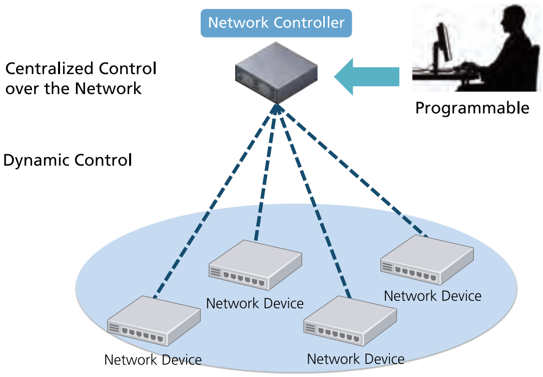
Architecture of SDN-enhanced MPI Framework

Cybermedia Center, Osaka University, Japan

Architecture of SDN-enhanced MPI Framework

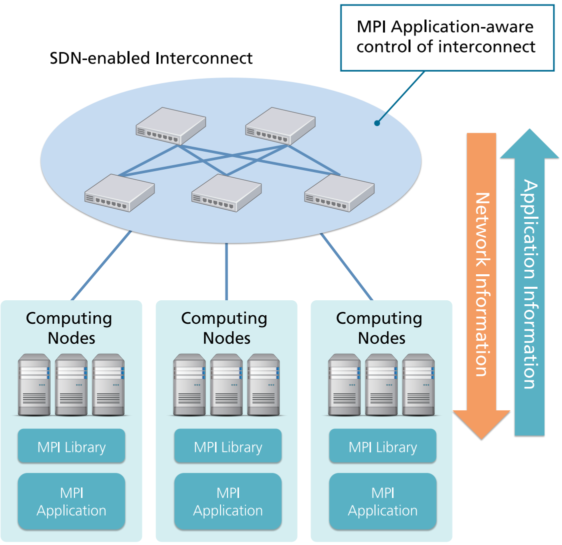
1. Software-Defined Networking (SDN)

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.

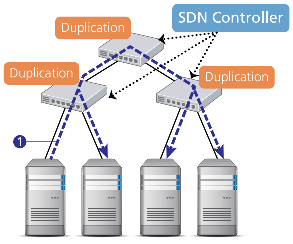


2. Basic Idea Behind the SDN-enhanced MPI Framework

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 controllability of SDN.

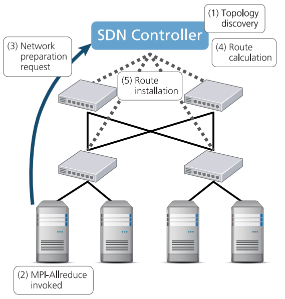


3. SDN-enhanced MPI Communication Functions



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.

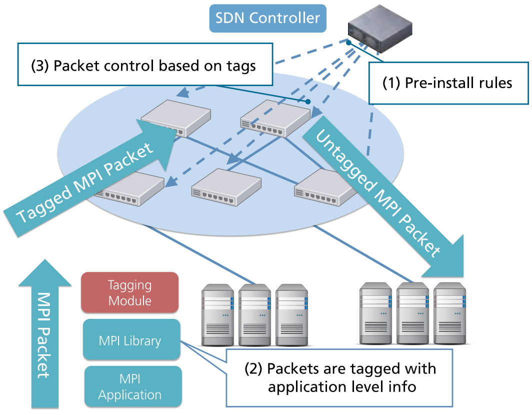


B. SDN_MPI_Allreduce

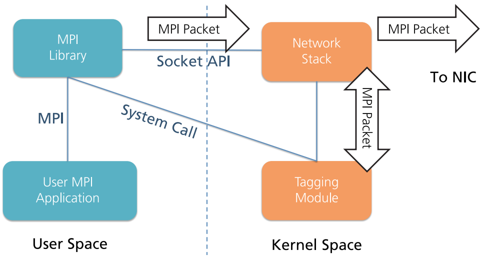
SDN_MPI_Allreduce is an SDN-enhanced version of MPI_ Allreeduce. Since MPI_Allreduce requires multiple simultaneous communication between nodes, congestion may happen on a interconnect without full bisection bandwidth. We employ a real-time traffic load balancing method using SDN to solve this problem.

4. Architecture of SDN-enhanced MPI Framework

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 a tag into the L2 header. The figure below illustrates the packet tagging mechanism.



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