Concept and Design of SDN-enhanced MPI Framework

Keichi Takahashi[1], Dashdavaa Khureltulga[1], Baatarsuren Munkhdorj[1], Yoshiyuki Kido[1], Susumu Date[1], Hiroaki Yamanaka[2], Eiji Kawai[2] and Shinji Shimojo [1]

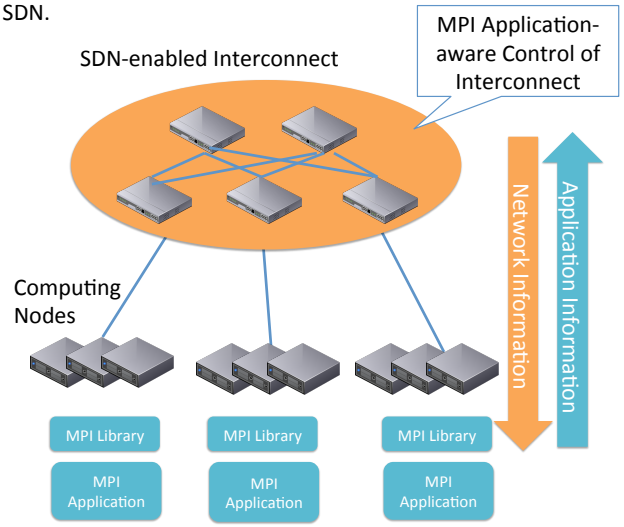
Osaka University[1], National Institute of Information and Communications Technology[2]

1. Abstract

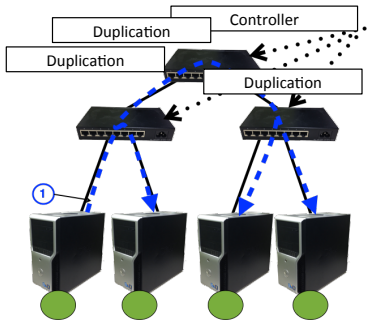
In general, modern high-performance computing systems are built as cluster systems. We have been investigating the feasibility of optimizing MPI communications by integrating the dynamic network control realized by SDN. In this work, we present a concept of a generic SDN enhanced MPI framework; an application-aware network control mechanism optimized for MPI applications.

2. The Concept of SDN-enhanced MPI

Practical HPC systems are often deployed with an exceedingly low-latency and high-throughput network. However, this approach is getting increasingly difficult and expensive as a result of the recent rapid scale-out in node number. We have been developing SDN-enhanced MPI based on the idea that a mechanism that configures and controls the network of a cluster system depending on the requirement of each application is essential. The key concept of SDN-enhanced MPI is to utilize the underlying network of a computer cluster to its maximum capacity by fully leveraging the flexible network control ability of SDN.



3. Proof-of-Concept Implementations

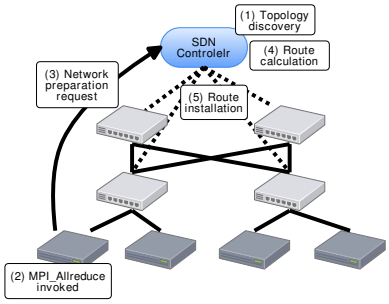


A. SDN_MPI_Bcast

SDN_MPI_Bcast is an SDN-enhanced version of MPI_Bcast, which is the broadcasting function in MPI. SDN_MPI_Bcast offloads packet duplication operations during the broadcast onto SDN switches.

As a result, SDN_MPI_Bcast has successfully decreased the number of communications and communication latency of MPI_Bcast.

[1] Khureltulga Dashdavaa, Susumu Date, Hiroaki Yamanaka, Eiji Kawai, Yasuhiro Watashiba, Kohei Ichikawa, Hirotake Abe and Shinji Shimojo. Architecture of a High-speed MPI_Bcast Leveraging Software-Defined Network. In UCHPC2013: The 6th Workshop on UnConventional High Performance Computing 2013, Archen, Germany, Aug. 2013. DOI: 10.1007/978-3-642-54420-0_86



B. SDN_MPI_Allreduce

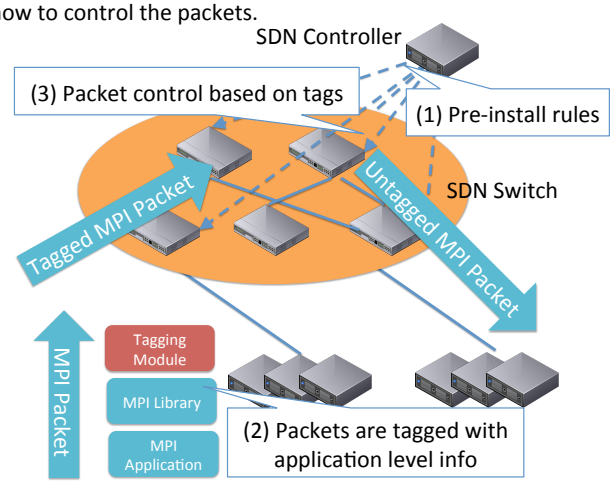
SDN_MPI_Allreduce is an SDN-enhanced version of MPI_Allreduce.

Since it requires multiple simultaneous communication between nodes, congestion may happen on a non full bisection bandwidth interconnect. We employ a real-time traffic load balancing method to solve this problem.

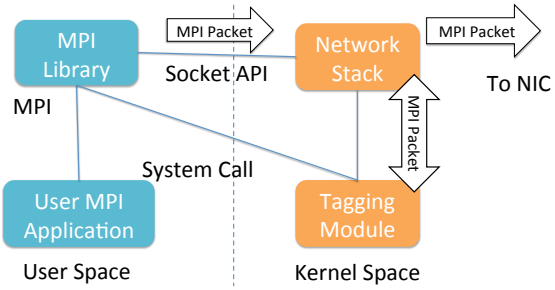
[2] Keichi Takahashi, Dashdavaa Khureltulga, Yasuhiro Watashiba, Yoshiyuki Kido, Susumu Date, Shinji Shimojo, "Performance Evaluation of SDN-enhanced MPI_Allreduce on a Cluster System with Fat-tree Interconnect", The International Conference on High Performance Computing and Simulations (HPCS2014), Bologna, Italy, Jul. 2014. DOI: 10.1109/HPCSim.2014.6903768

4. Proposed Architecture of SDN-enhanced MPI

We propose an integrated framework to combine SDN-MPI components that we have developed in our previous works. In this framework, MPI packets are tagged with MPI-layer information which are used by the SDN switches to determine how to control the packets.



Our implementation embeds tags into the L2 header. Below figure illustrates the packet tagging mechanism.



5. Conclusion and Future Issues

In this work, we presented a concept of our envisaged SDN-enhanced MPI framework. Our focus is to realize an application-aware network control specifically for MPI applications. As a next step, we implement this design and evaluate how it can improve the performance of MPI applications.

Contact: Keichi Takahashi <takahashi.keichi@ais.cmc.osaka-u.ac.jp>