A Variable Length Packet Switch Fabric for ATM / IP Switching in ATM-PON Access System

Chan Kim*, Yeong Ho Park°, Peter Yan°, Hossein Saidi°,
Kyeong Soo Han*, Je Soo Ko*,Tae Whan Yoo*,Hyeong Ho Lee*

* ETRI, 161 Gajeong-dong, Yuseong-gu, Daejeon, 305-350 Korea
° Erlang Technology, Inc.,345 Marshall Avenue, Suite 300, St. Louis MO. 63119 USA

Abstract

A switch board for ATM-PON OLT system was designed and implemented. It has 32 ports of 800Mbps and switches packets or cells among ports according to the routing tags and priority. It uses two shared memory switch ASICs in parallel for port group switching and four FPGAs for port mux, demux, buffering and format conversion. The ASIC is a shared memory switch designed to switch variable length packets as well as ATM cells with multicast, priority, backpressure, port grouping functions. The switch board can be used to IP switching system. (*Thorough tests were not done yet at this time of writing due to system conditions.)

1 Introduction

This paper describes the implementation of a shared memory variable length packet switch fabric for a 20 Gbps ATM-PON OLT system. Fig. 1 shows the ATM-PON system being developed in ETRI. IP services can be provided using IPOA or MPOA protocols. But Ethernet, POS or Ethernet PON line cards or ATM cards with SAR and routing functions can be attatched to the system in which case the switch needs to switch variable length packets with routing functions at the line cards.

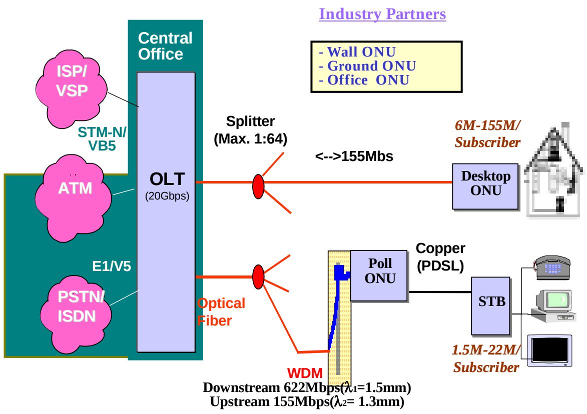

Fig. 1 ATM-PON based access network

The switch chip for the fabric was developed through collaborative work between ETRI and Erlang Technology. This chip has 16 input and output ports at over 622Mbps speed thus can process 10 Gbps data. By using 2 switch chips in parallel, a 20 Gbps variable length shared memory packet switch card was designed and implemented. ATM cells are switched as a general packet with specific length value. FPGAs were used for port multiplexing/demultiplexing, buffering, and format conversion.

The main idea for the switch chip, SE-1, came from Erlang Technology, and the specifications, coding and verification were performed jointly by ETRI and Erlang Technology. The back-end work and fabrication was done by Erlang Technology using 0.18um CMOS technology.

2 The SE-1 Switch Chip

2.1 The Switch Architecture

This switch architecture was designed with the following requirements.

- 10 Gbps throughput per chip
- Expandable to 20 or 40 Gbps throughput using parallel arrangement
- Any port can process variable length packets and fixed length with same scheme
- Full support for multicast traffic
- Channel grouping of up to 4 ports into one channel for higher bandwidth traffic
- Ingress and egress back-pressure for global, or for each port/class queues
- 8 priority class scheduling

Like any other shared memory switch, the data packet should first be segmented into basic units. Each unit is stored at the shared buffer locations and these addresses for these locations are used as common resources in the chip. There are 512 storage locations in the chip. With 16 input/output ports, the incoming and outgoing data packet has the 10 bit wide and 17 word long unit format with start of unit and backpressure. The unit format contains RGA (requested group address), RTAG (routing tag), PRI (priority), NOU (number of units), valid, EOP (end of packet) values. BOP (beginning of packet) is indicated by the extension of start of packet pulse width to two clocks.

For each valid unit received from the input ports, a free location is allocated from the free address list and these free addresses are sent to the data path so that the data can be stored at the newly assigned location. The units of a packet are linked using NEXT field in a memory. At the same time, whenever BOP unit arrives, the RGA and PRI values are extracted and combined with its CSSBA (common storage shared buffer address) and pushed into the ISQ (ingress sub queue). The data in the ISQ are read out and linked to the corresponding ESQ sub-queues according to the RGA/PRI values and multicasting occurs by linking the packet pointer to many ESQs. Also, to enable the multicasting, the FO (fan-out) value extracted from the RGA is written in the memory and this represents how many readings should be done for the location for multicast. The scheduler monitors the ESQ levels to initiate a new packet service when needed. The ESQ itself is a linked list queue having head and tail pointers for each subqueues and free list.

The scheduler also keeps the number of times a packet has been read and if the value is equal to its final fan-out value, it returns the packet’s addresses to the free list.

Every unit-time of 17 clocks, the scheduler determines 16 CSSBA values to read the data from and supplies them to the data path as well as the CSSBA values to write the data at. For each port, if it should begin to read a new packet, it reads the selected ESQ header to know out the CSSBA and the value is supplied to the data path. The FO, NEXT, and Nread values are also read. The NEXT address is kept for later use and the Nread value is compared to FO and either incremented or cleared. When it is not a beginning of a packet, the NEXT address for the port is used as the CSSBA value. Fig. 2 shows the conceptual data flow in this switch chip.

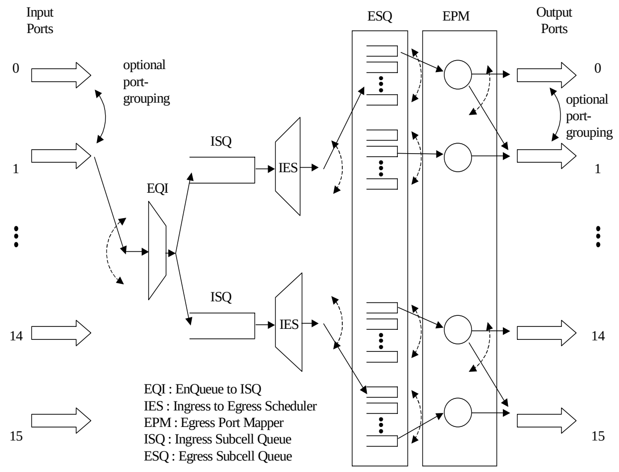

Fig. 2. SE-1 data flow

This chip has the port grouping function by which 2 or 4 input or output ports are grouped to handle higher speed ports like Gigabit Ethernet or STS-48 signals.

2.2 Switch Chip Implementation

The chip is divided into data path and pointer path. The shared buffer has 512 unit locations. The block diagram of the chip is as shown in Fig. 3.

The pointer processing is done in serial manner with many pipelining to increase operating clock frequency. Some techniques were devised to handle coherency problem occuring from consecutive writing and reading with memory access latency. In the datapath, it was possible to write and read the data for all the ports at the same time with small latency without increasing the clock frequncy using techniques similar to those in [1],[2]. Before being written, the write data is aligned with cyclically shifted offset between ports, and the write address is shifted together so that at each clock, different parts of units are written to each physical memory but a specific unit is written to the same address for all the memories. The read process is the same as the write operation.

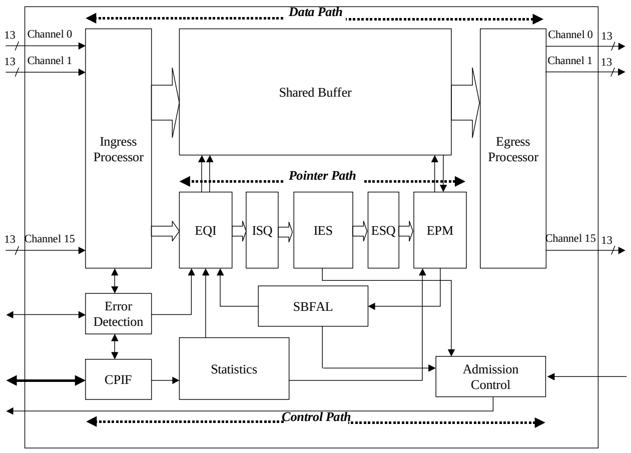

Fig. 3. Chip block diagram

The chip handles parity errors, interface violations, backpressure violations by discarding the packets and forcing EOP for errored packets. Buffer levels are monitored to generate the backpressure and to hold pushing the packet into the ISQ. All the operations are done taking the port goupr into account.

3 The Switch Fabric

The switch card has 32 ports at 622Mbps input and output speed so that the throughput of the board is 32 x 622Mbps = 20 Gbps. It switches packets or cells among ports according to the routing tag and priority of the packet or cell set at the line cards. The card runs at 100 MHz clock frequency and the system uses 850Mb/s serial back plane technology. As shown in Fig. 4, to implement the 20 Gbps switching card, the packets from each two ports are multiplexed at the source SE-0 chip and the multiplexed packet is split into two parallel streams of units for processing by the two SE-1 chips. The split packet in the format of units are independently switched at each SE-1 chip and output to the same SE-2 chip at the same time. The original 32 bit routing tag is split and carried in each stream and combined back after the switching to be used at the destination SE-2 chip for final de-multiplexing to destination port. In short, the SE-0 chip converts the variable length packets into stream of parallel units and SE-2 chips convert each parallel unit streams back into packets. The SE-0 and SE-2 chips were implemented using Lucent's ORT8850H FPSC chips. This FPSC provides the 850Mb/s serial back plane interface with internal pseudo STM framer and PLL/CDR(clock and data recovery).

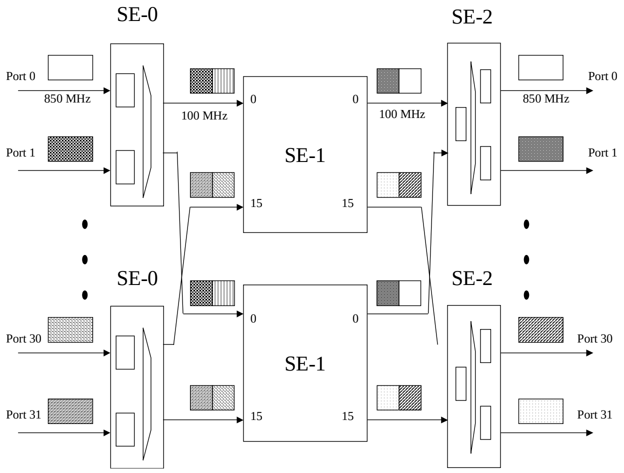

Fig. 4. The architecture of the switching fabric card

At the center of the board are the two SE-1 chips and 4 FPSCs are located between the SE-1 chips and the back plane. One FPSC contains four SE-0 and SE-2 pairs and in one pair, SE-0 processes the ingress traffic for two ports and SE-2 processes the egress traffic for the same ports. Therefore, one FPSC processes 8 ports.

Since there are over 850 interconnection lines on the PCB that are clocked by 100MHz clock, and more than 30 clocks with 100 MHz frequency, designing the switch board is not an easy task. All the high speed lines are impedance controlled. The signal traces were kept as short as possible and damping resistors were used for all the high speed signals(over 850 lines) to reduce the signal reflection and thus to reduce RF emission. Ground vias were placed near the high speed signal vias to provide the RF return current paths.

Fig. 5 shows the picture of the board.

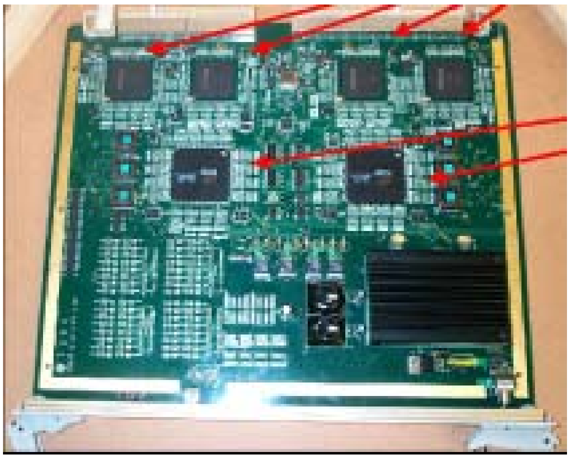

Fig. 5. Photo of the switch board

4 Summary

A variable length packet shared memory switch chip was developed through a collaborative work between ETRI and Erlang Technology and a switch fabric for 20 Gpbs switching was implemented. Together with internal shared buffer size meeting the cell loss requirement [3], the backpressure mechanism plays a major role in reducing the cell loss ratio because it makes the line cards to work closely with the switch card so that the whole system effectively has much larger memory than those in the switch card itself. The ATM layer processor at the line card has much larger egress buffer with backpressure support[4].

The switchboard for the OLT system uses two of this SE-1 chips in parallel configuration for 20 Gbps switching. With serial back-plane technology and optimal block design for less FIFO counts, it was possible to design a switchboard for 32 ports at 800Mbps with four FPSC chips and two SE-1 chips. Using this switch fabric, it is possible to provide pure ATM services, IP services using IPOA, MPOA or pure IP services like Gigabit Ethernet or POS (Packet over SONET).

5 References

[1] Sherry X. Wey, Vijay P. Kumar, "On the Multiple Shared Memory Module Approach to ATM Switching" INFOCOM'92, pp. 116-123, 1992.

[2] W.E.Denzel, et. al, "A Highly Modular Packet Switch for Gb/s Rate," ISS'92, vol.2, pp.236-240.

[3] Yasuro Shobatake, et. al, "A One Chip Scalable 8x8 ATM Switch LSI Employing Shared Buffer Architecture," IEEE JSAC vol. 9, No.8, Oct. 1991.

[4] Chan Kim, et. al, "Implementation of a QOS buffering in ASAH-L4 ASIC with Weighted Round-Robin and Maximum Delay Threshold ," ICACT99, Muju, Korea, Feb. 1999.