An MPI Framework Enhanced for SDN Architecture Cluster

Cybermedia Center, Osaka University, Japan

An MPI framework Enhanced for SDN Architecture Cluster

Message Passing Interface

Message Passing Interface (MPI) has been a *de facto* standard programming model in parallel computing for around two decades. MPI offers two types of communications:

1. *One-to-one communication:* used for low-level description of communication pattern.
2. *Collective communication:* used for high-level and human-friendly description.

Software-Defined Networking

Software-Defined Networking (SDN) is a concept of network architecture that decouples conventional networking function into a programmable control plane (network controller) and a data plane (physical network).

Problem

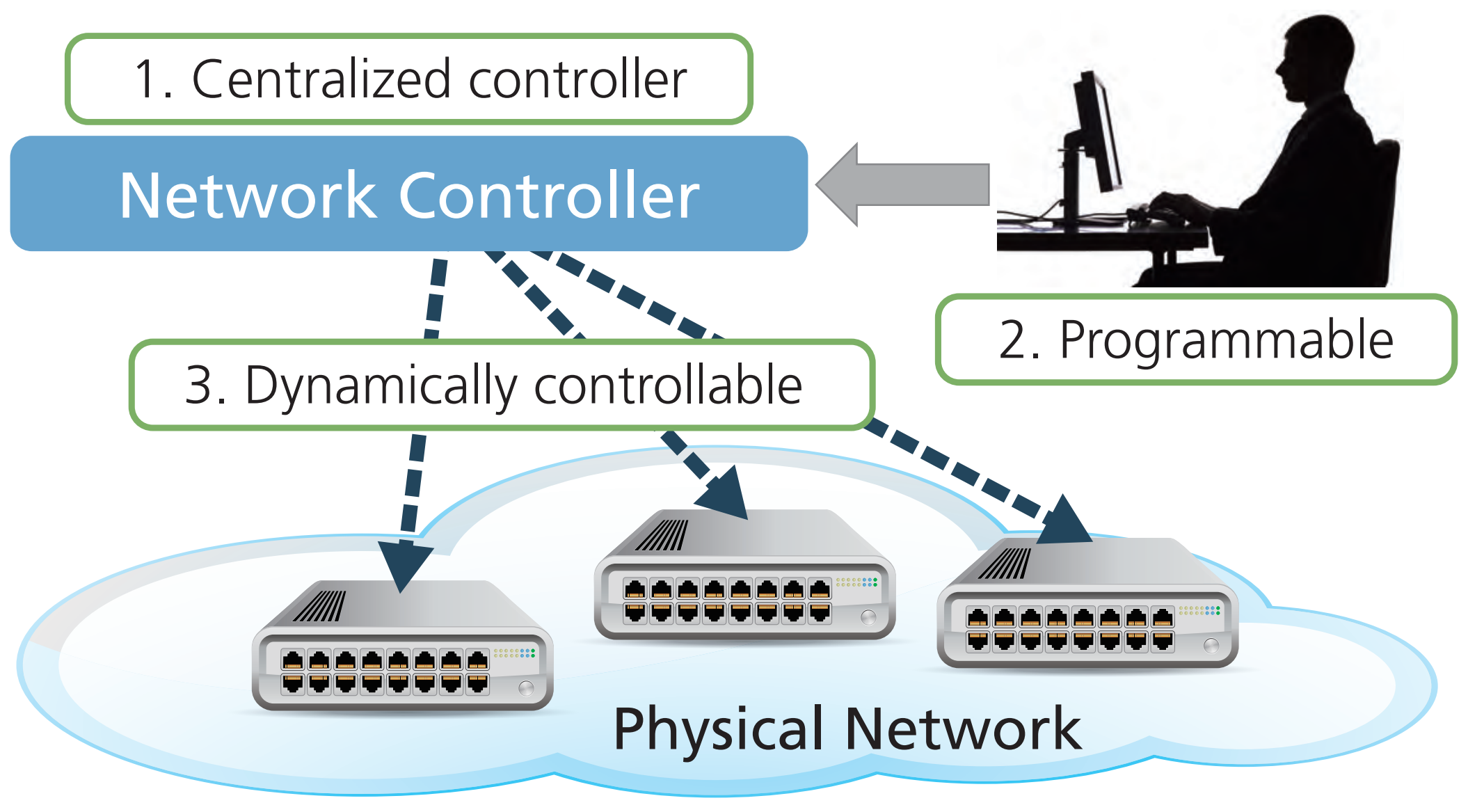
However, collective communication of MPI often suffers from performance degradation. One of the main causes is that most of the MPI implementations are designed on the assumption that:

1. MPI programs are not able to acquire network topology information.
2. Network resources are statically allocated.

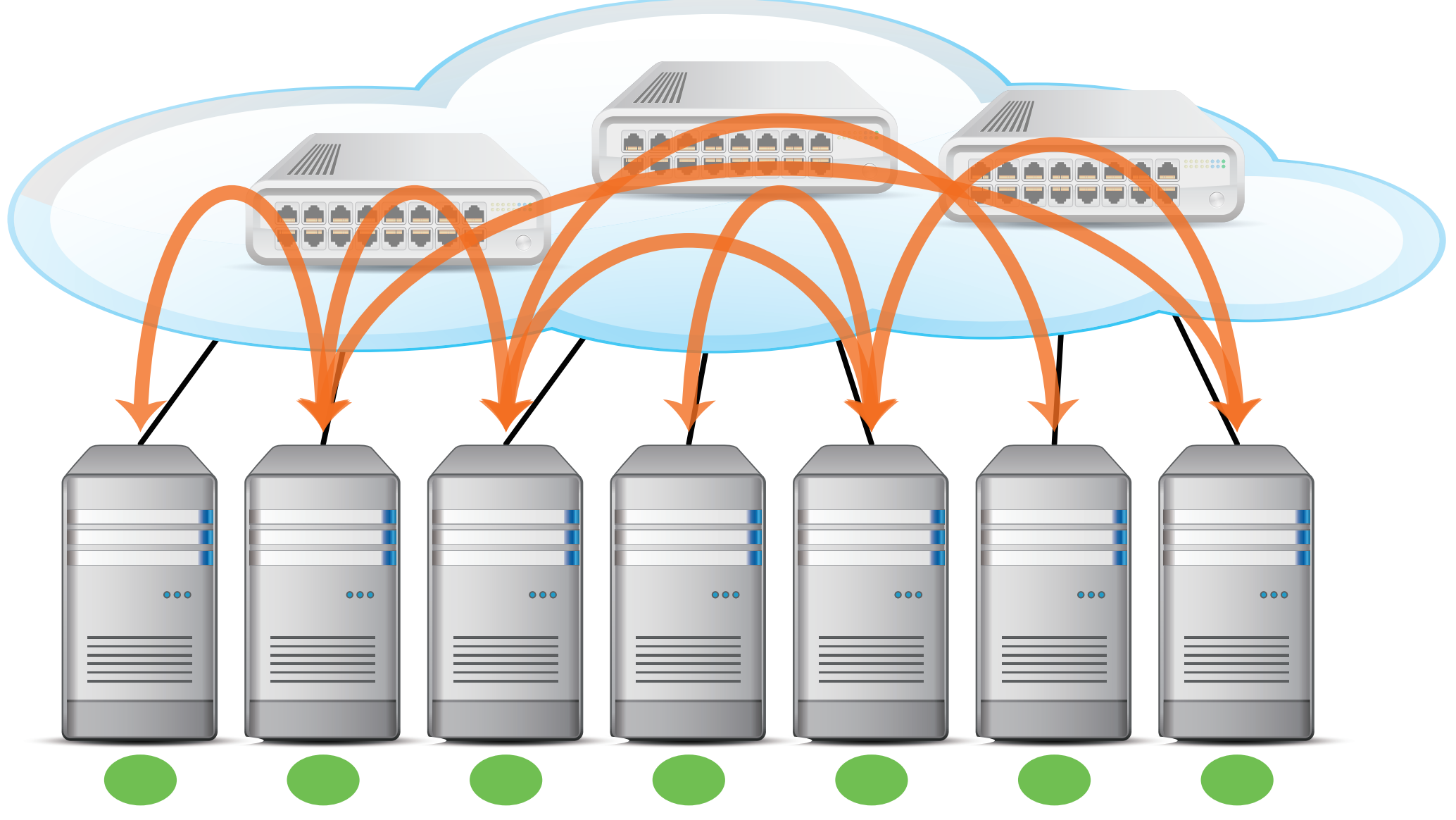
Research Goal

Integrate the *dynamic* controlling ability of Software-Defined Networking into MPI in order to optimize *collective communication* by overturning the assumption that network is a static resource. In the future, we'd like to implement a new MPI library which cuts down communication latency and traffic amount by programmatically controlling the underlying network.

Software-Defined Networking

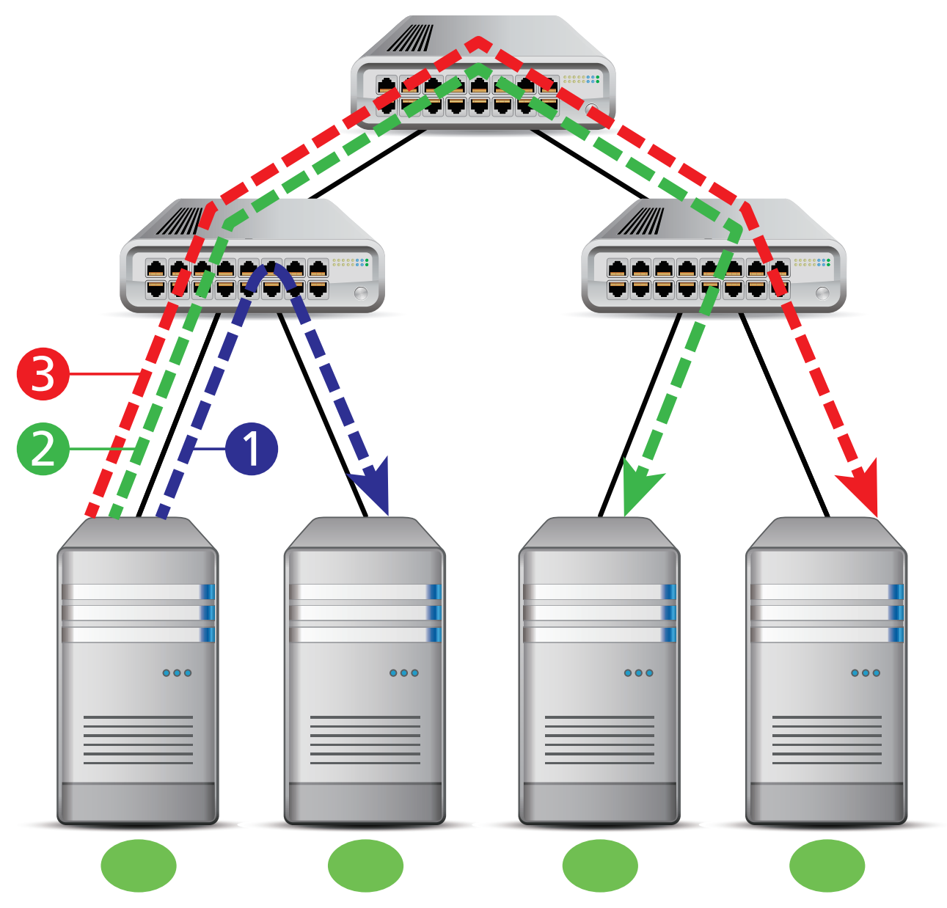


Conventional parallel computing's communication method



Fast MPI_Bcast Design Enhanced with SDN

MPI_Bcast in Conventional Methods

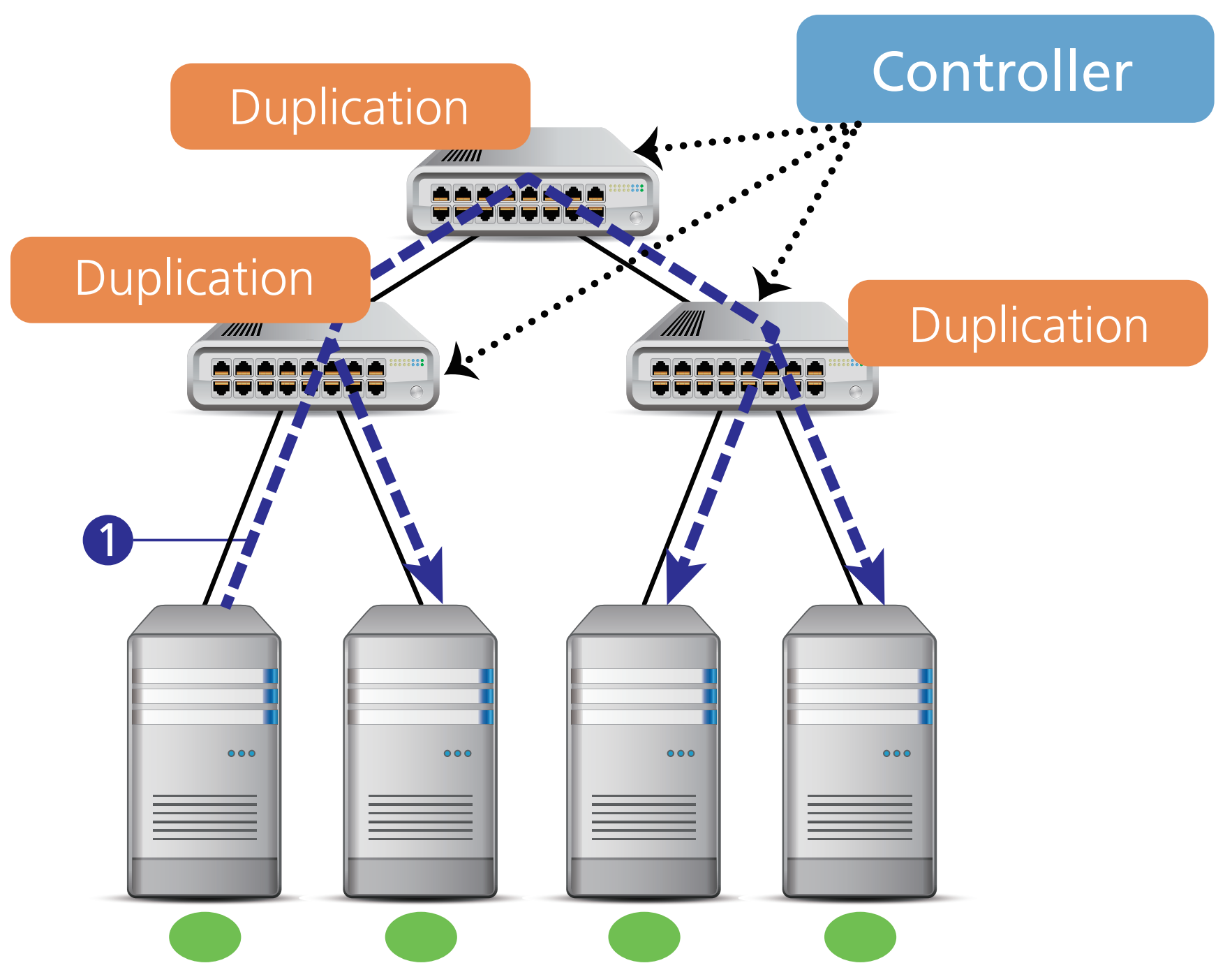


MPI_Bcast uses combination of multiple one-to-one communciations.
Intra- and Inter-node communications does not overlap with each others well.

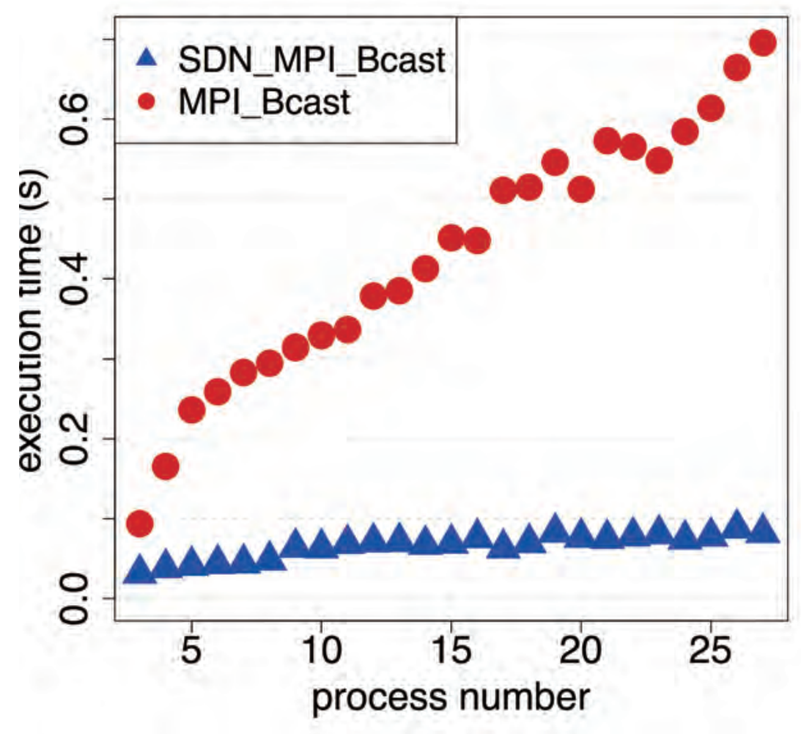
SDN MPI_Bcast Method

SDN_MPI_Bcast uses SDN-enabled switches as packet copier.

Stage 1 – Broadcast Data



Execution time of SDN_MPI_Bcast

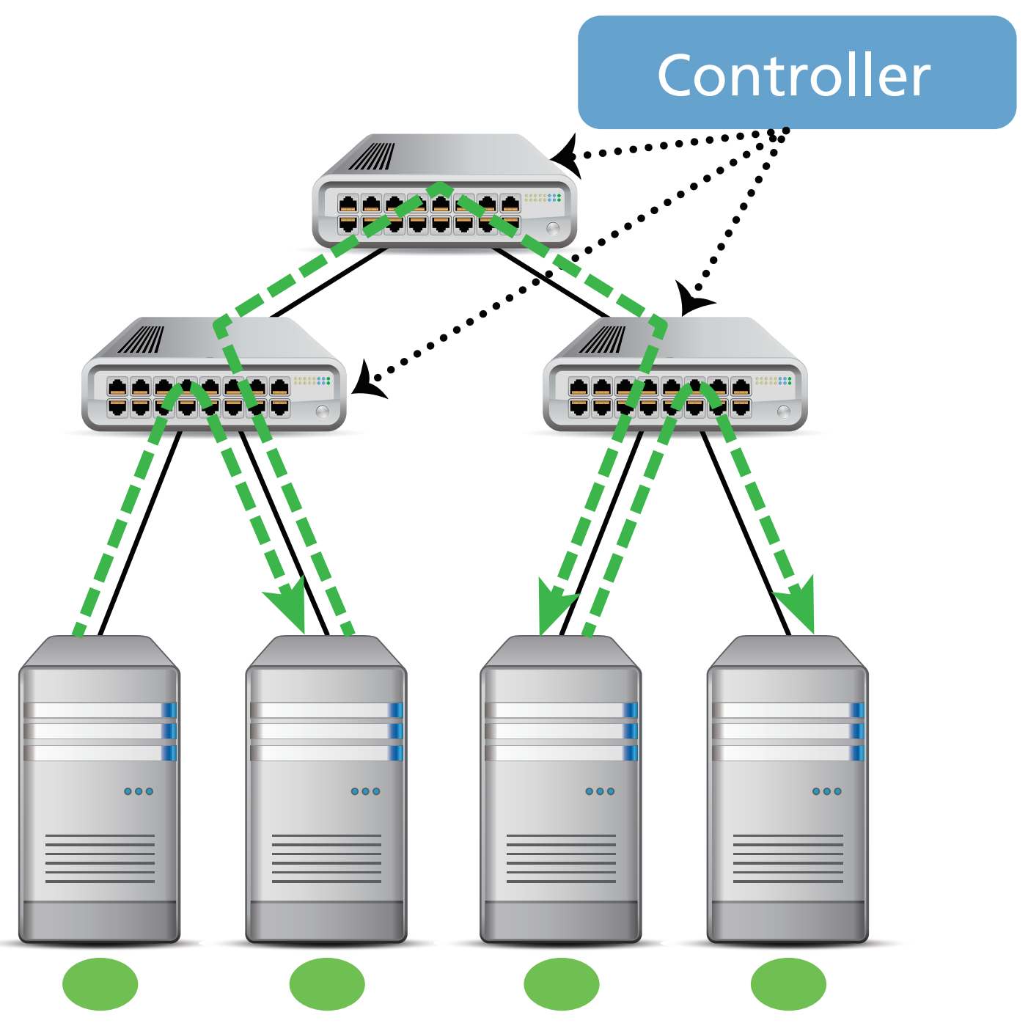


Stage 2 – Assure Reliability

MPI processors send data to "next" processor using one-to-one communication.

SDN Controller calculates "next" processor considering network topology and physical location of processes.

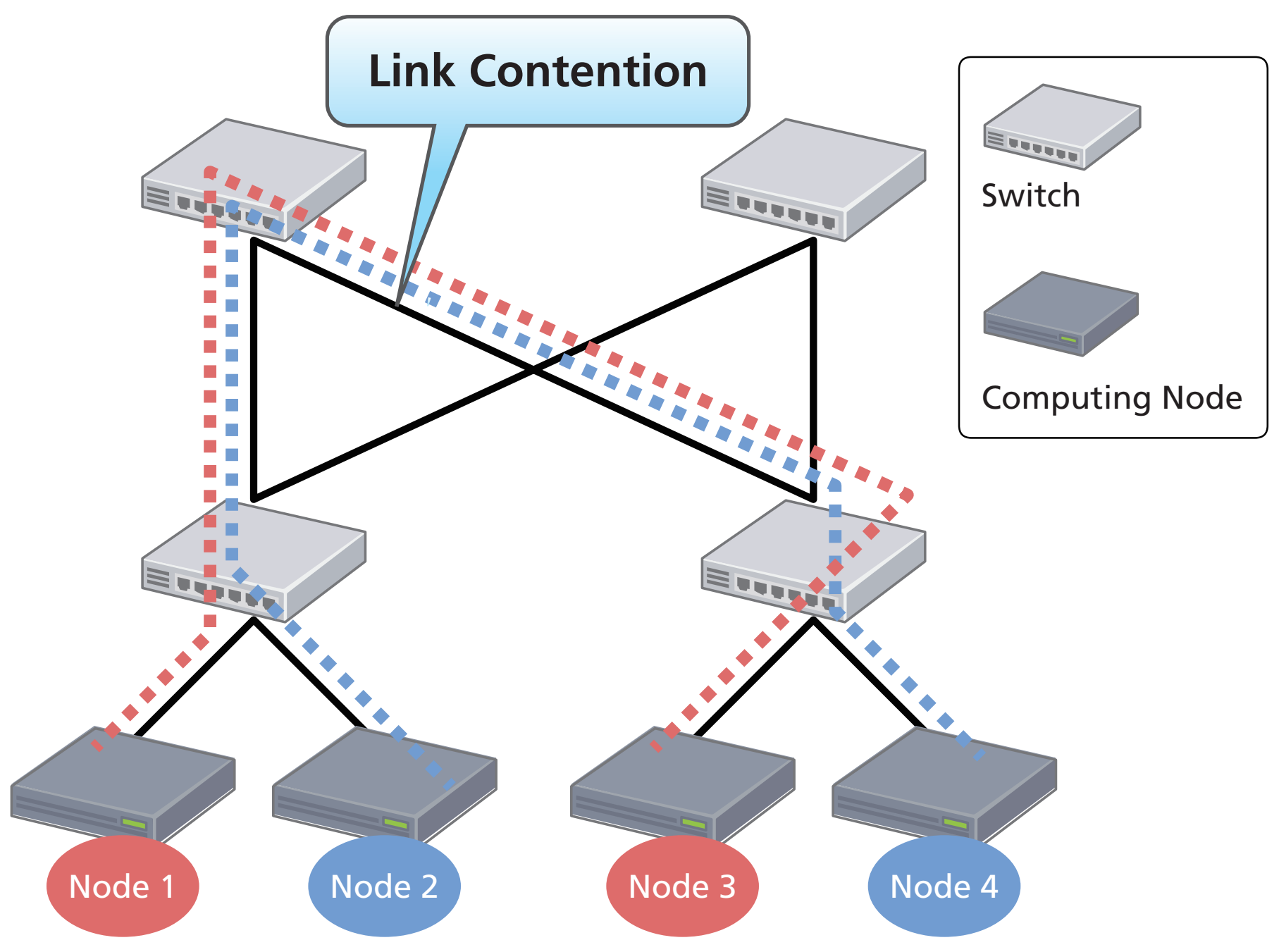
Intra-node communications occur during Stage 2.



Efficient Bandwidth Utilization of Clusters

Link Contentions in Conventional Clusters

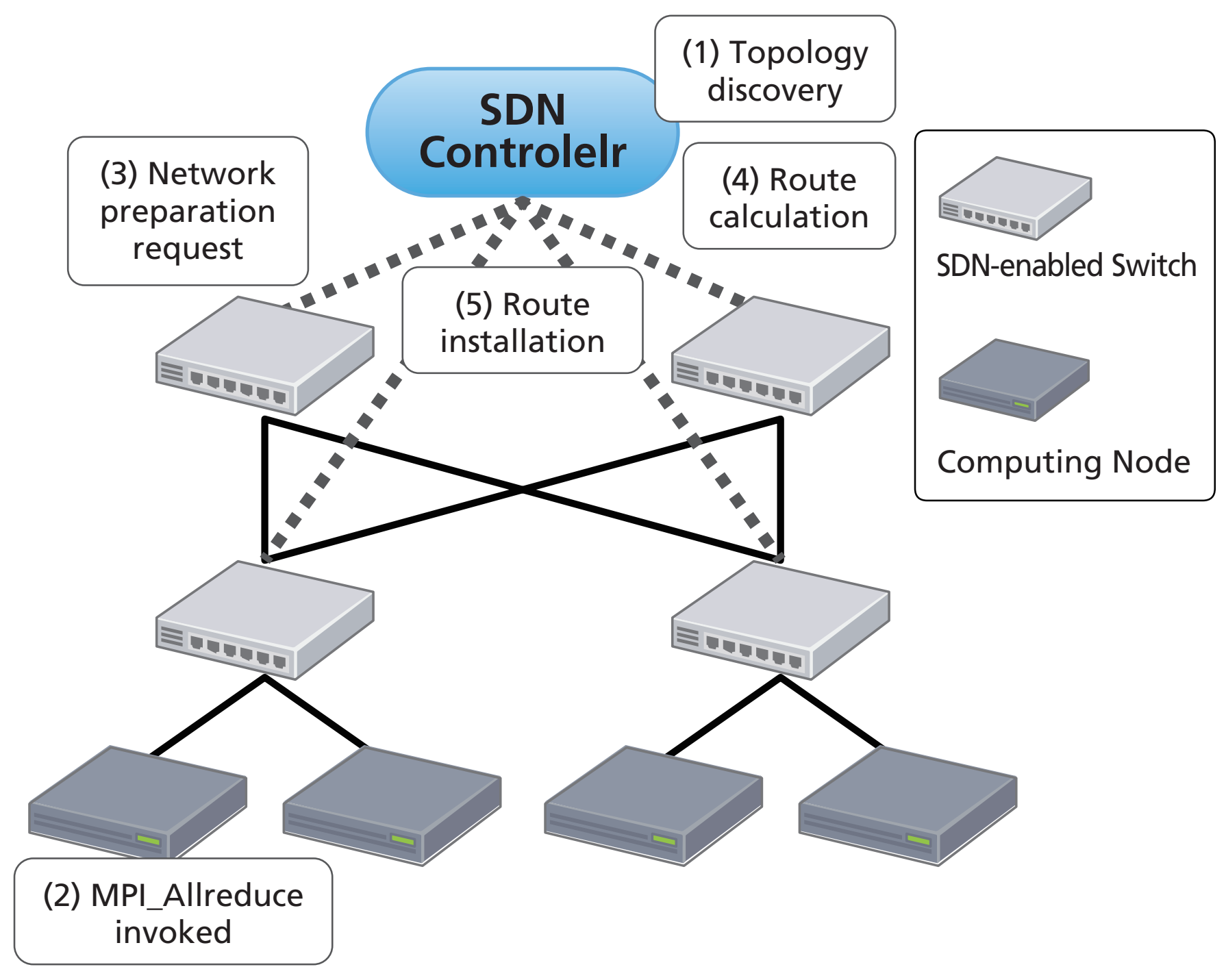
In conventional computer clusters without full bisection bandwidth interconnect, link congestion could happen due to the deviation of packet flow.

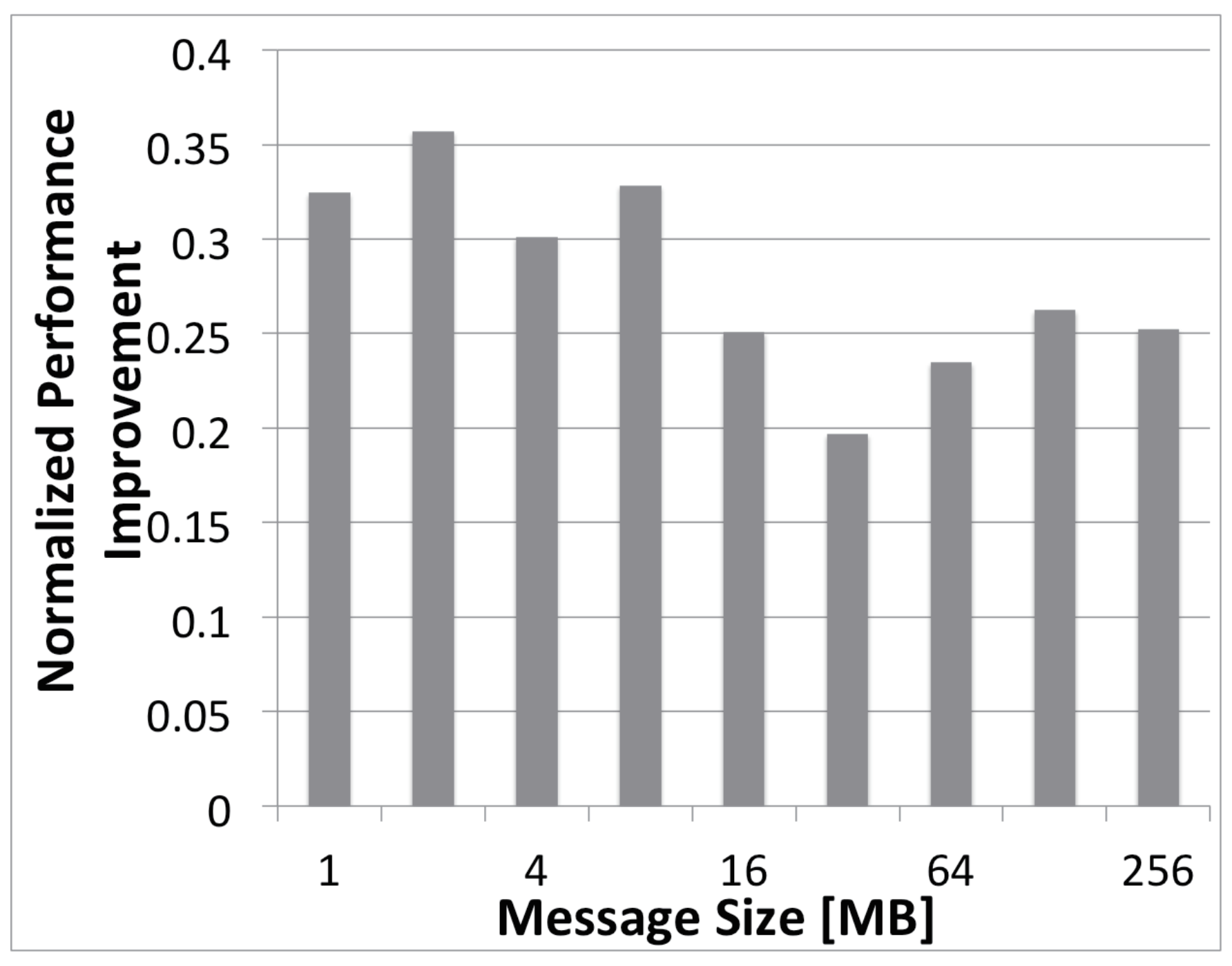


Cooperation of MPI Library and SDN Controller

We propose a new computer cluster architecture consisting of a customized MPI library, SDN controller and LLDP daemon. This system dynamically reconfigures the underlying network depending on the MPI communication request so that link contention will not happen.

An experiment conducted on a computer cluster with fat-tree interconnect revealed that our method increased the performance of MPI_Allreduce for 36 % at maximum.





[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.

[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.

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