MPICH- GX: Message Passing Interface CHameleon- Grid eXensible

Kwon, Oh-kyoung
KISTI Supercomputing Center
Agenda

- Motivation
- What is MPICH- GX?
  - Private IP Support
  - Fault Tolerance Support
- Experiment
- Conclusion
Running on a Grid presents the following problems:

- Standard MPI implementations require that all compute nodes are visible to each other, making it necessary to have them on public IP addresses
  - Public IP addresses for assigning to all compute nodes aren't always readily available
  - There are security issues when exposing all compute nodes with public IP addresses
  - At developing countries, the majority of government and research institutions only have public IP addresses
Running on a Grid presents the following problems: (cont.)

- What if a node is broken or a running process is die in geographically distributed Grid environments?
  - Difficult to manually find the broken node and the failure process among many compute nodes.
What is MPICH-GX?

- MPICH-GX is a patch of MPICH-G2 to extend following functionalities required in the Grid.
  - Private IP Support
  - Fault Tolerant Support
**Private IP Support**

- **MPICH-G2 does not support private IP clusters**

1. **Job submission**
2. **Job distribution**
3. **Job allocation**
4. **Computation & communication**
   - How to do it with private IP?
5. **Report result**

Diagram:
- Cluster Head (public IP)
- Compute node (private IP)
- Public IP
- Members

Diagram shows the process flow with arrows indicating the sequence of steps.
User-level proxy
- Use separate application
- It is easy to implement and portable
- But it causes performance degradation due to additional user-level switching
- Kernel-level proxy
  - Generally, it is neither easy to implement nor portable
  - But it can minimize communication overhead due to firewall
  - NAT (Network Address Translation)
    - Main mechanisms of Linux masquerading

![Diagram of private networks with IP addresses and masquerading]

Private Network

Private Network
NAT Hole Punching (1/2)

- Easily applicable kernel-level solution
  - It is a way to reach otherwise unreachable hosts with a minimal additional effort
  - All you need is a server that coordinates the connections
  - When a client registers with server, it records two endpoints for that client
NAT Hole Punching (2/2)

Cluster Head
(public IP: 150.183.234.100)

(request hole punching)

Get a Hole

Compute node
(private IP)

192.168.1.1
150.183.234.100:30000

Cluster Head
(public IP: 210.98.24.130)

(request hole punching)

Get a Hole

Compute node
(private IP)

10.0.5.1
210.98.24.130:30000
Fault Tolerant Support

- We provide a checkpointing-recovery system for Grid.
- Our library requires no modifications of application source codes.
  - affects the MPICH communication characteristics as less as possible.
- All of the implementation have been done at the low level, that is, the abstract device level of MPICH
Experiment (1/2)

- Experiment of MPICH-GX using Atmospheric application (MM5/WRF)
- Collaboration efforts with PRAGMA people (CICESE in Mexico, SDSC)
- Testbed
  - Geographically distributed 5 Linux Clusters: Daejeon, Seoul, Busan, Gwangju, Pohang
  - Network bandwidth between nodes is 1Gbps
Experiment (2/2)

HHurricane Marty Simulation

SSantana Winds Simulation
12 nodes on single cluster (orion cluster): 17.55min

12 nodes on cross-site
- 6 nodes on orion + 6 nodes on eros, where all nodes have public IP: 21min
- 6 nodes on orion + 6 nodes on eros, where the nodes on orion have private IP and the nodes on eros have public IP: 24min

16 nodes on cross-site
- 8 nodes on orion + 8 nodes on eros, where all nodes have public IP: 18min
- 8 nodes on orion + 8 nodes on eros, where the nodes on orion have private IP and the nodes on eros have public IP: 20min
Conclusion

- MPICH- GX is a patch of MPICH- G2 to provide useful functionalities for supporting the private IP and fault tolerance.
- The application of WRF model work well with MPICH- GX at geographically distributed Grid environments.
- The functionality of the private IP could be usable practically, and the performance of the private IP is reasonable.
Thank you