Performance Evaluation of Routing Algorithms in RHiNET-2 Cluster

Michihiro Koibuchi, Konosuke Watanabe,
Kenichi Kono, Akiya Jouraku,
and Hideharu Amano

Keio University, Japan

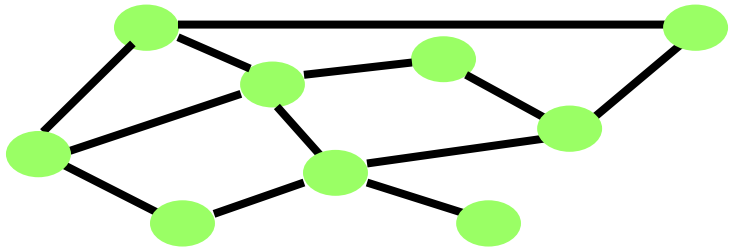
Overview

- Introduction
  - What is the System Area Networks (SANs)?
- Deterministic routing
  - Up*/Down* routing
  - Structured buffer pools (SBP)
  - The DL routing (developed by Keio University)
- The RHiNET-2 cluster system
- Performance evaluation
- Conclusion

Large-scale parallel computing

- Traditional parallel computers (multi-computers)
- PC clusters
  - Beowulf cluster
    - PC + LAN(Ethernet) with TCP/IP
  - High-performance cluster
    - PC + SANs(TCP/IP off-loading) with free topologies.

(e.g.Myrinet,InfiniBand, RHiNET)



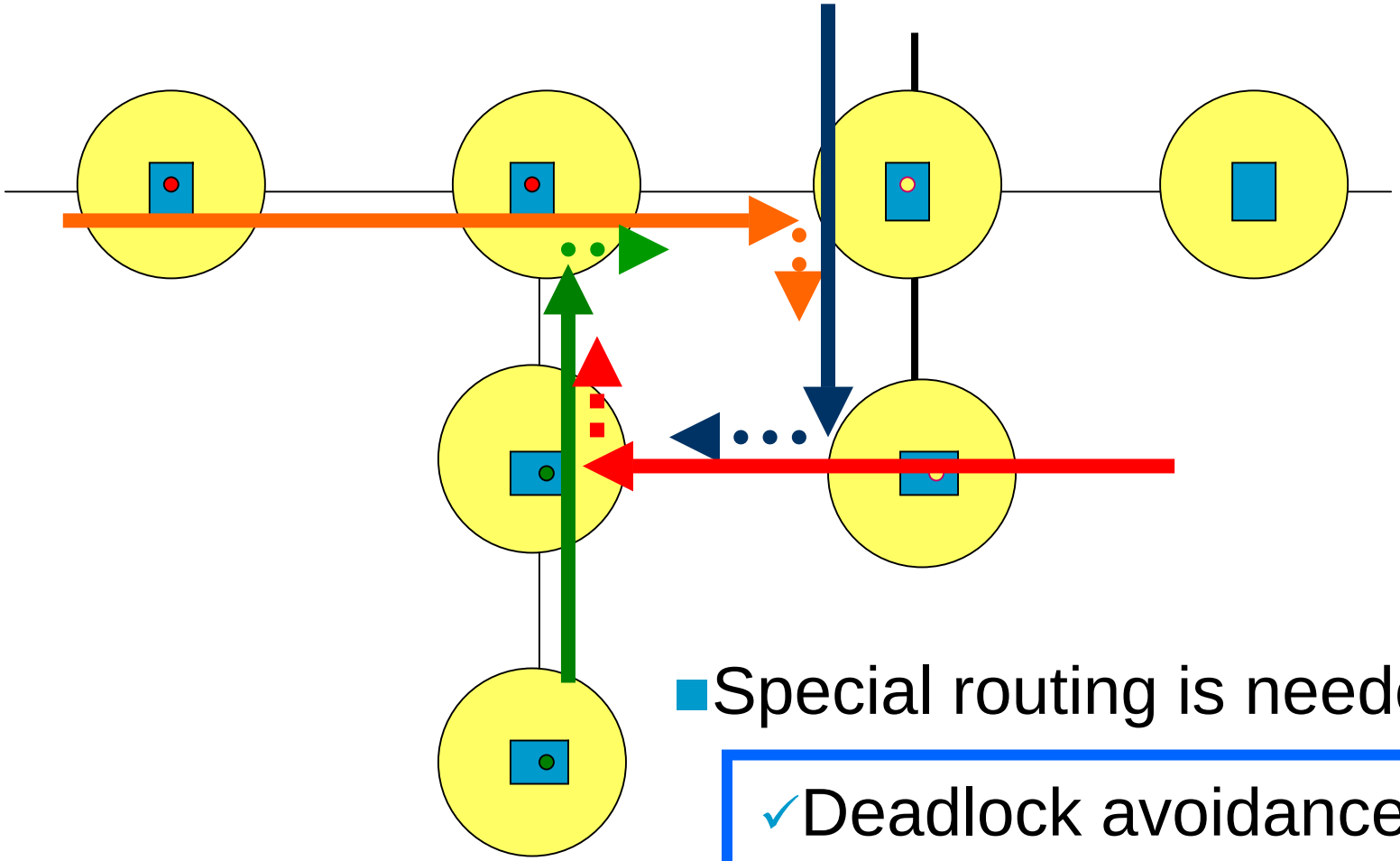
SAN for PC clusters

- Network Interface on PC + Switch + Link
- High-speed direct-memory-communication is required.

↓

- Switching tech. is VCT/WH (Go & Stop): Not Store-and-Forward manner

Deadlock problem on communication



- Special routing is needed.
  - ✓Deadlock avoidance
  - ~~✓Deadlock recovery~~

Overview

- Introduction
  - What is the System Area Networks (SANs)?
- **Deterministic routing**
  - **Up*/Down* routing**
  - **Structured buffer pools (SBP)**
  - **The DL routing (developed by Keio University)**
- The RHiNET-2 cluster system
- Performance evaluation
- Conclusion

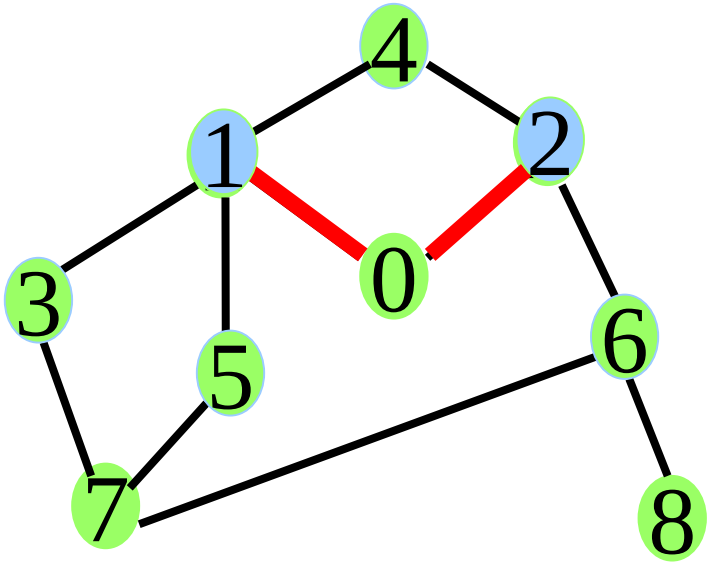
Deadlock-free deterministic routings

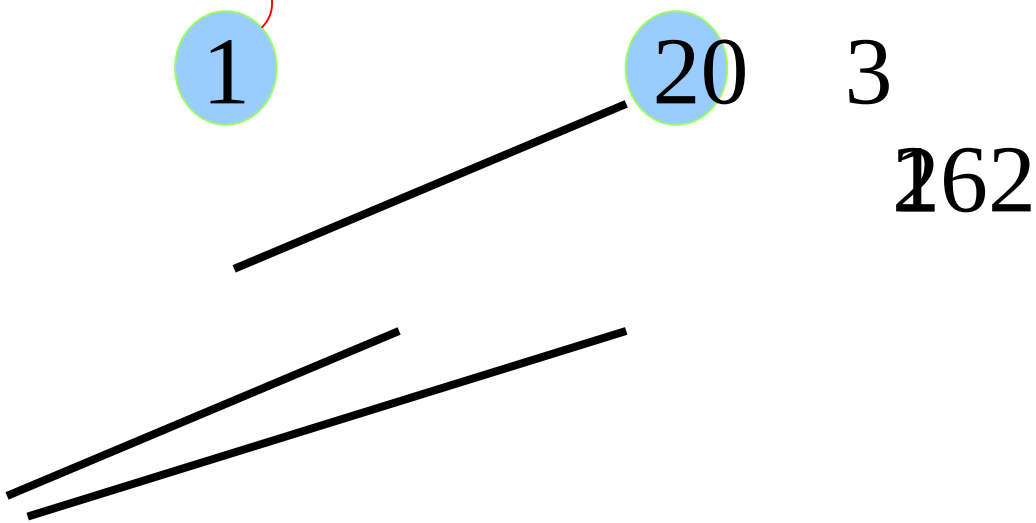
Deterministic routing is preferred to adaptive routing in SANs.

- Up*/down* routing(1990,Autonet)
  - requires no vchs, non-minimal paths
- Structure buffer pools(SBP) based routing
  - (Diameter+1) vchs, minimal paths
- DL routing (2002, developed by us)
  - More than 1 vch
  - shorter paths than Up*/Down* routing

graph

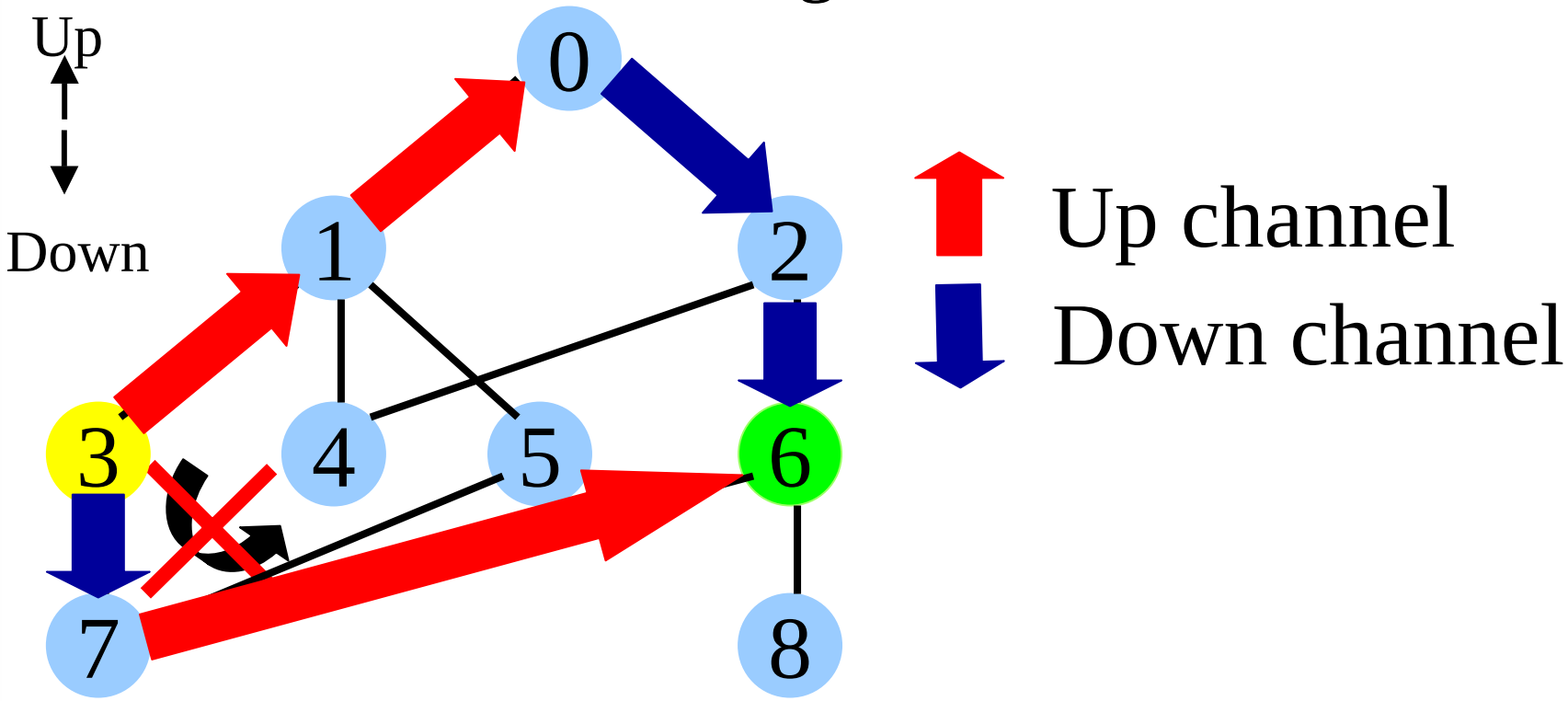
2. Add the rest node



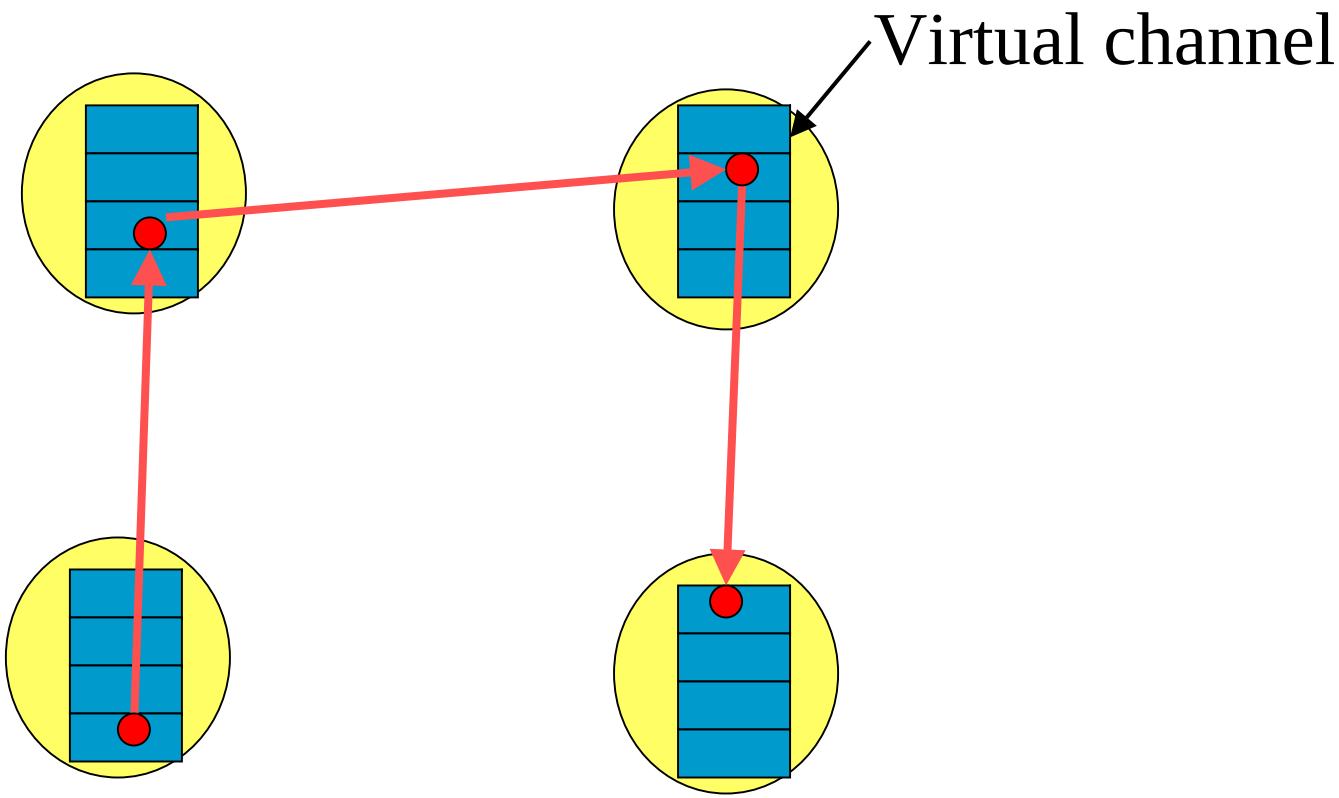


Up*/down* routing (2)
-Restrict routing paths

- After using up channel, use down channel.
- Non-minimal routing



Structural buffer pools (SBP)

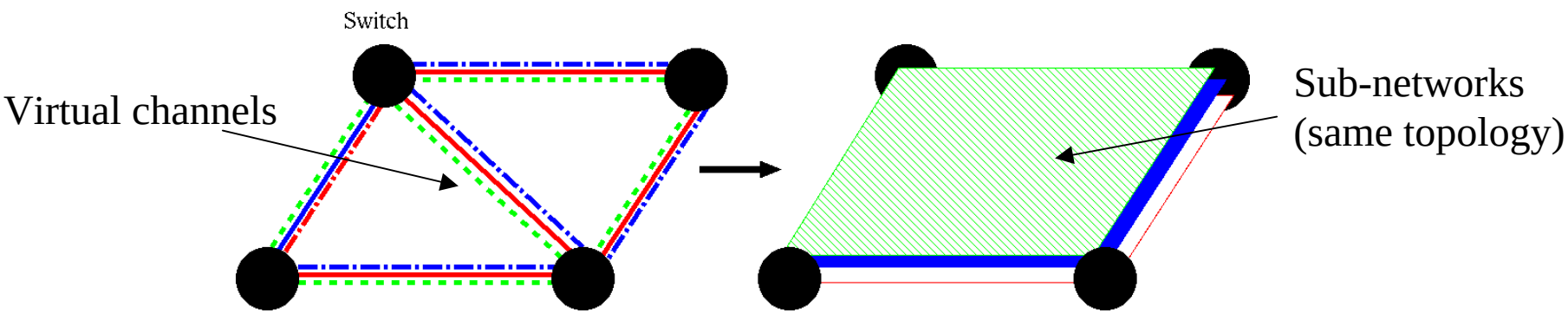


- sends a packet to (the channel number +1).
- takes (Diameter+1) virtual channels.

The DL routing

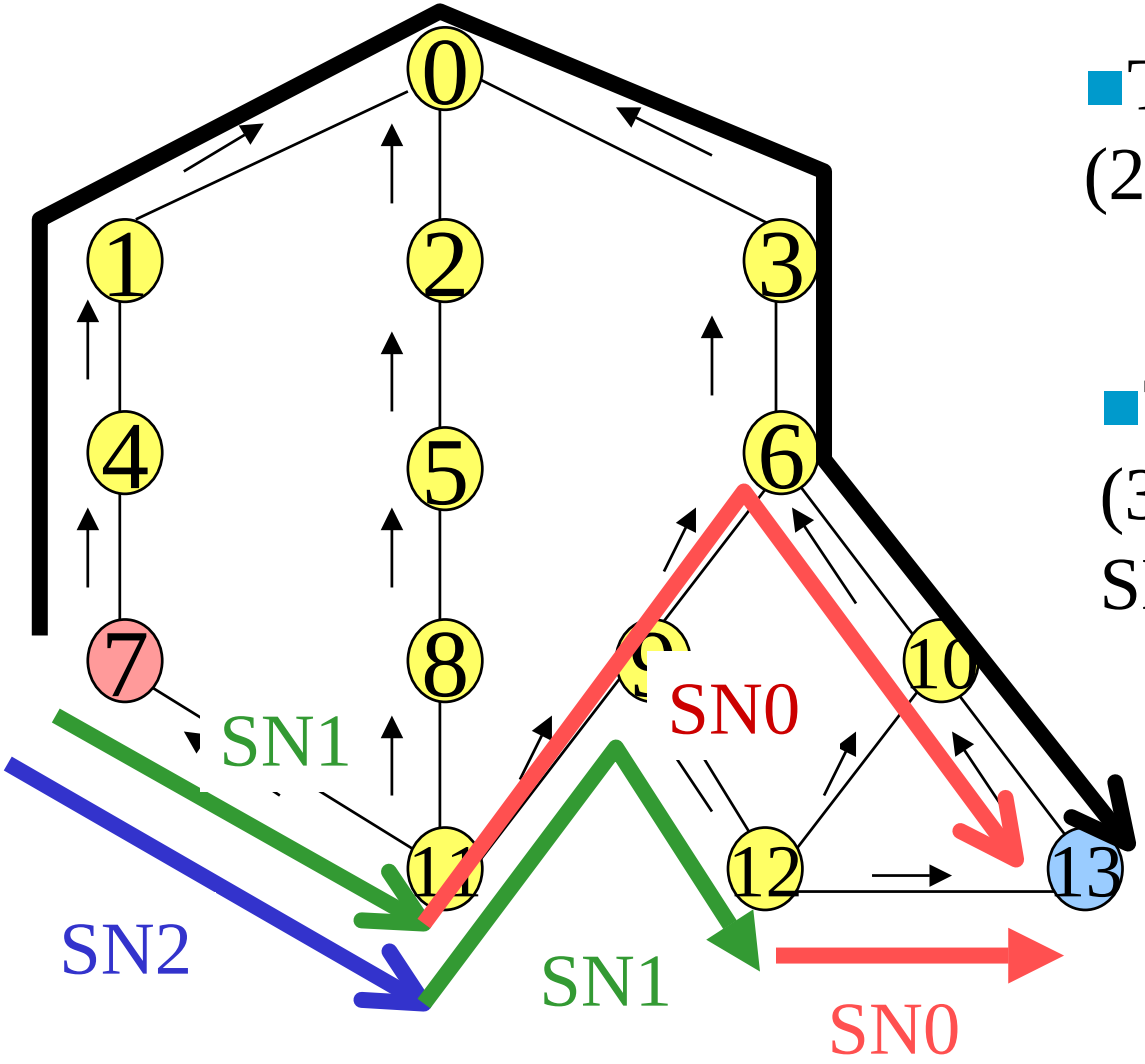
1. Divide the network into the multiple sub-networks.
   - Sub-network is the same topology as the target network.
2. Avoid dead-locks.
   - Impose up*/down* routing within each sub-net..
   - Use some sub-networks in the descending order.
3. Establish a single shortest path.
   - Descend a sub-network when forwarding from down direction to up direction.

Shorter paths than up*/down* routing



Routing example

- **Switch 7 → Switch 13**



- Up*/down* routing

7hops

- The DL routing (2 sub-networks)

5hops

- The DL routing (3 sub-networks), SBP

4hops

Characteristics of deterministic routings

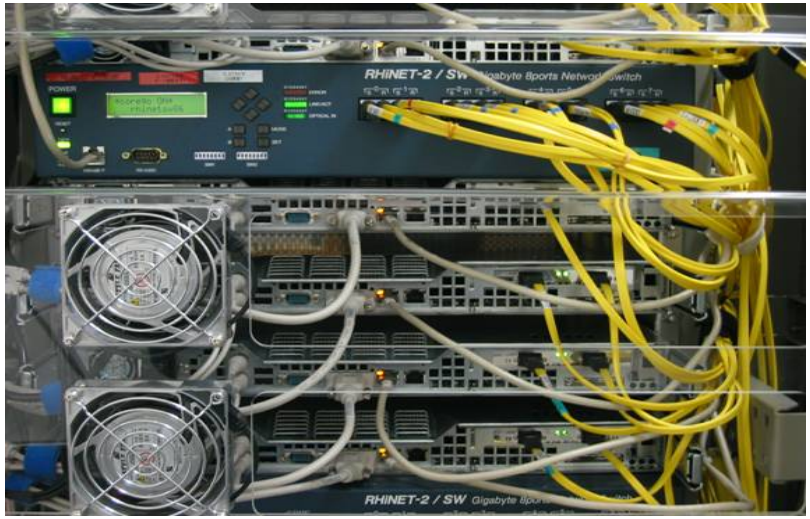
	Size limit ?	Minimal path ?	Virtual channel ?
Up*/Down*	No	No	No
SBP	Yes	Yes	Yes
DL	No	No	Yes

Previous simulation, analysis works indicate that SBP >= DL >= up*/down*.

Overview

- Introduction
  - What is the System Area Networks (SANs)?
- Deterministic routing
  - Up*/Down* routing
  - Structured buffer pools (SBP)
  - The DL routing (developed by Keio University)
- **The RHiNET-2 cluster system**
- **Performance evaluation**
- Conclusion

The RHiNET-2 cluster system

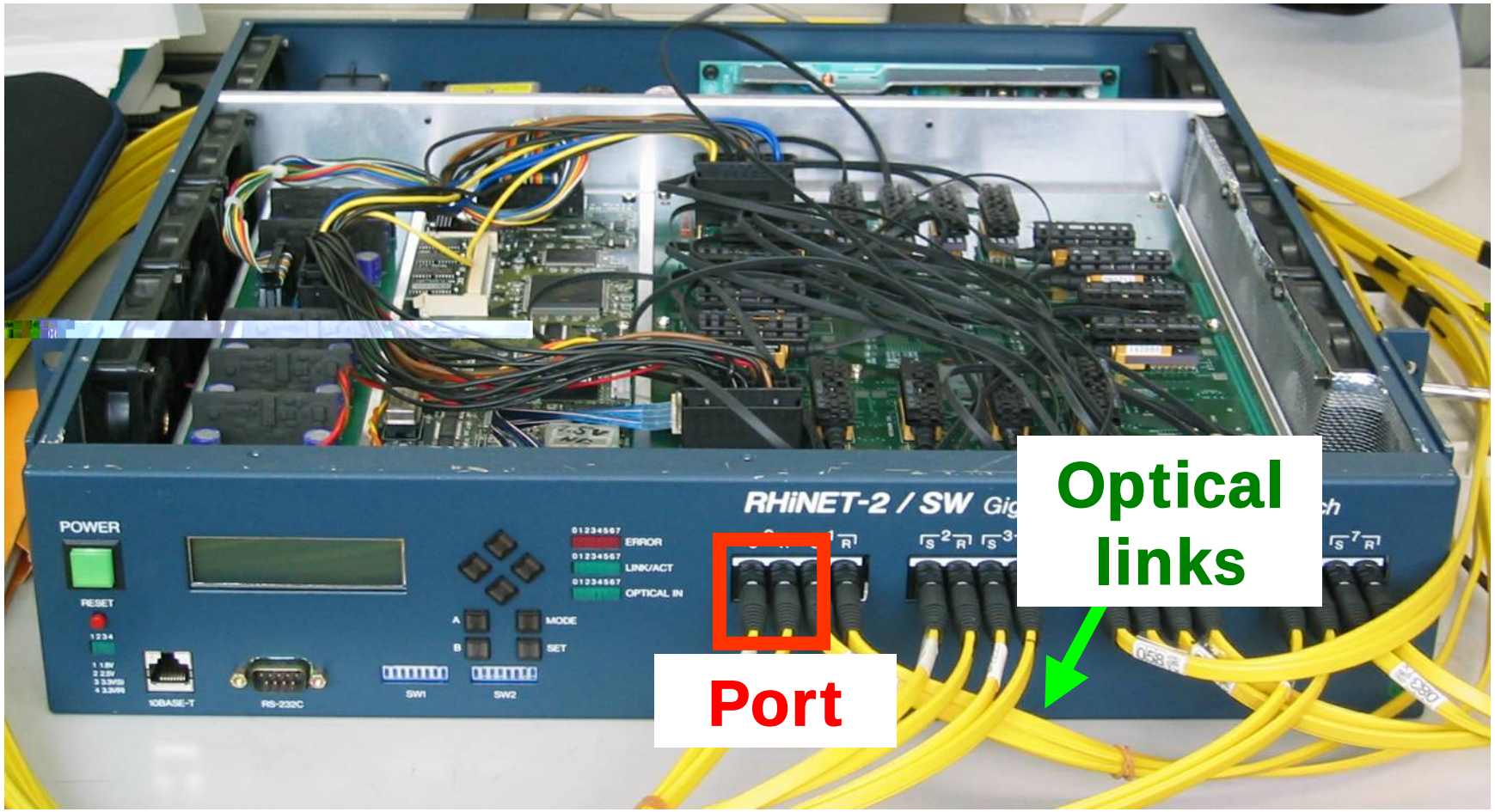


- Developed by RWCP, Hitachi Co, and Keio University
- 64 hosts + 16 switches + optical link
- Deterministic routings with table look-up manner.
- Supporting arbitrary topologies and routing algorithms by rewriting the routing table.

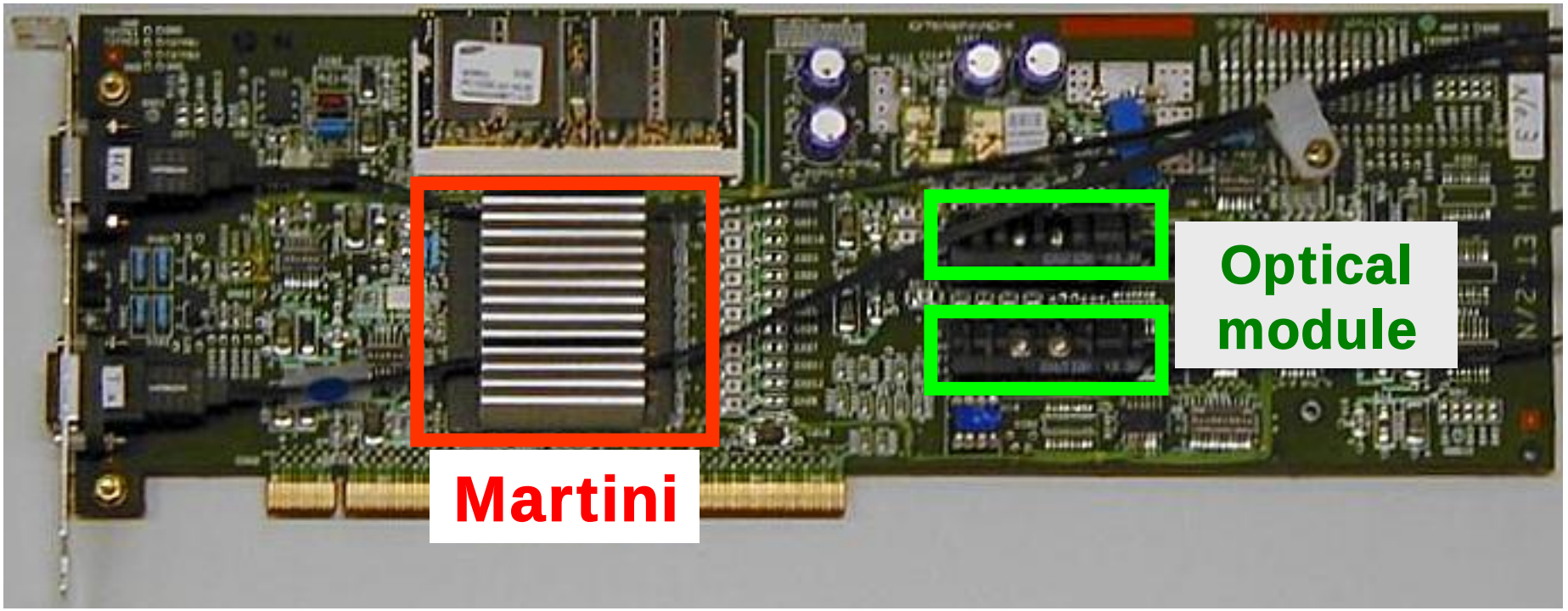
The RHiNET-2 cluster system

- RHiNET-2/SW (network switch)
  - Throughput: 64Gbps
  - 16 virtual channels
- RHiNET-2/NI (network interface)
  - User-level Zero-copy communication
  - Remote DMA and PIO based transfer
- Host PC
  - CPU: Intel Pentium III 933MHz (x2)
  - Memory: 1Gbyte SDRAM
  - PCI bus: 64bit/66MHz
  - OS: RedHat Linux 7.2、kernel-2.4.18

RHiNET-2/SW


RHiNET-2 / SW
POWER
RESET
ERROR
LINK/ACT
OPTICAL IN
MODE
SET
10BASE-T
RS-232C
SW1
SW2
Optical links
Port

RHiNET-2/NI



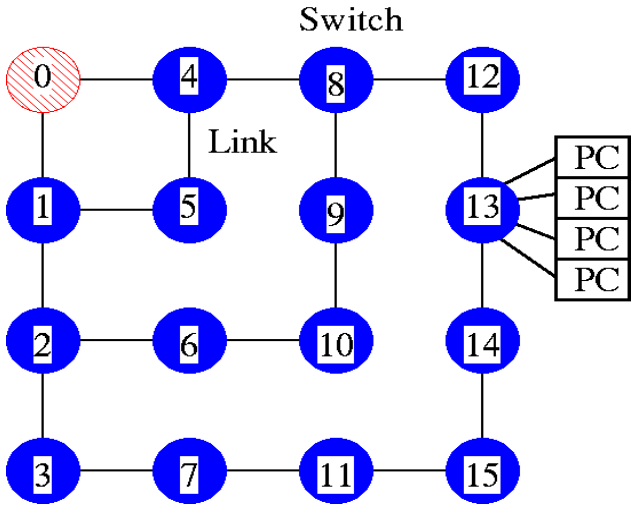
Measures

- Latency of barrier synchronization on 64 hosts[usec]
  - NIC based method: 12(2*log 64) steps
  - Using the PIO-based low-latency transfer
- Bandwidth
  - Typical traffic patterns
    - Bit rev.,matrix transpose, butterfly, complement
  - Using R-DMA transfer

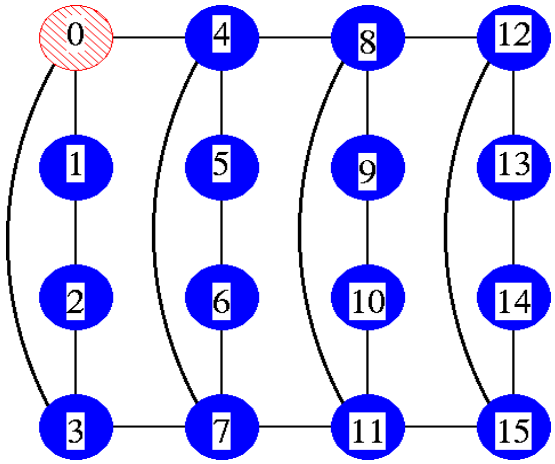
Routings and topologies

- Three routings x Three topologies

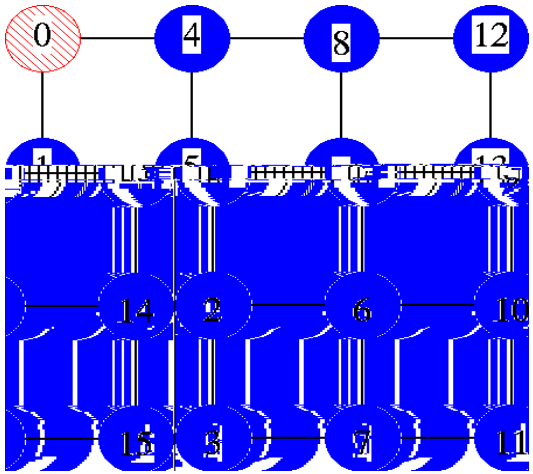
Topology A



Topology B

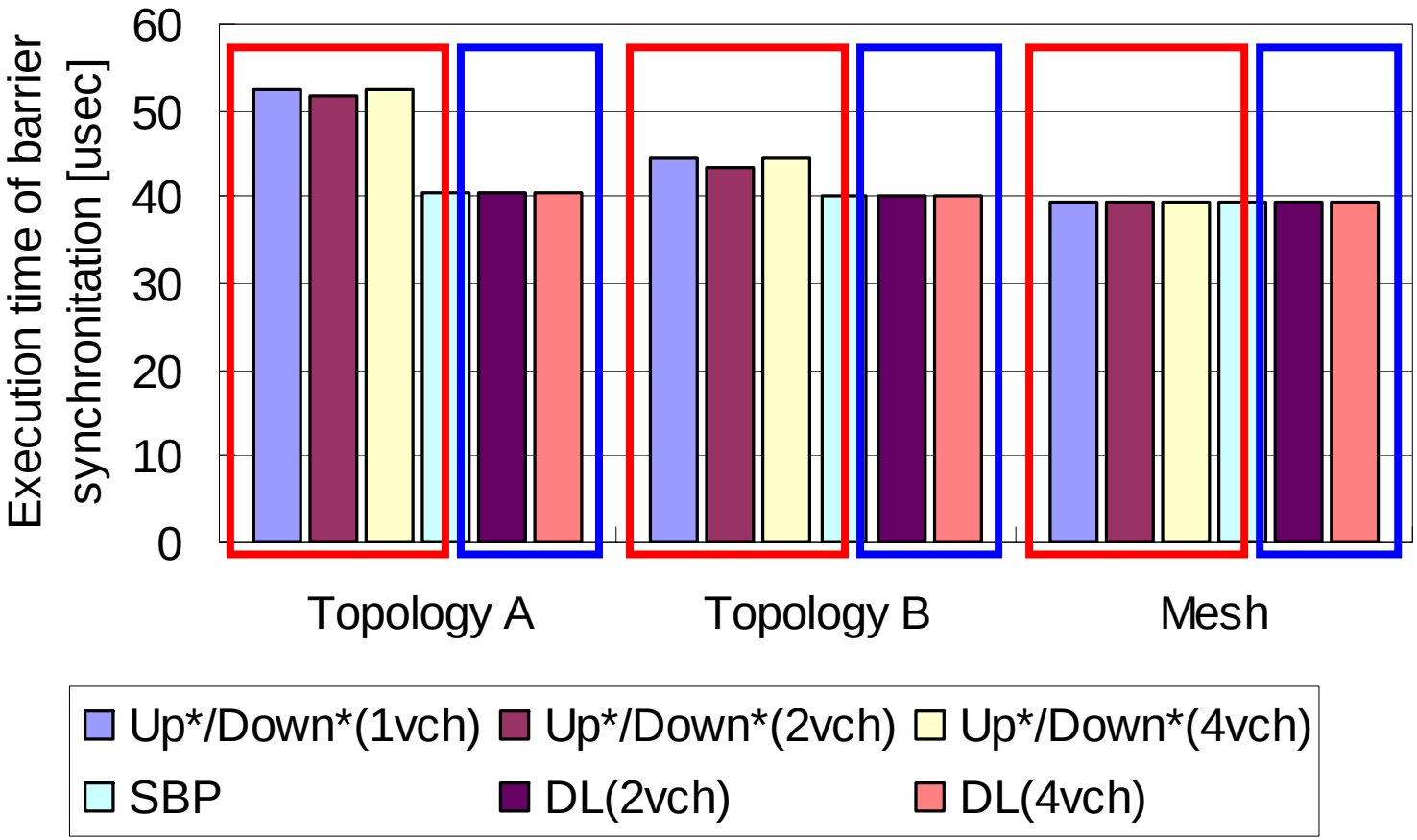


Mesh



- Up*/Down* routing with 1, 2, or 4 vchs
  - Non-minimal paths in T-A, T-B
- SBP with 5 or 6 vchs
- DL routing with 2 or 4 vchs
  - Minimal paths in all the topologies

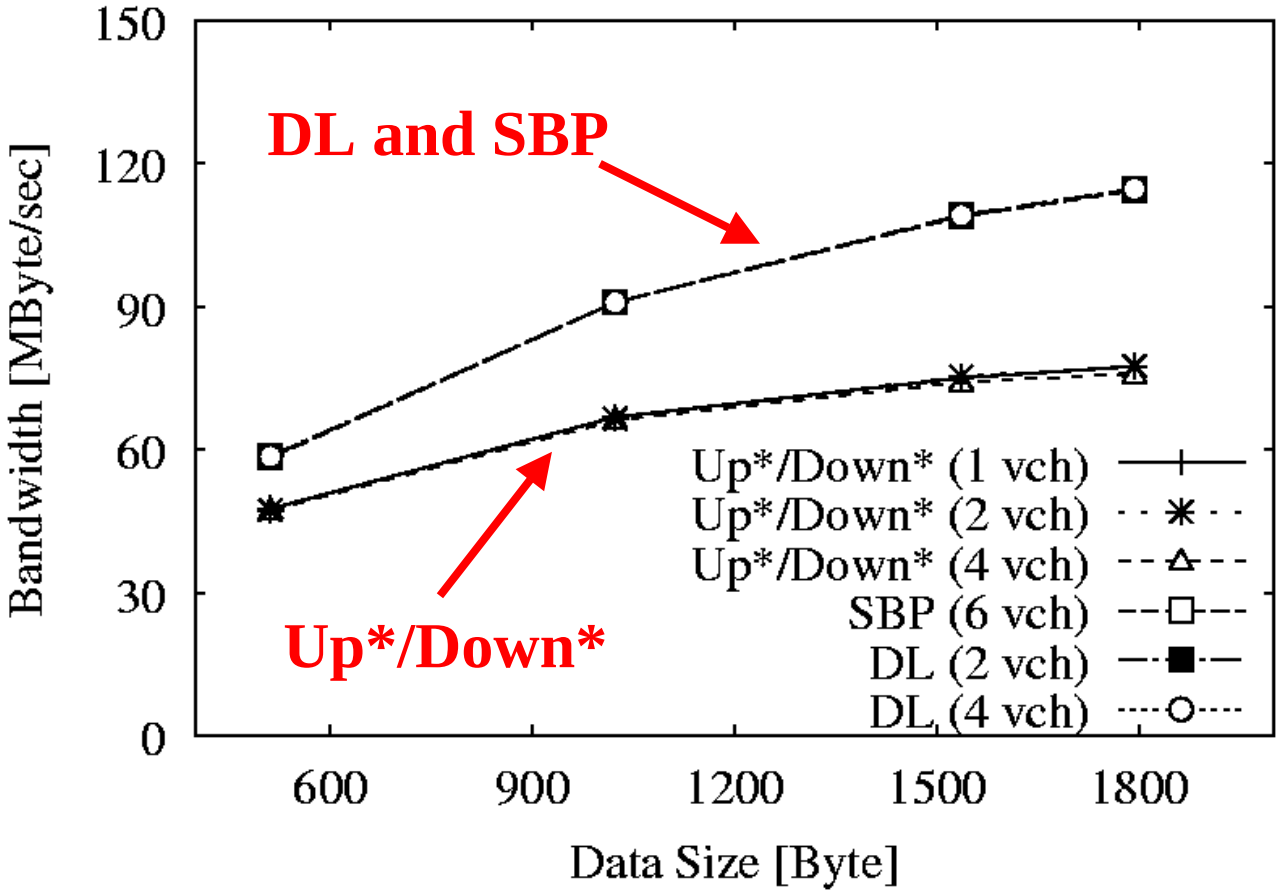
Barrier synchronization time



Only Packet hops are crucial.

(The number of vchs, and traffic balance are un-important.)

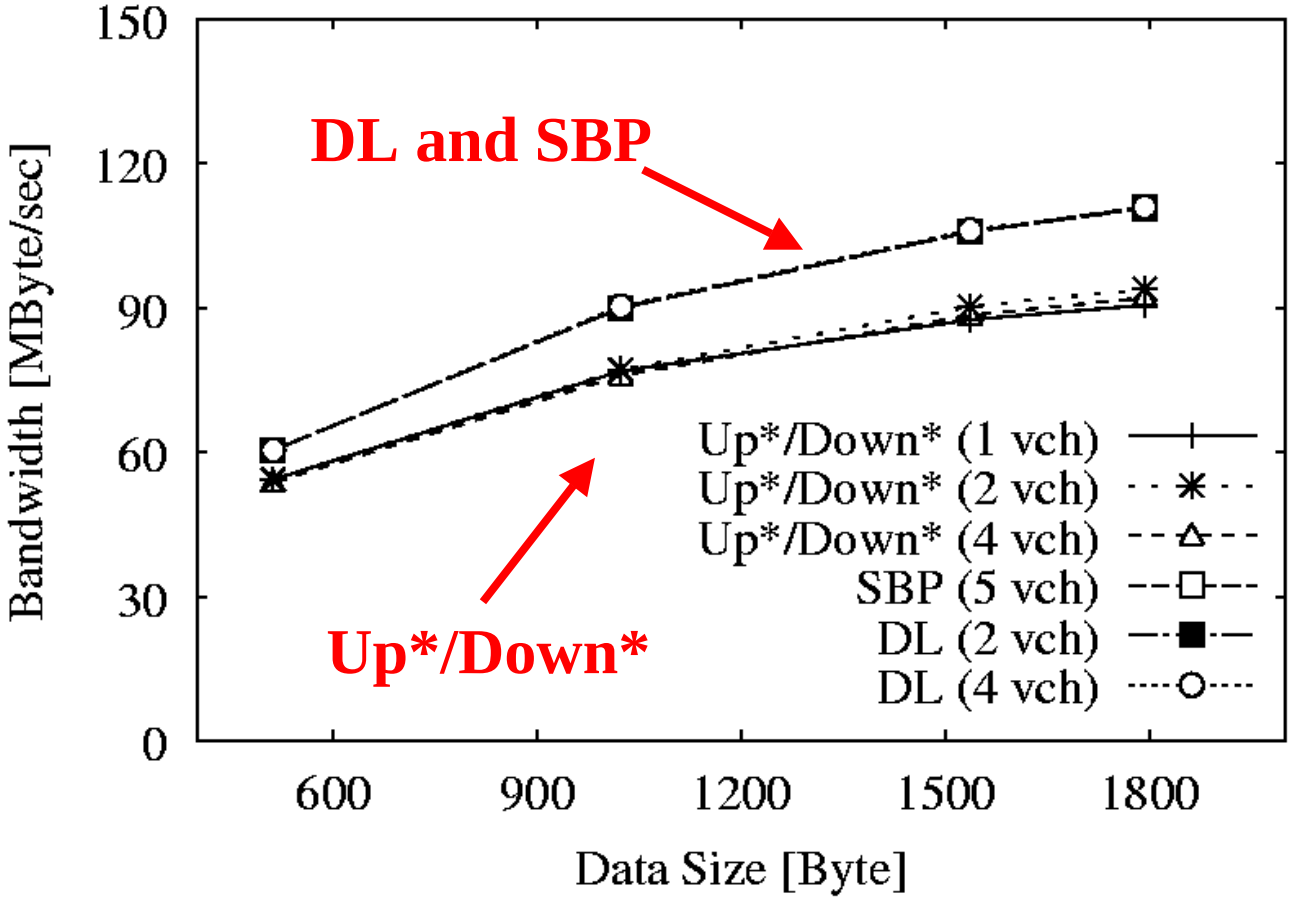
Bandwidth (Topology A, bit rev)



1) Shorter paths make larger bandwidth.

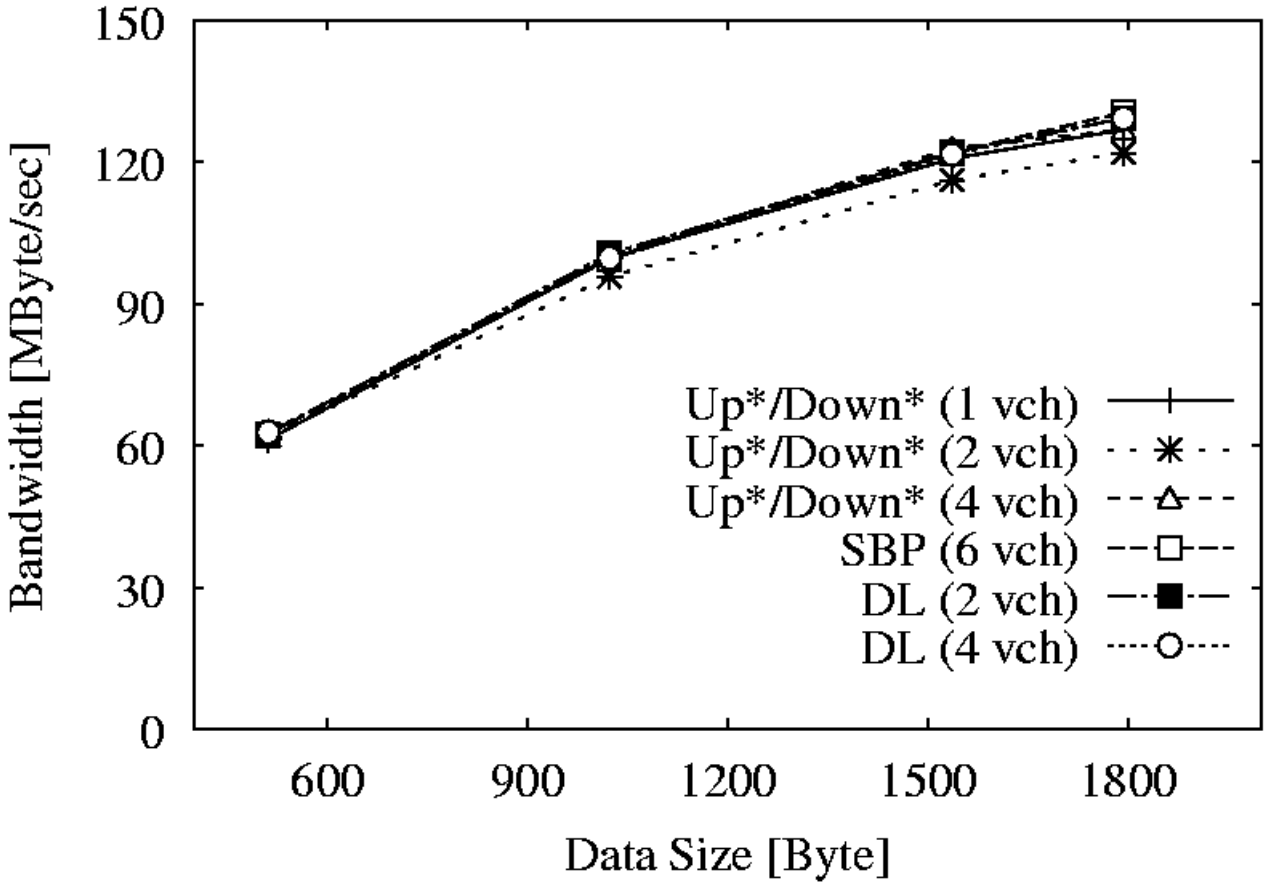
2) DL is better (minimal-paths, only 2-vchs).

Bandwidth (Topology B, bit rev)



Routing impact is smaller than that of Topology A.

Bandwidth (in Mesh, bit rev)



All the routings take minimal paths.

→ **They have the similar bandwidth.**

Bandwidth

- Packet hops are crucial.
- The number of vchs are un-important.
- The DL routing and SBP have almost the same bandwidth.

Conclusion

- We evaluated the performance of routing on the RHiNET-2 cluster.

- Packet hops are crucial to routing bandwidth, and latency.
- The DL routing and SBP have the almost same performance in each topology.
  - Up to 29% improvement (barrier synchronization)
  - Up to 51% improvement (bandwidth)

Future work on the RHiNET-2 cluster

- Routing impact under using smaller buffer
  - Unlike this eval, VCHs may work well.
- Uni-cast based Multicast algorithms
  - Host ID order V.S. random V.S. others
- Topology
  - Myrinet Clos, Fat-tree,torus, mesh

END