A Single Chip Shared Memory Switch with Twelve 10Gb Ethernet Ports

Takeshi Shimizu, Yukihiro Nakagawa, Sridhar Pathi, Yasushi Umezawa, Takashi Miyoshi, Yoichi Koyanagi, Takeshi Horie, Akira Hattori

Hot Chips 15 - August 19, 2003

Fujitsu Laboratories of America, Inc.

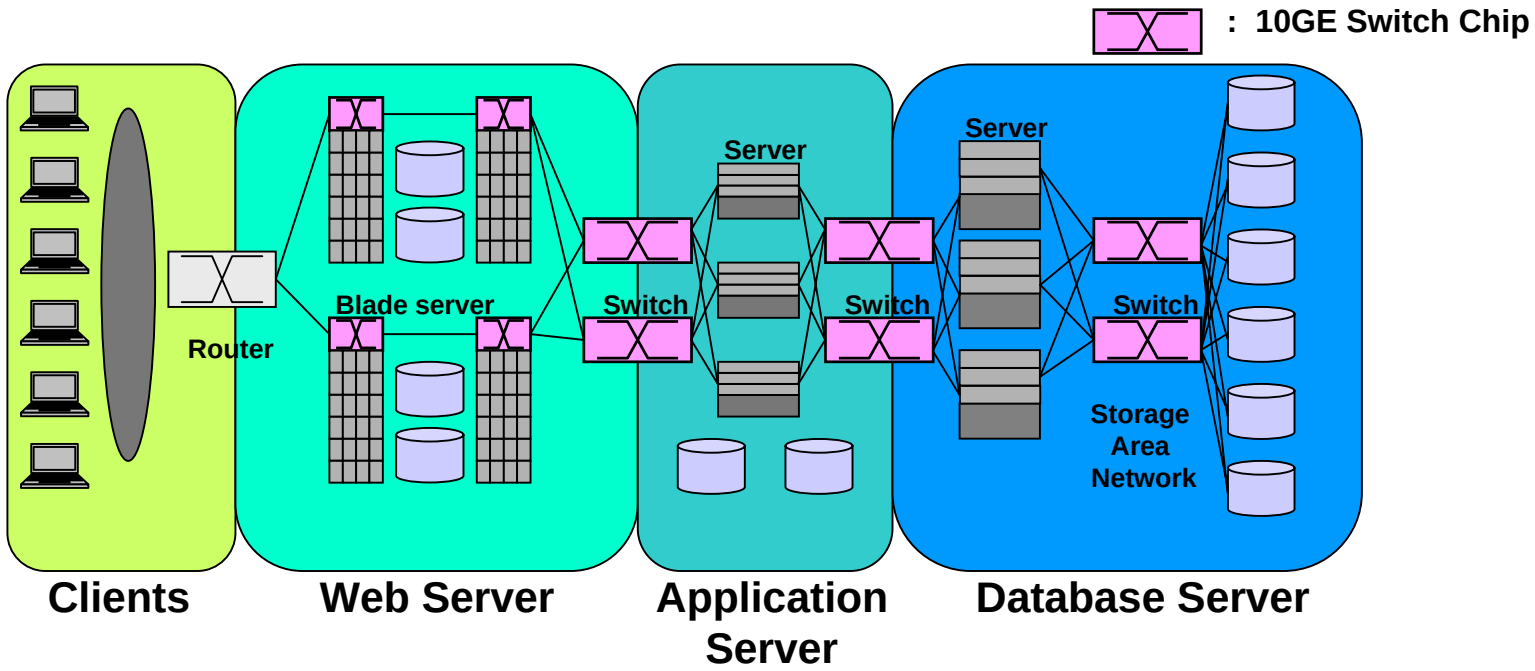
Advanced Interconnect Technology Department

Outline

- Background
- Overview and Features
- Switch Implementation
- Evaluation
- Summary

Background

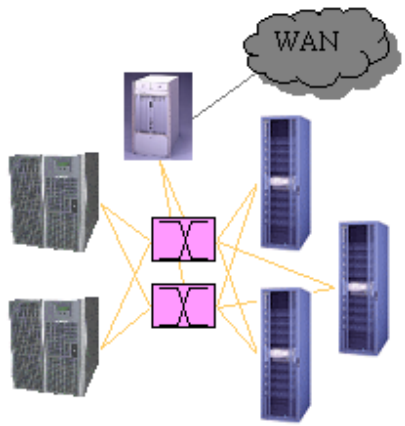
- IP-based networks connect all the computing resources (All-IP).
- Ethernet protocol is commonly used for IP networks.
- 10Gb Ethernet is a promising solution for unified, fat pipe between servers and storage systems.



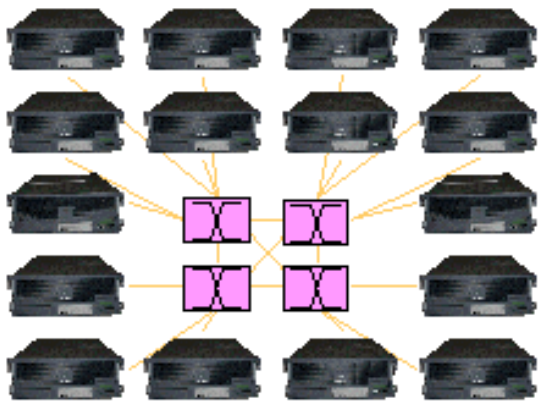
Primary Target Applications

- Our primary target is infrastructure for computing platform, such as SAN, Clusters, Blade Servers.
- Those applications require short latency, low cost and high density.
  - Motivation to develop a dense 10Gb Ethernet switch.
- So, the design strategy is set as follows.
  - Focus on layer-2 switching.
  - High-throughput/low latency switch core.
  - SerDes integration for copper solution.

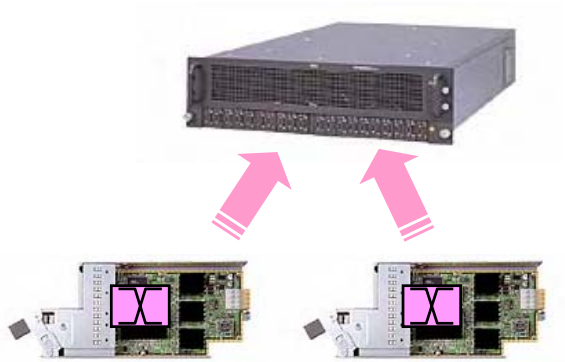
Storage Area Network



Cluster Computing

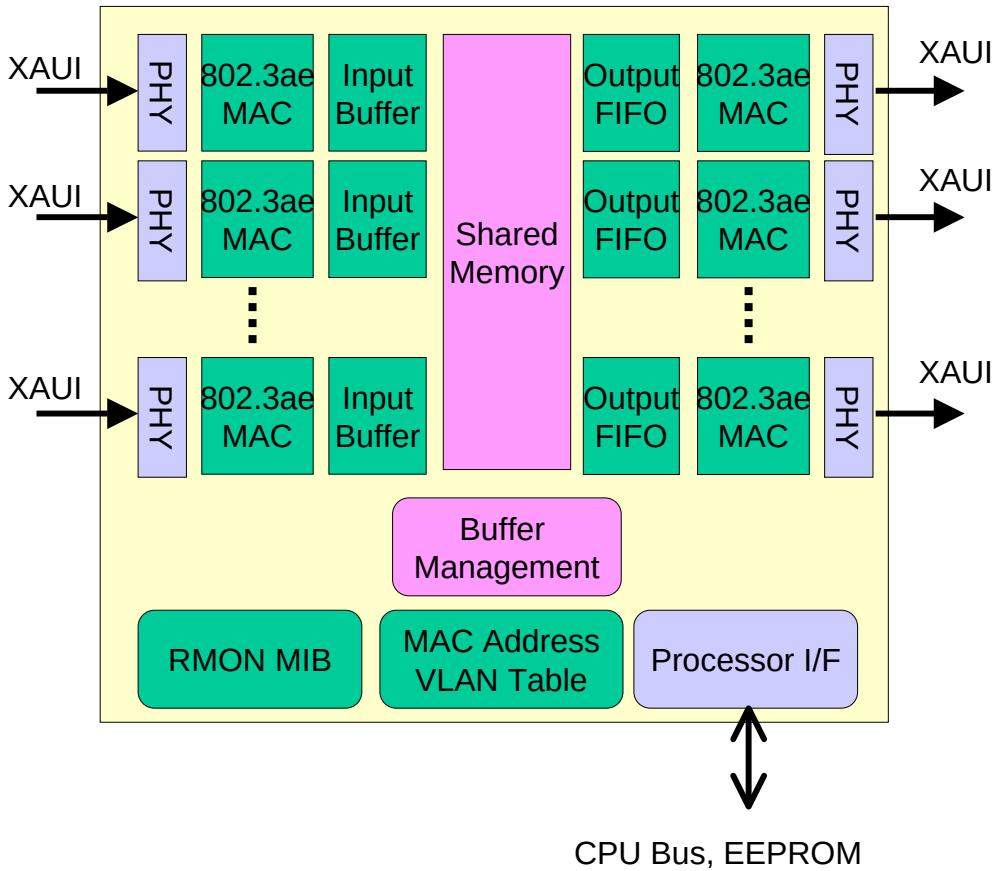


Blade Server



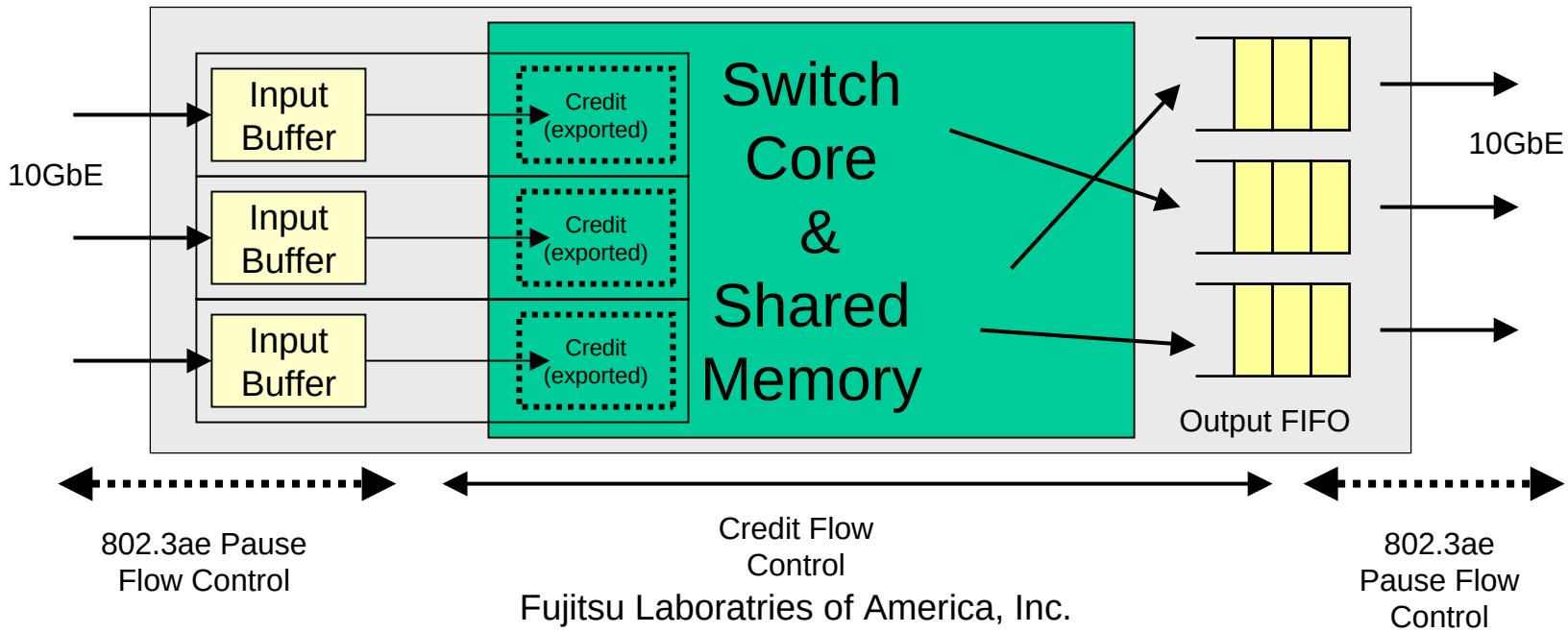
MB87Q3050 Overview

- Twelve 10 Gb Ethernet ports.
- Integrated SerDes for XAUI.
  - 3.125Gbps x 4 lane/direction
- Layer-2 switching with 802.1Q VLAN.
- Output queue switching using shared memory.
- 240 Gbps shared memory bandwidth.
- Cut-through forwarding.
- Statistics counters for RMON.



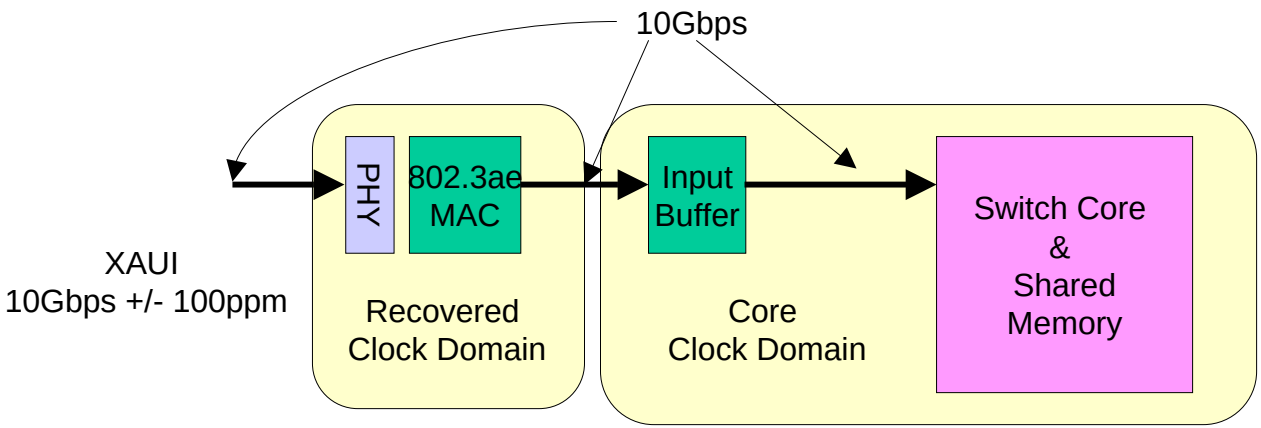
Designing a Switch Core

- A protocol independent switch core was developed.
  - Credit based flow control.
  - Packet by packet operation, allowing variable length from 64B to 9KB.
    - This is not a fixed-length cell switch.
  - Four level priority queue, with simple distributed arbiters per output port.
  - Cut-through forwarding (fall-through latency of the core: 150ns)
  - Multicast support for “single copy, multiple read”.



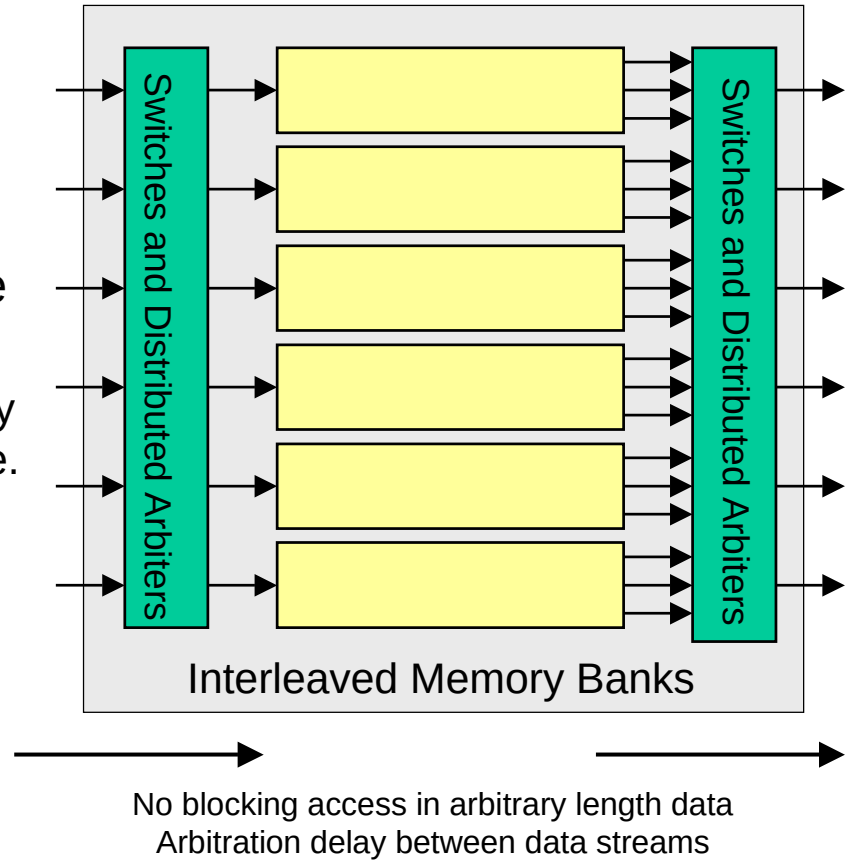
Input Buffering for Cut-Through

- The switch core is designed for cut-through forwarding of variable length frames.
  - The switch core has NO speed-up in terms of the bandwidth per port.
    - All the data path in the switch is running at 10Gbps data rate.
    - Popular implementation with crossbar switches may have increased B/W per port.
  - The switch core assumes NO fragmentation.
    - The entire packet should be transferred at 10Gbps rate.
  - These features are desired to handle variable length frames in Ethernet protocol, to minimize store&forward operation.
- Input buffering is simply for speed-matching between clock domains when cut-through mode is chosen.



On-chip Interconnect for Shared Memory

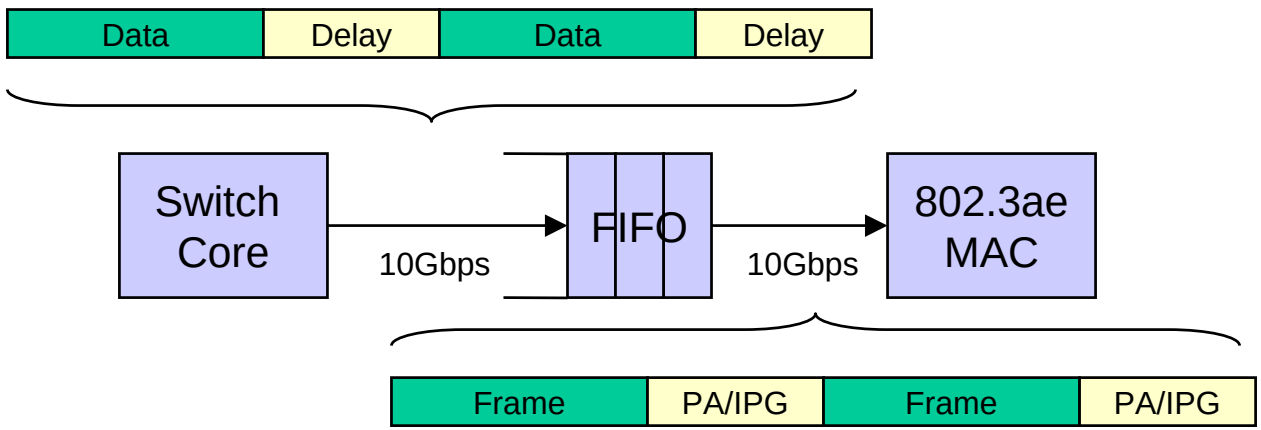
- An on-chip memory sub-system is designed for 12 10Gbps reads AND 12 10Gbps writes.
  - "Multi-port Stream Memory"
- Deep interleaved memory banks are connected via distributed, multi-stage interconnection network.
  - "No global arbitration" is a good policy for chip integration and timing closure.
- Switching is rescheduled at every packet boundary for access path arbitration.
- Arbitration delay exists between packets (as shown later).



Output Buffering for Short Latency

- Frames come out of the switch core without fragmentation, at 10Gbps rate.
- However, arbitration delay (caused by the on-chip interconnect) exists between frames.
- The average value is 3.2 cycle (assuming random, full-loaded traffics).

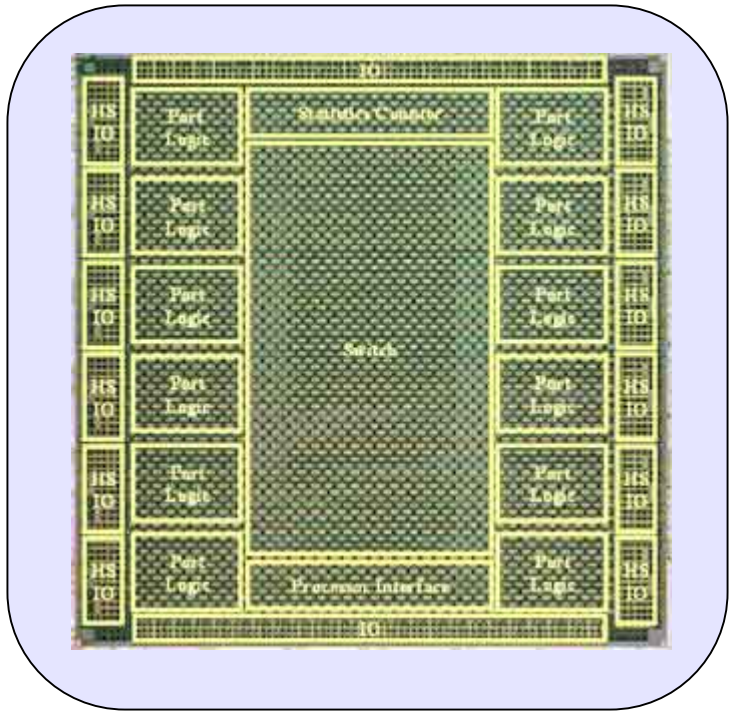
- On the other hand, outgoing packets has 8 byte preamble and 12 byte IPG as defined in the Ethernet protocol.
  - 5 cycles between frames in the switch core.
- Thus, with a small amount of FIFO, it is possible to sustain wire-speed throughput.
  - We do not need a large FIFO.The size is much smaller than the maximum packet size.
  - It also helps to reduce the fall-through latency in a loaded condition.



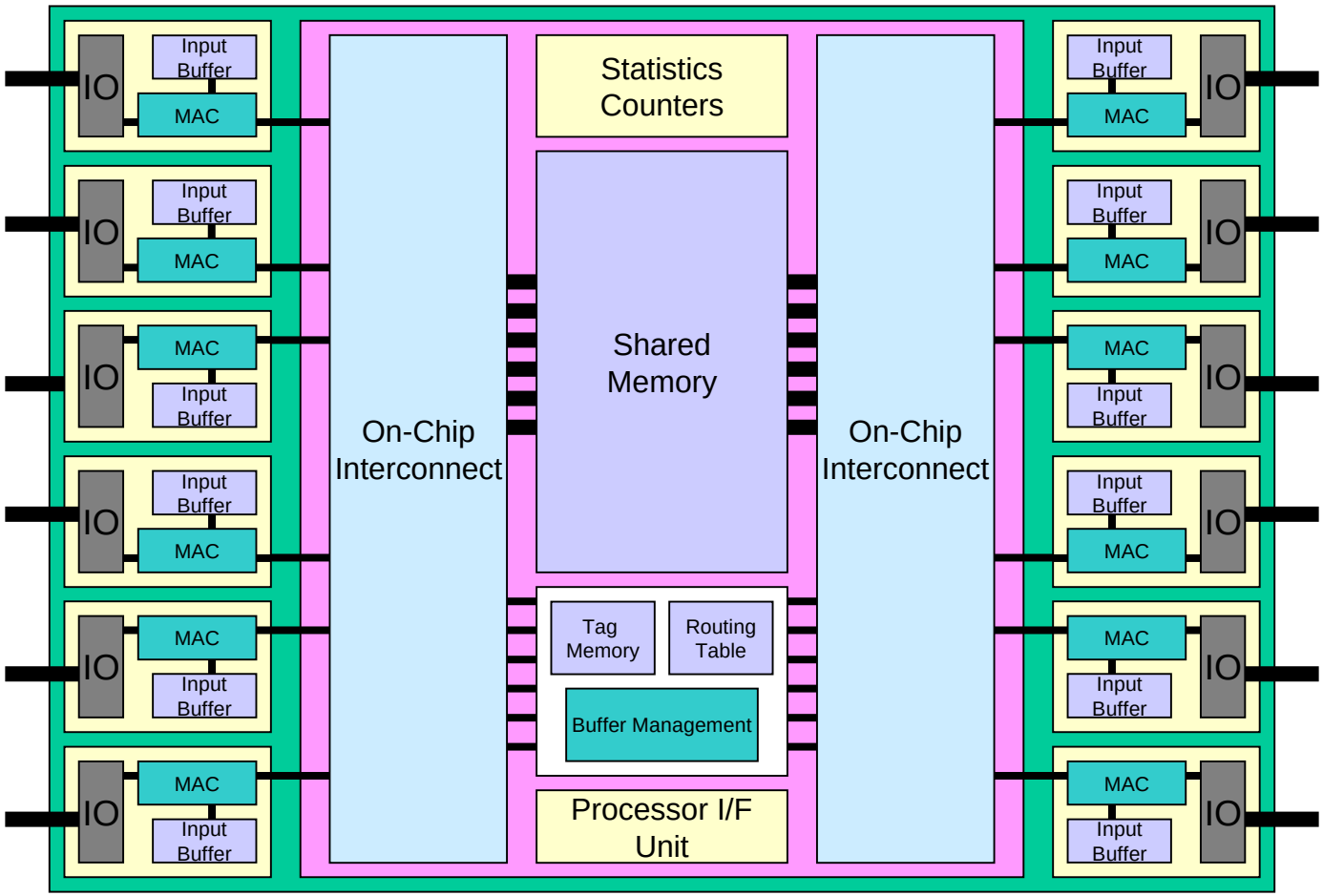
Chip Implementation

Items	Specification
Technology	Fujitsu CS91: 0.11um CMOS ASIC
Logic	6.3M gates (total)
SRAM	897K Bytes (total)
Core Logic Frequency	312.5MHz
Package	FCBGA-728
Signals	336
High Speed IO	XAUI (3.125Gb/s x 4) X 12
Power Consumption	15.3 W (typical), full-loaded
Die Size	16mm x 16mm

MB87Q3050
12-port, 10Gbps Ethernet Switch Chip



Floorplan Image

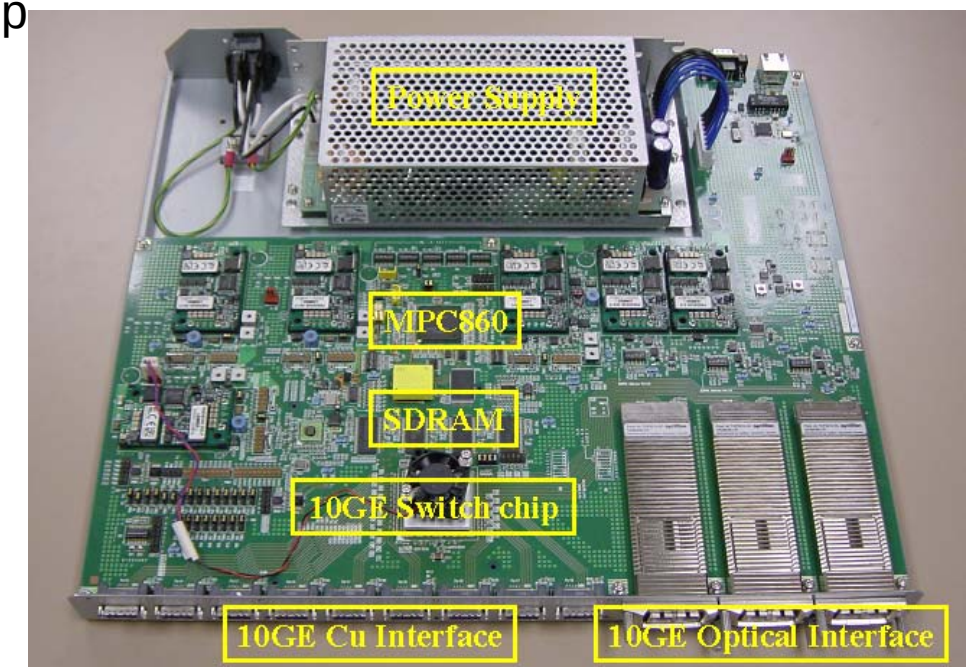


Fujitsu Laboratries of America, Inc.

Evaluation Board

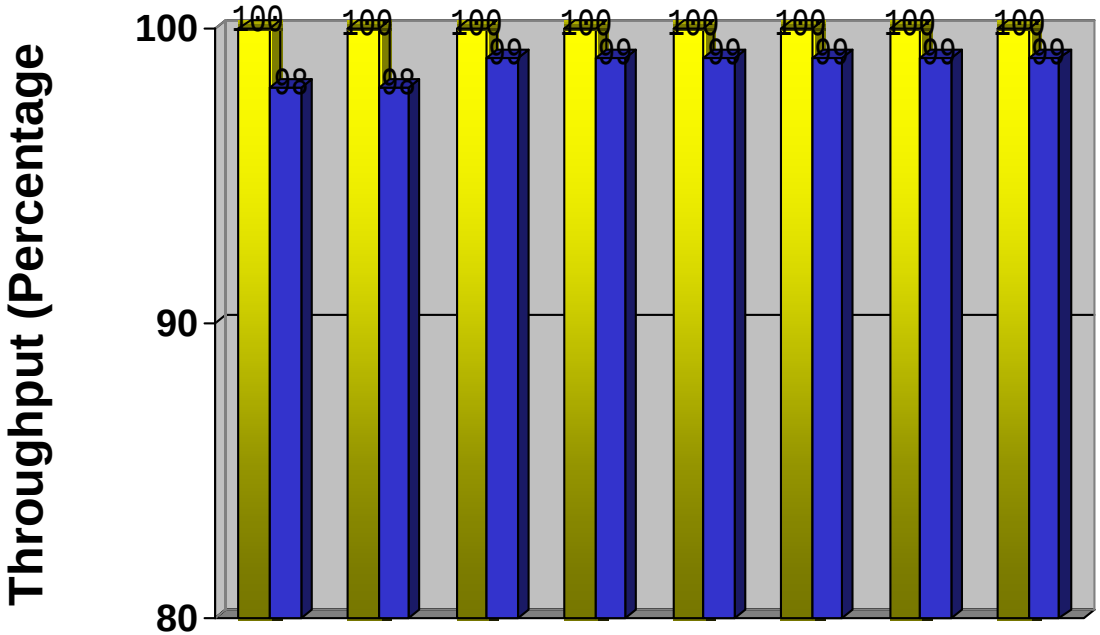
- For functional validation
  - Hardware
  - Firmware
- Nine copper cable connectors.
  - Tested with copper cables up to 5m.
- Three optical Interfaces.
  - XENPAK modules.

Evaluation Board Top View



Performance: Zero-loss Throughput

- Zero-loss Throughput is measured using the *first-silicon*.
  - Optical links are used for 2-port paring configuration.
  - Optical links and Cupper cables are used for full mesh configuration.
    - Slight packet loss is observed (to be fixed in future silicon).



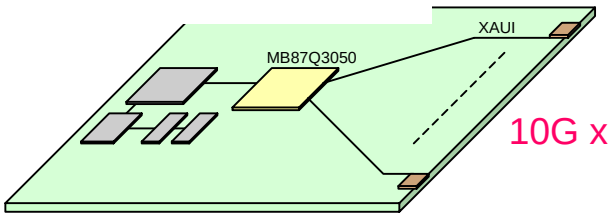
Performance: Latency

- Fall-through latency at 100% throughput workload is measured with the first silicon.

	Latency(nsec)	
Packet size bytes	Cupper Cable	Optical Cable (w/ XENPAK modules)
64	750	1160
128	670	1090
256	660	1190
512	630	1230
1024	670	1280
1280	640	1140
1518	630	1190
9216	660	1220

Application Prototype:1Gb Switch Box

10G Switch Board



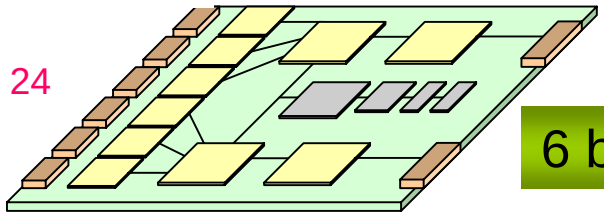
Twelve 10G ports on switch backplane

10G x 12

Switch

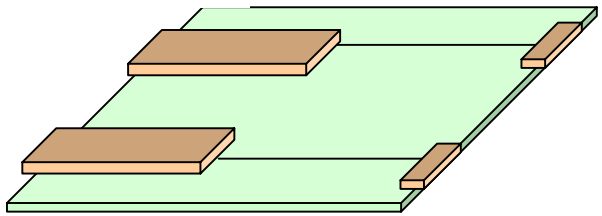
1G Board

1G x 24



6 boards: 1G or 10G board

10G Board



Layer-2 switch for cluster systems
by Fujitsu Laboratories Ltd. (Japan)

Flexible port configuration
Ex) 1G x 144 ports
10G x 12 ports

Summary

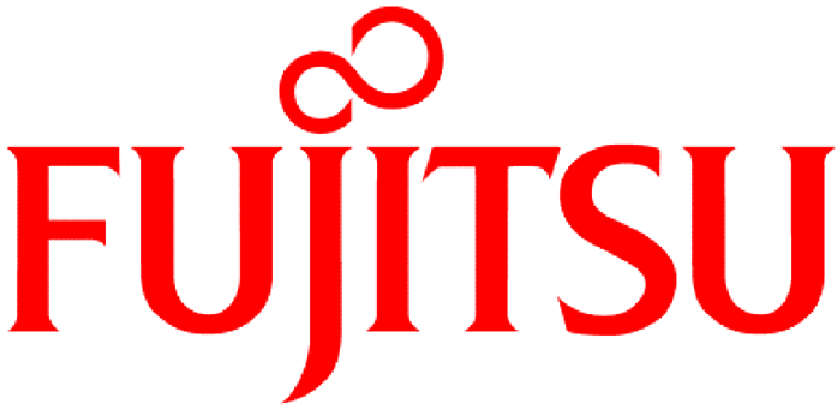
- MB87Q3050 was designed for 10Gb Ethernet-based interconnection for servers and storage equipment.
  - Low cost, dense integration.
  - high-throughput, low latency.
- The switch achieves wire-speed operation at each port, by 240Gbps shared memory bandwidth.
- It also achieves 450ns latency with cut-through forwarding.

Acknowledgement:

The chip was jointly developed by Fujitsu Laboratories of America, Inc., Fujitsu Laboratories Ltd., Japan, and Fujitsu Ltd., Japan.

The backend design was supported by Fujitsu Microelectronics America, Inc.

The development was partially funded by the New Energy and Industrial Technology Development Organization (NEDO), which is one of theJapanese Governmental Agencies.



THE POSSIBILITIES ARE INFINITE