Optical Switching System for MPP, LAN, or WAN Systems

W.A. Castellano
R.D. Chamberlain
R.R. Krchnavek

W.A. Castellano, R.D. Chamberlain, and R.R. Krchnavek, “Optical Switching System for MPP, LAN, or WAN Systems,” in *Proc. of the 1997 IEEE Pacific Rim Conf. on Communications, Computers and Signal Processing*, August 1997, pp. 260-264.

Department of Electrical Engineering
Washington University
St. Louis, Missouri

Optical Switching System for MPP, LAN, or WAN Systems

William A. Castellano Roger D. Chamberlain Robert R. Krchnavek
bill@ccrc.wustl.edu *roger@ccrc.wustl.edu* *rrk@ee.wustl.edu*
Dept. of Electrical Engineering, Washington University, St. Louis, Missouri

Abstract

Whether intended for use in massively parallel processing (MPP), local-area networking (LAN), or wide-area networking (WAN) systems, optical switching technology has not yet seen significant use in commercial practice. This paper presents the design of an optical switching system suitable for all of the above uses. The data path is all optical, with electrical-to-optical signal translation taking place only at the message source and optical-to-electrical signal translation taking place only at the destination. The control path, however, makes extensive use of electrical technology, using a unique control algorithm appropriate to the mixed-technology design. We refer to this dual (optical and electrical) design as the Gemini *interconnection network.*

1 Introduction

During the past ten years, the computer networking industry has experienced tremendous growth in the performance of local- and wide-area networks. The wide proliferation of the Ethernet LAN set a de facto standard data transfer rate at 10 Mb/s. Fast Ethernet, FDDI, and ATM switching architectures increased network performance by an order of magnitude or more as 100 Mb/s to 622 Mb/s data rates became realizable. The second generation of ATM switching networks have increased network data rates by another order of magnitude as 2.4 Gb/s systems are now making the transition from the research labs to the commercial marketplace. At the same time, the interconnection networks linking processing elements in MPP systems have mirrored the above growth in data rates, often using the same technological advances to achieve performance improvement.

A central component in all modern WANs and many high-speed LANs is optical fiber. As a link technology, it provides very high bandwidth (25 THz) and very low loss over long distances (< 0.5 dB/km). As a switching technology, however, optics leaves much to be desired. Currently, optical switching techniques (e.g. $LiNbO_3$ electrooptic switches, semiconductor optical amplifier gates, self-electrooptic effect devices) are generally expensive, have high insertion loss, and have low density in comparison to electronic switching techniques. However, recent trends indicate economies of scale are reducing the price of such technologies. The result is that current systems use electrical technology to perform the switching function. Optical signals are converted to electrical form, routed through the switch, and converted back to optical form for insertion into the outbound optical link. As the electronic switching elements cannot operate at anywhere near the data rate of the optical link, the system bandwidth limits are determined primarily by the data rate within the switches.

This paper describes the design of a switching system that enables the use of optics throughout the data path. High data rate translation between the electrical and optical domains occurs only at the two endpoints of a connection, not at the ends of every link. One implication of this is that the network will scale up to higher data rates as the end-point technology improves, without having to replace network switching elements.

To minimize the use of optical switching elements, a Benes topology interconnect is used. This allows the switching system to be non-blocking while using only $O(N \log N)$ switching elements, rather than the $O(N^2)$ required for a crossbar topology. A novel contribution of this work is the electronic control used for output arbitration and routing within the Benes topology. The separation of control and data path yields a system where the operating frequency and functional requirements for each technology are closely matched to their capabilities. We refer to this dual (optical and electrical) design as the *Gemini* switch. With appropriate alterations to the same basic design, the system can be used in an MPP environment, a LAN environment, or a WAN environment.

2 Optical Switching Fabric

Electrooptical 2×2 switching elements are the key devices in the fabrication of the *Gemini* $N \times N$ optical data path. These switching elements rely on the electrooptic effect (i.e., the application of an electric field to an electrooptical material changes the refractive index of the material). The result is a 2×2 optical switching element whose state

[1]The research described here was performed at the Computer and Communications Research Center, Department of Electrical Engineering, Washington University, Campus Box 1115, One Brookings Dr., St. Louis, MO, 63130-4899. William Castellano is currently with Comdisco, Inc., Rosemont, IL.

is determined by an electrical control signal. This is illustrated in Figure 1, which shows a switching element in the pass through state as well as in the crossover state. These switching elements can be fabricated using $LiNbO_3$ as well as other materials.

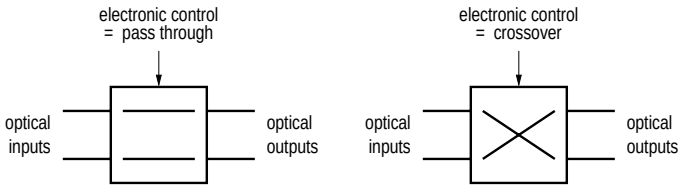

Figure 1: Electrooptical switching elements

A Benes topology is used in the *Gemini* switch. An $N \times N$ Benes topology is constructed in a recursive fashion as illustrated in Figure 2. The first and last stages are each composed of $N/2$ 2×2 switching elements. The center stages are constructed from two $N/2 \times N/2$ middle stage subnetworks (MSS) that also use a Benes topology. The first and last stages are connected to the MSSs using a perfect shuffle interconnect. The complete network has $2 \log_2 N - 1$ stages of switching elements that must be traversed in any path from input to output.

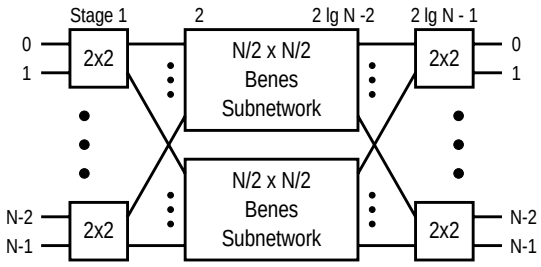

Figure 2: $N \times N$ Benes topology

Physical interconnection between switching elements on the same circuit board is accomplished using polymer channel waveguides and between boards optical fiber is used. This results in a highly manufacturable and ultimately low-cost design.

A brief power budget analysis is pertinent. We begin with the source and receiver. High-speed laser diodes are routinely fabricated with output powers of 0 dBm. Receivers designed to operate in the 1-5 Gb/s data rate have been fabricated with minimum receiver sensitivities of -30 dBm [8]. Therefore, the power budget allows for a total loss of 30 dB. We next consider $LiNbO_3$ electrooptic switching elements. Interferometric $LiNbO_3$ switches exhibit significant polarization dependence and polarization preserving fiber is required to connect them. This presents a difficulty in board-level applications. Polarization preserving waveguides are not readily fabricated and excessive losses and reduced extinction ratios would result from using standard (non-polarization preserving) waveguides. However, Y-branch $LiNbO_3$ switches demonstrate a high degree of polarization insensitivity [7]. Recent results have shown 2×2 $LiNbO_3$ Y-branch switches that demonstrate an insertion loss of < 5 dB, a polarization dependent loss of < 1 dB, and crosstalk of < -35 dB [6]. Therefore, an 8×8 Benes topology (consisting of 5 stages) would have a total insertion loss due to the switches of 25 dB. In addition to switch loss, additional losses arise due to the waveguides connecting the switches. Organic polymers are the likely candidate for the interconnection medium because of the relatively large size of substrate required to hold the $LiNbO_3$ switches. Current polymer materials can exhibit relatively low loss. For example, photochemically-set, multifunctional acrylate monomers/oligomers from [4] have a reported loss of 0.05 dB/cm at 1550 nm. Using this number, a loop diameter of 15 mm for curved waveguides, 0.1 dB for crossover losses, and 0.5 dB for fiber/waveguide connector loss [1], the total insertion loss for the longest path length in the 8×8 Benes topology is 28.5 dB. This is less than the total allowed loss of 30 dB, but clearly does not allow room for an additional stage.

Finally, while improvements in receiver sensitivity, polymer waveguide losses, and 2×2 $LiNbO_3$ insertion losses are likely to increase the switch size, it is expected that such improvements will not dramatically increase the size and optical amplification will be required. This amplification could be provided by semiconductor optical amplifiers or fiber amplifiers.

3 MPP, LAN, and WAN Systems

Using the optical switching fabric described above, an $N \times N$ switch optimized for an MPP system is shown in Figure 3. Each processor attached to the switch has two parallel input links, an electrical control path and an optical data path, and two output links. The use of parallel (electrical and optical) connections makes sense in a tightly-coupled MPP system environment, where the physical distances involved are small (≈ 1 m), and pairing an optical fiber with a copper cable would not be a significant cost.

Adapting the MPP system to a LAN environment involves altering the design to support a greater distance between the devices connected to the network. This is accomplished by merging the parallel control path and data path into a single optical link between the *Gemini* switch and the devices connected to the network. The control information is transmitted at a lower data rate than the payload, using a distinct wavelength (this is feasible since the header is significantly smaller than the payload). This is illustrated in Figure 4, where λ_c represents the control wavelength and λ_d represents the data wavelength. Optical-to-electrical and electrical-to-optical translation takes place on the lower bandwidth control wavelength, but not on the high bandwidth data wavelength. Multiplexing and demultiplexing the control and data wavelengths can be integrated

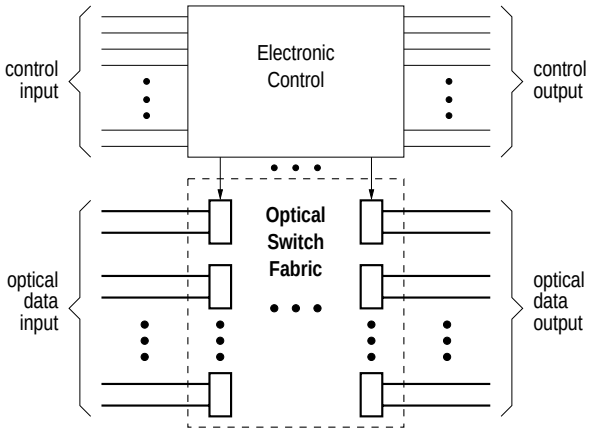

Figure 3: MPP optical switching system

with the optical switch fabric using planar diffraction grating techniques [5].

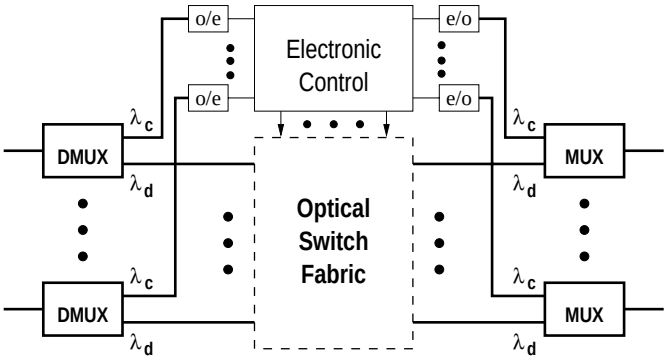

Figure 4: LAN optical switching system

The major distinction between a LAN switch and a WAN switch is the need to perform buffering of the payload within the switch. In a WAN environment, packet switching has significant performance advantages over circuit switching, and when two incoming packets are destined for the same output link, one of them must be buffered. In the *Gemini* data path, this is accomplished by using an $M \times M$ switch fabric to implement an $N \times N$ switch, where $M = kN$ for integer k of approximately 2 to 10. The $M - N$ unused ports form recirculation paths (of one packet length) from the output port back to the input port. In this way, any packet delivered to a recirculation path is buffered (in optical memory) until the next packet time. The WAN data path is illustrated in Figure 5. The recirculation paths need to provide both delay and optical amplification. An optical fiber amplifier is appropriate for this function.

4 Controller Design

An electronic controller associated with each $N \times N$ switch is responsible for output arbitration, routing, sig-

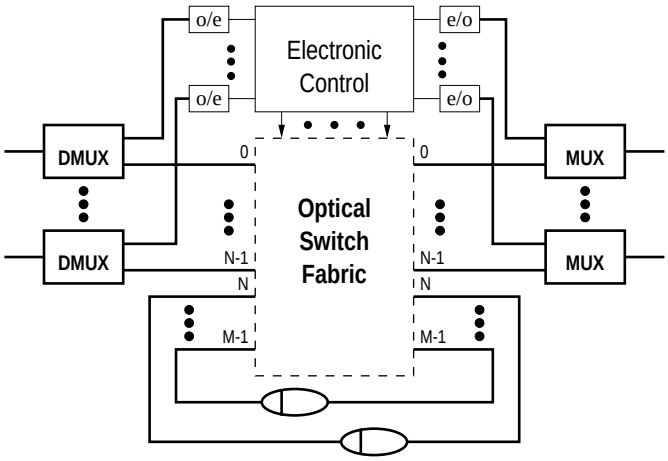

Figure 5: WAN optical switching system

naling, and flow control decisions both within the network and within an individual switch. The design of the routing functionality deserves special mention. Routing in a Benes network involves the implementation of a graph coloring algorithm that, although straightforward in concept, executes serially, requiring significant execution time for large graphs. In this paper, we describe a novel technique that is implemented entirely in combinational logic, significantly decreasing the time required to perform the Benes network routing function. A VHDL model of the control path has been implemented and shown to operate as expected [2].

The detailed description of the controller design is centered around a WAN system, since it is the most involved of the three configurations. In this configuration, an $M \times M$ switching fabric is used to build an $N \times N$ switch with $M-N$ recirculation paths.

4.1 Controller Data Path

The data path within the electronic controller consists of M input ports, M output ports, and a crossbar interconnect between the input and output ports. This is illustrated in Figure 6. Note that this data path is distinct from the optical switching fabric which handles actual payload data. The electronic crosspoint complexity is significantly less than that of the electrooptical switching elements in the data path, making an $O(N^2)$ controller a reasonable match with on $O(N \log N)$ data path. The N external output ports are arranged in ascending order whereas the $M-N$ recirculation output ports are arranged in descending order. This arrangement was chosen in order to implement the priority scheme described below. For an MPP or LAN system, the N external output ports can be arranged in either order and the recirculation ports are unnecessary.

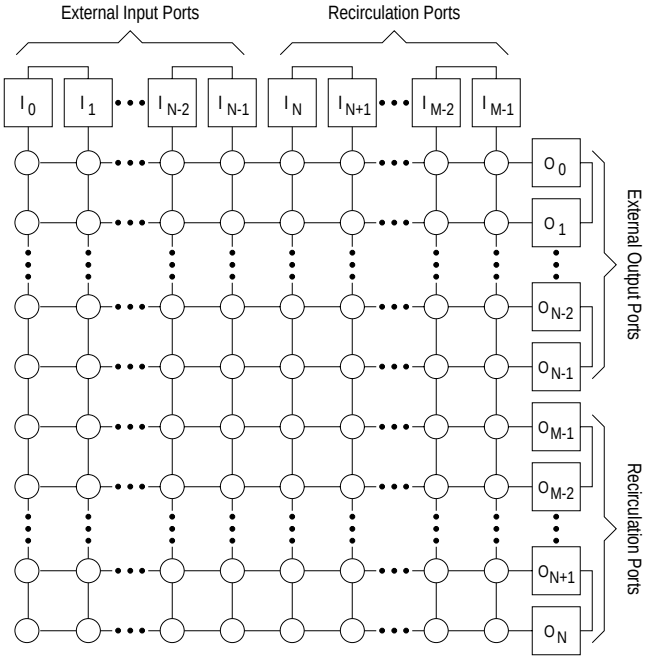

Figure 6: Controller data path

4.2 Output Arbitration

The output arbitration procedure is designed to give the recirculation ports priority over the external input ports when contending for the external outputs. The recirculation ports are prioritized in descending order of output arbitration. I_{M-1} is given first choice of the N external output ports, I_{M-2} is given second choice, etc. Packets which have lost the contention process are sent to the recirculation ports in descending order. A losing packet is first sent to O_{M-1}, if O_{M-1} is already occupied, the packet is sent to O_{M-2}, etc. This dual-priority system ensures that all packets will eventually reach the proper output port in the order that they have arrived, even if all packets in the recirculation fabric are addressed to the same output. Packet loss occurs when all $M - N$ recirculation ports are occupied and two cells contend for the same external output port. (In an MPP or LAN system, an upstream grant signal is used to preclude the delivery of packets that would be lost due to contention.)

The arbitration begins with the input ports asserting their target output port address on the vertical buses and the output ports issuing a grant signal to the left on the horizontal buses. The crosspoint logic (CL) compares the target output port address with its row address and when a match occurs waits for the grant signal from the right. On each row, the rightmost CL with a match receives the grant signal, winning the arbitration. It blocks the grant signal from traveling further, replacing it with a release signal that is sent to the right. A matched CL that receives a release signal lost the arbitration. It replaces the target output port address on the vertical bus with the address of the highest priority recirculation output port (port $M - 1$).

A similar arbitration technique is used for the recirculation output ports. When a CL detects a match, it wins the arbitration if it receives a grant from the right and loses if it receives a release signal from the right. Losing CLs place the next lower priority recirculation output port address on the vertical bus to contend for another recirculation port.

All of the above operations can be implemented in combinational logic, and if the delay associated with one CL is t_{CL}, the maximum delay for output arbitration to complete is $2Mt_{CL}$. Once complete, the winning CLs put themselves in a "corner" configuration, connecting their top port to their right port in an L. Losing CLs put themselves in a "pass-through" configuration.

4.3 Graph Coloring

The routing of signals through the outermost stages of a Benes network requires the solution of a graph coloring algorithm. In an $M \times M$ fabric, the first and last stages are each comprised of $M/2$ switching elements. Each switching element (pair of input ports or pair of output ports) forms a vertex in a bipartite connection graph. The edges in the graph (two for each vertex) represent the connections that result from the output arbitration described above. The edges are colored such that no two edges that connect to the same vertex can share the same color. The color is a single bit that indicates whether the packets at a switch should be routed through the upper subnetwork or the lower subnetwork of the Benes switching fabric. The two ports in a vertex must route to different subnetworks so that they do not contend for a common switch port.

Traditional graph coloring algorithms execute serially, traversing the graph one vertex at a time and assigning color values. Our design improves upon this technique by doing the graph traversal in combinational logic, significantly decreasing the execution time of the algorithm. The connection graph edges are represented in the controller data path by the "corner" connections present in Figure 6. Input ports and output ports are paired to represent graph vertices.

The graph coloring algorithm uses a global priority-based resolution scheme in which each color bit has a priority value. At each vertex (port pair), when two color bits are the same, the lower priority color bit is flipped and its priority updated to that of the higher priority bit. This is an entirely combinational function. If t_{edge} is the time required to perform the above color bit resolution and propagate the results along a graph edge, the time required for the color bits to settle is $vt_{edge}/2$, where v is the number of vertices in the connection graph ($v = M/2$ for the initial iteration).

4.4 Iteration

The execution of the graph coloring algorithm described in the above section determines the switch settings (pass

through or crossover) for the first and last stages of the Benes network. This algorithm must be recursively executed to determine the switch settings for the interior subnetworks.

The connections between the outermost stages and the interior subnetworks are via a perfect shuffle.

$$PS(x, v) = (x \bmod 2) \times v + \lfloor x/2 \rfloor \quad (1)$$

Here, x is the output port number of the current stage and v is the the number of 2×2 switching elements in the current stage (equal to the number of vertices in the connection graph). $PS(x, v)$ returns the input port number of the middle stage subnetwork to which port x connects. Equation (1) expects x to be between 0 and $2v$ and returns a value in the same range. The next stage equation maps port numbers over the complete range of 0 to M.

$$NS(x, v) = PS(x \bmod 2v, v) + x - (x \bmod 2v) \quad (2)$$

Since v is constrained to be a power of 2, the implementation of (2) is straightforward and can be area efficient.

Once (2) has been applied to the output ports of the first stage and the input ports of the last stage, the desired input-output connections are known for the next interior stage. The output arbitration mechanism is executed to set the CLs (there should be no output conflicts), the graph coloring algorithm is repeated to determine the switch settings, v is divided by 2, and (2) is applied again.

After $\log_2 M$ iterations of the above sequence, the middle stage of switches are all that remain. This consists of $M/2$ 2×2 switching elements that are set to pass through if the output port is connected to the same-numbered input port and are set to crossover otherwise.

5 Summary and Conclusions

Based upon a structural VHDL model of the controller electronics and our past experience fabricating semi-custom VLSI chips, we estimate the following chip area requirements and packet cycle times for a 0.5 μm fabrication process [2, 3],

M	est. area	est. cycle time
32	100 mm^2	< 900 ns
64	400 mm^2	< 3400 ns

which indicates that a controller for a 32×32 fabric is currently practical to implement on a single chip. To construct larger systems, a pipelined approach is necessary, where the controller data path is replicated and multiple iterations of the graph coloring algorithm are executed concurrently. This has the ability to decrease the per packet cycle times listed above (although the latency for an individual packet would not be diminished).

The above analysis indicates that an 8×8 *Gemini* switch can be constructed without the need for optical amplification and a 32×32 switch is feasible with amplification. The data rate achievable through the system is limited only by the speed of the endpoints (i.e., electrical-to-optical and optical-to-electrical translation). We are currently constructing a limited scale (4×4) prototype of the optical data path, and will report on its performance in the near future.

References

[1] T. S. Barry, D. L. Rode, and R. R. Krchnavek. Highly efficient coupling between single-mode fiber and polymer optical waveguides. *IEEE Transactions on Components, Packaging and Manufacturing Technology, Part B: Advanced Packaging*, August 1997.

[2] W. Castellano. Design and implementation of the controller hardware for a high performance 8×8 optical switching system using VHDL. Technical report, Computer and Communications Research Center, Washington University, St. Louis, MO, 1997.

[3] Compass Design Automation. ESCB52SY140 0.5-micron 3-volt optimum silicon (OS) library. Technical report, Compass Design Automation, San Jose, CA, June 1996.

[4] L. Eldada, C. Xu, K. Stengel, L. Shacklette, R. Norwood, and J. Yardley. Low-loss high-thermal-stability polymer interconnects for low-cost high-performance massively parallel processing. In *Proc. of the Third International Conf. on Massively Parallel Processing Using Optical Interconnections*, pages 192–205, 1996.

[5] R. A. Livingston and Robert R. Krchnavek. Planar diffraction grating for board-level WDM applications. In *Proc. of the Third International Conf. on Massively Parallel Processing Using Optical Interconnects*, pages 77–84, 1996.

[6] E. J. Murphy, T. O. Murphy, R. W. Irvin, R. Grencavich, G. W. Davis, and G. W. Richards. Enhanced performance switch arrays for optical switching networks. In *Proc. of ECIO*, 1997.

[7] Y. Silberberg, P. Perlmutter, and J. E. Baran. Digital optical switch. *Appl. Phys. Lett.*, 51:1230, 1987.

[8] L. D. Tzeng, O. Mizuhzra, T. V. Nguyen, K. Ogawa, I. Watanabe, K. Makita, M. Tsuji, and K. Taguchi. A high-sensitivity APD receiver for 10-Gb/s system applications. *IEEE Photonics Technology Letters*, 8(9):1229–1231, September 1996.