



RHiNET-3/SW: an 80-Gbit/s high-speed network switch for distributed parallel computing

S. Nishimura¹, T. Kudoh², H. Nishi², J. Yamamoto², R. Ueno³, K. Harasawa⁴,
S. Fukuda⁴, Y. Shikichi⁴, S. Akutsu⁴, K. Tasho⁵, and H. Amano³

¹RWCP Optical Interconnection Hitachi Laboratory,

²RWCP Tsukuba Research Center,


³Keio University,

⁴Hitachi Communication Systems, Inc.

⁵Synergetech, Inc.

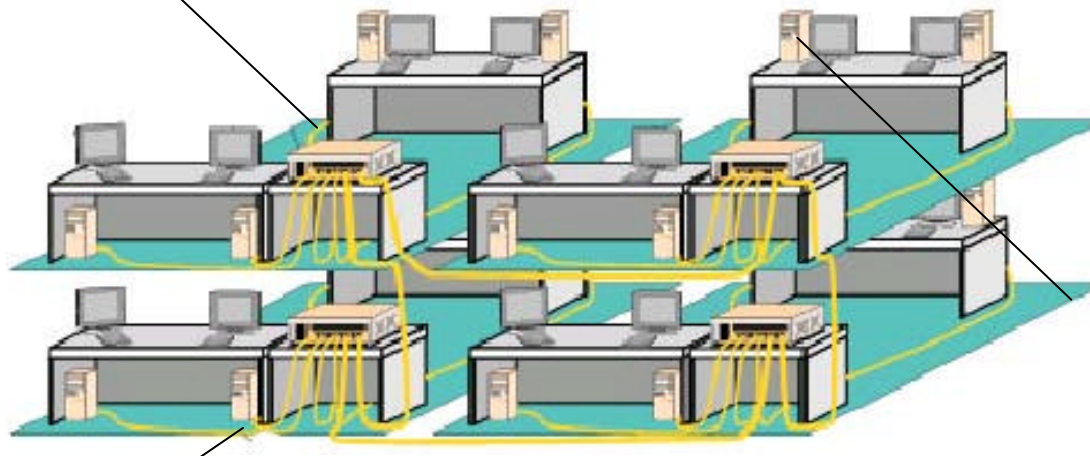
E-mail: nisimura@crl.hitachi.co.jp

Contents

- 
- RHiNET concept
(RWCP high-performance network)
 - Concept and architecture of RHiNET-3/SW
 - Key components in RHiNET-3/SW
 - switch-LSI, deskew-LSI, parallel optical link, board
 - Evaluation test results, on LSIs
 - bit-error-rate, deskew function

RHiNET concept

RHiNET switch: 8x8 high-speed crossbar switch



RHiNET-3/SW



PCI-bus based NIC

High-speed parallel optical link

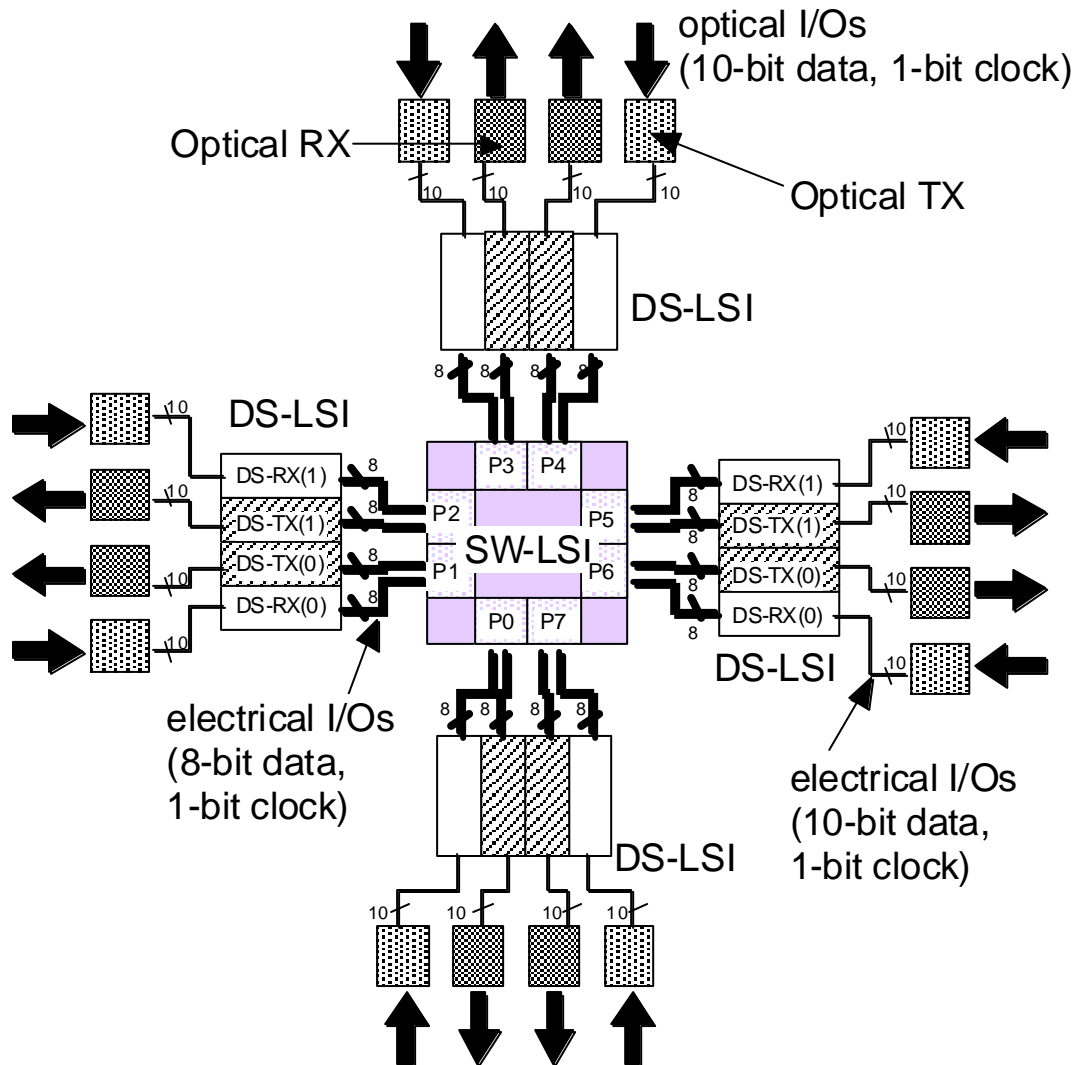
Targets:

- Low-cost, and high-performance parallel computing through the combined computational power of PCs
- Connecting computers distributed within one or more floors of a building

Features:

- Reliable low-latency communication, no upper layer
- Long links (- 1 km), free topology design
- Large bi-section bandwidth (- 10 Gbit/s)

Structure of RHINET-3/SW (schematic structure)



Switch: 10-Gbit/s x 8-port

- Aggregate throughput: 80 Gbit/s
- 8B10B encoded data with clock

I/O: 1.25-Gbit/s x 12-channel optical links

- Transmission length: < 1km

DS-LSI: skew compensation for long transmission length

Electrical I/O: CML or LVDS

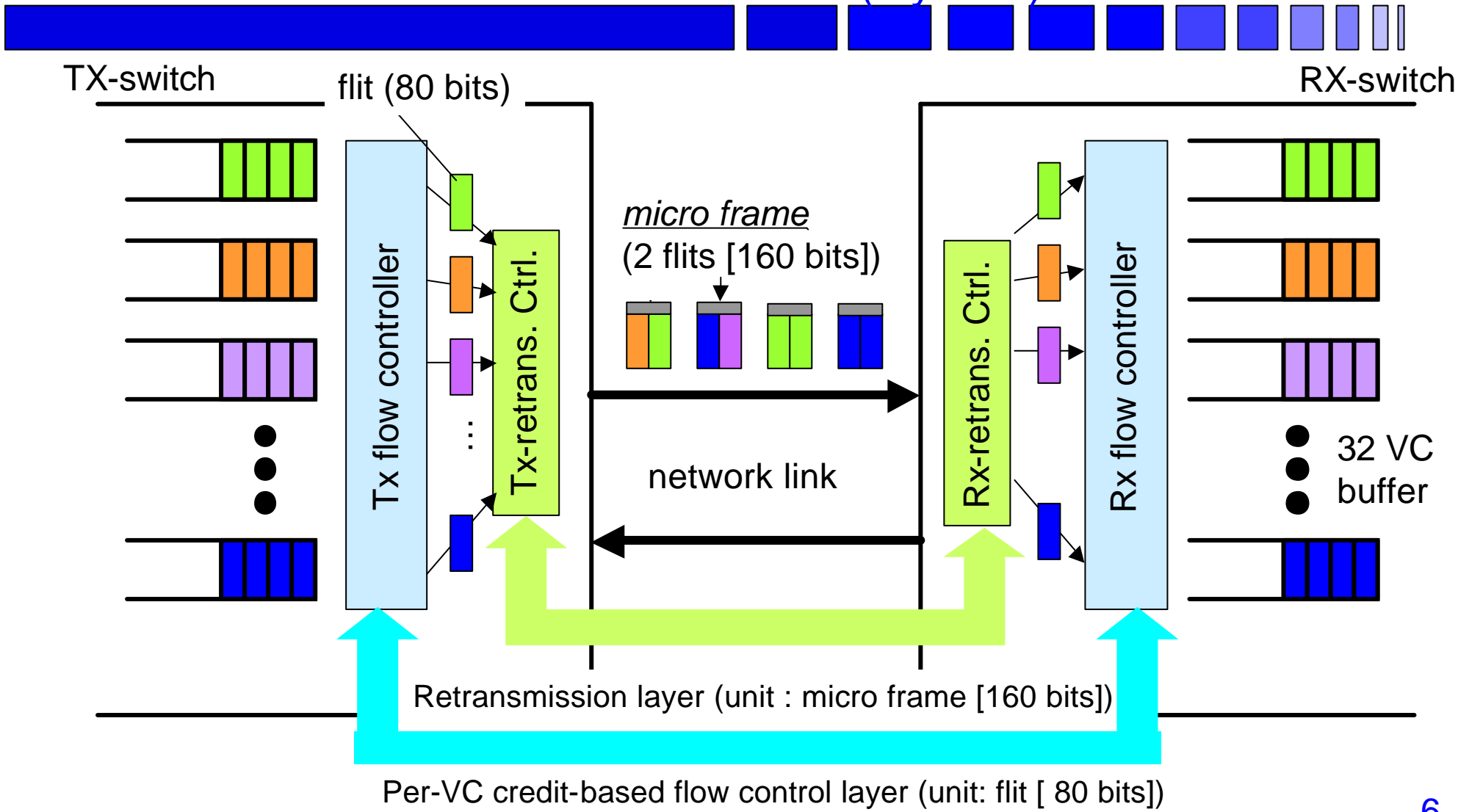
Design concepts of RHiNET-3

- Hop-by-hop retransmission
 - Low-cost optical link module
 - Retransmission: need for error-free data transmission
 - Simple procedures and compact circuits
 - Retransmission unit: *micro frame* (160 bits)

- Credit-based flow control
 - For long transmission length
 - Effective use of packet buffer

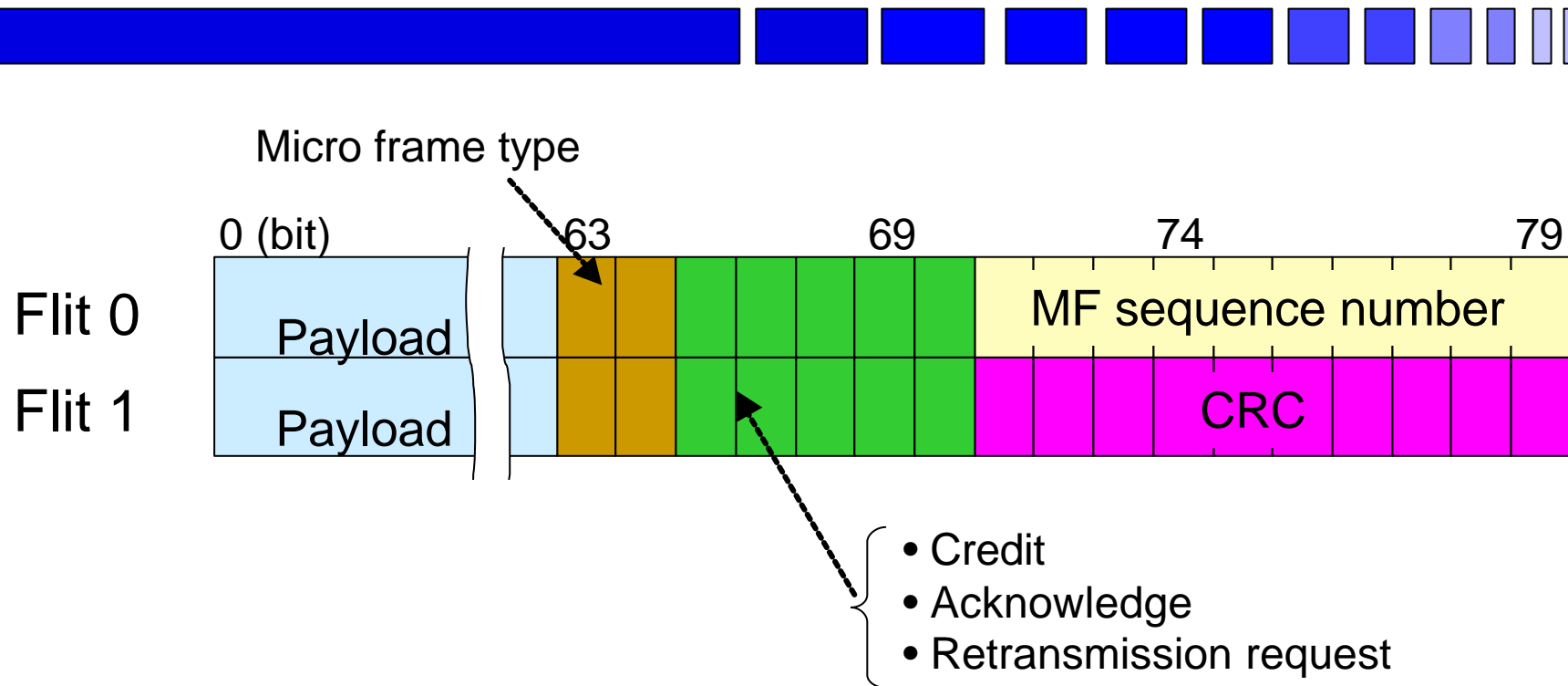
- 32 Virtual channels (VCs) - Virtual lane -
 - Deadlock-free and topology-free

Flow control and retransmission (layered)



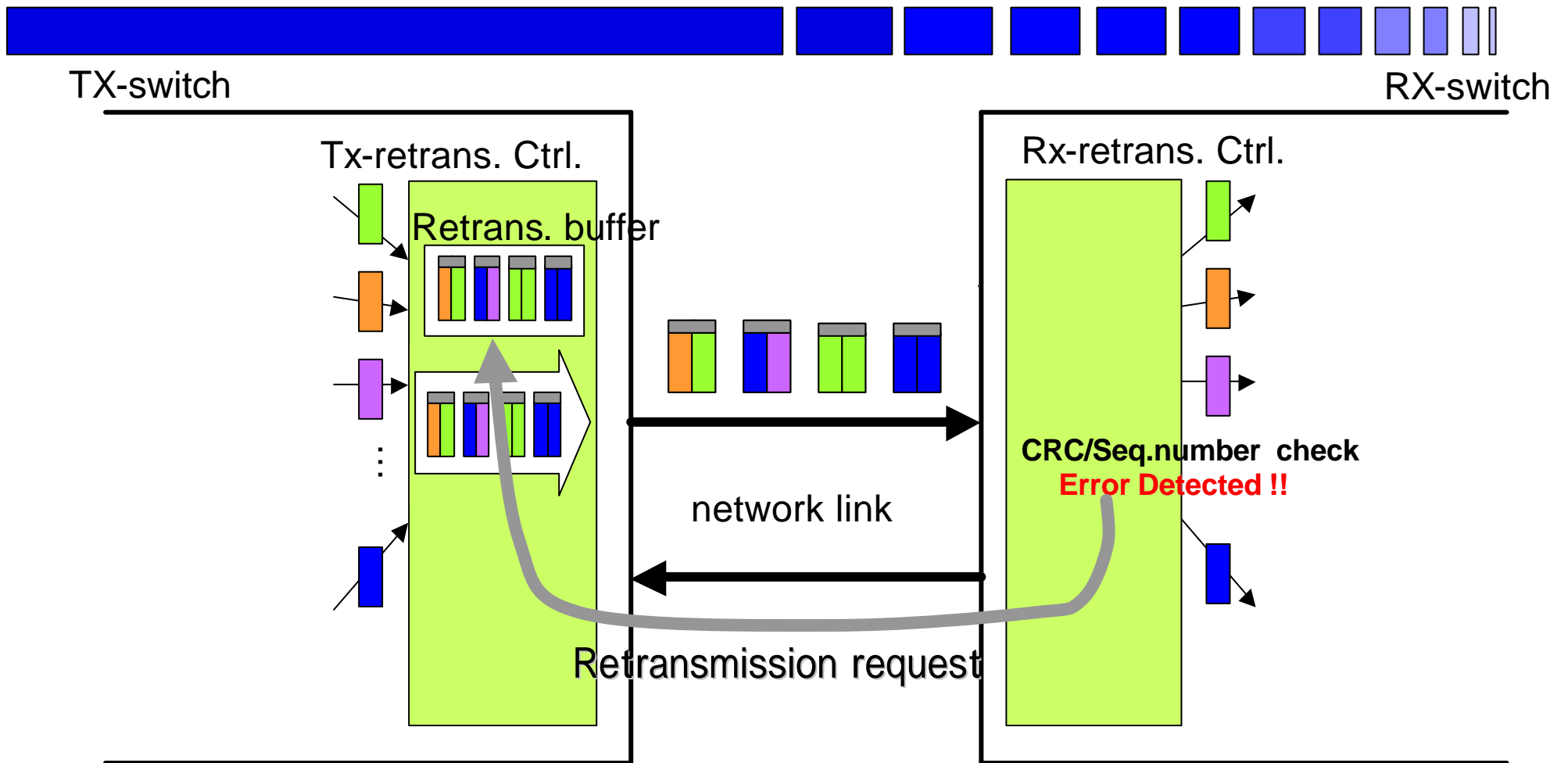
■ Small data size: reduce overhead (latency and bandwidth)

Format of micro frame (MF)



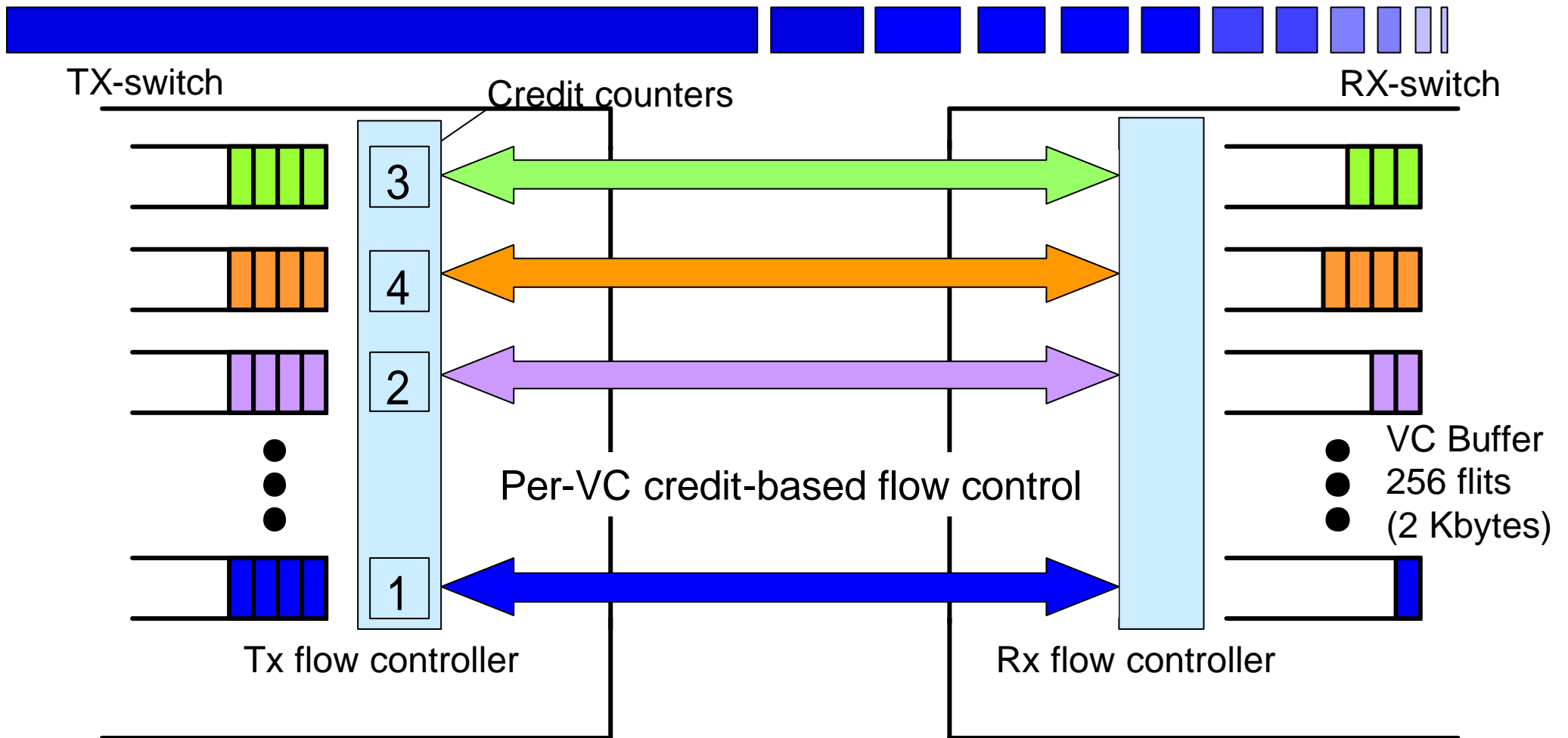
- CRC and sequence-number based retransmission mechanism
Retransmission unit: micro frame (128 bits payload / 160 bits)
- Acknowledge: sequence number of successfully received MF
- Credit, acknowledge and retransmission request use the same field
⇒ Small retransmission overhead

Retransmission mechanism (behavior)



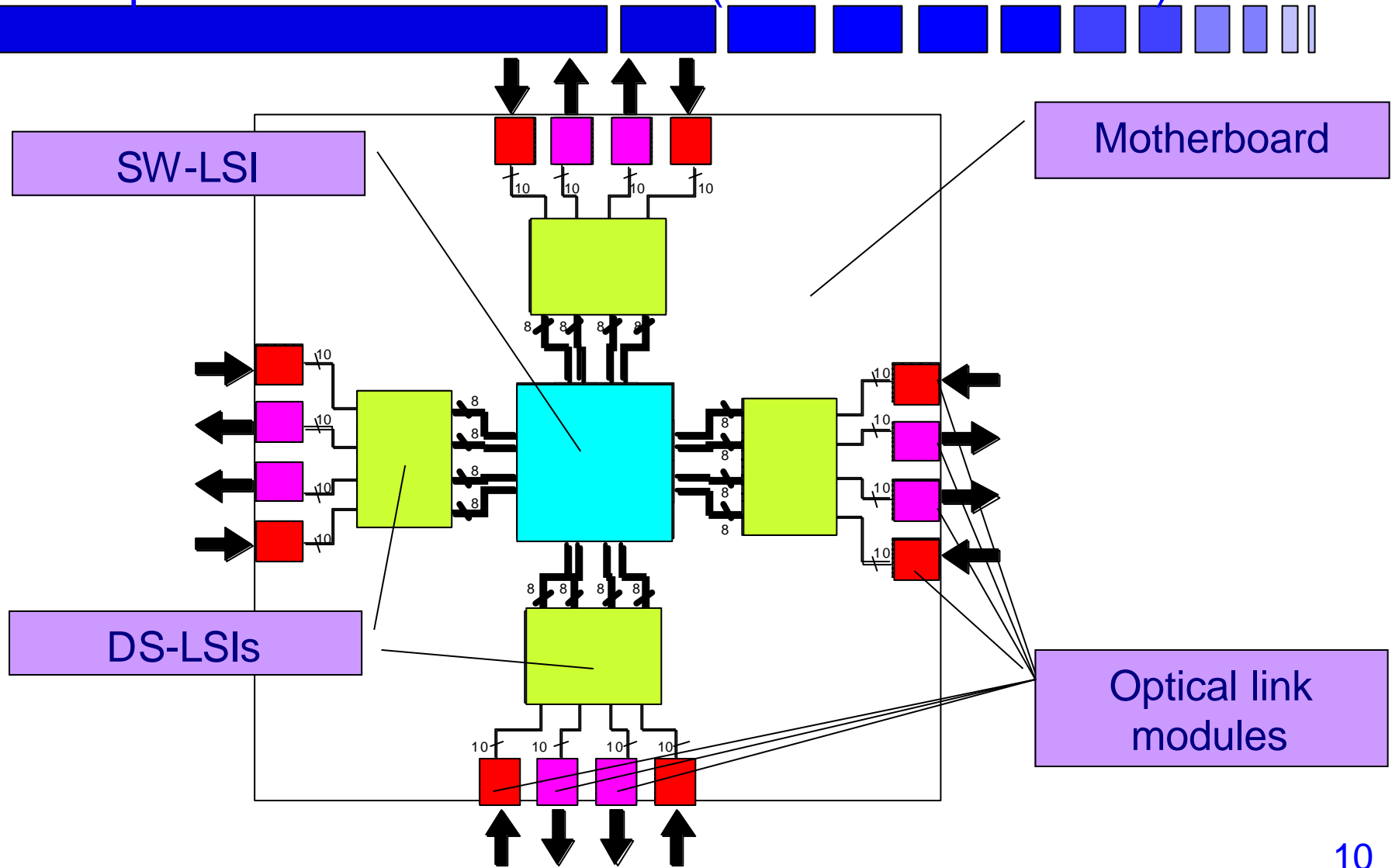
Hop-by-hop retransmission : error-free transmission, and small overhead

Credit-based flow control

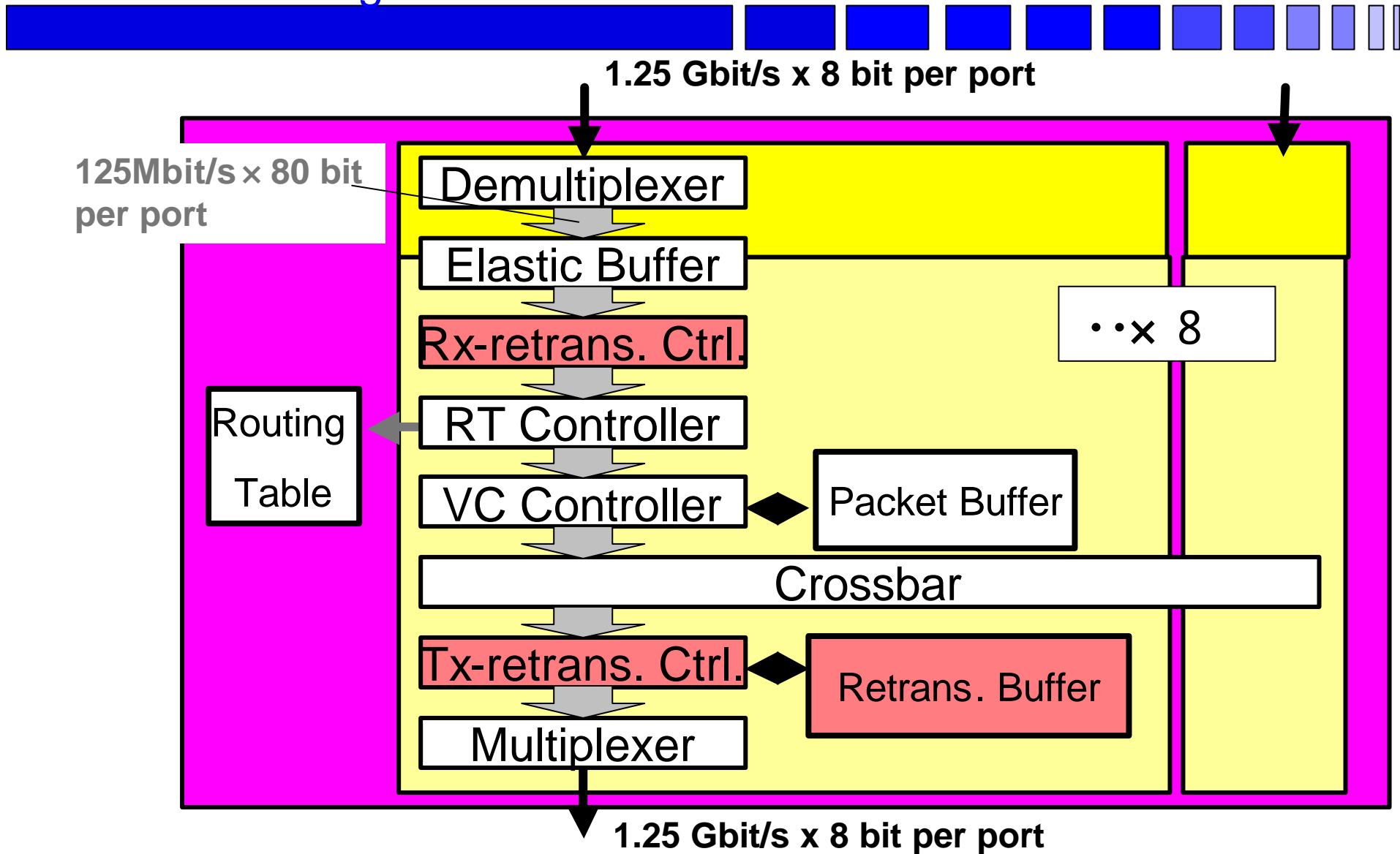


Credit-based flow control mechanism enables long data transmission and uses VC buffer effectively

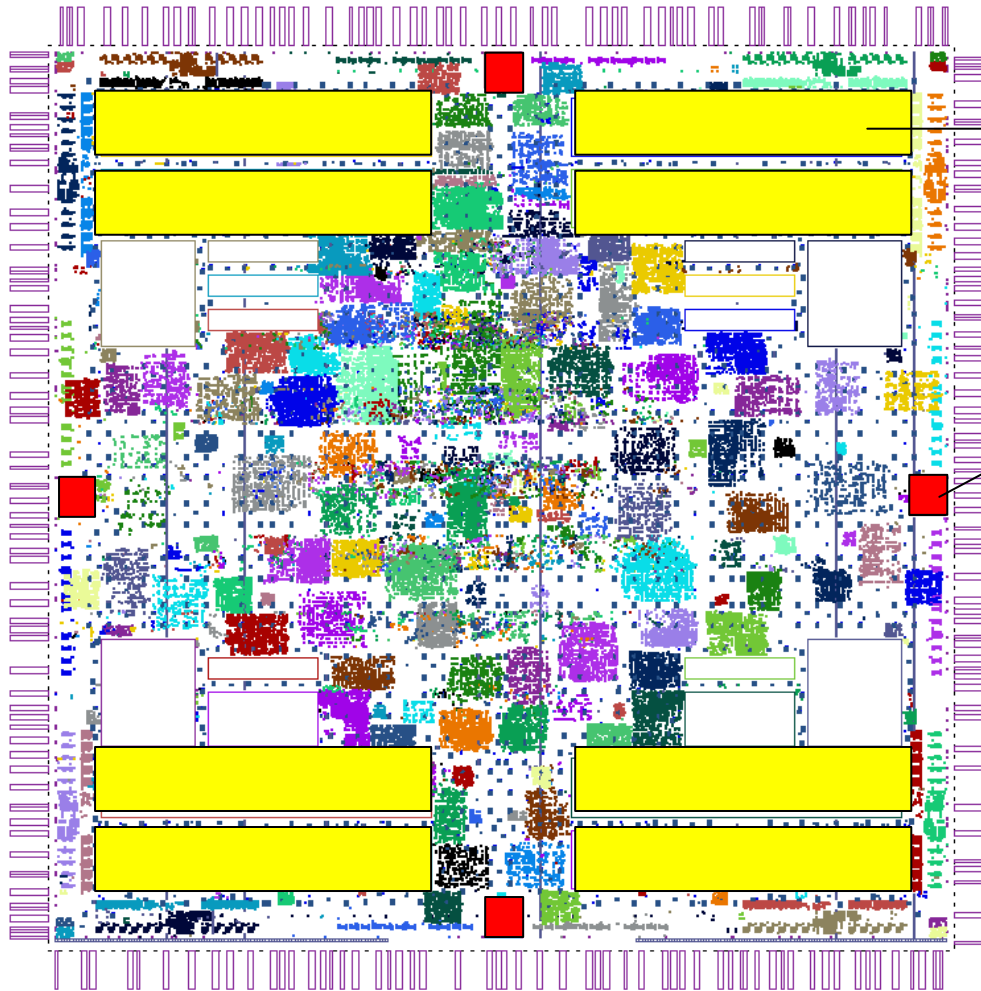
Components of RHINET-3/SW (schematic structure)



Blockdiagram of SW-LSI



Floor plan of SW-LSI (1st cut)

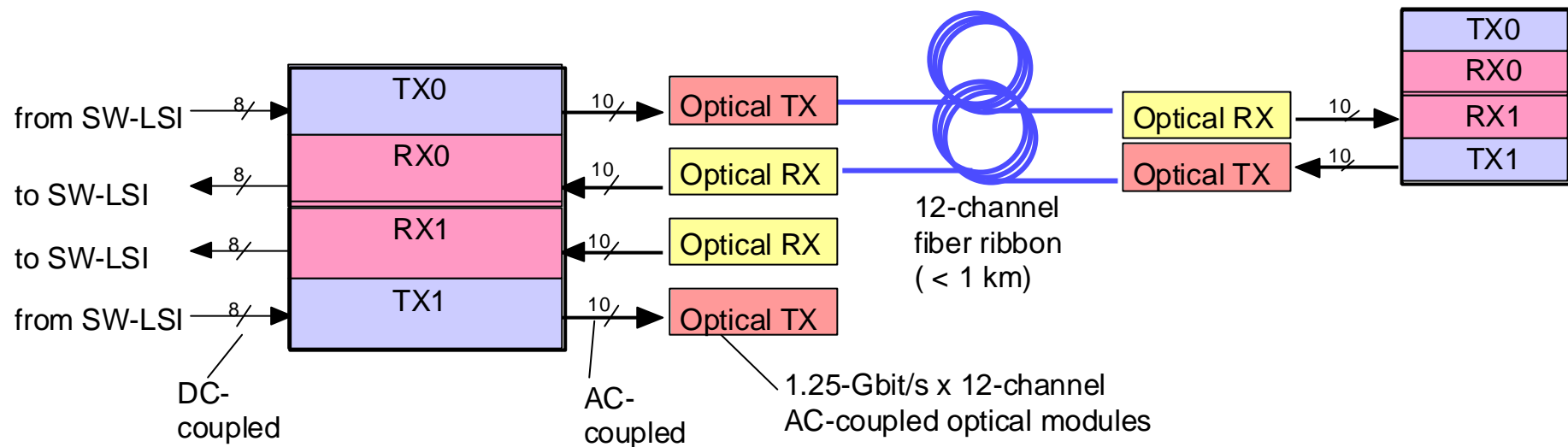


VC buffer memory

PLL

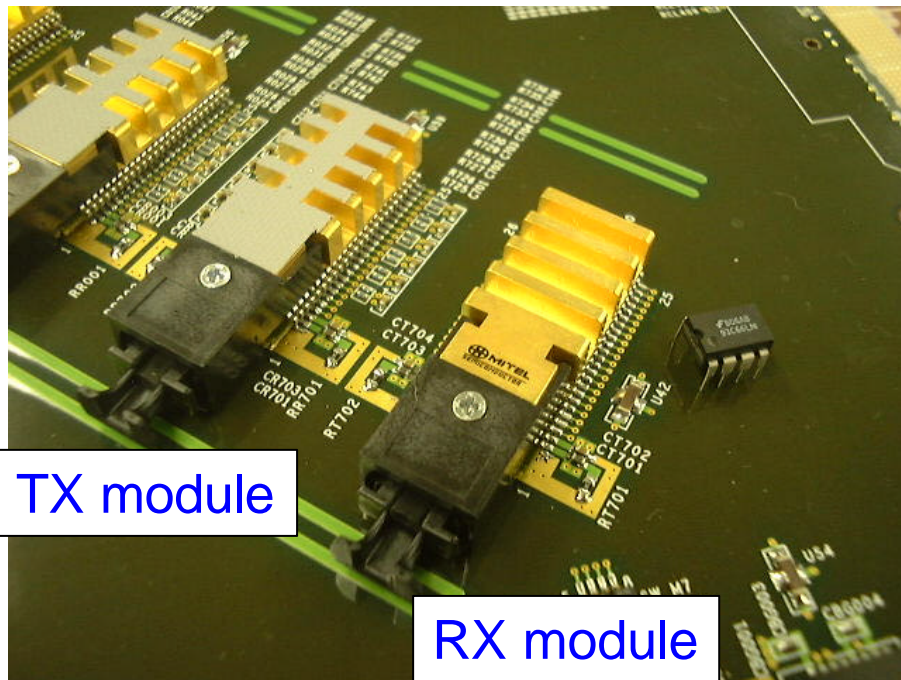
0.14-um CMOS ASIC
Die size: 16.5 mm x 16.5 mm
Number of gates: 1502 k
Buffer memory:
 a total of 640 kbytes
I/O: 1.25 Gbit/s per pin
Package: 784-pin BGA

DS-LSI (LSI for skew compensation)



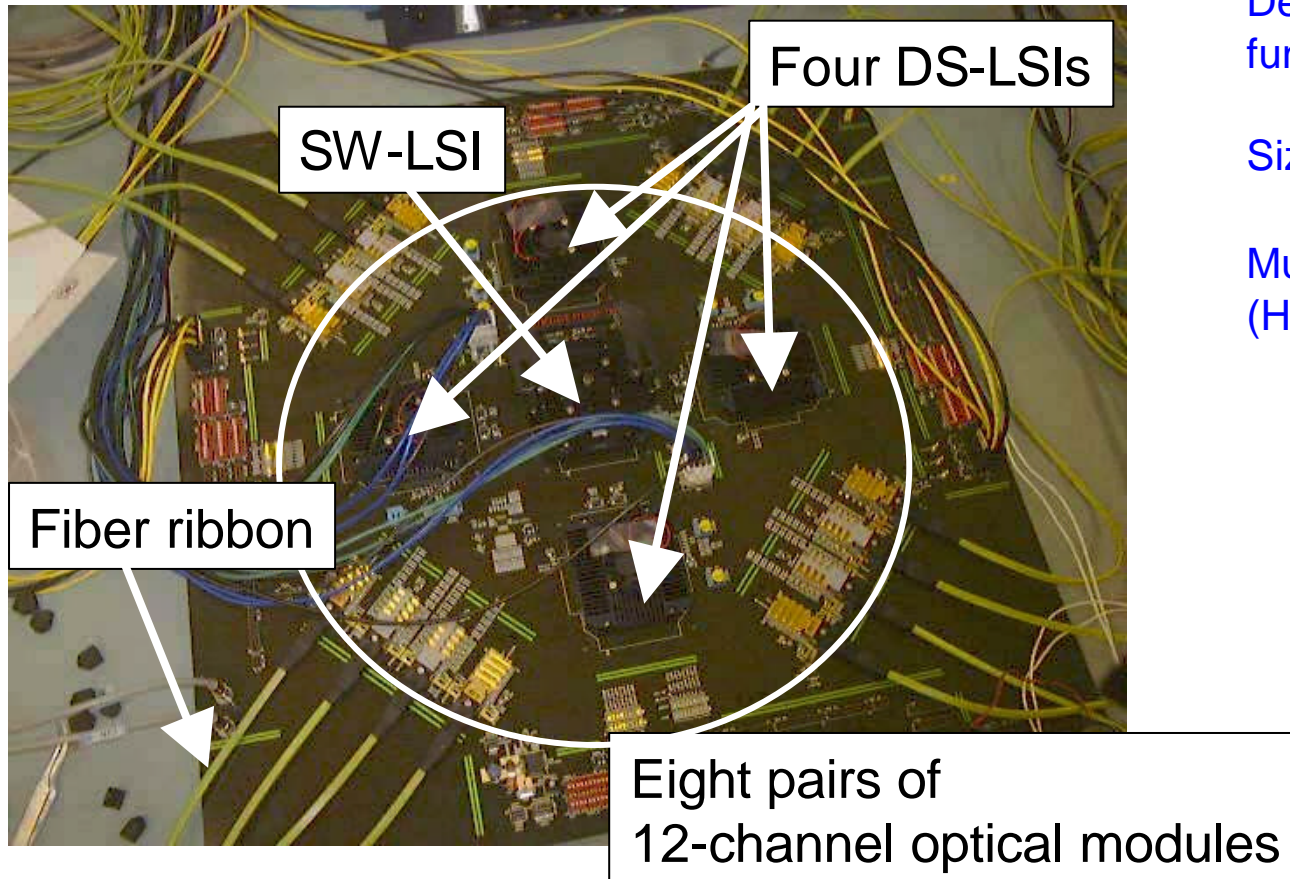
- DS-LSI has 8B10B encoder and decoder
 - For high-speed (1.25 Gbit/s per pin) AC-coupled optical data transmission
- DS-LSI compensates skew between 10-bit data and 1-bit clock
 - Maximum skew: +/- 256 ns
 - larger than a skew of 1-km MMF fiber ribbon (+/- 64 ns)
 - Initial data pattern consists of 64 8B10B special characters

12-channel parallel optical link



- 12-channel parallel data transmission (products of ZARLINK™ semiconductor)
- 850-nm VCSEL
- 12-channel CML interfaces
- 155 Mbit/s - 2.5 Gbit/s (AC-coupled)
- GI 50/125 12-channel MMF fiber
- Up to 300-m data transmission at 2.5-Gbit/s
- BER: 10^{-12}

Structure of motherboard (1st test-bed)



Designed to evaluate switching function

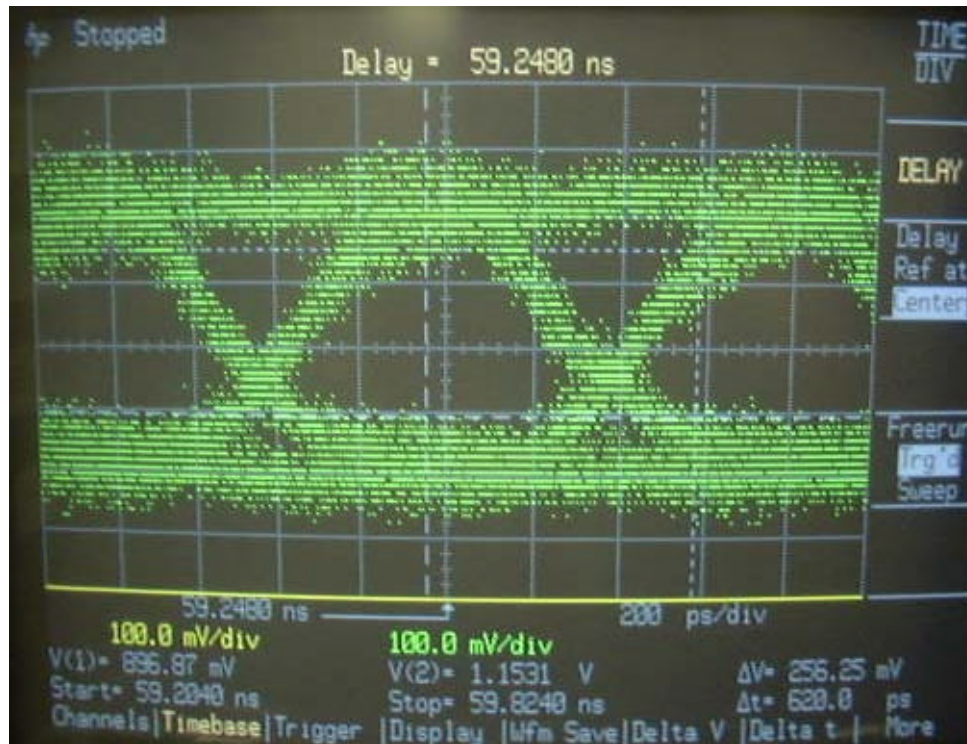
Size: 550 x 550 mm

Multi-wire interconnection board TM
(Hitachi Chemical, Ltd.)

To overcome crosstalk, skew,
and propagation loss

Layout is optimized according
to experimental results

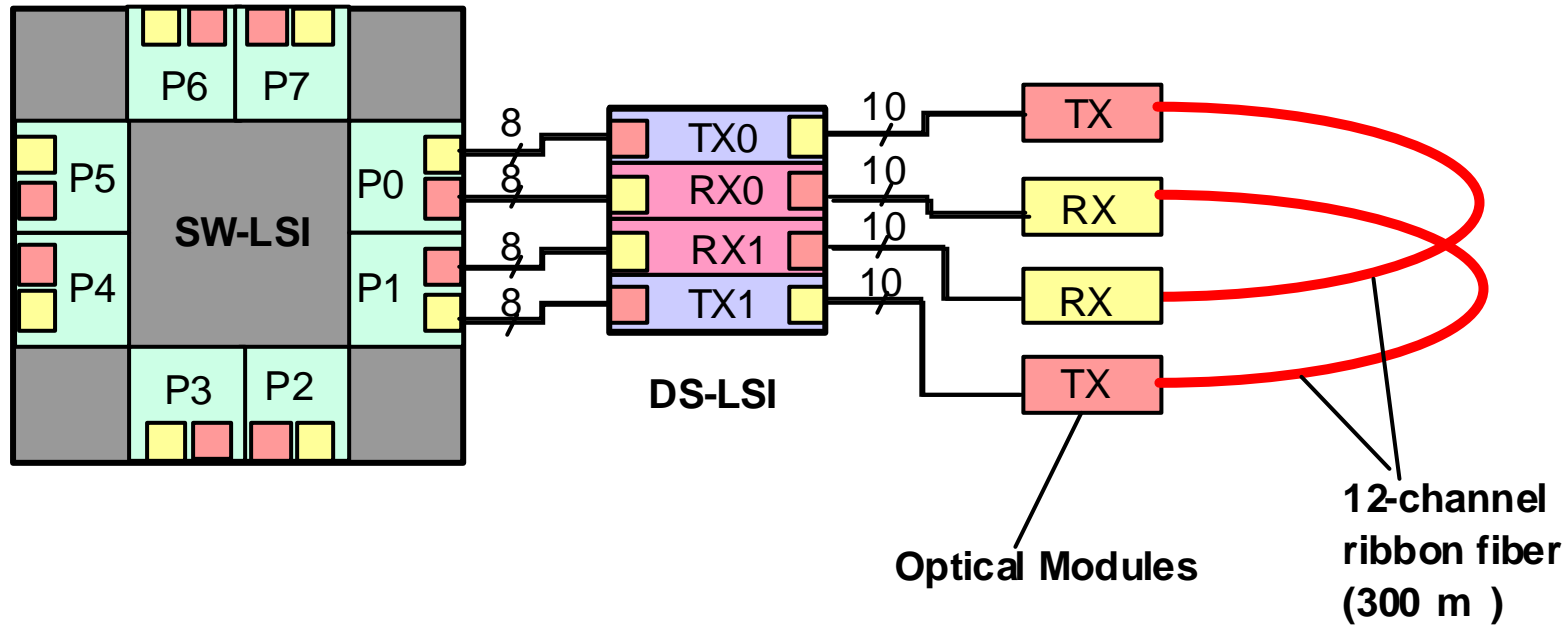
Evaluation results (bit error rate)



SW-LSI output from channel D0 of port0

- BER (bit error rate): $< 10^{-11}$ at data rate of 1.25 Gbit/s per pin
- Timing budget margin: about 400 ps

Evaluation results (deskew function)



Port 0 ↔ Port 1

- Deskew Function works successfully.

Summary

- A prototype network switch, RHiNET-3/SW, for a RHiNET high-performance distributed parallel computing environment
- Specifications
 - 10 Gbit/s x 8 ports
 - Parallel optical data transmission over a distance of up to 1 km
 - Aggregate throughput is 80 Gbit/s per board
- Architecture
 - Hop-by-hop retransmission mechanism
 - Credit-based flow control
 - reliable and long-transmission-distance data communication
 - For 8-nodes parallel computing
- RHiNET-3/SW
 - High-throughput, long-distance and flexible-flow-control
 - In a distributed parallel computer system using commercial PCs