

Design and Demonstration of an All-Optical Hybrid Packet and Circuit Switched Network Platform for Next Generation Data Centers

Howard Wang, Ajay S. Garg, and Keren Bergman

Department of Electrical Engineering, Columbia University, 500 West 120th Street, New York, New York 10027
howard@ee.columbia.edu

Madeleine Glick

Intel Labs Pittsburgh, 4720 Forbes Avenue, Suite 410 / CM2, Pittsburgh, Pennsylvania 15213

Abstract: The design of a reconfigurable 4×4 photonic switch simultaneously supporting both packet and circuit switching is proposed and experimentally validated. Error-free operation and routing correctness is verified for pseudorandom 4×10-Gb/s packets and a 10-Gb/s stream.

©2010 Optical Society of America

OCIS codes: (060.4253) Networks, circuit-switched; (060.4259) Networks, packet-switched

1. Introduction

Due to the growing trend towards the utilization of commodity microprocessors in implementing clusters and data centers, large-scale computing systems consisting of up to tens of thousands of nodes can be found with increasing frequency in both commercial and research environments. As a result, the performance of these systems is increasingly reliant on the capabilities of the communications infrastructure. Furthermore, as various applications characterized by communication intensive workloads in areas including finance, scientific computing, data mining, and cloud computing begin to emerge, the demand placed on the underlying interconnect is reaching an intensity such that traditional electronic networks can no longer meet the aggregate bandwidth needs demanded by these systems. Although, current generation electronic fabrics can nominally support up to 10 Gb/s per port [1], it is becoming restrictively costly to maintain large-scale data centers at this level of performance. Power, cooling, space, and component cost have become critically limiting figures of merit in the design of future data centers and clusters. Moreover, as the overall performance of commodity compute nodes continues to scale, the resulting stress on the data center interconnection network will only serve to increase the need for resources in order to maintain acceptable performance.

Photonic interconnect-based solutions can provide significant improvements over traditional electronic means in terms of capacity, power consumption, and bandwidth-distance product [2]. By leveraging the bit-rate transparency of optics and the capacities afforded by wavelength division multiplexing (WDM), the energy per switched bit can be greatly reduced, potentially enabling scalability to hundreds of thousands of compute nodes. As such, optical technologies represent a promising solution space to this emerging communications bottleneck.

However, the wide range of applications classes found in data centers generate a variety of traffic patterns characterized by both short messages with random destinations and large extended flows, the latter corresponding to a significant proportion of the network capacity [3,4]. Therefore, an interconnect that can support both types of traffic would represent a highly advantageous solution in terms of performance, energy, and cost. To that end, we present the design and implementation of an all-optically switched network platform featuring the ability to route both packet and circuit switched traffic *simultaneously* in a configurable manner. Our experimental platform is implemented as a hybrid packet/circuit 4×4 SOA-based switching node, leveraging our previous work in optical packet switching design [5]. We demonstrate the capacity and traffic adaptability of the proposed platform by concurrently validating the error-free routing and transmission of short, 4×10 Gb/s wavelength-striped optical packets and the successful establishment of a high-bandwidth circuit transporting a 10 Gb/s data stream.

2. Hybrid Switching Node Architecture

Fig. 1a depicts the basic architecture of the 4×4 optical switch. Sixteen semiconductor optical amplifiers (SOAs) organized in a gate-array configuration serve as photonic switching elements. The broadband capability of the SOA gates facilitates the organization of the transmitted data onto multiple optical channels, exploiting WDM to achieve extremely high capacities. The SOA array can be dissected into subsets of four, with each group corresponding to one of the four input ports. Similarly, one SOA gate in each subset corresponds to one of four output ports, enabling

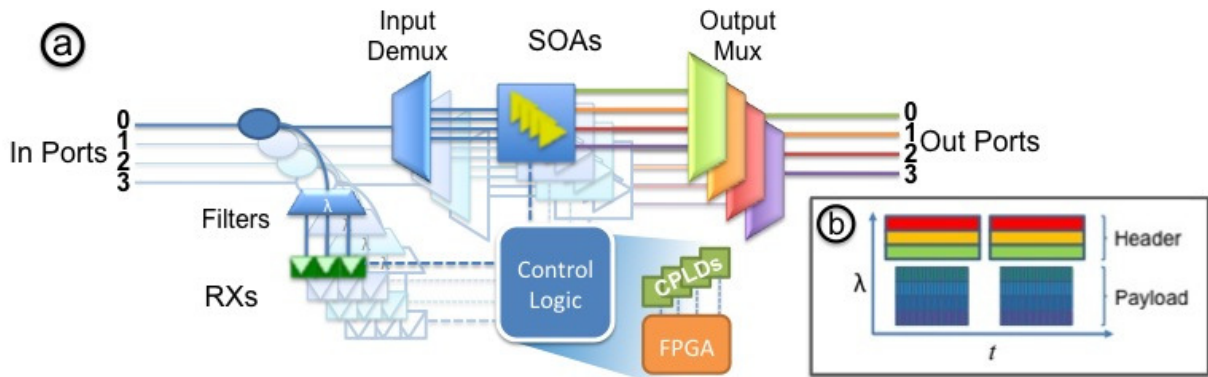


Fig. 1. a) 4x4 hybrid SOA-based switching node architecture. b) Wavelength-striped packet structure

non-blocking operation of the switching node. The SOA gate matrix is controlled via electronic signaling originating from programmable logic elements.

Each input of the switching node can be configured on a port-by-port basis to operate in either packet or circuit mode, resulting in a switching decision based on the current configuration of each input port. Ports configured to accept packetized traffic support messages adhering to a wavelength-striped format (Fig 1b). These packets utilize the wavelength domain to not only segment high bit-rate payload data onto multiple wavelengths, but also to encode routing information in the form of a low-latency protocol using single-bit per wavelength headers. Each input port is associated with three low-speed receivers used to recover the header information of each packet. One receiver retrieves a dedicated frame bit used to denote the presence of a packet while the remaining two receivers detect addressing bits. If a port is configured to serve circuit traffic, the control logic is set to ignore the incoming header bits and establishes an externally assigned path. Passive optical couplers and filters serve to divide the incoming optical packet between the receivers for header detection and the SOA array for switching. The gain of the SOAs is set such that packets experience zero net insertion loss through the switch. Output contentions are managed within the control logic and resolved by dropping packets and prioritizing circuit switched paths. However, due to the programmability of the nodes, this behavior can be easily modified to adapt to applications that may benefit from a different policy.

3. Experimental Test Bed

In the laboratory, the 4x4 hybrid optical switch is implemented as four sub-modules (on custom electronic boards) each corresponding to a specific input port (Fig 1a). Each module is functionally equivalent to a 1x4 switch and contains a subset of the functional components of the node. Specifically, any given module contains four SOA switching elements controlled by a complex programmable logic device (CPLD). Each CPLD is connected to three low-speed 155Mb/s PIN-TIA-LA receiver assemblies for header detection. Each module is managed by a central field programmable gate array (FPGA)-based controller implemented on a multipurpose evaluation board connected

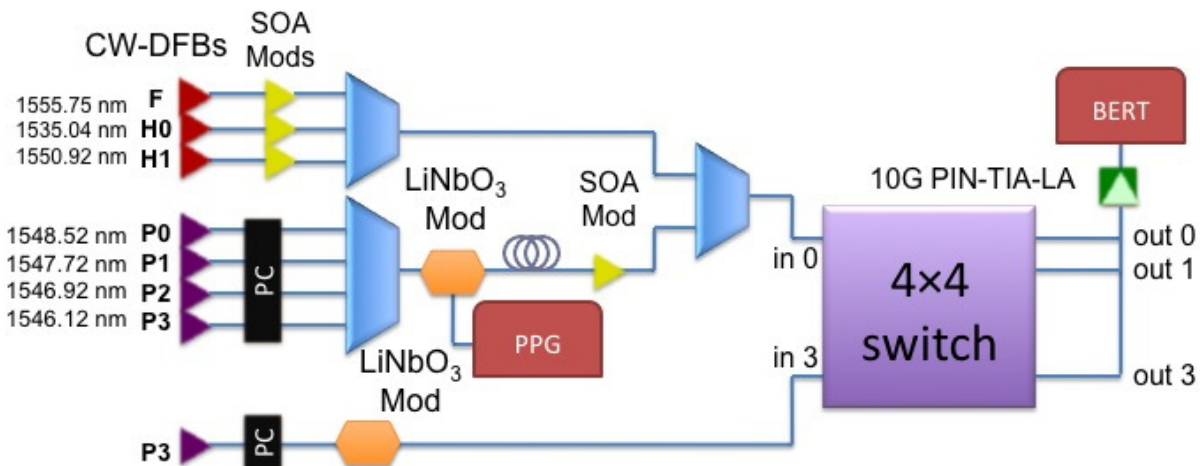


Fig 2. Experimental setup for the evaluation of hybrid switch demonstrating simultaneous packet and circuit switching functionality

to each of the four CPLDs. The FPGA controller manages output contentions and delivers the appropriate routing information to each sub-module according to their input port configurations (i.e. packet or circuit switched). In the current implementation, the FPGA-based controller is programmed to accept incoming configuration information from an onboard Ethernet port. Configuration frames are generated from a computer terminal by the user via a graphical user interface.

4. Demonstration and Results

In order to verify the capability of the proposed architecture as a hybrid-traffic network platform, both packetized and circuit switched data originating from distinct sources are simultaneously injected into the test-bed. Fig 2 schematically illustrates the experimental setup used in this evaluation. Four CW laser sources (P0-P3) are simultaneously modulated with pseudorandom data at 10 Gb/s via a LiNbO₃ modulator, decorrelated, and gated by a SOA to form 204.8 ns packets. The header signals (F, H0, H1) are generated in a similar manner and combined with the payload data to form wavelength-striped packets. The packets are injected into input port 0 and routed to the

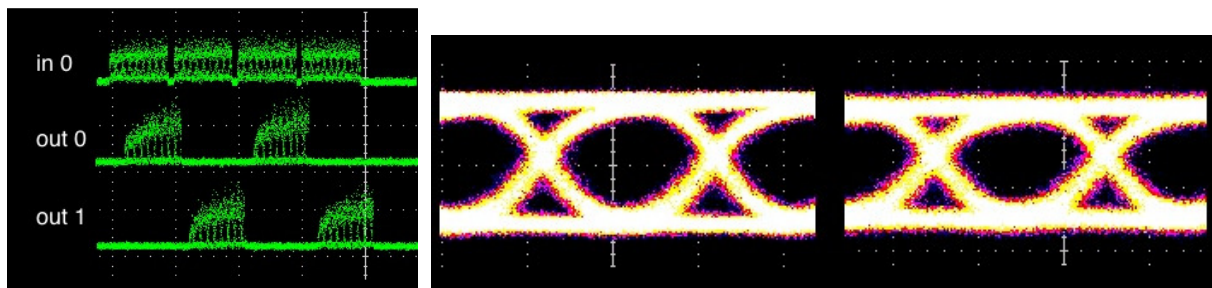


Fig. 3. Left: Waveforms corresponding to packets routed from input port 0 to output ports 0 and 1. Middle: Input eye diagram of circuit switched stream. Right: Output eye diagram of circuit switched stream.

correct destination ports appropriately while a circuit path is concurrently established from input port 3 to output port 3 and injected with a 10 Gb/s circuit switched data stream.

Fig 3. illustrates the results of the experimental evaluation, depicting correct routing of packets and the input and output eye diagrams of the 10 Gb/s circuit switched stream. All packets were detected at the output via a 10Gb/s pin-tia-la assembly and received by a bit-error rate tester to validate error-free data.

4. Conclusions

We have presented and experimentally implemented a design for a 4×4 all-optical network switch concurrently supporting both packet and circuit switched traffic in a configurable manner. As a result, the switch represents a functional experimental platform for the evaluation of data center applications characterized by varying traffic demands. Future plans are in place for a more integrated and modular node design, which will serve to minimize latency, expand functionality, and optimize footprint. Additionally, plans for the evaluation of the node in a practical computing environment are in place for the near future.

The authors gratefully acknowledge support for this work from the Intel Corporation under grant SINTEL CU08-7952 and the NSF FIND under grant CNS-837995.

5. References

- [1] IEEE 802.3 LAN/MAN CSMA/CD (Ethernet) Access Method. <http://standards.ieee.org/getieee802/802.3.html>
- [2] Glick, M., "Optical Interconnects in Next Generation Data Centers: An End to End View," *HOTI '08. 16th IEEE Symposium on High Performance Interconnects*, 2008. vol., no., pp.178-181, 26-28 Aug 2008.
- [3] Fred, S. B., Bonald, T., Proutiere, A., Régnié, G., and Roberts, J. W. 2001. "Statistical bandwidth sharing: a study of congestion at flow level." in *Proc. of the 2001 Conf. on Applications, Technologies, Architectures, and Protocols For Computer Communications* (San Diego, California, United States). SIGCOMM '01. ACM, New York, NY, 111-122.
- [4] Al-Fares, M., Loukissas, A., and Vahdat, A. 2008. "A scalable, commodity data center network architecture." in *Proc. of the ACM SIGCOMM 2008 Conf. on Data Communication* (Seattle, WA, USA, August 17 - 22, 2008). SIGCOMM '08. ACM, New York, NY, 63-74.
- [5] A. Shacham, K. Bergman, "An Experimental Validation of a Wavelength-Striped, Packet Switched, Optical Interconnection Network," *Journal of Lightwave Technology* **27** (7) 841-850 (Apr 1, 2009).