RI2N: High-Bandwidth and Fault-Tolerant Network with Multi-link Ethernet for PC Clusters

Today's most of HPC PC clusters are equipped with SAN as the interconnection network to provide high bandwidth. However, the cost of such a network is very high especially for a large-scale system with thousands of processors. Another serious issue of the interconnection network is the reliability or dependability in such a large-scale PC cluster.

We introduce a sophisticated network trunking technology based on multiple links of LAN (mainly Ethernet) to provide the following features:

- High bandwidth by network link trunking keeping high cost/performance
- Dynamic scheduling of used link against to the link failure for high dependability

The system is named RI2N (Redundant Interconnection with Inexpensive Network).

Concepts of RI2N

To support HPC applications on a large-scale PC cluster, we utilize a set of multiple Ethernet links both for expansion of bandwidth and fault tolerance.

- In normal case; Transmit all packets on multiple links in round-robin manner
- When fault is detected; recovers the communication without the broken link

RI2N provides high bandwidth and dependability very high cost/performance, which is required for very large-scale cluster systems.

Design of RI2N/DRV

RI2N/DRV is implement as a pseudo network interface on Linux

- User can use this system transparently
- The User application can perform TCP/IP communication without performance degradation

RI2N/DRV provides a packet sequencing function to avoid the misinterpretation on TCP/IP layer for non-consequent packets on each channel to reduce the demand of retransmission

- Improve throughput
- Reduce CPU load