Experimental 5-Tb/s Packet-by-Packet Wavelength Switching System Using 2.5-Gb/s \( \times 8\lambda \) WDM Links

Kimihiro YAMAKOSHI†, Nobuaki MATSUURA†, Kohei NAKAI†, Eiji OKI†,††, Naoaki YAMANAKA†††, Takaharu OHYAMA††††, and Yuji AKAHORI††††, Regular Members

SUMMARY We have developed an experimental 5-Tb/s packet-by-packet wavelength switching system, OPTIMA-2. This paper describes its hardware architecture. OPTIMA-2 is a non-blocking 3-stage switch using optical wavelength division multiplexing (WDM) links and dynamic bandwidth-sharing. A new scheduling algorithm for variable-length packets is used for the receiver ports of WDM links and simulation results show that it can suppress short-packet delay while keeping high throughput. An implementation of the WDM link using field programable gate arrays and a compact planar lightwave circuit platform is described. Experimental results for the basic operation of optical wavelength switching are also presented.

key words: packet, switch, WDM, AWG

1. Introduction

Recently, the broadband access services such as ADSL, FTTH, wireless LAN, have been spread and the volume of Internet Protocol (IP) data traffic across the access networks are growing year-by-year. The growth of the access networks will also drive the increase of data traffic of backbone network. Moreover, new services using contents delivery network (CDN) and peer-to-peer have started and these services will also increase the data traffic of the backbone network. As a result, the volume of Internet Protocol (IP) data traffic is growing 2–3 times by year, as shown in Fig. 1. On the other hand, the size of commercial-base switch system for the backbone network is now hundreds of Gb/s. This means that a Tb/s-class switching system [1], [2] for the backbone network will be required in the near future if the data traffic are going to increase at the same pace. The throughput of conventional switches has increased as the process technology for VLSI has progressed according to Moore’s law, but IP traffic has increased even faster. Therefore, some breakthrough technologies are required.

To increase the switch size, there are two approaches. One is to enlarge the size of a basic switch-element by using VLSIs fabricated by deep sub-micron process technology and the high density packaging technologies such as ball grid array (BGA), chip-scale package (CSP), and multi-chip module (MCM) to assemble switch VLSIs. Another approach is to scale the switch size by using the basic switch-elements as many as required using a scalable multi-stage architecture such as a clos-network.

Both approaches have limitations as shown in Fig. 2. In the first approach, a cooling limitation will arise. High-density packaging technologies has resulted in high power consumptions over 1 kW, so cooling systems using a cooling fan or a liquid coolant with a radiator have been required. To deal with much higher power consumption, we propose a distributed switch architecture in order to disperse the heat of the basic switch-element.

Another limitation is the interconnection between basic switch-elements in a scalable architecture. High-speed interconnections are required if the multi-stage switch architecture is used. Electrical signals are likely to suffer from noise or reflections when the signal speed exceeds 1 Gb/s over tens of cm. Moreover, it becomes difficult to deal with the large number of electrical cables used for interconnection between switch-elements. An optical WDM interconnection is superior to an electrical interconnection in that it can propagate high-speed signals over long distances. There-
before we have proposed a distributed switch architecture combined with optical WDM interconnections for Tb/s-class switch system in order to break the above limitations.

As a first step towards Tb/s switch system, we have already developed a 640-Gb/s switching system called OPTIMA-1, which stands for optically interconnected distributed multi-stage Tb/s switching network architecture. OPTIMA-1 has a non-blocking 3-stage switch architecture composed of eight 80-Gb/s multi-chip module (MCM)-based basic switch-elements at each stage, interconnected by optical wavelength division multiplexing (WDM) links [3].

To expand the OPTIMA-1 architecture, we have proposed a 5-Tb/s switching system, OPTIMA-2, that can better handle IP packets [4]. OPTIMA-2 also has a 3-stage switch architecture and employs optical WDM links and dynamic bandwidth sharing to expand the total throughput to 5 Tb/s. The WDM links enable the switching system to expand throughput decreasing the number of cables needed to interconnect switch-elements. OPTIMA-2 has only 1/8 the number of cables that a system using conventional links would need. Moreover, the 2.5-Gb/s maximum bandwidth of each wavelength enables dynamic bandwidth sharing over a total bandwidth of 10 Gb/s at the input or output ports of a switch-element. This makes it possible to expand the total switch throughput to 5 Tb/s without decreasing the statistical multiplexing gain. The bandwidth of each wavelength is controlled dynamically according to the traffic so that packet-by-packet optical wavelength switching is performed.

Most high-speed packet switching systems including IP routers adopt a fixed-sized cell in the switch fabric. Variable-length packets are segmented into several fixed-sized cells when they arrive, are switched through the switch fabric, and are reassembled into packets before they depart [5]. On the other hand, the cost of reassembling cells into packets is considerable high, because the cells switched based on cell-based scheme are likely to be interleaved. To reduce the reassembly cost, OPTIMA-2 has a packet-based scheduling algorithm for variable-length packets, along with simulation results showing the efficiency of the scheduling algorithm. Section 4 describes the architecture of the WDM link system of optical wavelength switching. Finally, Sect. 6 concludes the paper.

2. 5-Tb/s Switch Architecture

This section briefly describes the architecture and operation of OPTIMA-2 [9], [10]. Figure 3 shows the architecture of OPTIMA-2, which has a 3-stage switch architecture eight times larger than that of OPTIMA-1 [3]. Each stage has 64 electrical switch-elements in an 8 × 8 matrix. A switch-element has 8 input × 8 output ports of 10 Gb/s each, giving it a total through-
put of 80 Gb/s \cite{7,8}. The switch-elements in the 1st stage are interconnected to those in the 2nd stage by optical WDM links using eight wavelengths. The switch-elements in the 2nd and 3rd stages are similarly interconnected. An optical WDM link consists of sender ports, receiver ports, and a wavelength arrayed waveguide grating (AWG) router. To expand the total throughput of the switching system, we divide the 10-Gb/s bandwidth of the output port of a switch-element into one of the eight WDM wavelengths at the sender port.

The link speed for each wavelength, $S_{\lambda_i}$ ($\lambda_i$ : wavelength), should be set between 1.25 Gb/s (1/8 of 10 Gb/s) to 10 Gb/s so as to keep the statistical multiplexing gain. On the other hand, the bandwidth for each wavelength, $B_{\lambda}$, is limited by $S_{\lambda_i}$. Therefore the $B_{\lambda}$ satisfies the following formula.

$$B_{\lambda} \leq S_{\lambda_i}, \quad i = 1, \cdots, 8,$$

$$\sum_{i=1}^{8} B_{\lambda} \leq 10 \text{ Gb/s}.$$

$B_{\lambda}$ is dynamically changed from 0 to $S_{\lambda_i}$ as the data traffic varies. In this way, the total 10-Gb/s bandwidth of a switch port is shared among the eight wavelengths.

The bandwidths of these wavelengths are fed to the receiver port via the wavelength AWG router. The scheduler in the receiver port at the input port of a switch-element at the next stage arbitrates cells with or without concatenation for all eight wavelengths. $S_{\lambda_i}$ should be chosen to be as slow as possible under the worst traffic condition \cite{9}. In OPTIMA-2, $S_{\lambda_i}$ is cost-effectively set to 2.5 Gb/s.

To explain the switching operation, we define a switch-element as follows. A switch-element belongs to a particular group numbered from 1 to 8, as shown in Fig. 3. Within a group, each switch-element is a member numbered from 1 to 8. Therefore, we denote the $m$-th member of the $g$-th group at the $n$-th stage as $S_{g,m}(g,m)$. Routing bits of a cell header, $RB_i$ (1 $\leq i \leq 5$) are used for switching the cell inside the system. Here, we explain the case when a cell is switched from switch-element $S_1(g,m)$ at the 1st stage to $S_2(g',m')$ at the 2nd stage. The $g'$-th output ports of all members of the $g$-th group at the 1st stage are connected to the $g$-th input ports of all members of the $g'$-th group at the 2nd stage via the wavelength AWG router. As shown in Fig. 4, thus a cell is switched to the $g'$-th output port at the 1st-stage switch-element in order to switch it to a switch-element belonging to the $g'$-th group at the 2nd stage. First, the cell is electrically switched at $S_1(g,m)$ to the $g'$-th output port according to the routing bit $RB_1$. Then, an appropriate wavelength $\lambda_{m,m'}$ is selected to route the cell from the $m$-th member at the 1st stage to the $m'$-th member at the 2nd stage according to the routing bits $RB_2$. Table 1 shows the wavelength rules for routing from member $m_i$ at the 1st stage (i.e. sender side) to member $m'_j$ at the 2nd stage (i.e. receiver side). The wavelengths vary from 1536 to 1564 nm at a spacing of 500 GHz. In this way, cell-based switching from the $g$-th group to the $g'$-th group is performed electrically by the 1st-stage switch-element, and wavelength routing from $m_i$ to $m'_j$ is performed optically by the optical WDM link using wavelength $\lambda_{m,m'}$. Switching from a member at the 2nd stage to one at the 3rd stage is performed using the routing bits, $RB_3$, $RB_4$.

When a connection request is made, such as from $S_1(g,m)$ to $S_3(g'',m'')$, the system selects an appropriate 2nd-stage switch-element so as to equalize the average load in each wavelength and establishes the con-
nection from the 1st stage to the 3rd stage. To monitor the data traffic in each wavelength, the system uses operations, administration, and maintenance (OAM) cells [10]. Figure 3 shows a cell being switched from $S_1(1,1)$ to $S_3(1,8)$ via $S_2(8,1)$. This represents the system selecting $S_2(8,1)$ at the 2nd stage. The cell is first electrically switched to the 8th output-port in $S_1(1,1)$ at the 1st stage and then optically switched to $S_2(8,1)$ by using wavelength $\lambda_{m1,m'_1} = \lambda_2$. At $S_2(8,1)$, the cell is electrically switched to the 1st output-port and then optically switched to $S_3(1,8)$ using wavelength $\lambda_{m1,m'_8} = \lambda_1$.

3. Design of WDM Link

Cells switched by a switch element at the third stage are reassembled into an IP packet at a line-interface card. In a cell-based switching system, cells belonging to a packet are switched in an interleaved manner. This results in a high reassembly cost at the line interface card. In OPTIMA-2, a variable-length packet is segmented to variable number of fixed-sized cells. Here, we say the variable number of fixed-sized cells as variable-length concatenated-cells. The WDM link is designed to switch the cells belonging to a variable-length packet as variable-length concatenated-cells to reduce the re-assembly cost at the line-interface card.

The optical WDM link consists of sender ports, receiver ports, and an AWG router as shown in Fig. 5. The cells from an output port of a switch element are fed to the sender ports. The sender port assigns a definite wavelength to a cell according to the routing bit in the cell header in order to send the cell to the destination receiver port. Cells are distributed through an address filter into a speed conversion buffer corresponding to the specific wavelength.

Optical cells of eight wavelengths are multiplexed at the output of the sender port and each cell is optically routed at the AWG router. At the receiver port, optical cells of eight wavelengths are converted to electrical signals and arbitrated to be fed to the 10 Gb/s input port of the switch element at the next stage. Arbitration is done at each receiver port independently in order to output variable-length concatenated cells to an input port of the next-stage switch-element. In this way, packet-by-packet wavelength switching is performed in the WDM link.

3.1 Sender Port

The sender port has an 8-channel structure of address filters, speed conversion buffers, and an electro-optical (E/O) module with eight wavelengths. Each channel in the sender port corresponds to one wavelength, and the eight wavelength signals are multiplexed onto a single optical fiber by the AWG coupler in the E/O module. After the routing bits have been checked by the address filters, the cells are accumulated in the speed conversion buffers in order to absorb the link-speed difference between 10 and 2.5 Gb/s, and the cells are sent from the sender port with the specific wavelength corresponding to the routing bit. For optical transmission, scrambling is also employed to avoid long sequences of the same bit in the bit stream of the optical signal. The discontinuity of the same bit sequences ensures operation of the clock data recovery (CDR) circuit in the receiver port. The 2.5-Gb/s × 8-ch electrical signals are converted by the E/O module into optical signals that have a wavelength range from 1536 to 1564 nm with a spacing of 500 GHz.

3.2 Receiver Port

The receiver port consists of an opto-electrical (O/E) module with eight wavelengths, buffers for received cells, and a scheduler for concatenated-cells in eight queues. The eight-wavelength WDM signals are first optically demultiplexed into individual wavelengths by a decoupling-AWG and then converted into electrical signals by the O/E module. Data electrically converted from optical signals of each wavelengths are deserialized, and cells accumulated in the buffers are arbitrated in order to be sent out to a 10-Gb/s bandwidth input port of the next-stage switch-element.

Figure 6 shows a DRR-based arbiter for variable-length packets. As described in Sect. 1, a conventional DRR cannot suppress short-packet delay while keeping...
high-throughput. Figure 7 shows the average packet delay dependency on packet length in the case of DRR with granularity of 5, 10, and 100 cells. Here we assumed that the packet length was normalized to an integer number of cell times. The length of arriving packets was assumed to have an exponential distribution with an average length of ten cells. The probability of packet arrival was assumed to have a Poisson distribution. The short packets are delayed as the granularity becomes large ($G = 100$). It is necessary to use a small granularity ($G/10$) in order to suppress the delay for short packets. On the other hand, the number of hops traversing queues needed to decide the output queue becomes large as the granularity becomes small. This causes the throughput to become low, because there is an upper limit of the hop number at a given line speed. In Fig. 8, though the average number of hops is just 1.43 in the case of $G = 10$, the maximum number of hops is 6 for $G = 10$. If $G$ is set to more than 50 cells, the maximum number as well as the average number of hops is merely one. But the short packets have a large delay than long packets, as in Fig. 7. Therefore, it is difficult to suppress short packet delay while keeping high throughput when we use a conventional DRR.

To resolve the above bottleneck, we used a scheduler with a variable granularity counter value for arbitration. Its algorithm is as follows.

**Step 1.** Calculate the difference $m_i = l_i - d_i$, where $l_i$ and $d_i$ are the packet length at the head of channel $i$ (1 ≤ $i$ ≤ $N$) and the counter value for the channel, respectively.

**Step 2.** Determine the minimum value $m_0$ from $m_i$.

**Step 3.** Add the minimum value $m_0$ to the counter for each channel other than the selected channel.

All $d_i$ are initialized to 0 before these steps. In step 1, $m_i$ is calculated only when there is a packet at the head of the queue in channel $i$. Three steps must be done within one packet-time in order to meet the line speed. When the number of channels is $2^N$, only $N$ operations determining the smaller of two values $m_i$, $m_j$ are required to decide $m_0$ in step 2. For example, when the number of channels is eight, three operations are enough to determine the minimum value among any $m_i$ (1 ≤ $i$ ≤ 8). This resolves the timing margin bottleneck of DRR with small granularity ($G = 10$).

We estimated the packet delay by simulation for the case of an 8-channel scheduler. The conditions for the simulation are the same as in the case of Fig. 7. Figure 9 shows the simulation results for average packet delay for our new scheduler and the conventional DRR scheduler. Our scheduler has a smaller packet delay than the conventional DRR scheduler especially under a heavy load larger than $\rho = 0.9$. Here, the granularity of DRR was assumed to be 100 cells. As Fig. 7 shows, packet delay decreases linearly with decreasing packet length in our scheduling method. Moreover, the delays of packets shorter than 15 cells are smaller than that in DRR for any value of granularity. Therefore, our scheduler can satisfy both high throughput and small delay for short packets especially under a heavy load.

4. Implementation of WDM Link

We experimentally implemented the functions of the
Fig. 10  Block diagrams of the electrical boards for the sender/receiver ports.

Fig. 11  Overview of electrical boards for the sender/receiver ports.

4.1 Sender Port
The interface of the electrical board can handle a total of 10 Gb/s from an output port of the switch-element in the form of 622 Mb/s × 16. The 622-Mb/s signals are first down-converted to 311 Mb/s and then reduced again by a factor of 4 in order to enable a FPGA to operate at 78 MHz, as shown in Fig. 10(a). Signals processed by the FPGA are up-converted to 622 Mb/s. Figure 11 shows an overview of the electrical boards of the sender and receiver ports. The lengths of parallel interconnects on the boards are designed to have variation of less than 2 cm to reduce skews between parallel signals and to ensure a timing margin.

The four parallel 622-Mb/s signals are converted to a 2.5-Gb/s optical signal of the specific wavelength and optically multiplexed onto a single fiber at the E/O module. Figure 12 shows an overview of the E/O and O/E modules of the sender and receiver ports. The distributed feed-back laser diode (DFB-LD) and optical multiplexer provided by the AWG for eight wavelengths were fabricated on a compact PLC platform by hybrid integration [12]. The PLC platform was maintained at a constant temperature of 25°C using a Peltier device in order to prevent wavelength shifts of the DFB-LD.

4.2 Receiver Port
As with the E/O module of the sender port, the AWG-based optical module of demultiplexer, photodetectors (PDs), and pre-amplifiers for O/E module of the receiver port are fabricated on a PLC platform [12], and the temperature of the PLC platform is also maintained at a constant 25°C. Each 2.5-Gb/s electrical signal is converted to the form of 622 Mb/s × 4 parallel by the CDR circuit and then down-converted to 78 Mb/s, just as for the sender port to ensure the operation of the FPGA. The processed data in the FPGA are again up-converted to 622 Mb/s × 16 and input to a switch-element at the next stage.

5. Measurements
To demonstrate the feasibility of wavelength switching of the WDM link, we performed an experiment in the environment shown in Fig. 13. The output of the sender port was connected to the input port of the receiver port via an optical attenuator, which had an optical...
loss corresponding to that of the AWG router. In our system, the AWG router has an average loss of about 7 dB. Therefore we set the loss of the optical attenuator to 10 dB to represent the worst-case scenario for the optical loss of the AWG router.

To estimate the performance of the WDM link, we measured the electrical waveform of the output of a receiver port. We examined the wavelength switching of cells for two wavelengths. Cells corresponding to $\lambda_1$, $\lambda_2$ and two dummy cells were periodically sent to the input port of the sender port. Figure 14 shows the measured waveforms for the output of the receiver port. Figure 14(a) corresponds to the received cells of wavelength $\lambda_1$, which had a routing bit of 0, while Fig. 14(b) corresponds to wavelength $\lambda_2$ with a routing bit of 1. In this way, optical wavelength switching was performed according to the routing bit in the cell header.

6. Conclusion

An experimental system of 5-Tb/s packet-by-packet wavelength switch, OPTIMA-2, was developed. Its hardware architecture and efficiency for handling variable-length packets was presented and the architecture of the optical WDM link was described. The dynamic bandwidth sharing technique of the WDM link enables a total throughput up to 5-Tb/s while keeping the statistical multiplexing gain. A packet-based wavelength switching scheme was used for the receiver port of the WDM link to reduce the cell-assembly cost at the line-interface card. A new scheduling algorithm for variable-length packets was used to resolve the bottleneck of conventional DRR. Simulation results showed that our algorithm can reduce the delay for short packets while keeping high throughput. We described as implementation of an experimental optical WDM link system, composed of electrical boards and optical (E/O and O/E) modules for the sender and receiver ports fabricated using FPGAs and compact PLC platforms. The basic operation of the WDM link for optical wavelength switching was demonstrated in experiments.

References

Kimihiro Yamakoshi  received the B.S. degree in physics from Waseda University, Tokyo, Japan, in 1988 and M.S. degree in physics from Tokyo Institute of Technology, Tokyo, Japan, in 1990. In 1990, he joined Nippon Telegraph and Telephone Corporation’s (NTT’s) LSI Laboratories, Kanagawa Japan. Since then he engaged in the research and development of CMOS digital logic LSIs. Since 1999, he has been engaged in the research and development of high-speed large-throughput switching system. He is now a research engineer at NTT Network Service Systems Laboratories, Musashino, Japan.

Nobuaki Matsuura  was born in Nagoya, Japan on April 18, 1967. He received the B.S. and M.E. degrees from the Tokyo Institute of Technology, Tokyo, Japan, in 1990 and 1992, respectively. In 1992, he joined Nippon Telegraph and Telephone Corporation (NTT), Tokyo. Since joining NTT, he has been engaged in research on the optical interconnection modules and systems. His recent research interests are in signaling and routing in photonic IP networks.

Kohei Nakai  received the B.E. and M.E. degrees in electronic engineering from the University of Tokyo, Japan, in 1995 and 1997, respectively. In 1997, he joined Nippon Telegraph and Telephone Corporation’s (NTT) Network Service Systems Laboratories, Tokyo, Japan. He has been involved in the research of high-speed ATM switching system. Currently, he is engaged in the research and development of the Media Gateway Controller (MGC) in the integrated PSTN - IP network at NTT Network Service Systems Laboratories. He is a member of IEEE Communication Society.

Eiji Oki  received B.E. and M.E. degrees in Instrumentation Engineering and a Ph.D. degree in Electrical Engineering from Keio University, Yokohama, Japan, in 1991, 1993, and 1999, respectively. In 1993, he joined Nippon Telegraph and Telephone Corporation’s (NTT’s) Communication Switching Laboratories, Tokyo Japan. He has been researching multimedia-communication network architectures based on ATM techniques, traffic-control methods, and high-speed switching systems in NTT Network Service Systems Laboratories. He was a Visiting Scholar at Polytechnic University, Brooklyn, New York, from 2000 to 2001. He is now engaged in researching and developing high-speed optical IP backbone networks as a Research Engineer with NTT Network Innovation Laboratories. Dr. Oki received the Switching System Research Award and the Excellent Paper Award from the IEICE in 1998 and 1999, respectively. He co-authored a book, “Broadband Packet Switching Technologies,” published by John Wiley, New York, in 2001. He is a member of the IEEE.

Naoaki Yamanaka  graduated from Keio University, Japan where he received B.E., M.E. and Ph.D. degrees in engineering in 1981, 1983 and 1991, respectively. In 1983 he joined Nippon Telegraph and Telephone Corporation’s (NTT’s) Communication Switching Laboratories, Tokyo Japan, where he was engaged in research and development of a high-speed switching system and high-speed switching technologies for Broadband ISDN services. Since 1994, he has been active in the development of ATM base backbone network and system including Tb/s electrical/optical backbone switching as NTT’s Distinguished Technical Member. He is now researching future optical IP network, and optical MPLS router system. He is currently a senior research engineer, supervisor, and research group leader in Network Innovation Laboratories at NTT. He has published over 112 peer-reviewed journal and transaction articles, written 82 international conference papers, and been awarded 174 patents including 17 international patents. Dr. Yamanaka received Best of Conference Awards from the 40th, 44th, and 48th IEEE Electronic Components and Technology Conference in 1990, 1994 and 1998, TELECOM System Technology Prize from the Telecommunications Advancement Foundation in 1994, IEEE CPMT Transactions Part B: Best Transactions Paper Award in 1996 and IEICE Transaction Paper award in 1999. Dr. Yamanaka is Technical Editor of IEEE Communication Magazine, Broadband Network Area Editor of IEEE Communication Surveys, Editor of IEICE Transaction as well as TAC Chair of Asia Pacific Board at IEEE Communications Society. Dr. Yamanaka is an IEEE Fellow.
Takaharu Ohyama received a B.E. degree from Kyusyu Institute of Technology, Fukuoka, in 1992 and an M.E. degree from Kyusyu University, Fukuoka, in 1994. Since joining NTT Opto-electronics (now Photonics) Laboratories, Ibaraki, Japan, in 1994, he has been engaged in research on the hybrid integration using silica-based planar lightwave circuits.

Yuji Akahori was born in Tokyo, Japan, on May 6, 1954. He received B.E., and M.E. degrees from Waseda University, Tokyo, Japan and a Ph.D. degree from the University of Tokyo, Tokyo, Japan, all in electronics, in 1978, 1980, 1991, respectively. In 1980 he joined the Electrical Communication Laboratory, Nippon Telegraph and Telephone Public Corporation, where he engaged in the research and development of superconductive logic circuits using Josephson junctions. From 1985 to 1993 he worked on the development of InP based monolithically integrated photoreceivers. From 1993 to 2001 he worked on the hybrid integration of opto-electronic devices on silica-based planar lightwave circuits. He joined NTT Electronics Corporation in 2001. Dr. Akahori is a member of the IEEE/LEOS.