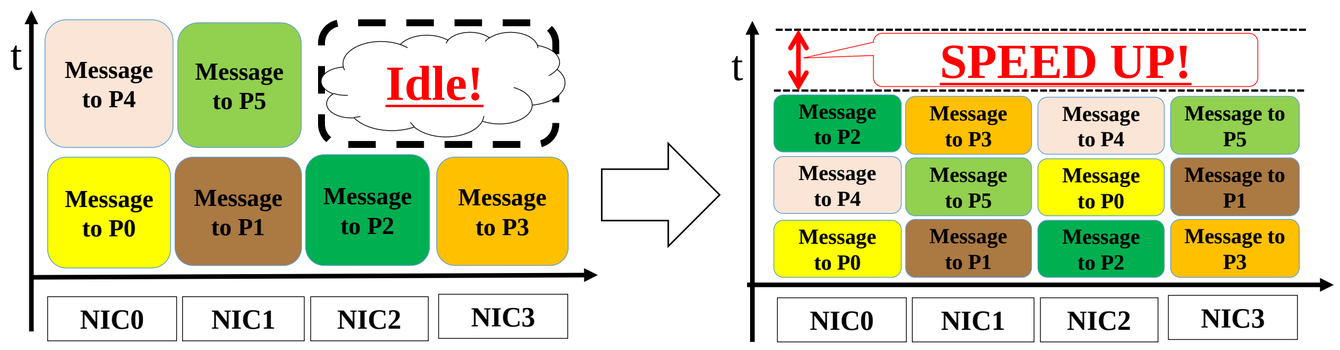
Neighbor collective communication algorithm with effective using multiple NICs on mesh/torus

Yoshiyuki morie and Takeshi Nanri
contact: morie.yoshiyuki.404@m.kyushu-u.ac.jp

The concept of the proposed neighboring communication algorithm.

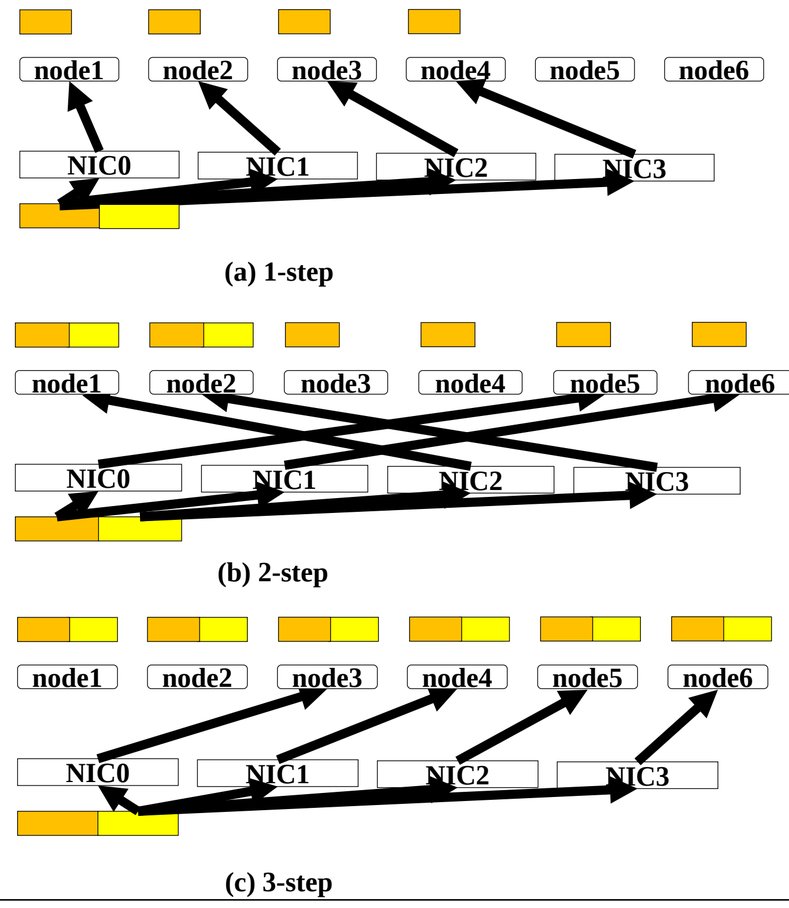
- ***Algorithm for sending messages to N-neighbors with C-NICs.***
- **Minimize idle NICs by dividing message into several segments.**



The example: The proposed algorithm on 6-neighbors with 4-NICs.

- ***A communication bandwidth can be used up by transmitting 3/2 messages per each NIC.***
- **To transmit 3/2 messages by each NIC, data are divided to half segment, and these data are transmitted in three steps.**
- The prediction of communication time of the proposed neighbor communication algorithm is shown by a following equation.

$$time_{exec} = 3(L + \frac{3M}{2B})$$



(a) 1-step

(b) 2-step

(c) 3-step

Evaluation experiment.

- This experiment that compares the performance of the proposed neighbor communication algorithm with the existing one is conducted.
- Both of these algorithms are implemented by MPI functions and RDMA interfaces.
- The evaluation experiment is executed in FX10 which is family of “K-computer” at Tokyo university (Table 1).

Table 1 Experimental environment

Machine name	Fujitsu PRIMEHPC FX10
CPU	SPARR64TM Ixfx 1.848GHz (16 cores)
Memory	32GB
# of nodes	4800
MPI library	Fujitsu Technical Computing Suite v1.0
Network	5GB/sec
Topology	6 dimensional mesh/torus

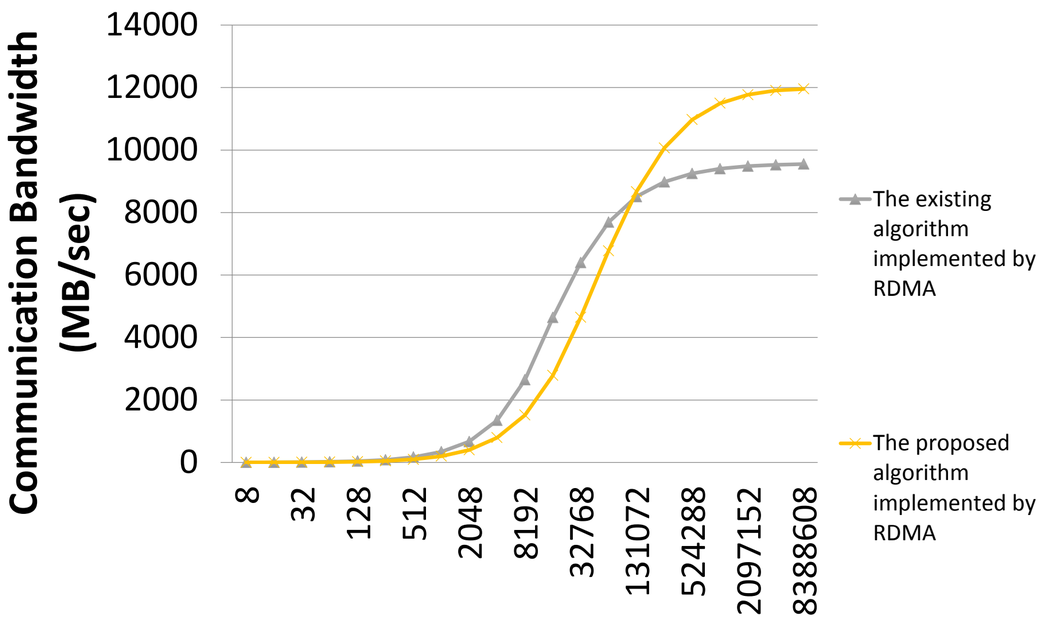


Figure 1 Communication Bandwidth

Research Institute for Information Technology., Kyushu University