Integrated Fault Tolerant Architecture and High-Performance Network

Cuckoo FT-MPI, RI2N and VFREC-NET

MegaScale

http://www.para.tutics.tut.ac.jp/megascale/

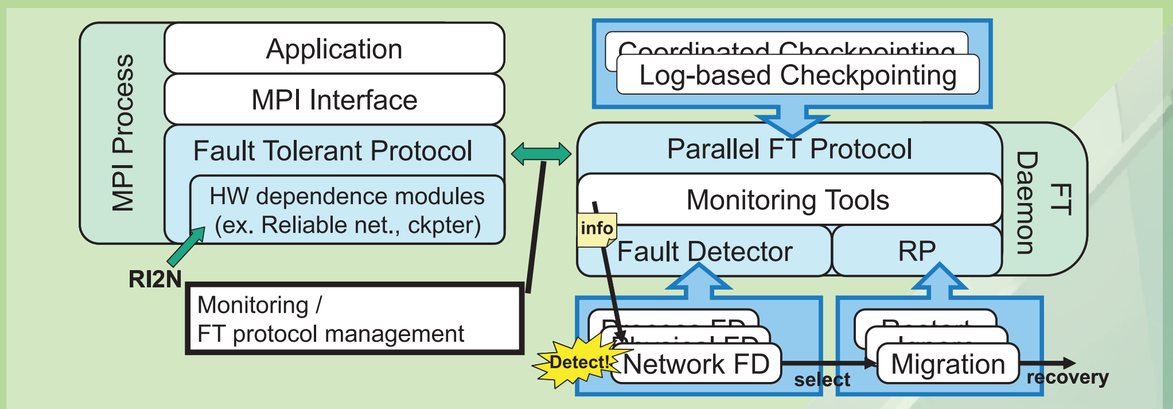
Objective

To obtain high level dependability for large scale parallel computing on MegaScale cluster, different concepts of technology on several software layers are required. We provide fault-tolerant MPI and efficient checkpoint system as well as reliable, scalable and high-bandwidth interconnection network, for this purpose.

Cuckoo FT-MPI

Fault/Recovery Model Aware Component-Based FT MPI

- Cuckoo FT-MPI is a Fault/Recovery model-aware fault tolerant component framework for MPI. Users can customize MPI fault detection and recovery algorithms according to their applications and execution environmental requirements by merely selecting appropriate fault/recovery components.



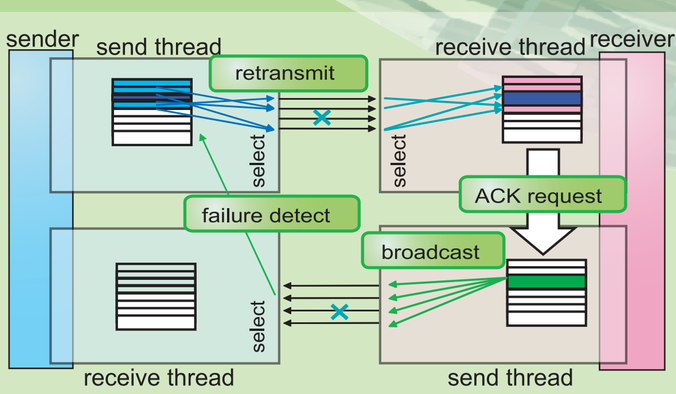
- Fault Detector Components (FD)
  - Selects an appropriate fault recovery protocol (e.g., ignore/restart/migrate) at each fault occurrence
  - Facilitates multiple recovery models to adapt to different applications and computing environment
    - *Example: repeated occurrences of network faults may be due to external causes ← upon reaching repetition threshold other fault detectors are activated & decision delegated*
- Parallel FT Protocol Components (PFTP)
  - A suitable PFTP (e.g., PML/CIC ...) can be selected per each application and computing environment
  - Also useful for evaluating parallel fault tolerant algorithms

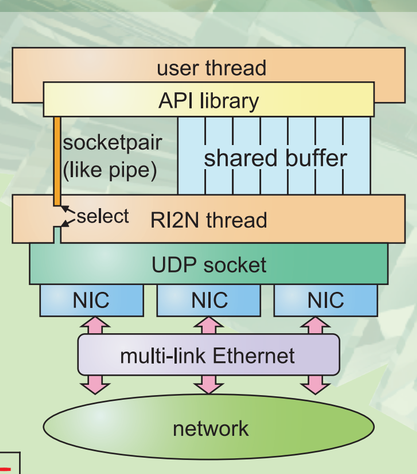
RI2N & VFREC-Net System

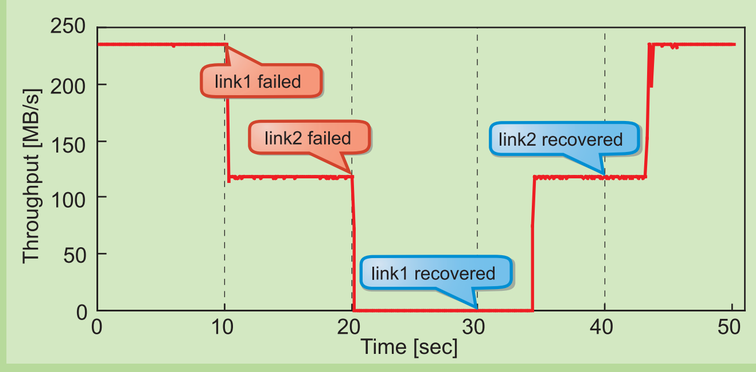
Network System for Clusters with High-Bandwidth, High-Scalability and Fault-Tolerance

RI2N (Redundant Interconnection with Inexpensive Network)

- Utilizes multiple links of Ethernets (e.g., GbE) to achieve both high-bandwidth and high-dependability
- Aggregates bandwidth of multiple links with trunking, and enhances link failure detection with broadcasting of ACK packets
- Completely software-layer implementation; does not depend on IEEE 802.3ad thereby avoiding single point of failure on switches
- RI2N/UDP is an implementation of RI2N on UDP/IP to provide TCP-like streaming



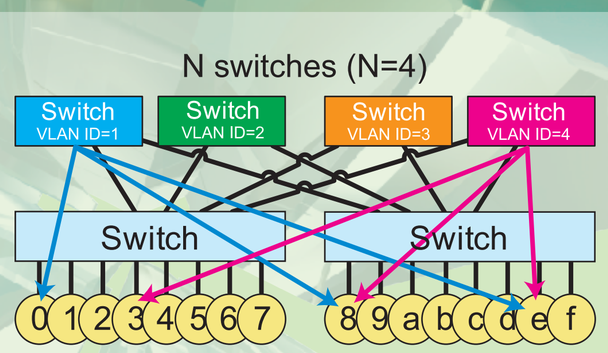




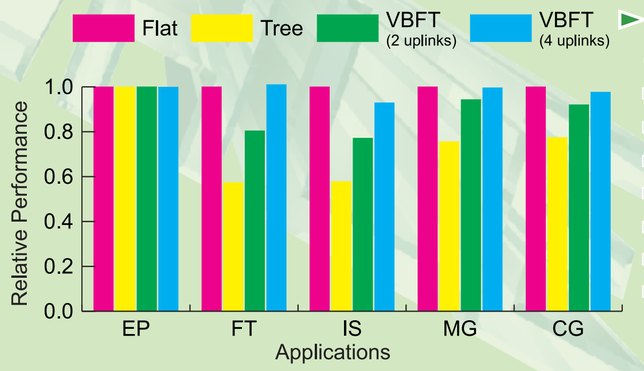
- Throughput on single-side burst transfer using RI2N/UDP with 2-links GbE
  - Intel Xeon 3.0GHz EM64T with HT
  - 1GB DDR2/400MHz Main Memory
  - Linux kernel 2.6.15
  - Intel PRO1000MT Dual port
  - Dell Power Connect 5224

VFREC-Net (VLAN-based Flexible, Redundant and Expandable Commodity Network)

- An interconnection network system based on multi-path Ethernet links to provide high-scalability and wide-bandwidth with inexpensive Layer-2 switches
- Tagged-VLAN technology controlled by a dedicated pseudo device driver makes an explicit routing on VLAN-ready Layer-2 switches
- Various topologies are available including fat-tree and traditional MPP networks



- VLAN Based Fat Tree (VBFT) Network is an example of VFREC-Net, and it provides wide bisection bandwidth and high scalability on the system.



- NPB 3.2 Kernels (Class B, 16 processors)
  - Intel Xeon 3.0 GHz 1-way
  - 1 GByte DDR2/400Hz Main Memory
  - Intel 82541EI Gigabit Ethernet
  - Dell PowerConnect 5224 GbE Switch
  - Linux kernel 2.6.13
  - GCC 4.0
  - LAM 7.1.1