

#### High Performance Relay Mechanism for MPI Libraries Run on Multiple Private IP Address Clusters

Ryousei Takano<sup>1</sup>, Motohiko Matsuda<sup>2</sup>, Tomohiro Kudoh<sup>1</sup>, Yuetsu Kodama<sup>1</sup>, Fumihiro Okazaki<sup>1</sup>, Yutaka Ishikawa<sup>2</sup>, and Yasufumi Yoshizawa<sup>3</sup>

<sup>1)</sup>National Institute of Advanced Industrial Science and Technology (AIST) <sup>2)</sup>University of Tokyo <sup>3)</sup>Tokyo University of Agriculture and Technology

CCGrid 2008, May 22 2008, Lyon France







### Agenda

- Background
  - GridMPI
  - Supporting private address clusters
- IMPI Relay Mechanism
- Evaluation
- Conclusion & Future Work







- Enable execution of single MPI program using multiple clusters connected by wide area network
  – MPICH-G2, PACX-MPI, StaMPI, MC-MPI, and so on
- Users can seamlessly deploy applications from a laboratory to a Grid environment
- We focus on metropolitan-area, high-bandwidth network: more than 10 Gbps, less than 10 msec of latency (≒ 1000 km)
  - We have already demonstrated that it is feasible to run large-scale applications over distances up to 1000 km [Cluster2007]





### GridMPI

- GridMPI is an open-source implementation of the MPI-1.2 and MPI-2.0 standards developed from the scratch by AIST and University of Tokyo.
  - Project homepage: http://www.gridmpi.org/
- Full standard conformance
  - GridMPI passes 100% of the conformance test suites from Intel and ANL (MPI-1.2) even in heterogeneous setting
- Interoperability
  - GridMPI complies with IMPI (Interoperable MPI) standard for the inter-cluster communication
- High performance
  - GridMPI achieves high performance optimized for high bandwidth networks



# IMPI (Interoperable MPI)

- IMPI standard realizes interoperability among different MPI implementations
- IMPI standard specifies:
  - Startup/Shutdown protocol
  - Data transfer protocol and data format
  - Collective algorithms





### Motivation (1/2)

- Not all nodes may be able to communicate with each other directly in a Grid environment
  - Private IP address and/or Firewalls
- IMPI design does not support communication from/to private address clusters
- Make communication from/to private address clusters compliant with the IMPI standard





### Motivation (2/2)

7/25

- When single front-end node is used, performance is limited by the NIC bandwidth of the front-end node
  - NIC: GbE, WAN-link: 10 Gbps:

The link utilization achieves only 1/10 of the inter-cluster bandwidth

Use multiple front-end nodes at each cluster for improving the inter-cluster communication







- Background
- IMPI Relay Mechanism
  - Forwarding packet scheme in manner of the IMPI standard
  - Trunking for high performance communication
- Evaluation
- Conclusion & Future Work







# IMPI Relay: Design Goal

- Support communication from/to private address clusters
- 1. Make communication from/to private address clusters compliant with the IMPI standard
  - Packet forwarding scheme in the manner of the IMPI standard
- 2. Use multiple front-end nodes at each cluster for improving the inter-cluster communication
  - Trunking of relay communication





#### **Basic Idea**

- Process identifier
  - MPI: rank
  - IMPI: Host ID (IPv6 address) + Proc ID (64bit integer)
- An IMPI Relay has two host IDs (e.g., AR<sub>P</sub> and AR<sub>G</sub>), one for the intra-cluster, and one for the inter-cluster communication





## Packet forwarding (1/2)

 IMPI Relay forwards packets according to the Host-Proc ID mapping table



AR's Host-Proc ID mapping table

rank	0	1	2	3	
Global ID	AR <sub>G</sub> :0	AR <sub>G</sub> :1	B0:0	B1:0	
Private ID	A0:0	A1:0	AR <sub>P</sub> :0	AR <sub>P</sub> :1	



CCGrid2008, May 22 2008, Lyon France



## Packet forwarding (2/2)

• Forwarding between two private address clusters:



AR's Host-Proc ID mapping table				BR's Host-Proc ID mapping table					
rank	0	1	2	3	rank	0	1	2	3
Global ID	AR <sub>G</sub> :0	AR <sub>G</sub> :1	BR <sub>G</sub> :0	BR <sub>G</sub> :0	Global ID	AR <sub>G</sub> :0	AR <sub>G</sub> :	BR <sub>G</sub> :0	BR <sub>G</sub>
Private ID	A0:0	A1:0	AR <sub>P</sub> :0	AR <sub>P</sub> :1	Private ID	BR <sub>P</sub> :0	BR <sub>P</sub> :1	B0:0	B1:0
CCGrid2008, May 22 2008, Lyon France									



### Trunking



CCGrid2008, May 22 2008, Lyon France





### Agenda

- Background
- IMPI Relay Mechanism
- Evaluation
  - All-to-all communication performance
  - NAS Parallel Benchmarks in a 10 gigabit emulated WAN environment
- Conclusion







#### **Experimental Setting**



Execution Time of All-to-all <sup>16/25</sup> Communication (Latency: 0 msec)

• The execution becomes faster as the number of trunks increases



Execution Time of All-to-all <sup>17/25</sup> Communication (Latency: 0 msec)

- The "direct" suffers from the heavy congestion
- IMPI Relays reduce congestion: the inter-cluster bandwidth is limited at the number of IMPI Relays. There is no congestion at the inter-cluster link





- Problem size: Class B
- #Process: 32 (16 per a cluster)
- One-way latency: 0 10 msec
- IMPI Relay: direct, 1, 2, 4, 8 trunks
- Here, the results of the following benchmarks are shown:

IS (Integer Sort)	Communication-bound		
MG (Multi-Grid method)	Medium		
LU (LU factorization)	Computation-bound		





### IS Benchmark

- For communication-bound programs, trunking has a large impact
- Performance improves as the number of trunks increases







### MG Benchmark

- Latency has a large impact on the performance
- The overhead of the IMPI Relay becomes smaller as the latency increases





### LU Benchmark

• For computation-bound programs, even the "1-trunk" performs as fast as the "direct"







### Agenda

- Background
- IMPI Relay Mechanism
- Evaluation
- Conclusion & Future Work







- We have proposed a high performance relay mechanism for MPI libraries run on multiple private address clusters
  - Packet forwarding in the manner of the IMPI standard
  - Trunking for high performance communication
- The experimental results show that trunking is effective and efficient for running MPI programs over high bandwidth-delay product networks







- Performance evaluation on a multi-site (more than three) setting
- Interoperability test between GridMPI and the other IMPI implementation (LAM/MPI, HP-MPI, ...) via the IMPI Relay
- More IMPI implementation
  - Porting our IMPI implementation to Open MPI





- GridMPI: http://www.gridmpi.org/
- GtrcNET: http://projects.gtrc.aist.go.jp/gnet/



Part of this research was supported by a grant from the Ministry of Education, Sports, Culture, Science and Technology (MEXT) of Japan through the NAREGI (National Research Grid Initiative) Project.

"GridMPI" is a registered trademark in Japan

