MegaScale

Integrated Fault Tolerant Architecture and High-Performance Network

Cuckoo FT-MPI, MPI Migration, RI2N and VFREC-NET

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

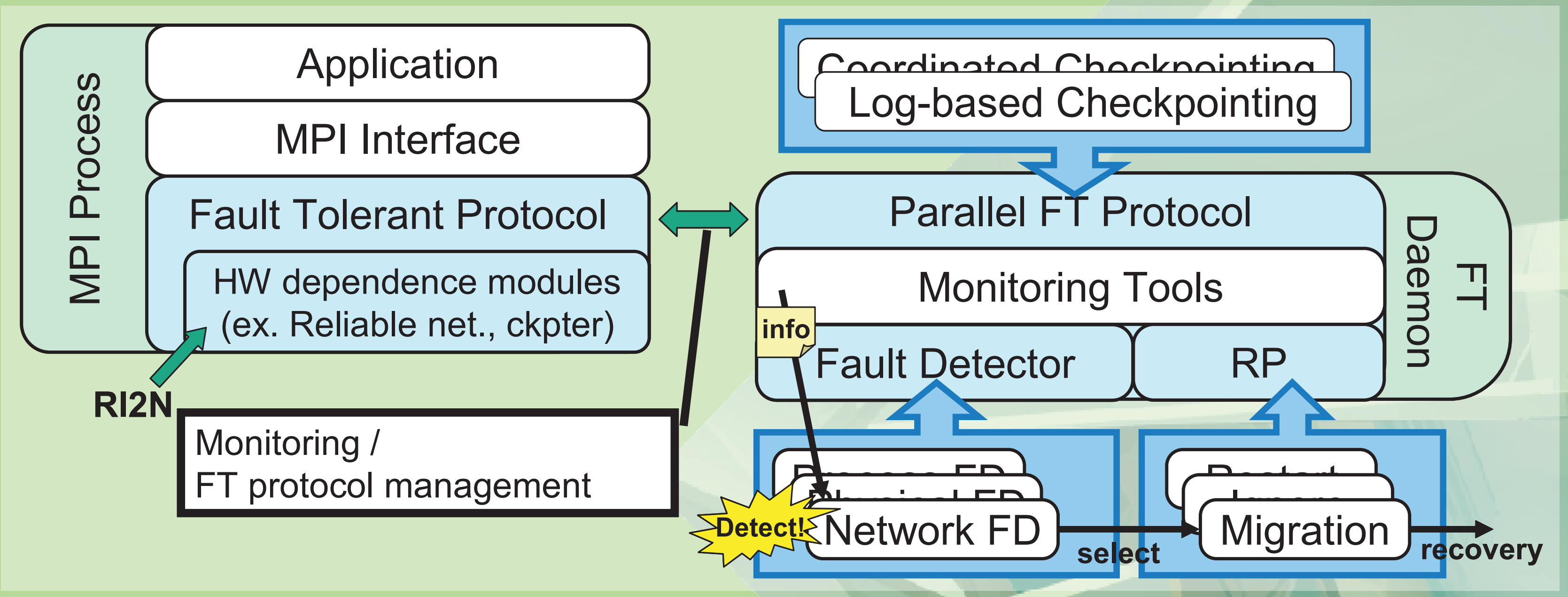
Objective

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 and high-bandwidth interconnection network, for this purpose.

Cuckoo FT-MPI

Fault/Recovery Model Aware Component-Based FT MPI

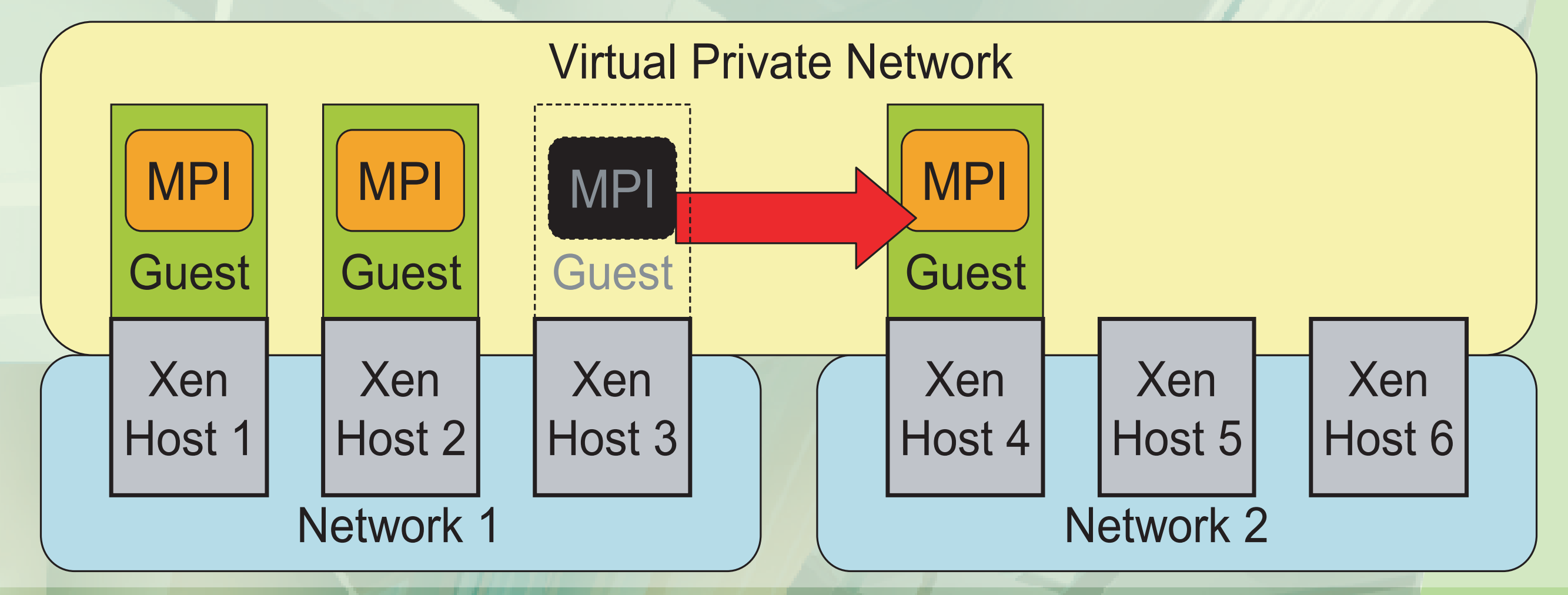
- Fault Detector (FD)
  - Selects an appropriate recovery protocol (e.g., ignore/restart/migrate) at each fault occurrence
  - Facilitates multiple recovery models to adapt to different computing environments
  - Example: repeated occurrences of network faults may be due to other reasons ← upon threshold other fault detectors are activated & delegated
- Parallel FT Protocol (PFTP) Components
  - A suitable PFTP (e.g., PML/CIC ...) can be selected per each computing environment
  - Can evaluate PFTP more accurately
    - Same implementation but exclude PFTP



MPI Migration

Migratable MPI using Virtual Machines

- Motivation
  - Migrating MPI processes for Grid computing, load balancing, fault tolerance, etc.
- Architecture
  - Utilizes Guest OS migration of Virtual Machines (Xen [Pratt et al. '03], VMWare)
  - The Guest OS and the MPI process therein are migrated, with appropriate forwarding
  - Automatically configures VPN to allow migration in wide-area settings
- Benchmark Results
  - Migration overhead is very low (approx. 1 sec. over entire execution)
  - Xen overhead ranges from 0 to 50% depending on network I/O

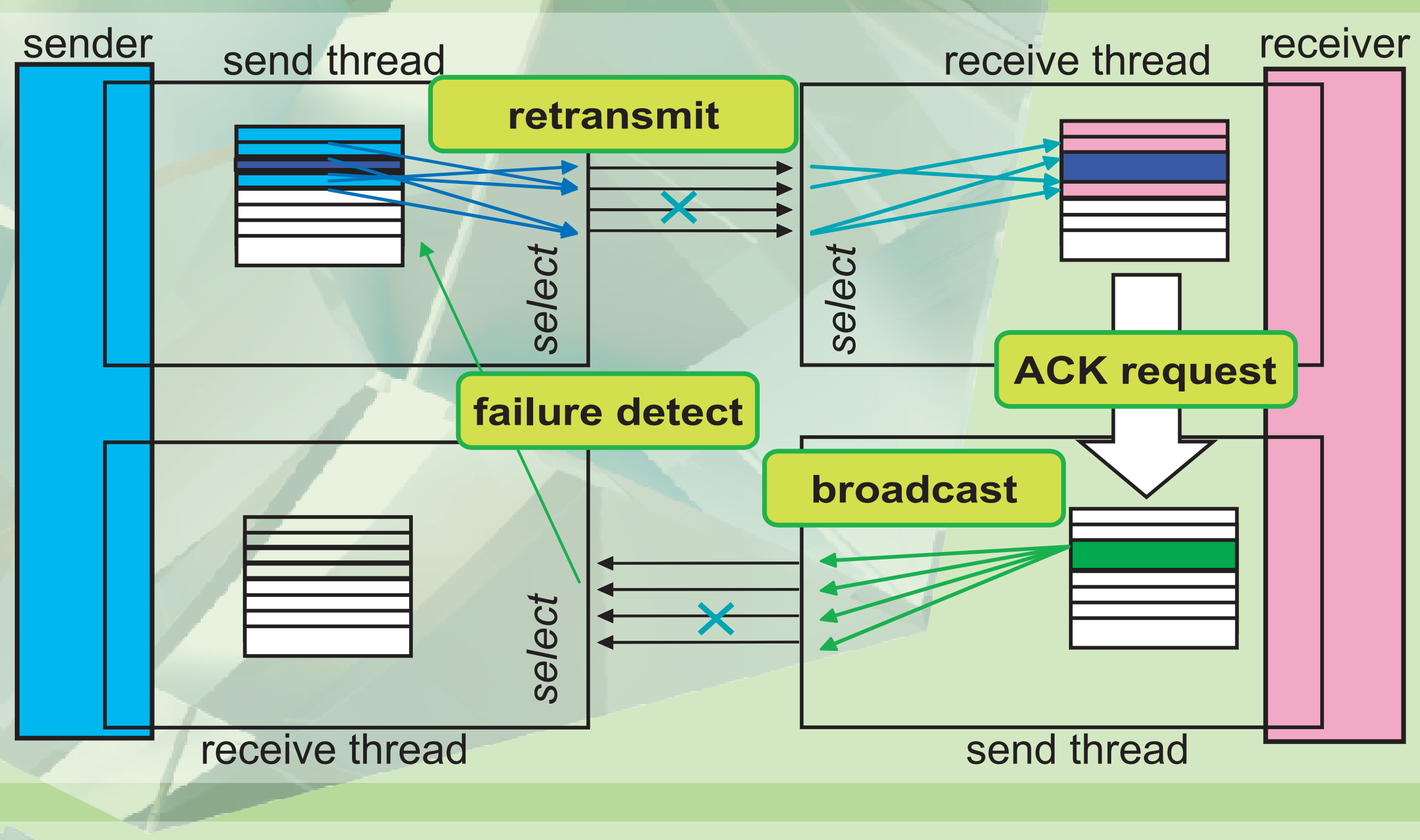


RI2N & VFREC-NET System

Network System for Clusters with Wide-Bandwidth and Fault-Tolerance

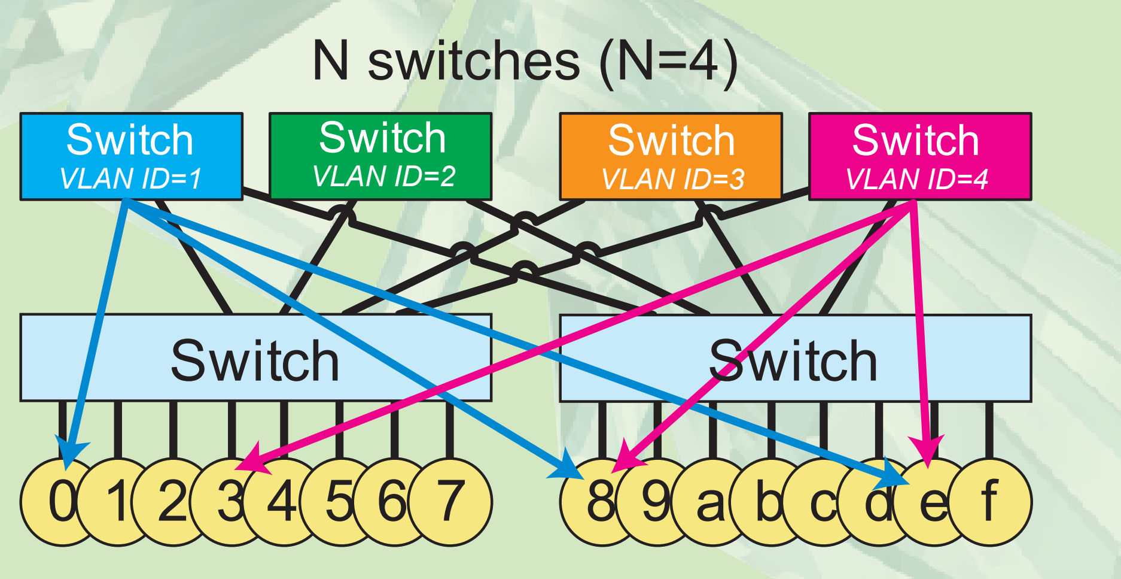
RI2N (Redundant Interconnection with Inexpensive Network)

- Utilizes multiple links of commodity networks (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 failure of switches

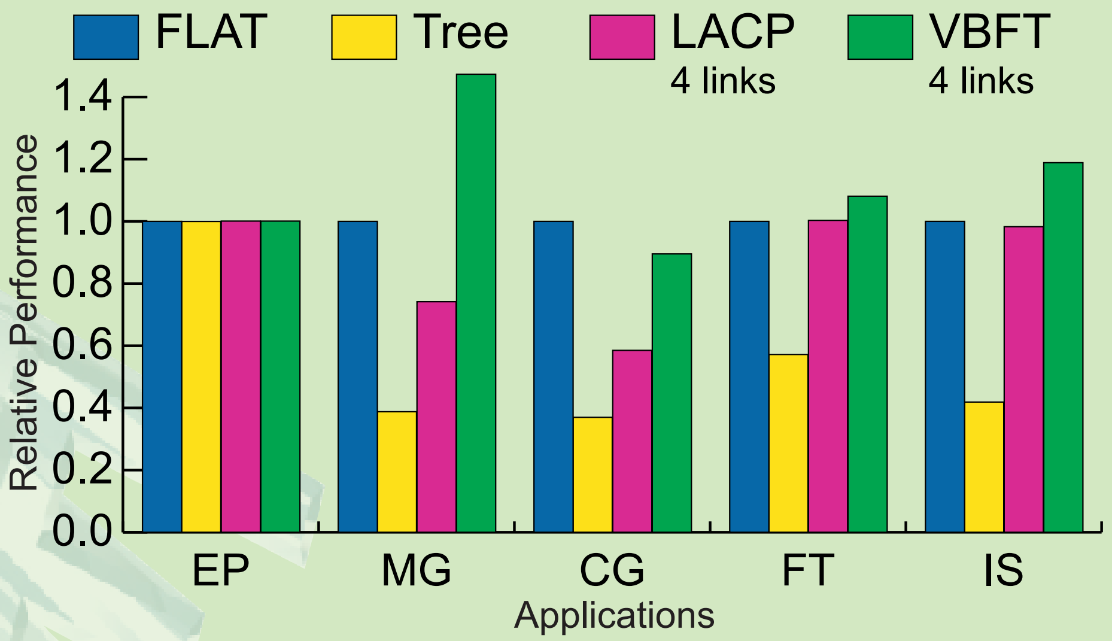


VFREC-NET (VLAN-based Flexible, Redundant and Expandable Commodity Network)

- VFREC-NET provides
  - multi-path interconnection over multiple stages and switches for wide bandwidth on Ethernet
  - variety of network topologies including traditional dedicated MPP networks
- Message routing based on tagged-VLAN technology controlled by a dedicated pseudo device driver to handle VLAN-ID with explicit reference of MAC addresses



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



NPB 3.2 Kernels (Class B, 16 processors)

- Intel Xeon 3.0 GHz 1-way
- 1 GByte DDR2/400 Memory
- Intel 82541EI Gigabit Ethernet
- Dell PowerConnect 5224 GbE Switch
- Linux kernel 2.6.13
- GCC 4.0
- LAM 7.1.1