Scalable and Low-latency Communication Method for Reliability Improvement of SDN MPI_Bcast

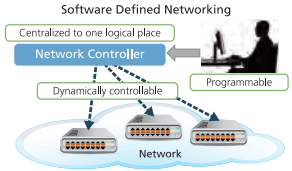
Cybermedia Center, Osaka University, Japan

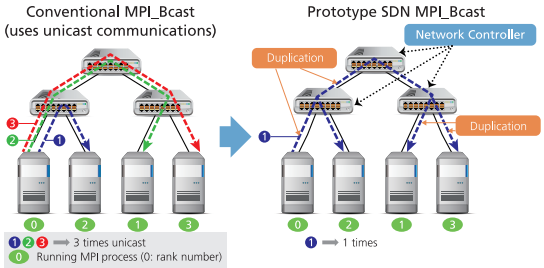
Scalable and Low-latency Communication Method for Reliability Improvement of SDN MPI_Bcast

Communication time of MPI_Bcast collective tends to get longer on a large-scaled cluster.

Previous Work

Our previous work implements MPI_Bcast through duplication of broadcast data on the fly from source process to others leveraging SDN. As the result, source process sends data only once for broadcasting data.





Problem of Previous Work

Data delivery from source process to others is not guaranteed in prototype SDN MPI_Bcast.

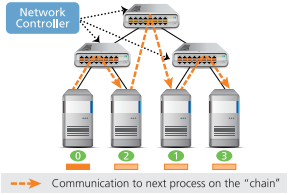
Research Goal

To implement scalable and low-latency "chain" communication method for improving reliability of prototype SDN MPI_Bcast.

- All receiving processes need to let source process know they received data.

Proposal

Each process sends data to next process on the "chain" for the acknowledgement of data receiving.



"chain": series of all processes placed in a line.
Eg. 0→2→3→1, 0→1→2→3

Reliable SDN MPI_Bcast has two stages.

1. Source process sends data using prototype SDN MPI_Bcast.
2. Each process sends data to next process on the "chain" as soon as receives it.

Low-latency: Network controller generates the "chain" considering network topology and process placement
Scalable: Each process responsible for only one process' data delivery

Khureltulga Dashdavaa*, Munkhdorj Baatarsuren†, Keichi Takahashi*, Susumu Date*, Yoshiyuki Kido*, and Shinji Shimojo* *Osaka University, Japan, †The University of Tokyo, Japan

Contact : Khureltulga Dashdavaa huchka@ais.cmc.osaka-u.ac.jp