Bit-Parallel Message Exchange and Data Recovery in Optical Packet Switched Interconnection Networks

Odile Liboiron-Ladouceur, *Student Member, IEEE*, Carl Gray, *Student Member, IEEE*, David C. Keezer, *Senior Member, IEEE*, and Keren Bergman, *Member, IEEE*

***Abstract*—Multiwavelength optical messages encoded in a bit-parallel fashion are successfully routed through five switching nodes of a 12-port optical packet switching interconnection network. The data payloads are entirely recovered and processed at the destination node using an embedded clock signal with a measured clock-to-data skew tolerance window of 150 ps.**

***Index Terms*—Bit-parallel, interconnection networks (multiprocessor), optical packet switching (OPS), synchronization, wavelength-division multiplexing (WDM).**

I. Introduction

In large-scale high-performance computing systems with physically separated processing and memory elements, one of the most critical challenges is achieving low-latency data exchange among potentially thousands of interconnected elements [1]. The insertion of an optical interconnection network provides ultrahigh capacity communications as well as the opportunity to reduce the overall latency by utilizing the wavelength domain for bit-parallel message transmission. Data traffic in these interconnection networks often consists of short and bursty message exchanges and one of the key challenges is resynchronization and recovery of the data at the destination node without the use of conventional phase-lock loop designs. In synchronous networks, this is achieved by distributing a global clock to every destination node that provides a timing reference for the incoming data [2]. In asynchronous networks, each destination node either has its own local clock reference [3] or incoming data are resynchronized through optical means [4], [5].

In this work, a clock synchronous to the data payload is embedded in the exchanged message and used as the timing reference at the destination node. The embedded clock alleviates the complexity of low-skew clock distribution through large-scale synchronous interconnection networks and simplifies the message recovery circuitry compared to asynchronous networks. This approach is demonstrated by recovering wavelength-division-multiplexing (WDM) bit-parallel messages routed through five switching nodes of an optical packet switching (OPS) interconnection network. Designed to directly interface with two optical network ports, a field programmable gate array (FPGA)-based message processor generates and maps data bytes onto sequences of multiwavelength bit-parallel messages and then detects errors on recovered incoming messages. The signal conversion interface between the message processor and the OPS network is transparent, fully exploiting the available transmission bandwidth while delivering low processing latencies in a scalable WDM bit-parallel message format.

Manuscript received September 13, 2005; revised December 22, 2005. This work was supported by the U.S. Department of Defense under Subcontract B-12-644.

O. Liboiron-Ladouceur and K. Bergman are with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: ol2007@columbia.edu).

C. Gray and D. C. Keezer are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, GA 30332 USA.

Digital Object Identifier 10.1109/LPT.2006.871651

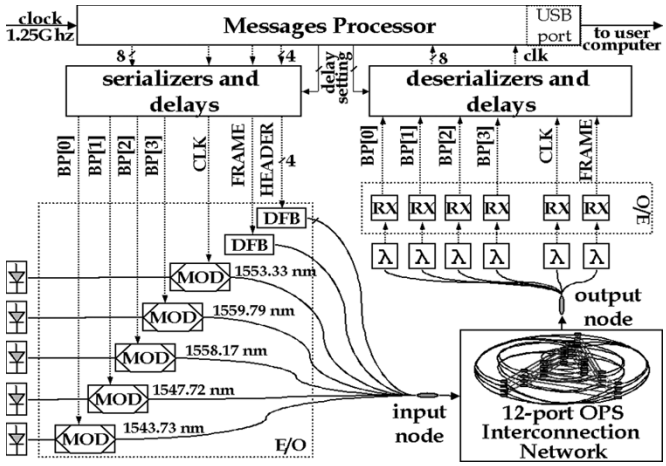

Fig. 1. Message processor connected to an OPS network. Solid lines are optical fibers and dashed lines are electrical signals.

II. System Overview

A complementary metal–oxide–semiconductor FPGA-based (XC2V1000) message processor [6] generates and recovers bit-parallel messages emulating data exchange between microprocessor and shared memory elements interconnected through an OPS network, as shown in Fig. 1.

Each bit of the bytes generated is encoded in a parallel form (BP[0:3] in Fig. 1) on four WDM channels across the ITU C-band to form four-wavelength bit-parallel messages in nonreturn-to-zero format. Messages of 400-ps duration are generated and captured by serializing and deserializing 8 bits of parallel data of 3.2-ns duration at the boundaries of the message processor. A fifth channel is similarly generated to obtain a 1.25-GHz embedded clock signal synchronous to the bit-parallel messages and encoded on a separate WDM channel. Lower bit-rate header and frame signals are additionally generated for proper routing through the OPS interconnection network.

The implemented bit-parallel message processor is a stand-alone design requiring a continuous 1.25-GHz differential clock signal and power supplies (Fig. 2). The clock from

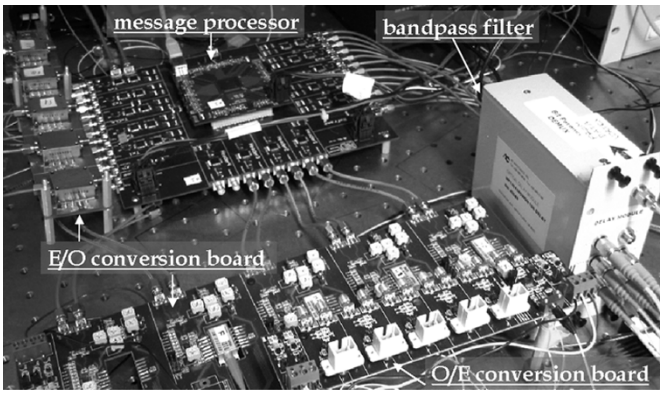

Fig. 2. WDM bit-parallel message processor related components.

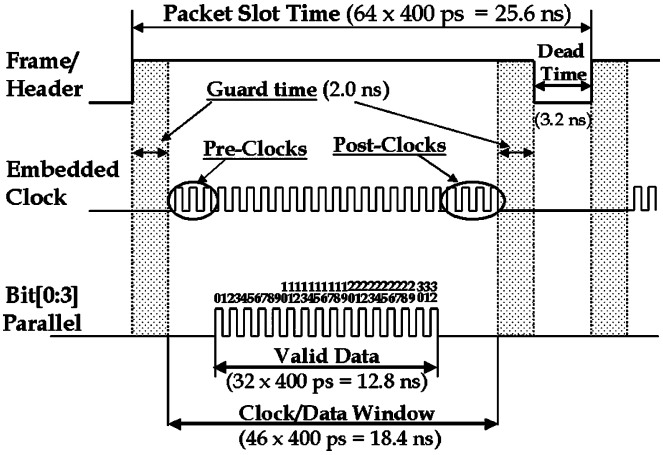

Fig. 3. WDM bit-parallel messages packet structure.

an external instrument provides a low-jitter timing reference to the message processor and serves as a reference for all timing-critical signals. High bandwidth RF cables connect the FPGA-based message processor to the electrical-to-optical and optical-to-electrical (O/E) signal conversion boards. The bit-parallel data payloads and the embedded clock are indirectly modulated using $LiNbO_3$ optical modulators to ensure an adequate extinction ratio required by the OPS interconnection network. The header and frame control signals are encoded by directly modulating cooled WDM distributed feedback lasers. The channel's power is set between -12 and -15 dBm to avoid saturation of the SOA-based switching elements [7]. All channels are multiplexed through optical passive couplers into a single fiber before injection at the input port of the network.

In this demonstration, the interconnection network is a 12-port data vortex [7], a time-slotted self-routing deflection architecture that enables transparent transmission of bit-parallel messages. The data vortex switched interconnection network requires a message slot time of 25.6 ns framed by 2 ns of guard times for routing transients associated with the SOA rise and fall times (Fig. 3). An additional 3.2 ns of dead time between packets is used to distinguish adjacent slots. A time window of 18.4 ns is allocated to the embedded clock due to additional clock transitions required for proper serialization and deserialization of the messages. The valid data portion of the

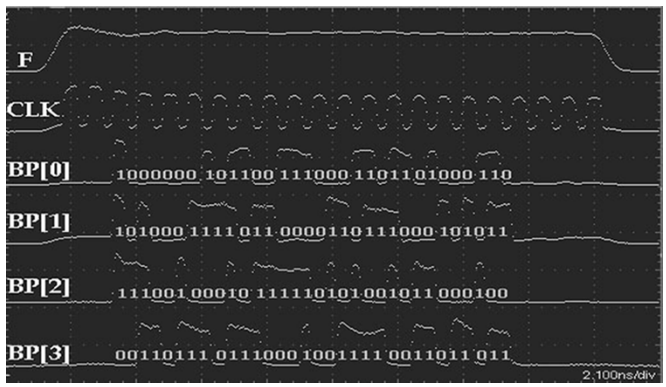

Fig. 4. Received optical bit-parallel messages at the destination node. Top to bottom: the frame signal (F), the embedded clock (CLK), and the 32 bit-parallel messages (BP[0:3]). The 32 messages are indicated.

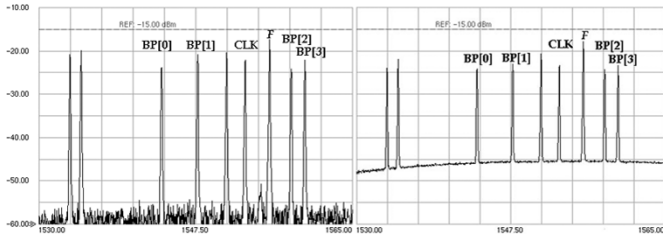

Fig. 5. Optical spectrum of the injected packet at the source (left) and at destination after propagation through five internal switching nodes (right). BP[0]:1543.73 nm; BP[1]:1547.72 nm; BP[2]:1558.17 nm; BP[3]:1559.79 nm; Clock:1553.33 nm; Frame:1555.75 nm. Nonlabeled wavelengths are the routing header signals.

packet structure is then 12.8 ns corresponding to 32 400-ps-long bit-parallel messages, as shown in Fig. 3.

The relative timing of the leading and the trailing edges of the bit-parallel data payloads, the embedded clock, and the frame/header signals are controlled to within 10-ps resolution over a 10-ns range through high-speed variable delay gates controlled by the FPGA. Communication with the message processor via a USB port interface enables a user computer to program values for the tunable delay gates. This timing controllability compensates for relative propagation delay differences in the optical and electrical components, meeting the strict timing constraints of the packet structure protocol.

III. Low Latency Bit-Parallel Message Exchange and Data Recovery

Packets containing a sequence of 32 randomized bit-parallel messages are routed in the data vortex OPS interconnection network (Fig. 4). Based on the encoded header address, the messages propagate through five internal switching nodes, the mean number of nodes required for a packet to reach its destination in this network [7].

The optical spectrum of the packet at the input and output ports of the interconnection network is shown in Fig. 5. The channels' power levels are maintained via amplification by the network's SOA-based switching elements which compensate for all passive splitting losses but also contribute to the accumulation of amplified spontaneous emission noise (Fig. 5). At the output port, the bit-parallel data channels, the embedded clock, and the frame signal are demultiplexed using optical bandpass

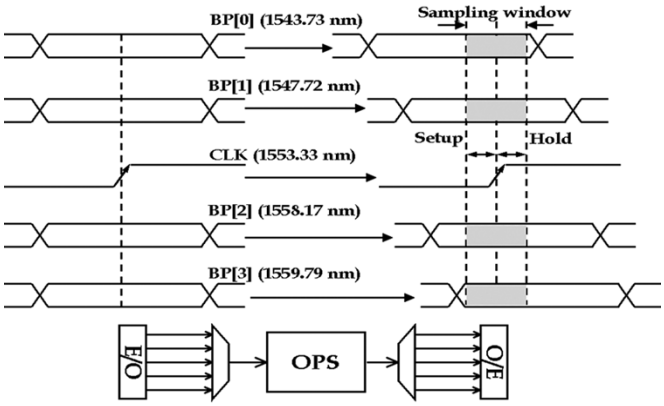

Fig. 6. Clock-to-data skew induced to the bit-parallel messages by the OPS interconnection network.

filters and converted back to electrical signals through 2-GHz p-i-n TIA receiver modules.

After the O/E signal conversion, all signals are passed through delay gates that deskew the channels individually to compensate for relative timing differences between receiver modules and ensure timing accuracy. The 400-ps data payloads are deserialized using the delayed embedded clock as the timing reference and the frame signal as the enabling signal. The deserializer also sends a parallel clock to the FPGA to sample the 3.2-ns parallel word. Via a user software interface with the FPGA, the incoming 32 messages are displayed and compared with the transmitted ones. The message processor thus functions as a bit-error detector. Twenty packets each with 32 randomly generated bit-parallel messages were routed and captured error free.

The data is recovered and deserialized using the embedded clock in 8.7 ns. This processing time is mainly due to the pipeline nature of the deserializer and remains constant as the network scales, thus becoming negligible compared to the communication latency of larger-scale OPS interconnection networks [1].

The clock-to-data skew, defined as the relative timing of the bit-parallel data payload to the embedded clock, must remain within the setup and hold time requirements of the deserializer (Fig. 6). One main contributor to clock-to-data skew is group velocity dispersion (GVD) [8], [9]. The maximum tolerable clock-to-data skew, therefore, limits the wavelength band available for use in the WDM bit-parallel messages packet structure as well as the scalability of OPS networks in terms of fiber length. Hence, a sampling window was defined as the tolerated clock-to-data skew at which the data is sampled and recovered by the message processor error-free. To measure this sampling window, a timing shift is artificially generated between the four bit-parallel data payloads and the embedded clock.

A valid sampling window of 150 ps is measured for the tolerated clock-to-data skew which is used as a metric for the ultimate network scalability. For a system with GVD dominated skew, for example, this sampling window corresponds to over 500 m of interconnected fiber. A very large scale OPS interconnection network such as a 10 k $\times$ 10 k data vortex would incorporate approximately 200 m of interconnected fiber [7].

IV. Conclusion

WDM bit-parallel messages emulating processor/memory data exchange were routed through five switching nodes of a 12-port OPS interconnection network. The message processor specifically designed to directly interface with the optical network transmits a sequence of 32 messages each consisting of four bit-parallel data channels. The messages are entirely recovered and processed at the destination node using an embedded clock signal with a measured clock-to-data skew tolerance window of 150 ps. The demonstrated low processing latency bit-parallel message exchange suggests that this approach can be used to reduce processor/memory access times in large scale computing systems by exploiting the high degree of parallelism afforded by WDM.

References

[1] A. K. Kodi and A. Louri, "Design of a high-speed optical interconnect for scalable shared-memory multiprocessors," *IEEE Micro*, vol. 25, no. 1, pp. 41–49, Jan./Feb. 2005.

[2] A. V. Mule', E. N. Glytsis, T. K. Gaylord, and J. D. Meindl, "Electrical and optical clock distribution networks for gigascale microprocessors," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 10, no. 5, pp. 582–594, Oct. 2002.

[3] I. D. Phillips, P. Gunning, A. D. Ellis, J. K. Lucek, D. G. Moodie, A. E. Kelly, and D. Cotter, "10-Gb/s asynchronous digital optical regenerator," *IEEE Photon. Technol. Lett.*, vol. 11, no. 7, pp. 892–894, Jul. 1999.

[4] M. C. Cardakli and A. E. Willner, "Synchronization of a network element for optical packet switching using optical correlators and wavelength shifting," *IEEE Photon. Technol. Lett.*, vol. 14, no. 9, pp. 1375–1378, Sep. 2002.

[5] E. Kehavas, G. T. Kanellos, L. Stampoulidis, D. Tsiokos, N. Pleros, G. Guekos, and H. Avramopoulos, "Packet-format and network-traffic transparent optical signal processing," *J. Lightw. Technol.*, vol. 22, no. 11, pp. 2548–2556, Nov. 2004.

[6] J. S. Davis, D. C. Keezer, K. Bergman, and O. Liboiron-Ladouceur, "Application and demonstration of a digital test core: Optoelectronic test bed and wafer-level prober," in *Proc. Int. Test Conf.*, Sept./Oct. 2003, pp. 166–174.

[7] A. Shacham, B. A. Small, O. Liboiron-Ladouceur, and K. Bergman, "A fully implemented 12 $\times$ 12 data vortex optical interconnection network," *J. Lightw. Technol.*, vol. 23, no. 10, pp. 3066–3075, Oct. 2005.

[8] L. A. Bergman, C. Yeh, and J. Morookian, "Advances in multichannel multiGbytes/s bit-parallel WDM single fiber link," *IEEE Trans. Adv. Packag.*, vol. 24, no. 4, pp. 456–462, Nov. 2001.

[9] A. P. Togneri and M. E. Vieira Segatto, "All optical bit parallel transmission systems," in *Proc. SBMO/IEEE MTT-S Int.*, vol. 1, Sep. 2003, pp. 367–372.

[10] M. L. Loeb and G. R. Stilwell, "High-speed data transmission on an optical fiber using a byte-wide WDM system," *J. Lightw. Technol.*, vol. 6, no. 8, pp. 1306–1311, Aug. 1988.