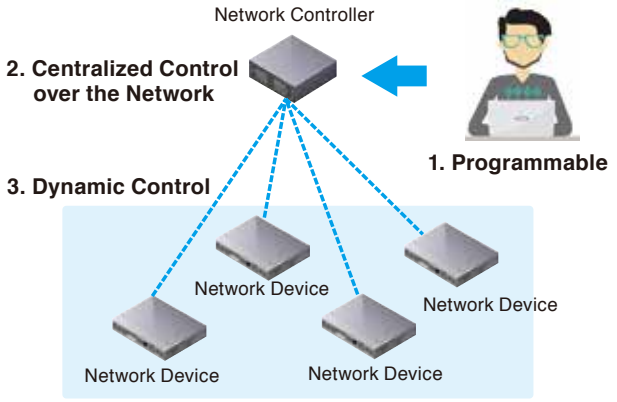
Cybermedia Center
Osaka University, Japan
SC16 BOOTH 1437

SDN-enhanced MPI: Towards Dynamic and Application-aware Interconnect Architecture

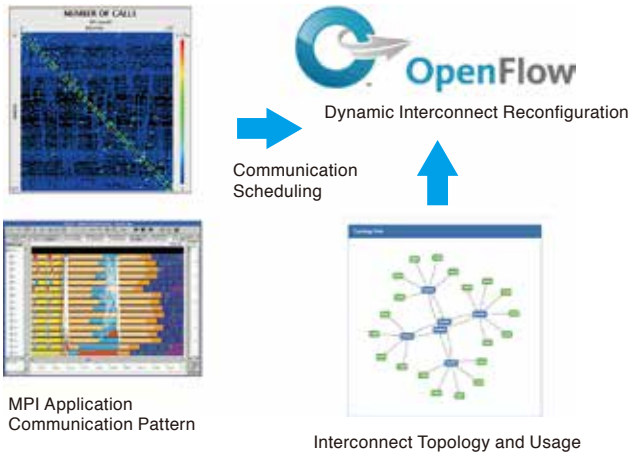
1. Software-Defined Networking (SDN)

Software-Defined Networking (SDN) is a new concept of network architecture that decouples conventional networking function into a programmable control plane (responsible for deciding how to control the packets) and a data plane (responsible for the actual packet delivery). Currently, OpenFlow is the most common implementation of SDN, which enables to dynamically control the forwarding functionality of network from a centralized controller.



2. Aim of SDN-enhanced MPI

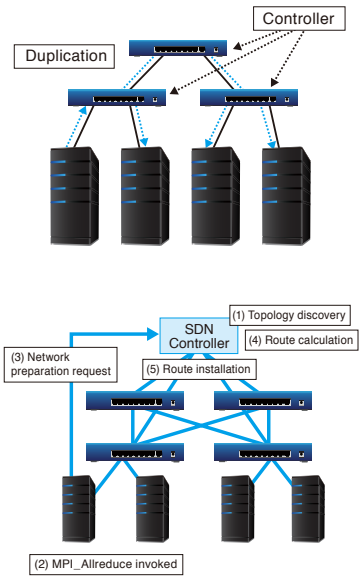
We have been developing SDN-enhanced MPI based on the idea that a mechanism that configures and controls the network of a cluster system depending on the requirement of each application is essential. The key concept of SDN-enhanced MPI is to utilize the underlying network of a computer cluster to its maximum capacity by leveraging the flexible network controllability of SDN.



Publications

[1] K. Dashdavaa et al., Architecture of a High-Speed MPI_Bcast Leveraging Software-Defined Network, in: UCHPC2013 6th Work. Unconv. High Perform. Comput., 2014: pp. 885–894.
[2] K. Takahashi et al., Performance Evaluation of SDN-enhanced MPI_Allreduce on a Cluster System with Fat-tree Interconnect, in: Proc. Int. Conf. High Perform. Comput. Simul. - HPCS 2014, 2014: pp. 784–792.
[3] K. Takahashi et al., Concept and Design of SDN-Enhanced MPI Framework, in: Proc. Fourth Eur. Work. Softw. Defin. Networks - EWSDN 2015, 2015: pp. 109–110.

3. SDN-enhanced MPI Communication Primitives



A. SDN_MPI_Bcast [1]

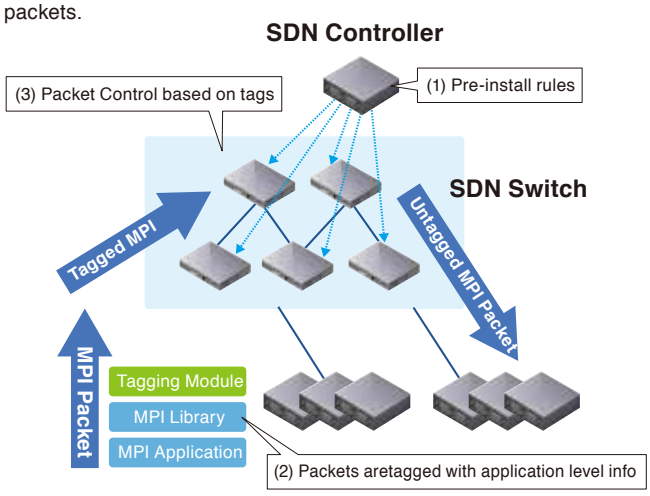
SDN_MPI_Bcast is an SDN-enhanced version of MPI_Bcast, which is the broadcasting function in MPI. SDN_MPI_Bcast offloads packet duplication operations during the broadcast onto SDN switches. As a result, SDN_MPI_Bcast has successfully decreased the number of communications and communication latency of MPI_Bcast.

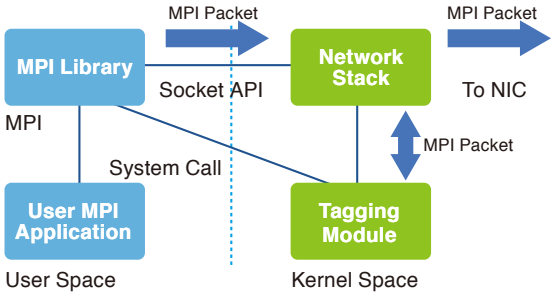
B. SDN_MPI_Allreduce [2]

SDN_MPI_Allreduce is an SDN-enhanced version of MPI_Allreduce. Since it requires multiple simultaneous communication between nodes, congestion may happen on an oversubscribed interconnect. We employ a real-time traffic load balancing method to solve this problem.

4. Coordination Mechanism of Communication and Computation

We propose an integrated framework [3] 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.





Contact
Keichi Takahashi : takahashi.keichi@ais.cmc.osaka-u.ac.jp
Khureltulga Dashdavaa : huchka@ais.cmc.osaka-u.ac.jp
Visit us at www.cmc.osaka-u.ac.jp