Network research group topics

- Scalable, high-performance and reliable cluster network
- Low-power and high-performance network link

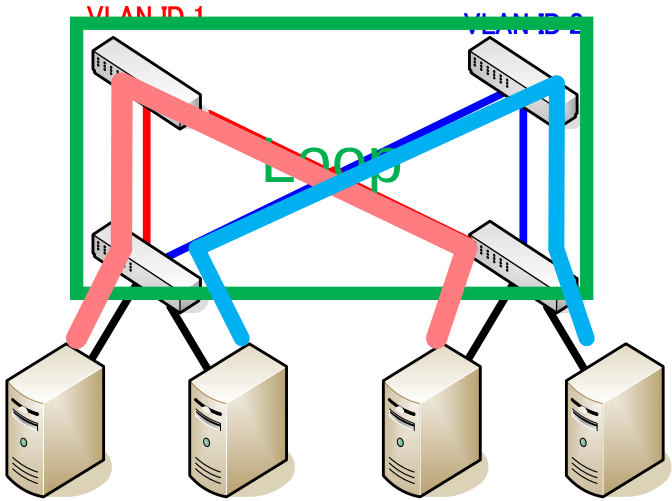
Taisuke Boku


Scalable, high-performance and reliable cluster network

- Background
  - Most of inexpensive PC clusters still use Ethernet (mainly GbE) for cost/effectiveness, but
    - Scalability is limited by its tree-configured network topology
    - Performance is limited by relatively poor peak performance
    - Reliability is limited by intermittent failure on switches (and cables)
  - How to provide scalability, performance and reliability on commodity Ethernet without much cost ?
- Solutions
  - Multi-link aggregation both for performance and reliability in the same manner with RAID in HDD
  - Multi-path utilization to configure fat-tree topology under conditions of Ethernet

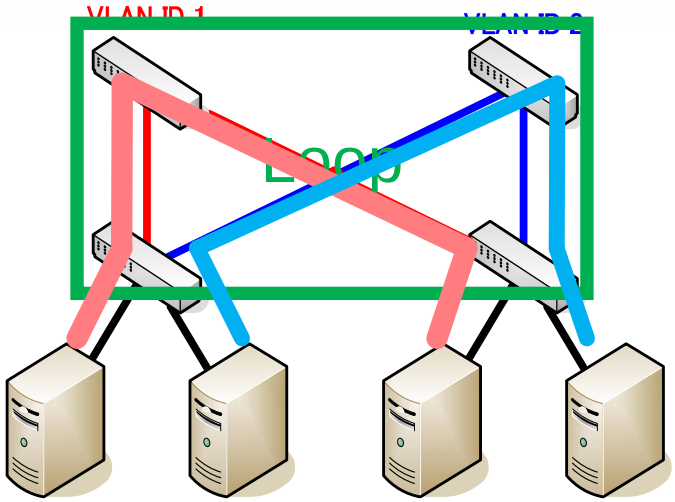
VFREC-Net: Scalable and high-performance network

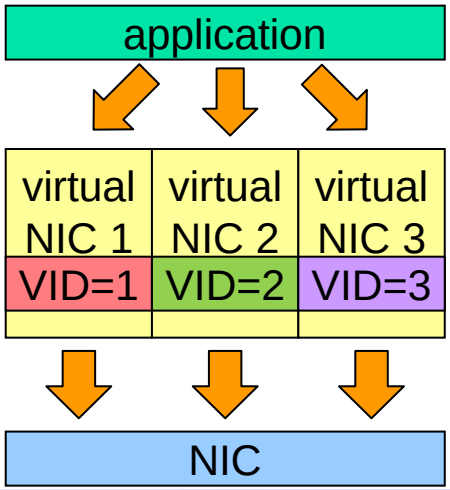
- Simple tree network topology largely degrades the total bandwidth on Ethernet with a large number of PC cluster nodes
- SAN (e.g. Inifiniband) supports Fat Tree to distribute the traffic load on multiple switches on upper-level in tree
- Fat Tree is not available with Ethernet because only “tree configuration” is allowed, but no cyclic graph
- Solution: If we can make “logically isolated” network on each logical tree ⇒ VLAN technology



VLAN routing method (by Kudo@AIST)

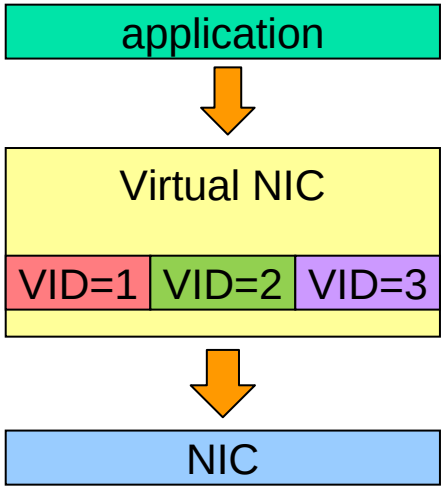
- Each VLAN configures isolated Tree
  - Eliminating the logical loop
- Each node belongs to one or more VLAN and Linux driver determines the VLAN-ID (tag)
  - Flat distribution of VLAN-ID distributes the load on upper-level
  - Possible to make Fat-Tree configuration
- Original method by Kudo depends on Linux virtual NIC
  ⇒ fixed and limited configuration, several detailed technical difficulty

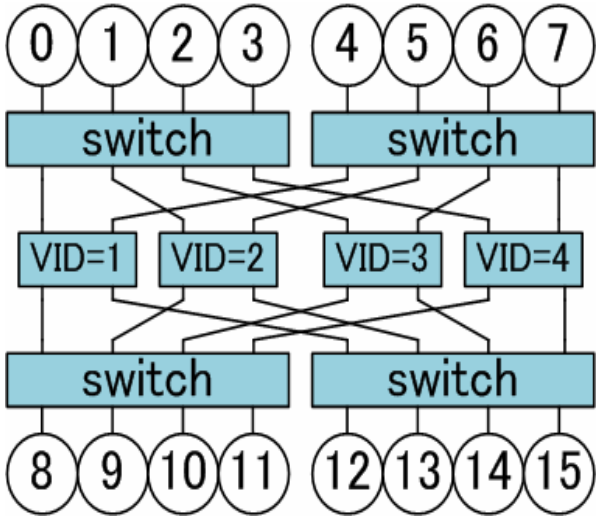




VFREC-Net

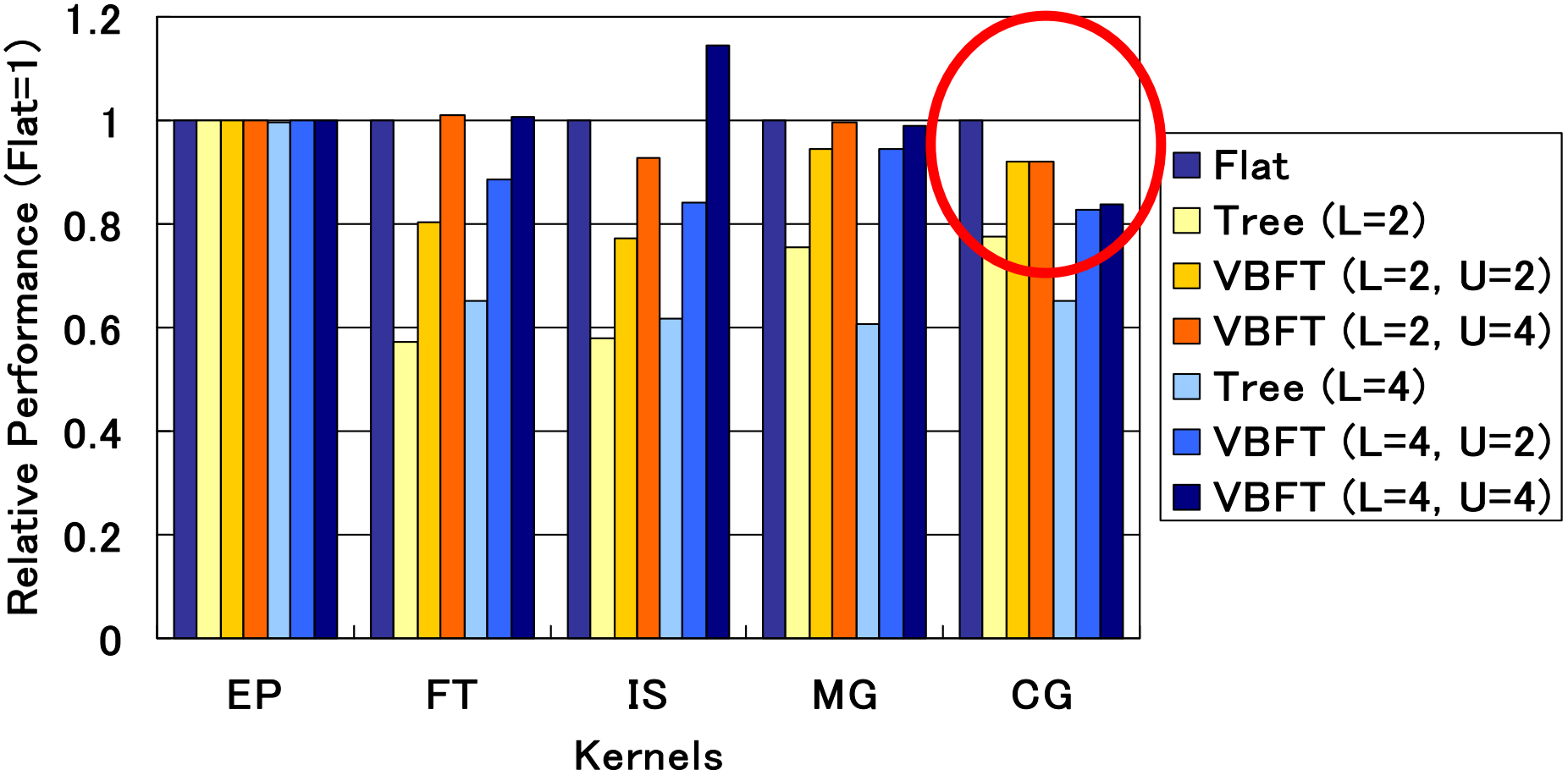
- VLAN-based Flexible, Reliable, Extendable Cluter Network
- Based on VLAN routing technology
- Developed new driver to handle multiple VLAN-ID by software on virtual NIC (VFN driver)
  - Table-based VLAN-ID decision
  - Can be referred as ordinary Ethernet device to support any protocol and communication library without modification
- VBFT (VLAN-Based Fat Tree) is supported





NAS Parallel Benchmark with Xeon cluster with 32 nodes

Why not improved ?



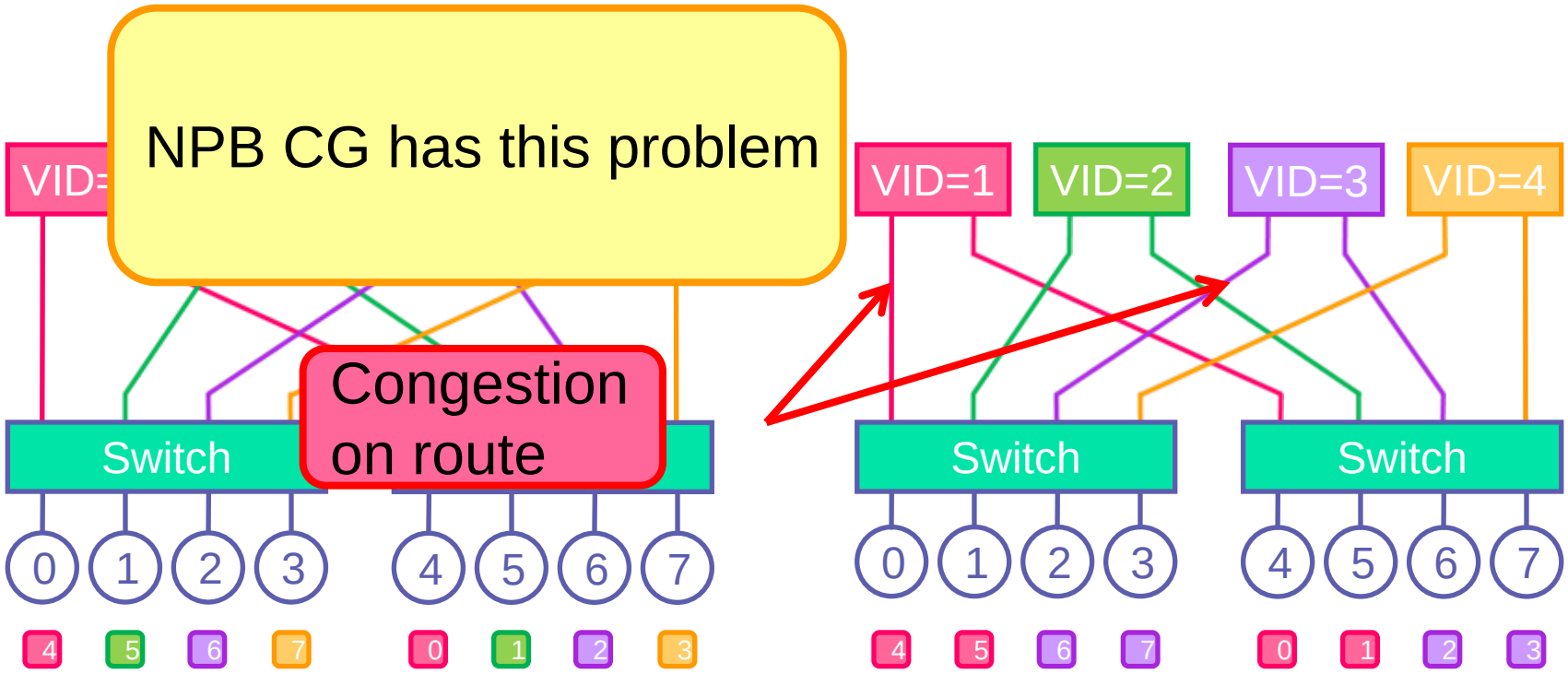
Traffic distribution on upper-level switch

Ideal case

- Flatly distributed on upper-level
  -> high performance

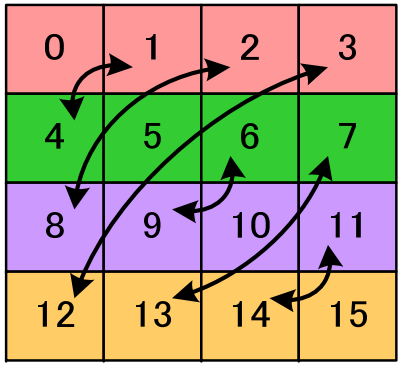
With imbalanced traffic

- Concentrated traffic on some switches
  -> low performance

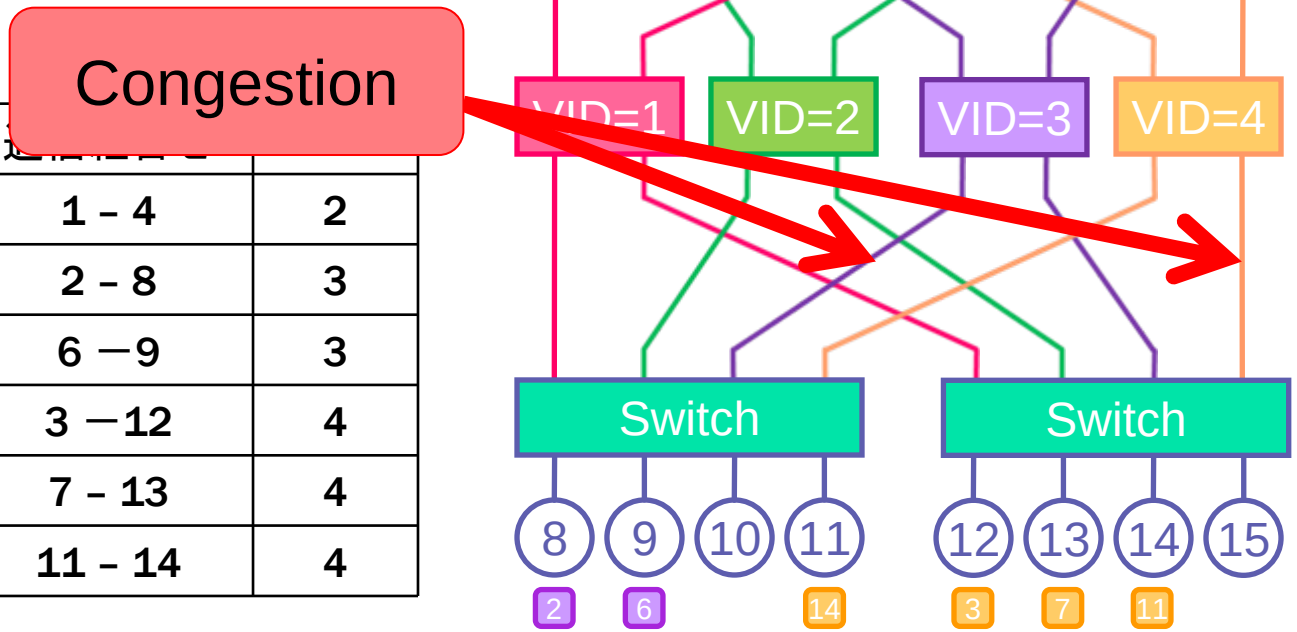


Traffic pattern in NPB-CG

- Matrix transposing is dominant
- Traffic congestion by default (flat) VID distribution



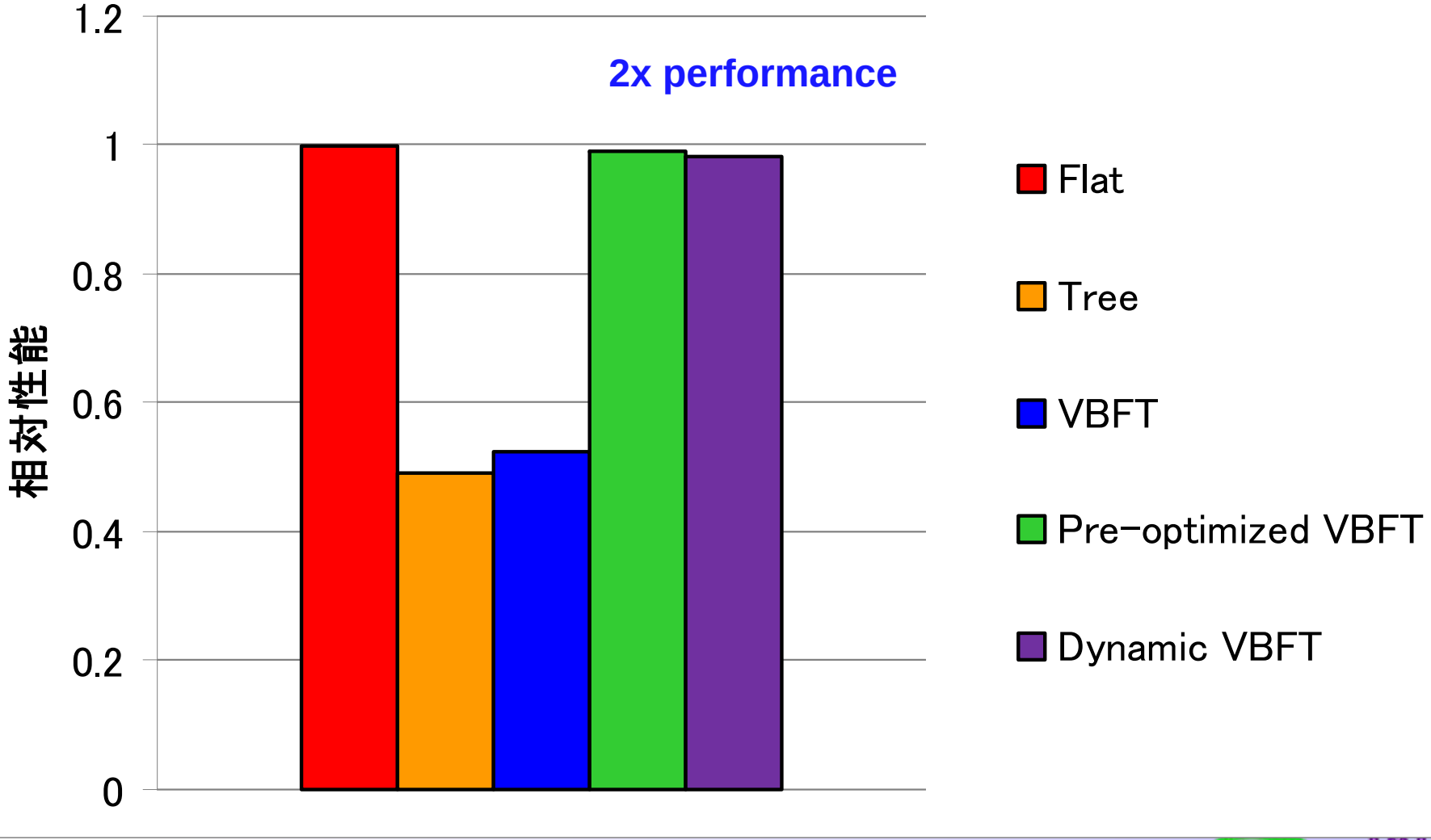
1 – 4	2
2 – 8	3
6 —9	3
3 —12	4
7 – 13	4
11 – 14	4



Dynamic routing on VBFT

- We modified the VBF driver (virtual NIC) for dynamic modification of VLAN-ID assignment table
  ⇒ It corresponds to “dynamic routing”
- User API is provided for modification of VLAN-ID table, which is available even in MPI programming
- “How to optimize the routing according to the traffic” is still open question, but we provide a generic platform for user/system level dynamic routing
- Applying it, NPB-CG performance is improved

NPB Kernel CG result (with dynamic routing)



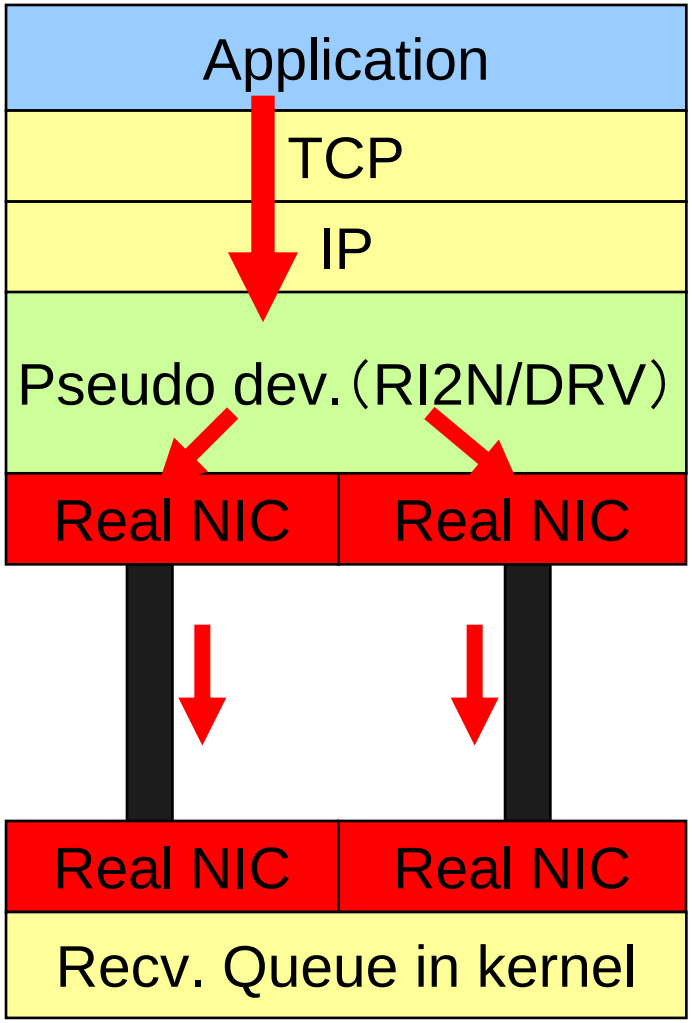
VFREC-Net with dynamic routing

- Enables very cost-effective high-performance network with any size of PC clusters
- Today's most of inexpensive L2 Ethernet switches are equipped with IEEE802.1Q VLAN technology, so it can be applied very widely
- Just adding a virtual NIC driver and several daemons
- Provides user-flexibility to tune the network routing according to the applications behavior
- Optimization is still open question, but we provide the way to apply it

RI2N: Reliable & high-performance network

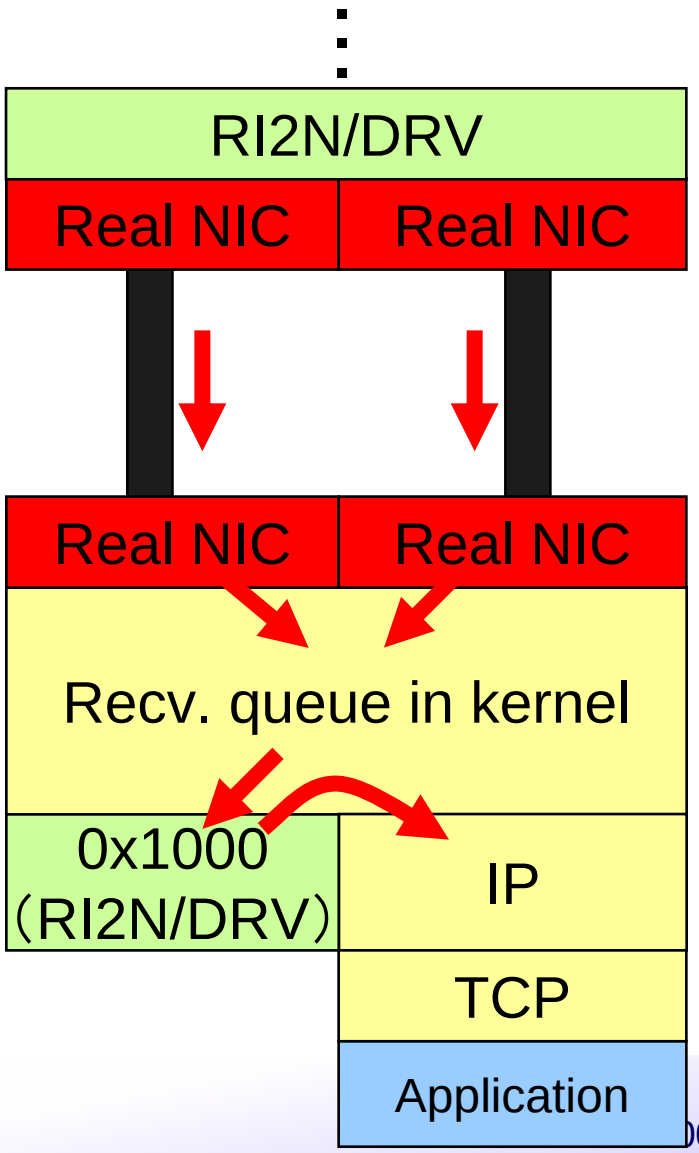
- RI2N: Redundant Interconnection with Inexpensive Network
  ⇔ RAID: Redundant Array of Inexpensive Disks
- Multiple (parallel) network links are used for
  - Multiplying the bandwidth (without failure)
  - Fault recovery network with live links (on link or switch failure)
- The network must be user-transparent, so we implemented it as a pseudo network driver on GbEhernet

Data sending



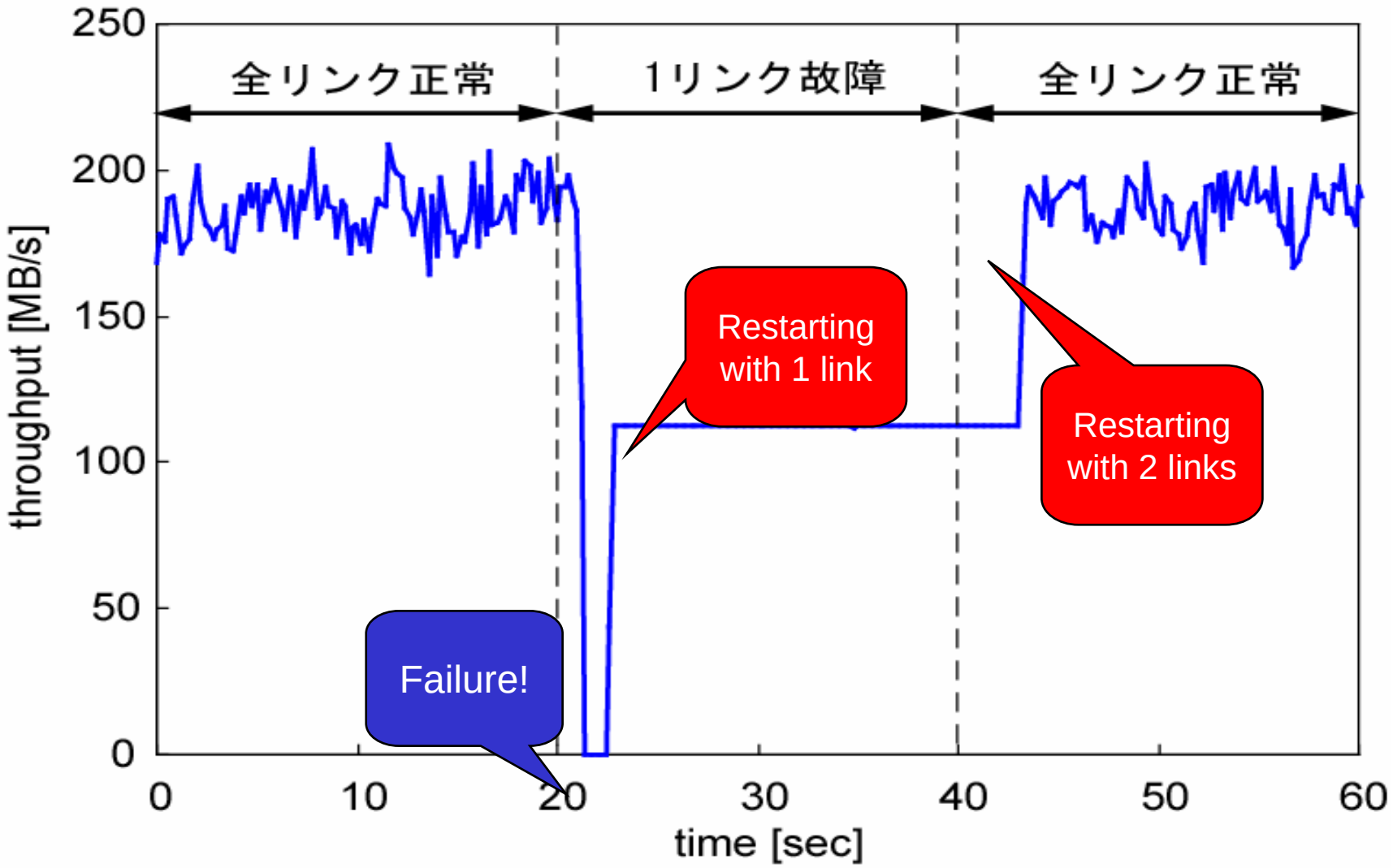
- Data transmission
  - Pseudo device is added
  - Device transmit function is called according to IP space
  - Rewrite the protocol ID for correct receiving on receive node
  - Select real-NIC in round-robin manner, then push the packet

Data receiving



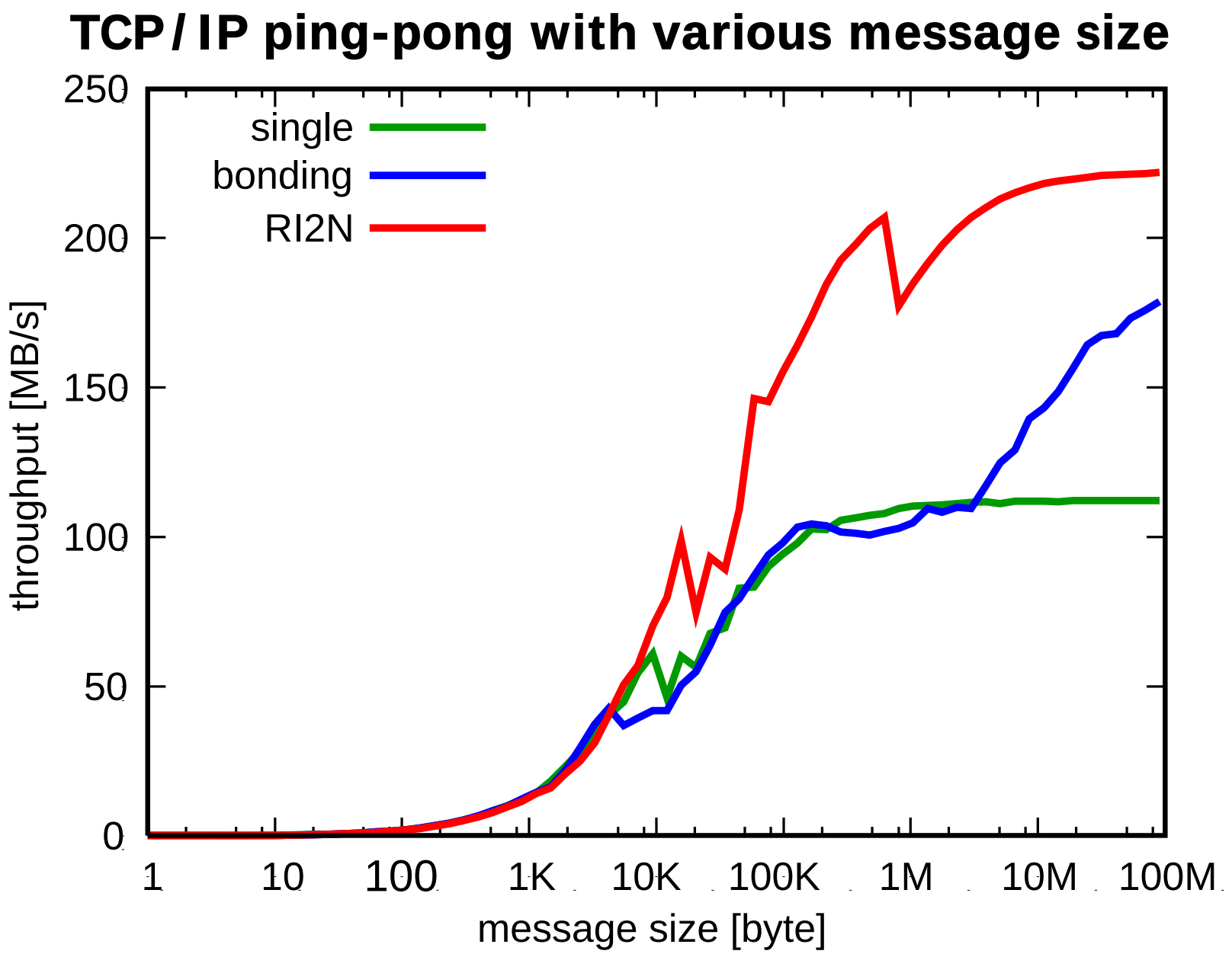
- Data receiving
  - Add the protocol handler
    - Handling pseudo protocol (0x100)
  - Rewrite the protocol ID and merge the split packets in multiple channels
  - Re-push the packet in to kernel
    - Real protocol handling is performed on higher layer

Throughput on failure



TCP/IP communication

Comparison with Linux “channel bonding”



Low-power & high-performance link

- Study just started in this year
- Preliminary study on next generation MPP system
  - CPU power consumption has been reduced while the network power consumption is increasing
  - For the future systems with millions of processors, “global” network is impossible
    ⇒ “local” network (something like “mesh”)
  - Network link power is mainly used for driving for long distance communication (e.g. 100m for Ethernet)
    ⇒ reducing the communication range much contributes to save the power for network link
  - “DVFS” + “Parallelization” ⇒ new generation network link
  - Reliability is also required for large scale network

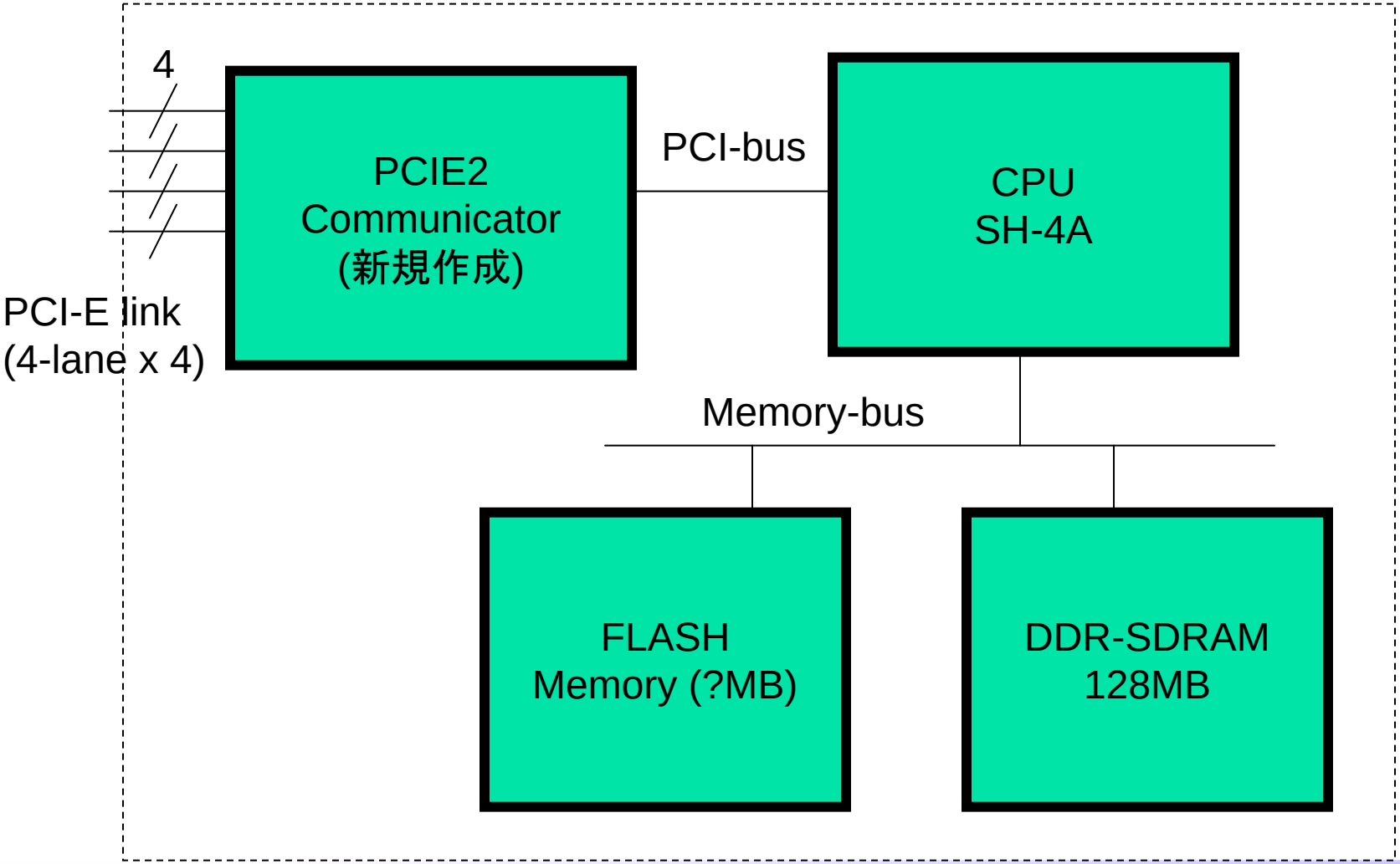
Reducing the network power on node

- Short-range & high-frequency network link
- Link-aggregation with relatively low performance (low power) rather than a single high performance link
  - Reducing power
  - Redundancy for fault tolerant on link disconnection or switch failure
- Dynamic frequency changing according to the application requirement
  - Low data rate is enough for CPU-bound applications
  - Real-time application requires “slow” mode or “rush” mode
  - Dynamic controlling is necessary for total power reduction

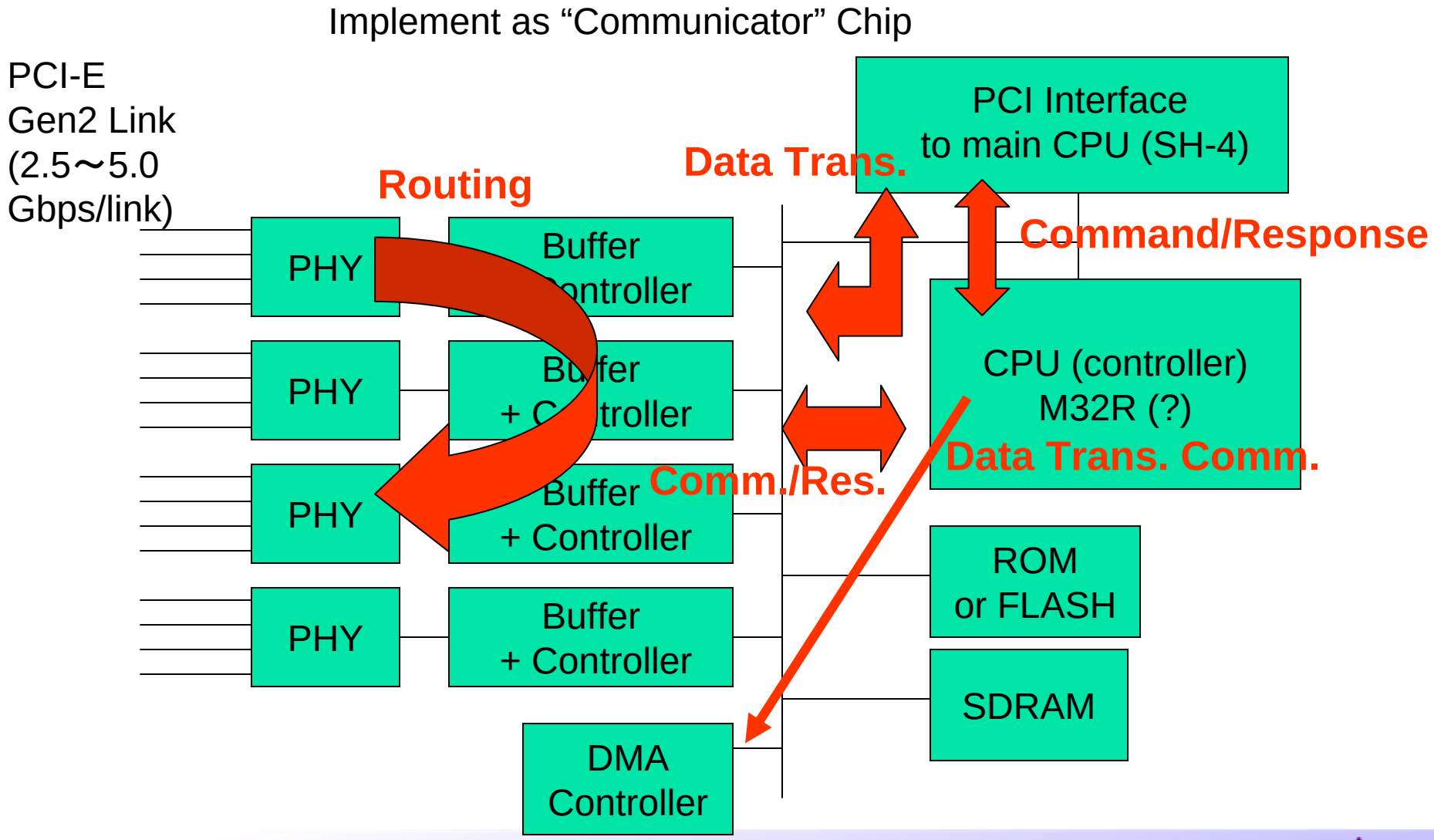
Our solution

- Using PCI-Express Gen2 network as the platform
  - Basic serial link can be used not just for I/O device connection but also for node-to-node connection (interconnect)
  - 2.5Gbps/5.0Gbps mode change by device control
  - Multiple lanes for various aggregated data rate
- Development of new network link device and controlling chip
  - Covering all features of PCI-E gen2
  - Dynamic lane# controlling to increase/decrease data rate
  - Link malfunction is automatically covered by othre links
  - Routing function to be an interconnection router on parallel processing system
  - Without big PHY for long distance communication
- Prototype development with Renesas Technology

Block diagram of SiP to develop



“Communicator” Block Diagram



Status

- In JST/CREST project, the platform is developed as a prototype of high-performance embedded sysem
- FPGA version of logic will be delivered on March 2008
- With FPGA board which corresponds to a single channel (4 lanes), system software on controlling processor (M32R) will be developed
- Communicator chip with 4-lanes x 4 channels will be shipped on March 2009
- SiP module with a compact module for embedded HPC will be available on March 2010 ?