

The OSMOSIS Optical Packet Switch for Supercomputers: Enabling Technologies and Measured Performance (Invited)

Richard R. Grzybowski¹, B. Roe Hemenway¹, Michael Sauer¹, Cyriel Minkenberg²,
François Abel², Peter Müller², Ronald Luijten²

1: Corning Incorporated, Science and Technology Division, Corning, NY, USA

2: IBM Research GmbH, Zurich Research Laboratory, Rüschlikon, Switzerland

Abstract – The OSMOSIS project explores the role of optics in large-scale interconnection networks for high-performance computing (HPC) systems. Its main objectives are solving the technical challenges to meet the stringent HPC requirements of high bandwidth, low latency, low error rates, and cost-effective scalability. We discuss the technologies and architectural innovations that enabled us to build a demonstration system meeting these targets. We demonstrate the optical performance for the 64 ports @ 40 Gb/s data paths across the semiconductor optical amplifier based optical crossbar, and report on the implementation of the electronic central controller.

Keywords: Optical Switch, SOA, 40G, low latency

Introduction

The rationale, architecture, and operation of Optical Shared MemOry Supercomputer Interconnect System (OSMOSIS) have been described in detail in [1], and more details on the implementation of the controller are available from [2]. Here, we present our latest advances in the implementation of the data and the control paths, as well as measurement results obtained on the actual demonstrator hardware.

2. Data-path technologies

The key requirements for an high-performance computing system interconnection network are low latency, high bisectional bandwidth, extremely low error rates, and scalability to thousands of nodes. In the Corning-IBM joint OSMOSIS project fast optical switching is accomplished using Semiconductor Optical Amplifiers (SOA). Electronic packet buffers at the inputs temporarily store packets when contention occurs. A low-latency scheduler coordinates the transmission of packets across the optical data path and the gate timing of the SOAs. The architecture is amenable to multistage scalability by means of electronic packet buffers between the stages with link-level flow control. The data path must scale to meet increased user throughput and port counts while remaining compact and dissipating acceptable power. The OSMOSIS data path is based on an optical broadcast and select architecture, wherein both space and wavelength division multiplexing at 40 Gb/s per port are used for broadcast and two stages of semiconductor optical amplifiers are used for selection switching. By activating just two SOAs per packet period at each receive node, one for “fiber select” and one for “color

select”, the entire packet switch is reconfigured packet by packet for all ports. The challenges of scaling, size synchronization, optical dynamics, power dissipation and cost are all determined by this architecture.

A data path requirement is to scale bit rate per port starting at 40 Gbp/s. This is achieved by designing an optical data path that is largely transparent to both bit rate and modulation format. In the demonstrator, we implement a DWDM channel map on 200 GHz spacing with binary, on-off amplitude shift keying (OOK) modulation of each optical carrier. The spectral efficiency is 0.2 Gb/s per hertz of optical spectrum, a relatively conservative figure of merit compared to the state of the art (at >1.2 bit/sec/Hz for 40Gb/s). The channel capacity is determined largely by the spectral bandwidth of the optical data path, from transmitter to receiver. Today, the path is limited by the concatenation of three identical optical DWDM filters. These filters have a 1 dB passband of 80 GHz which enables the OOK bit rate to increase directly by a factor of two. High speed electro-absorption modulators (EAMs) are used and these have been demonstrated at more than 80 Gbps, providing headroom in that design.

The architecture supports increasing port count in two ways. The first increases the wavelength dimension of the switch, presently at eight DWDM channels per bus fiber. The limit is determined by the gain bandwidth of the SOAs and by optical signal to noise ratio. Today the gain bandwidth of the amplifiers exceeds 4000 GHz, of which only 1600 GHz is used today. The minimum required OSNR for 10^{-12} uncorrected bit error ratio is >24 dB for OOK modulation while the present measured received OSNR is >35dB. A second approach to scale the port count increases the number of fiber buses. This doesn't induce additional optical impairment except by increasing the combiner loss in the passive collector stage following the fiber selectors.

In optical packet switches different paths will have slightly different loss characteristics, resulting in packet-packet power variations at the receiver. The challenge is to limit the power excursion of the optical packets to within the dynamic range of the receiver. Mitigation requires controlling the optical path loss variation, minimizing systematic loss variation in the optical components with wavelength or channel loading, and maximizing the dynamic range at the receiver. With OOK modulation, the optimum decision threshold tends to vary with received packet power also, so an electrical limiter amplifier prior to the decision point is used. In the present system the

component loss variation is controlled through standard specifications of the optical components. Typical variations are less than 0.5 dB worst case (over temperature and polarization) for the SOA and 0.3 dB for the multiplexers. The passive components such as splitters, combiners and passive interconnect matrices exhibit less than 0.1 dB port-port loss variation. The largest static contributor to path-dependent power variation arises from connector and splice losses throughout the system. Channel pre-emphasis and inverted gain tilt in the EDFA compensate a small, 0.7 dB, systematic gain tilt in the SOAs. The sum of these static variations is approximately 2 dB. The net result is a data path with maximum packet-packet power variations of approximately 3 dB. To match the channel dynamics, the receivers achieve a dynamic range of 5 dB, providing almost 2 dB margin.

3. Control-path implementation

The central controller's main job is to configure the data path on a cell-by-cell basis (i.e., every 51.2 ns) to minimize latency and maximize throughput. To this end, it executes a hybrid scheduling algorithm that allows speculative as well as pre-scheduled transmissions. The controller also performs flow control to avoid buffer overruns, relays end-to-end reliable delivery acknowledgments, and performs multicast scheduling to allow efficient one-to-many communications.

The central controller comprises a total of 48 FPGA devices. Of these, 32 implement the interfaces (i.e., two interfaces per device for a total of 64 ports) to the computing nodes (adapters) and 8 implement the interfaces to the optical switching modules (i.e., 16 switching modules per device for a total of 128). These 40 interface FPGAs are each located on a separate daughter board. The daughter boards plug into both sides of a large (57%42 cm²) midplane (see Fig. 1). In addition, this midplane hosts the remaining 8 FPGAs, which implement the scheduling logic: Three devices (SCH0-SCH2) for parallel unicast scheduling [3], one device (SCH3) for multicast scheduling [4], one (STX) for arbitration of speculative transmissions, one (MUX) to merge unicast and multicast schedules and perform multiplexing of the parallel unicast schedulers, one (ACK) to route end-to-end reliable-delivery acknowledgments, and finally one (CLK) for clocking and miscellaneous control functions. Figure 1 shows pictures of both sides of the populated midplane without the daughter boards. The design of this extremely complex board, which comprises 36 layers, 13,000 wires (including 4,000 differential pairs), and 45,000 vias, has been described in [5]. The board has been successfully powered up and is currently undergoing extensive testing.

5. Conclusion

OSMOSIS demonstrated that 40G optically switched system operation is possible with tremendous scalability options with respect to bit rate, number of wavelengths and use of fiber channels in the optical domain. Furthermore, 40G optics was shown to be a viable alternative to striped 10G systems in the future. While the development of custom ASIC circuitry could certainly optimize system

speed and packing density in the future, OSMOSIS was realized using only commercially available FPGAs. As of this report, OSMOSIS is the only system level 64-way, 40G test optical switch test bed in the world. It is also the largest installation of discretely packaged SOAs in a single system.

The implementation of the OSMOSIS demonstration system is approaching its final stage, in which the data path and the control path will be married to provide the full system functionality. Once completed, the system will undergo functional and performance testing to verify whether the requirements are fulfilled.

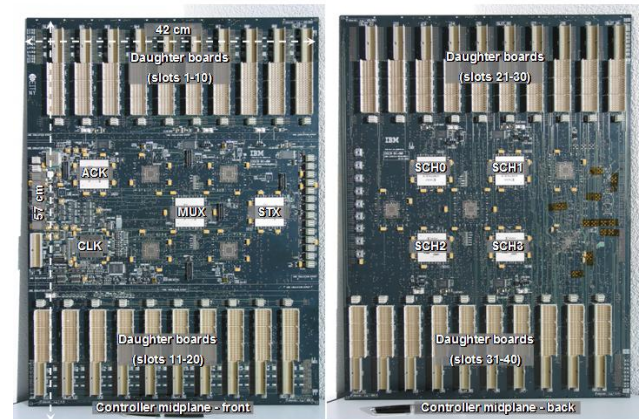


Figure 1 : Controller mid-plane, front and back.

6. Acknowledgment

This research is supported in part by the University of California under subcontract number B527064. We thank our sponsors at the University of California, acknowledge the contribution of their colleagues to this work.

6. References

- [1] R. Hemenway et al., "Optical Packet-Switched Interconnect for Supercomputer Applications," *OSA J. Optical Networks*, vol. 3, no. 12, Dec. 2004, pp. 900-913.
- [2] C. Minkenberg et al., "Designing a Crossbar Scheduler for HPC Applications," *IEEE Micro Special Issue on High-Performance Interconnects*, vol. 26, no. 3, May/June 2006, pp. 58-71.
- [3] C. Minkenberg, I. Iliadis, and F. Abel, "Low-Latency Pipelined Crossbar Arbitration," in *Proc. Global Telecommunications Conf. (Globecom 04)*, IEEE Press, 2004, pp. 1174-1179.
- [4] E. Schiattarella, C. Minkenberg, "Fair integrated scheduling of unicast and multicast traffic in an input-queued switch," in *Proc. 2006 IEEE International Conference on Communications (ICC '06)*, Istanbul, Turkey, Jun. 11-15, 2006
- [5] P. Dill, "Layout of the OSMOSIS Optical Switch Controller Board using Expedition," *2006 Mentor Graphics International User Conference (User2User '06)*, San Jose, CA, May 3-6, 2006