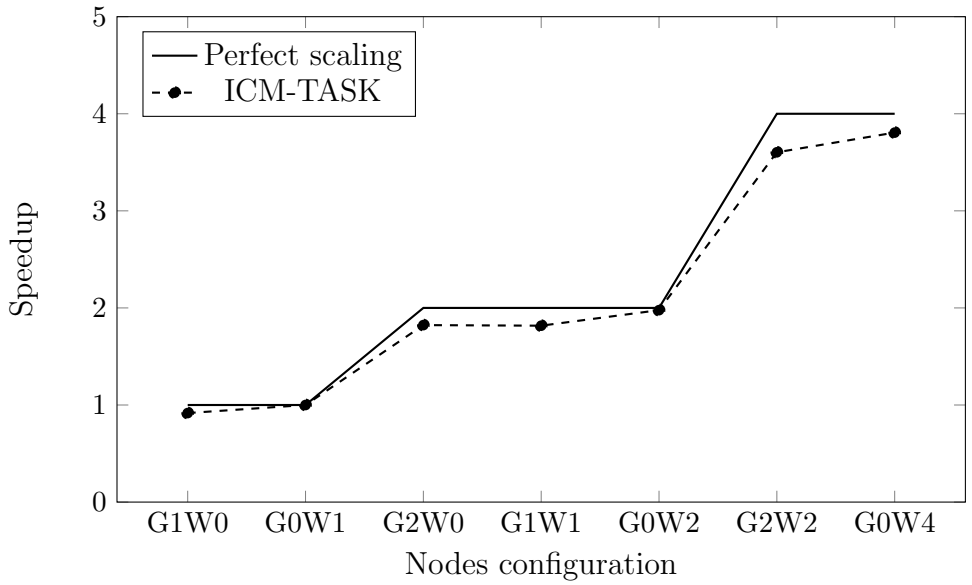


Figure 4. Measured speedup of madelon dataset analysis at ICM-TASK

It is clear that location of computations affects time of the analysis. Please see Tab. no. 1. Speedup changes are the outcome of analysis time changes (Tab. 2). Computations at ICM are faster (better performance) for the following reasons:

1. Jobs were invoked on the ICM side, therefore the excecution is delayed as well as receiving of the results is delayed. Globally the latency will be at least ∼9 ms because of the connection latencies.
2. Data was located on the ICM side, therefore nodes on the TASK side had to download dataset using NFS before analysis. Here again we observe minimum ∼9 ms latencies.
3. Nodes on the ICM side and the TASK side were equiped with different hardware (CPUs, RAM, Network card, etc.). This results in different computation times which are slower on the TASK side.

Nevertheless, location of the computations affect analysis time (performance) no more than 10 % and could be reduced by selection of optimal load balancing (less computations on TASK side). This brings us to conclusions that the differences of analysis time are not significant. Adventages (speedup) of computations on 'distributed' cluster overcome disadventages and can be beneficial in the future. We can observe linear scalability of MDFS method up to 4 nodes.

3. Simulation - Postprocessing - Visualization Distributed Workflow

We prepared simple simulation - postprocessing - visualization distributed workflow using the Gray-Scott MiniApp [28] and ADIOS 2 (version 2.4.0). ADIOS 2 (The Adaptable Input Output System version 2) is a framework dedicated for data I/O to write and read data when and where required. Its design introduces new approach to high level API that allows easy building of the data dependencies between components of applications. Important feature is possibility to build dependencies in distributed manner that makes ADIOS really interesting and powerful tool.

ADIOS2 remote IO between ICM and TASK was based on RDMA connection using SST files. This approach allows efficient reads of remote files and synchronous staging of sequences of simulation steps. Therefore none of the steps of the simulation data was omitted during postprocessing and visualization.

Distributed workflow is presented in Fig. 5 where we can see its several components written in C++ and python:

1. Gray-Scott (C++) – 3-D simulation of Gray-Scott reaction diffusion model [29] (using 4 mpi processes). Simulation and staging of data is run at the TASK site.
2. PDF Analysis (C++) – postprocessing of simulation data that prepares pdf images (using 1 mpi process). Run at the ICM site.
3. 2-D visualization (python) – 2-D cross section visualization of 3-D simulation (using 1 mpi process). Run at the ICM site. Example frame can be seen in Fig. 6a.
4. PDF ploting (python) – visualization of the plots from PDF Analysis (using 1 mpi process). Run at the ICM site. Example frame can be seen in Fig. 6b.

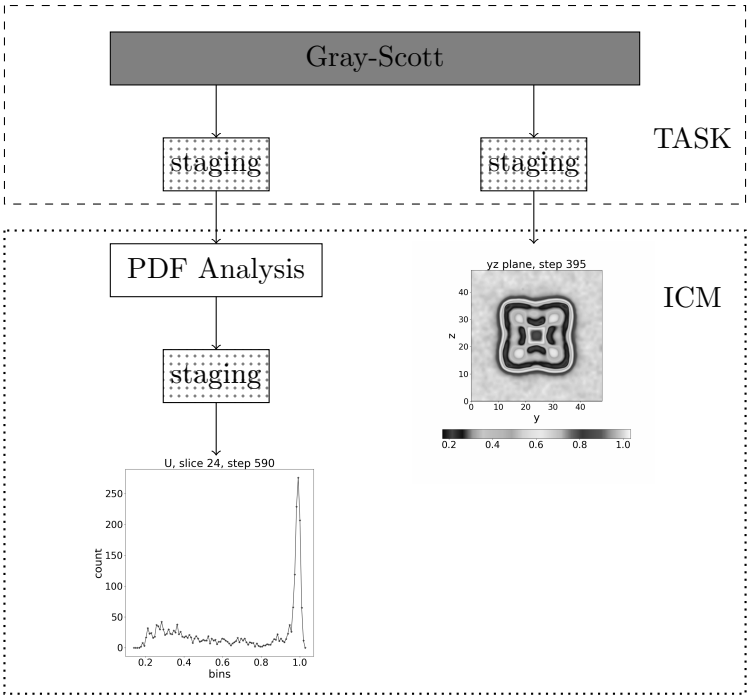
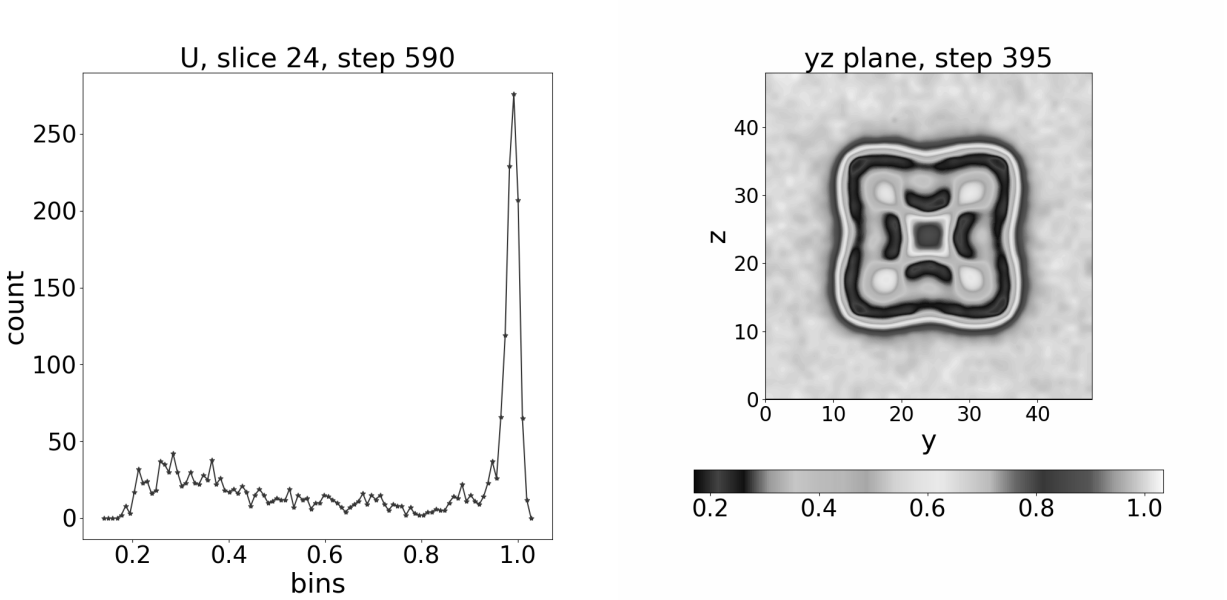


Figure 5. Workflow diagram



(a) Histogram of U in simulation of Gray-Scott reaction diffusion model

(b) 2-D cross section of 3-D simulation of Gray-Scott reaction diffusion model

Figure 6. Visualization examples

Conclusions

Our tests demonstrate that it is possible to perform computationally intensive data analysis on long distance geographically distributed computing cluster without substantial drop in performance. Additionally, we demonstrate that it is feasible to use high level abstraction libraries for distributed computing, such as High Performance Parallex, to develop software for

geographically distributed clusters and to maintain computational performance comparable to a cluster in a single location. Moreover our application has potential to be used in many domains such as genomics, economics or social sciences. Our approach is not limited to feature selection methods and can be applied to many other data analysis and machine learning workflows that have low communication requirements.

In second test we present capabilities of using simulation - post-processing - visualization distributed workflow to execute parts of application in geographically separated sites. As a consequence, it opens new ways for sharing of data and distributing various components of applications.

Our successful tests of the connection between ICM and TASK present new technical possibilities and potential benefits of future collaboration between computing centres and federated computing in general.

Furthermore, presented solutions can be widely used and are not limited to the two centres listed above. We envisage a Polish InfiniCortex federating all six top Polish HPC centres into the Polish National (Distributed, Concurrent) Supercomputer utilising the Pionier fibre-optic fabric and six new generation InfiniBand range extenders offering 100 Gbps bandwidth and unlimited range.

Acknowledgements

This research was carried at the Interdisciplinary Centre for Mathematical and Computational Modelling (ICM), University of Warsaw, and at the Centre of Informatics – Tricity Academic Supercomputer & networK (CI TASK) at Gdańsk University of Technology. We acknowledge support of Vcinity® Inc. and 2CRSI SA who provided us with the InfiniBand Range Extenders. This project would not be possible without infrastructure and support of personnel of Pionier, a Polish National Research and Education Network (NREN) – providing 100 G connectivity to all academical HPC centers in Poland.

References

1. Kaiser, H., Lelbach aka wash, B.A., Heller, T., Bergé, A., et al.: STEllAR-GROUP/hpx: HPX V1.3.0: The C++ Standards Library for Parallelism and Concurrency (2019), DOI: 10.5281/zenodo.3189323

2. The Adaptable Input Output System version 2, `https://github.com/ornladios/ADIOS2/`, accessed: 2020-02-08

3. Orłowski, Ł., Deng, Y., Michalewicz, M.: Galaxies of supercomputers and their underlying interconnect topologies hierarchies. In: International Supercomputer Conference, Leipzig, Germany (2014), DOI: 10.13140/2.1.4798.2728

4. Michalewicz, M., Southwell, D., Tan, T., Poppe, Y., et al.: InfiniCortex: concurrent supercomputing across the globe utilising trans-continental InfiniBand and Galaxy of Supercomputers. In: Supercomputing 2014: The International Conference for High Perfor-

mance Computing, Networking, Storage and Analysis, At New Orleans, LA, USA (2014), DOI: 10.13140/2.1.3267.7444

5. Michalewicz, M.T., Lian, T.G., Seng, L., Low, J., et al.: InfiniCortex: Present and Future Invited Paper. In: Proceedings of the ACM International Conference on Computing Frontiers, May 2016, Como, Italy. pp. 267–273. Association for Computing Machinery, New York, NY, USA (2016), DOI: 10.1145/2903150.2912887

6. Noaje, G., Davis, A., Low, J., Lim, S., et al.: InfiniCortex-From Proof-of-concept to Production. Supercomputing Frontiers and Innovations 4(2), 87–102 (2017), DOI: 10.14529/jsfi170207

7. Obsidian Strategics Inc., https://www.cybersecurityintelligence.com/obsidian-strategics-106.html, accessed: 2020-06-01

8. Obsidian Strategics Inc., https://obsidianstrategics.com/index.html, accessed: 2020-06-01

9. Vcinity Inc., https://vcinity.io/, accessed: 2020-06-01

10. Mellanox MetroX®-2 Systems, https://www.mellanox.com/products/long-haul, accessed: 2020-06-01

11. Obsidian Longbow Campus Solutions Extend Its Columbia Supercomputer across Multiple NASA Locations, https://www.militaryaerospace.com/home/article/16725502/obsidian-longbow-campus-solutions-extend-its-columbia-supercomputer-across-multiple-nasa-locations, accessed: 2020-06-01

12. Eikenberry, S., Lindekugel, K., Stanzione, D.: Long Haul InfiniBand Technology: Implications for Cluster Computing, Arizona State University (2006), https://obsidianstrategics.com/archives/2006/asu_stanzione_ccs.pdf, accessed: 2020-06-28

13. El-Harake, H.N., Gamboni, C., Gorini, S., Schoenemeyer, T.: Evaluation of infiniband range extension offered by obsidian (2011)

14. Richling, S., Kredel, H., Hau, S., Kruse, H.G.: A long-distance infiniband interconnection between two clusters in production use. In: State of the Practice Reports, November 2011, Seattle, Washington. Association for Computing Machinery, New York, NY, USA (2011), DOI: 10.1145/2063348.2063368

15. Ban, K., Chrzeszczyk, J., Howard, A., Li, D., Tan, T.W.: 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

16. Chrzeszczyk, J., Howard, A., Chrzeszczyk, A., Swift, B., Davis, P., Low, J., Tan, T.W., Ban, K.: InfiniCloud 2.0: distributing High Performance Computing across continents. Supercomputing Frontiers and Innovations 3(2), 54–71 (2016), DOI: 10.14529/jsfi160204

17. Antypas, K.: Superfacility: How new workflows in the DOE Office of Science are influencing storage system requirements? (2016), https://storageconference.us/2016/Slides/KatieAntypas.pdf, accessed: 2020-06-01

18. NERSC Superfacility, https://www.nersc.gov/research-and-development/superfacility/, accessed: 2020-06-01

19. Creating Super-facilities: a Coupled Facility Model for Data-Intensive Science, Internet 2 Global Summit 2015, http://meetings.internet2.edu/2015-global-summit/detail/10003679/, accessed: 2020-06-01

20. Bell, G.: The Energy Sciences Network: Overview, Update, Impact (DoE) - presentation, https://science.osti.gov/-/media/ascr/ascac/pdf/meetings/20150324/Bell_ESNet.pdf?la=en&hash=46C0168F7ADAB232EC32E4452C49A159453859C9, accessed: 2020-06-01

21. Fenix Research Infrastructure, https://fenix-ri.eu/about-fenix, accessed: 2020-06-01

22. Noaje, G.: InfiniCortex, InfiniBand nation-wide and world-wide, a talk given at Journee Scientifique ROMEO'2016, Reims, France (2016), https://romeo.univ-reims.fr/news/208/Journee_Scientifique_ROMEO_2016_le_9_juin_2016_a_REIMS, accessed: 2020-06-01

23. Proficz, J., Sumionka, P., Skomiał, J., Semeniuk, M., Niedzielewski, K., Walczak, M.: Investigation into MPI All-Reduce Performance in a Distributed Cluster with Consideration of Imbalanced Process Arrival Patterns. In: International Conference on Advanced Information Networking and Applications, 15-17 April, Caserta, Italy. pp. 817–829. Springer (2020), DOI: 10.1007/978-3-030-44041-1_72

24. Niedzielewski, K., Marchwiany, M.E., Piliszek, R., Michalewicz, M., Rudnicki, W.: Multidimensional feature selection and high performance parallex. SN Computer Science 1(1), 40 (2020), DOI: 10.1007/s42979-019-0037-5

25. Open MPI: Open source high performance computing, https://www.open-mpi.org/, accessed: 2020-02-08

26. Guyon, I., Gunn, S., Ben-Hur, A., Dror, G.: Result analysis of the nips 2003 feature selection challenge. In: Saul, L.K., Weiss, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems 17, pp. 545–552. MIT Press (2005), http://papers.nips.cc/paper/2728-result-analysis-of-the-nips-2003-feature-selection-challenge.pdf

27. Dua, D., Graff, C.: UCI machine learning repository (2017), http://archive.ics.uci.edu/ml

28. Application examples for the ADIOS2 I/O library, https://github.com/ornladios/ADIOS2-Examples, accessed: 2020-02-08

29. Pearson, J.E.: Complex Patterns in a Simple System. Science 261(5118), 189–192 (1993), DOI: 10.1126/science.261.5118.189