Large-scale IP router using a high-speed optical switch element [Invited]

Tom McDermott and Tony Brewer

Chiaro Networks, 269 W. Renner Parkway, Richardson, Texas 75080

tmcdermott@chiaro.com; tbrewer@chiaro.com

Received 2 April 2003; revised manuscript received 25 June 2003

The system design and architectural considerations for a large, high-performance IP packet router that uses a nonblocking optical switching fabric are presented. The objective of the router is to provide fully network-compatible routing of IP, multiprotocol label switching (MPLS), and Ethernet packets in a router with a very large number of high speed ports while maintaining the low-delay, low-jitter, low-packet-loss, and line-rate throughput characteristics of today's small port-count routers over a large scale. Such a large router is useful for the core of a packetized transport network capable of supporting various classes of real-time and best-effort service in a reliable and efficient manner. © 2003 Optical Society of America

OCIS codes: 060.0060, 060.4250.

1. Introduction

This paper describes the design considerations for an IP/multiprotocol label switching (MPLS) packet router. The router provides industry standard interfaces over a range of data rates from 155 Mbit/s (OC3/STM-1) to 10 Gbit/s (OC-192/STM-64) including 1- and 10-Gbit/s Ethernet. The inlet packet streams are independent of one another in both frequency and phase and are formatted in accordance with industry standards. The router provides electronic packet buffering and header processing for a large number (315) of 10-Gbit/s interfaces, and it uses an optical switch fabric for internal system interconnect and switching. This paper is organized as follows: In Section 2 we discuss the general operational requirements of the router and the optical switch fabric. Section 3 briefly discusses the technology employed in the optical switching element. In Section 4 we discuss the alternate approaches to implementing an optical packet fabric and discuss the architecture chosen. Section 5 discusses other considerations (such as synchronization and scheduling) for a large, centralized optical packet fabric. Section 6 summarizes the performance achieved, whereas Section 7 discusses lessons learned and areas for further research.

2. Requirements and Objectives

The requirements and performance objectives established for the IP router dictate the chosen characteristics and technology. The switching fabric provides the interconnection and switching between the various input–output (I/O) port cards that make up the router. Although much of the literature deals with the throughput and performance required from a switching fabric, there are additional requirements for a switching fabric that are needed for effectively making a router operational that is intended to grow modularly over time. This allows for starting small but increasing router capacity when it is necessary or desirable, in a manner that does not affect the operation of the router (i.e., the router remains in service while it is being grown).

The requirements established for the router that affect the design and architecture of the switching fabric used are as follows:

High availability. No single point of failure should be able to affect the performance of the router, including no degradation of throughput, delay, or performance. The approach chosen was to provide $1+1$ operation of the switch fabric—all modules, cables, and equipment common to the switch fabric were duplicated. In addition, maintenance actions on one copy of the switch fabric or cabling, or failures of one copy of the switch fabric or interconnecting cable, should not cause any packet loss. This facilitates performing upgrades, maintenance, or repair on the router without the need to schedule any down time, which is extremely important to some network operators. A 1:*N* fabric protection scheme was analyzed but rejected because it would make an in-service topology change very difficult (it would require many more upgrade steps).

Simple architecture. The switch architecture should be simple and topologically independent of the capacity, size, and fault situations. Some switch-fabric topologies (such as *n*-dimensional meshes) require rerouting or remapping of the connectivity when partial faults occur or when the fabric capacity is increased. Of concern was that the software algorithms or tables in the router would have to be changed in such a situation. In addition, it was believed that partial failure situations could make the response of the software very difficult to predict, thus leading to performance faults that might be difficult to discover completely during the testing phase. A nonblocking fabric architecture, combined with $1+1$ protection, can make the fabric routing independent of the fault or growth situation. In addition, a nonblocking fabric with sufficient capacity avoids the problem with hot spots, which may occur in attempting to assign packets dynamically to a distributed switch. Although the fabric itself is nonblocking, contention will occur. The scheduling and arbitration methods to resolve contention are described below.

Simple interconnect. A third requirement was that the interconnecting cables between the switch fabric and the line shelves be simple to understand and minimal in count. This makes growing the switch fabric capacity while the router is in-service a much more tractable proposition. The choice of an optical, nonblocking fabric absolutely minimizes the total system interconnect count. The use of $1+1$ protection means that the two copies of the fabric, and their interconnecting cables, are completely independent. In fact, they may be wired disjoint to avoid faults where both copies of an interconnecting cable might be accidentally disconnected or severed. In the final implementation the two working copies of cabling were color coded red and blue to prevent such a problem. Because an optical switch is used, data arriving at the optical switching element must be time aligned with the synchronous optical switching element. The requirement to cut optical cables to specific lengths to equalize delay was avoided through an automatic cable length range system that adjusts the timing of the line cards by means of a closed-loop feedback method to achieve timing synchronism of the optical data at the optical switch element. This makes changing cables very simple and avoids installation errors. This synchronization mechanism is described in Section 5.

Technology growth. A fourth requirement was that the router should be architecturally independent of the switching technology to the extent practical. The switching fabric should be able to be replaced with new technology (larger port count, higher capacity) while the router is in service. As will be discussed below, this difficult objective was achieved in a novel manner through the use of an optical switch.

Line-rate performance. The most important requirement was that the switch fabric permit complete line-rate performance of all the line cards in the system simultaneously re-

gardless of single faults, maintenance, or upgrade activities, for any implementable scale of the fabric. For practical purposes this means that a single copy of the switch fabric is capable of supporting the router at full performance, since the off-line copy may be undergoing upgrade or maintenance. In addition, the delay and delay jitter should not grow as the router and switch fabric scale in port count. Section 4 discusses the specific performance attributes of the optical switch fabric design.

In-service upgrade. The router must be able to be upgraded in service, that is, without the interruption of packet traffic. Ideally, this would happen without the loss of any packet traffic for upgrade of the common elements. Upgrade of any single line card may require brief (millisecond) outages for that one interface.

3. 64×64 Fast Optical Switching Element

Previous optical switch elements can be roughly characterized as small and fast or as large and slow. This means that nanosecond-speed optical switching is available only for small-scale optical switching elements (such as lithium-niobate–based switches), whereas large switching elements of dozens of ports switch in the millisecond range [such as moving mirror switches based on microelectromechanical systems (MEMs)]. There are a few technologies that span intermediate ranges (such as acousto-optical switches).

The optical technology employed for this project is based on the electro-optic effect and parallel waveguide arrays implemented in GaAs/AlGaAs. General construction details and performance measurements have been reported previously [1, 2]. The parallel waveguide arrays essentially form a phased-array emitter whose beam profile can be altered (or steered) to one of many output fibers by means of individually adjusting the applied bias voltage on each of the parallel waveguide devices. The switching element used is a 64×64 (64 input and 64 output fibers) nonblocking element, and it completely reconfigures to a new state in approximately 30 ns. Six of these are used as the center stage of a Clos switch to provide a 384-port nonblocking fabric. Section 5 provides more detail on the speedup factors and throughput.

The optical switch itself is composed of 64 beam deflectors, each of which contains 128 parallel optical waveguides. In the current switch, 72 beam deflectors are fabricated, but only 64 of them need to function properly (this enhances manufacturing yield). Thus there are 72×128, or 9216 optical waveguides contained within the switch. A custom 128-output digital-to-analog converter (DAC) integrated circuit (IC) was designed that provides all the bias voltages for one beam deflector. Seventy-two of these ICs are used in the switch to provide the 9216 analog voltages. The DAC ICs have a 30-ns output settling time, which accounts for the 30-ns switching stabilization time. The GaAs waveguides themselves have subnanosecond time constants. All 72 ICs are loaded with the next switch state over a period of ∼300 ns; then they are all clocked simultaneously to transfer all 9216 analog voltages to the new switch state values at the same time. Thus the switch completely changes state during a 30-ns period.

The optical switch module houses four GaAs die (each of which contains 18 beam deflectors), lenses, input fiber coupling v-groove arrays, and the output fiber array in a hermetic enclosure. In addition, the enclosure contains a circuit board that regulates the temperature of the switch assembly, shock mounts, and control electronics to program the switch state and the 9216 analog-to-digital converters. The assembly was environmentally tested for compliance with operational and nonoperational temperature, shock, humidity, and handling requirements derived from Telcordia GR-63.

The average fiber-to-fiber loss of the switch is ∼13 dB, the worst-case cross talk between any two inputs to one output is better than −24 dB. The 64-point cross talk is better

than −20 dB (defined as the sum of the energy from all 63 undesired inputs compared with the desired input measured at the output, for the worst-case switch state, for the worst-case optical path-loss variation). This cross-talk performance reduces problems with coherent cross talk between the switch channels. Each input to the switch is illuminated by a different directly-modulated laser located up to 100 m away on the line card connected to that switch input. Measurements of the directly modulated burst mode laser sources indicate that they have sufficient spectral widening (of the order of 0.2 nm) to reduce the cross-talk effect at the receiver by ∼6 dB. This was measured by temperature tuning of one such modulated source that was adjusted to cross talk by means of a fiber coupler to another such source and by observation of the bit error rate (BER) at a receiver. As the interferer was swept through the desired source's wavelength, the level at which noticeable BER degradation occurred changed from approximately −11 dB (at the point of noncoherence) to approximately −17 dB (at the worst case). This would normally occur between two sources at approximately −24 dB if narrow spectrum sources and modulation techniques were used [3,4]. In addition, forward error correction was used to minimize the effect of a low BER on the packet loss rate (this is discussed below).

4. Switching-Fabric Alternatives and Fabric Architecture

Several fabric topologies were considered during the investigation stage of the project [5]. The following paragraphs discuss the issues considered at each point in the decision process.

Centralized versus distributed. The first decision was whether the fabric should have a centralized core or be distributed. There were multiple considerations that all pointed toward a centralized core. The first consideration was based on performance. Distributed switching fabrics grow dimensionally as the size of the fabric increases. This results in an increase in the number of switching elements (decision points) that a packet must traverse to cross the fabric as the size of the system grows. In addition, buffering is generally required between each switching element to hold packets that are temporally blocked. The result is increased delay and jitter for larger-sized systems than for smaller systems. A second consideration was the difficulty in providing a distributed fabric that is nonblocking under all traffic situations. With a distributed fabric it is difficult to ensure that the fabric bandwidth is used in a manner that guarantees nonblocking behavior, because each switching element makes routing decisions independent of the other switching elements. Finally, as a practical matter, the optical switch element has approximately 15 dB of power loss and requires the input side to have single-mode polarization-maintaining fiber, and the output side must have multimode fiber. These technology constraints prohibit the optical switching element from being cascaded. A distributed fabric requires cascaded switching elements, whereas a centralized fabric does not.

Single-stage versus multistage fabric. The second major decision was whether the fabric would consist of a single switching stage or multiple stages. This decision was fairly simple to make. A single switch stage would limit the size of the router to the size of the largest crossbar switch that can be built with a given technology. The initial optical switch element is limited in size to 64 inputs and 64 outputs, each with 10 Gbit/s of data bandwidth. The resulting router would be limited to 640 Gbit/s of input and output bandwidth. Over time, the size of a router could increase as larger optical switch elements were developed. However, an initial starting point of a maximum of 640 Gbit/s of bandwidth is insufficient for a product that is to address the needs of a tier-one carrier's core network. Because of this scalability limitation, single-stage fabrics were quickly eliminated from consideration.

Multistage fabric topology. The third decision to be made was the topology for the multistage fabric. Considerations included the physical attributes of the optical switch, the performance characteristics expected of an IP router, and cost trade-offs. Because of the physical size of the optical switch and amount of bandwidth that it can handle, the decision was made to use the optical switch as the second stage of a three-stage enhanced Clos switching fabric. The first and third stages are physically located in each line shelf on modules referred to as internal optics modules (IOMs). The technology for the first and third stages was chosen to be digital complimentary metal–oxide semiconductor (CMOS) because of the compact size of the resulting solution. The IOMs contain the optical transmitter and receiver that are used to deliver and receive data from the second-stage optical switch. The optical switch is located in a centrally located shelf that is independent of the line shelves. Optical cabling of up to 100 m is used to connect the first and third stages in each line shelf to the centrally located core shelf. Figure 1 shows a block diagram of the three-stage enhanced Clos fabric (the redundant fabric copy is not shown).

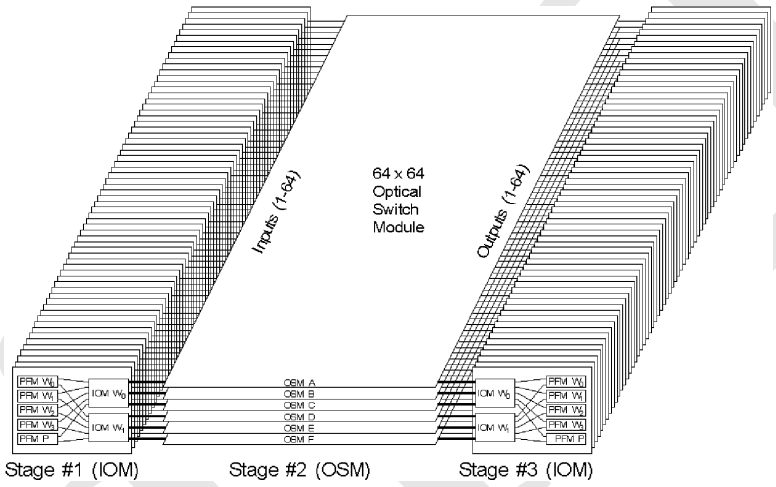

Fig. 1. Chiaro router switch fabric.

The enhanced Clos fabric uses a 5:6 expansion ratio to achieve the fabric speedup. This is required for achieving full line rate forwarding of IP packets simultaneously on all line interfaces. As shown in the diagram, the Clos fabric has been enhanced by means of providing two interfaces between the line cards (packet forwarding modules; PFMs) and the first and third fabric stages. This enhancement significantly increases the utilization of the second stage optical switch, at the cost of additional electrical connections between application-specific integrated circuits (ASICs) across a midplane.

Single point of control for entire fabric. A final decision was whether each of the fabric's three stages would have its own arbiter or whether to have a single arbiter for the entire fabric. This decision was coupled with whether buffering would be needed in the first and third stages. The decision was made to have a single point of control (central arbitration) for the entire three-stage fabric. The considerations used to make the decision were the desire to have low latency and jitter for high priority packets, optimizing throughput for the entire switch plane, and the availability of an efficient arbitration algorithm for the entire three-stage fabric. Simulations were performed on the basis of a centralized arbitration algorithm that resulted in very low latency and jitter for packets being routed through the system.

5. Other Consideration for Centralized Switch Fabrics

Determination of chunk size. A chunk period is defined to be the amount of time between optical switch configuration changes. This time includes the optical switch reconfiguration time plus the time a chunk payload passes through the switch. The chunk payload size is defined as the amount of information that is passed through the fabric between reconfiguration changes. A number of trade-offs were considered in setting the chunk size. The two main considerations were minimizing the loss of efficiency due to the optical switch reconfiguration time (30 ns) and maximizing the chunk payload utilization. The 30-ns reconfiguration time has less of an effect on fabric use the larger the chunk payload size becomes. However, the trade-off is that larger chunk payloads are more often partially filled. The chunk payload size was set to 400 Bytes. A chunk payload size of 400 Bytes plus the required chunk overhead is approximately 300 ns. This size is large enough that chunks pass through the optical switch approximately 90% of the time, and reconfiguration occurs only 10% of the time. Multiple packets are placed within a chunk's payload in order to minimize the number of partially filled chunks. Up to ten small packets can be packed into a single chunk payload. Minimal fabric speedup is required for allowing packets to be routed at line rate as a result of packets being packed in the chunk payload.

Fabric synchronization. One of the challenges of using a passive optical switch for the central stage of a fabric is synchronizing the arrival of chunks from all sources at the right point in time. Tighter synchronization results in a shorter chunk period and higher switch-fabric efficiency. Figure 2 shows a diagram of the fabric synchronization functionality.

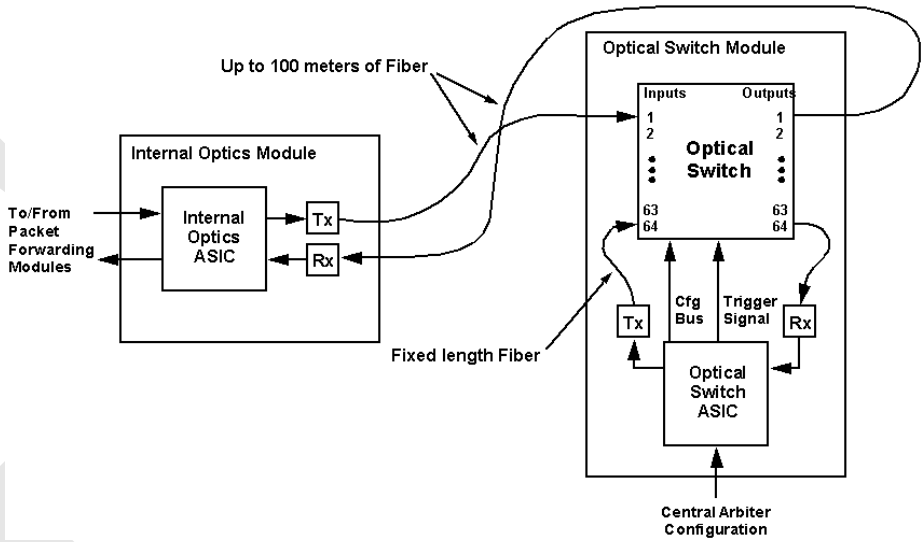

Fig. 2. Fabric synchronization functionality.

Synchronization is accomplished by use of the 64th input and 64th output ports of the optical switch to establish a chunk timing reference. The optical switch ASIC periodically sends an administrative chunk through the optical switch using the 64th input, with a target of the 64th output. The optical switch ASIC receives the administrative chunk and establishes a reference start-of-chunk. Received chunks from all IOMs are compared with the reference start-of-chunk. If the launch of a received chunk needs to be advanced or delayed, then the optical switch ASIC sends the amount of correction to the internal optics

ASIC with administrative chunks that are periodically sent from the optical switch ASIC to each internal optics ASIC. This mechanism is able to control the window in which all received chunks arrive to within 2 ns (less than 1% of a chunk period). The mechanism is independent of the length of the input fiber cables and compensates for component delay variations over time. The implementation allows input and output cables to range in length from 5 to 100 m in the same system. The packets are buffered electronically on the line cards, prior to being launched into the fabric; thus the inlet packet streams are decoupled from the internal system timing.

Burst-mode optics. A custom 12.5-Gbit/s burst-mode optical transceiver was developed to receive chunks with varying amplitude and bit phase alignment. Off-the-shelf transceivers are designed to receive an optical bit stream continuously from a single source transmitter. These transceivers rely on the knowledge that the optical power and bit phase alignment remain constant with only minor drift over time. The passive optical switch causes both optical power and bit phase alignment to shift dramatically at the beginning of each chunk period. The bit phase alignment can shift by $\pm 180°$ and the optical power can vary by up to 6 dB as a result of component variations (laser power, connector variation, cable length, and optical switch port variation). The custom transceiver uses standard transmitter circuitry with custom burst-mode receive functionality. The receive side relies on the presence of a preamble (alternating zeros and ones) on each chunk to establish the appropriate signal zero/one threshold and allow a fast-lock phase-locked loop (PLL) to acquire bit phase alignment in less than 200 bit times. The threshold circuit uses a fast sample-and-hold circuit to capture the center voltage level of the received preamble. The threshold is held for the duration of the chunk. The fast-lock PLL must be able to lock quickly on to the randomly aligned bit phase of the preamble and maintain lock through the duration of the chunk period with varying run lengths of ones and zeros in the chunk's payload. The fast- lock PLL functionality was accomplished by modification of an existing Vitesse semiconductor clock and data recovery device. Figure 3 below shows sample signal waveforms from the burst-mode receiver.

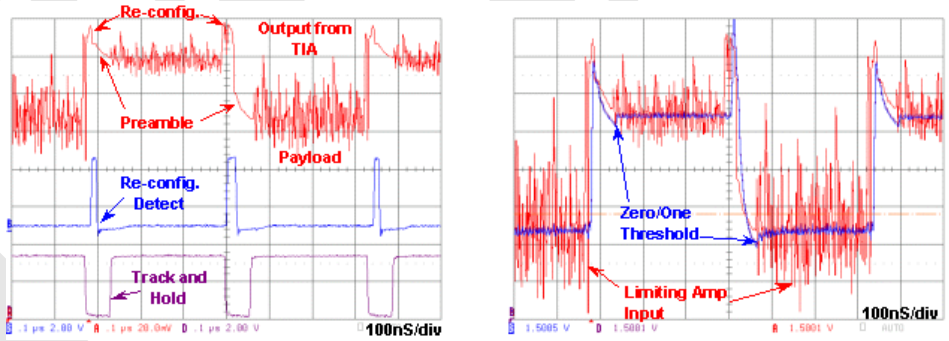

Fig. 3. Example burst receive signal with large chunk-to-chunk power variation.

The red waveform in Fig. 3 is the output of the transimpedance amplifier of the burst-mode receiver. Note that the waveform is inverted (higher values means less received light). During optical switch reconfiguration, minimal light is received. Increasing amounts of light is received as a new configuration settles. The track-and-hold circuitry tracks the increasing amount of light. The track-and-hold circuitry switches to hold mode after a fixed delay. The zero/one threshold from the track-and-hold circuit and the output of the transimpedance amplifier are fed into a limiting amplifier to produce the output digital bit

stream. An AGC approach to amplitude variation is described by Yamashita [6].

Arbitration and scheduling. The entire three-stage enhanced Clos fabric is controlled from a centralized arbitration controller [7, 8]. A proprietary algorithm is used to perform central arbitration. The algorithm is referred to as Clos parallel iterative matching (CPIM).

The central arbiter maintains a pool of requests for each line card. The CPIM algorithm performs multiple iterations to find as many matches as possible each chunk period. Once all iterations of a chunk period are complete, the new fabric configuration is sent directly to the optical switches and grants are sent to each line card indicating which requests were accepted. The line cards use the grant information to select packets to build chunks. The built chunks are then sent through the fabric.

A CPIM iteration consists of selecting up to two requests from each line card pool, submitting the selected requests to a fabric model, accepting requests that could be routed through the fabric model, and marking the resources for routed requests as unavailable for subsequent iterations. Up to two requests can be selected per line card pool each iteration due to the enhancement of the Clos fabric which provides each line card two inputs (and outputs) to the fabric. Routed requests of previous iterations mark line card inputs as unavailable for future iterations. These marked inputs cause fewer requests to be selected from each line card's request pool on subsequent iterations.

Each iteration's selected requests are sent to the fabric model. The fabric model includes the first-, second-, and third-stage crossbar switches and connections between the switches. When multiple requests require the same connection between crossbar switches, a sequence of decisions are used to make the selection. The selection criteria include the priority of the request and how long the request has been in the request pool. When all other selection criteria are equal, the selection decision is made by use of a pseudorandom number generator. At the end of each iteration, requests that have reached the output side of the fabric model are allowed to mark the connection resources that they used.

A single CPIM iteration is performed in four 155.52-MHz clock cycles, allowing 12 CPIM iterations per switch-fabric configuration. Figure 4 shows the sensitivity of the number of CPIM iterations per switch fabric configuration on the arbitration algorithm's efficiency.

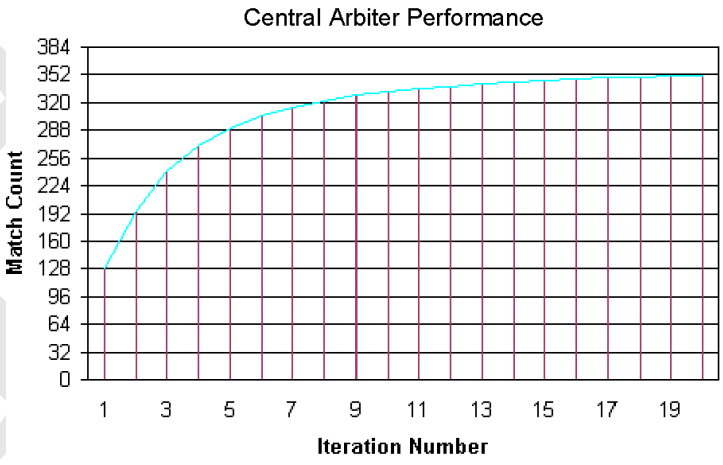

Fig. 4. Central arbiter performance.

As can be seen from Fig. 4, 12 iterations result in an average of 344 of the possible 384 paths through the second-stage fabric being utilized (89.6%). This utilization is sufficient to

allow full line rate routing for all 315 of the 10-Gbit/s router interfaces. The traffic pattern used to generate Fig. 4 is uniformly distributed from multiple inputs to multiple outputs.

Multistage delay from issued request to received grant. The minimum round-trip delay from the time a line card issues a request until the corresponding grant is received is eight chunk periods. This delay consists of the following components:

Line card request issue time	1	Chunk period
Fiber cable delay from line card to central arbiter	$1\frac{1}{2}$	Chunk periods
Central arbitration time	3	Chunk periods
Fiber cable delay from central arbiter grant to line card	$1\frac{1}{2}$	Chunk periods
Line card grant receive processing time	1	Chunk period

At any given time some number of requests will have been issued to the central arbiter from a line card without the corresponding grants having been received. The line-card hardware is designed not to exceed a maximum number of outstanding requests. The number must be large enough to cover the round-trip delay from the line card to the central arbiter and back, plus enough requests at the central arbiter to allow the CPIM algorithm to operate effectively. As described above, a subset of the requests in each line card's central arbiter request pool are selected for each iteration of the CPIM algorithm. Simulations have shown that the minimum number of requests in the central arbiter request pool for effective CPIM algorithm operation is four. A smaller number of requests available for the CPIM algorithm result in the algorithm finding significantly fewer paths through the fabric. A larger number of requests does not significantly increase the number of paths found. As described, the minimum number of outstanding requests from a line card is 12 (8 for request/grant round-trip delay plus 4 for the central arbiter request pool). The central arbiter has sufficient space to hold 16 requests per line card. With the maximum of 16 outstanding requests, there will typically be a total of 8 requests plus grants in flight and 8 requests at the central arbiter participating in the CPIM arbitration algorithm. The maximum cable length of 100 m between each line shelf and the central arbiter was set on the basis of the amount of round-trip delay between a line card and the central arbiter. Longer cable lengths would have increased the maximum number of outstanding requests required for maintaining high arbitration efficiency.

Multistage pipelining for fan out of centralized arbitration results. The results of the arbitration algorithm are used to instruct the source line cards of each constructed chunk's destination and are used to configure all three stages of the fabric. The fabric configuration is accomplished with two different communication paths. The second-stage optical switch receives its configuration information directly from the central arbiter. The first and third stages receive their configuration information in the header of each chunk that is routed through that stage (source routing). There are two communication paths that the arbitration information must travel in order to reach the line cards and the fabric's second stage. The first path is from the central arbiter to each line card. This information is in the form of grants that indicate that a previously issued request has made it through the central arbiter. The distance between the central arbiter and each line card can be up to 100 m. The second communication path is from the central arbiter directly to the second-stage fabric. This information consists of a list of second-stage optical switch input to output connections. The amount of time between completion of an arbitration cycle and applying the configuration to the second stage of the switch fabric is approximately seven chunk period pipeline stages. These pipeline stages are broken down as follows:

Central arbiter grant issue time	1	Chunk period
Fiber cable delay from central arbiter to line card	1½	Chunk periods
Line card chunk construction time	2	Chunk periods
First stage fabric time	1	Chunk periods
Fiber cable delay from line shelf to optical switch	1½	Chunk period

The seven pipeline stages affect the performance of the router as end-to-end latency but not as end-to-end jitter.

Ensuring high availability. Two goals for the router were to achieve high scalability and high availability. Previous experience has shown that these two do not go hand in hand without considerable forethought. High availability of the fabric was achieved by means of having two copies of the fabric and central arbitration. Both copies of the fabric are working copies. The source line shelf sends the same chunk information on both sides of the fabric simultaneously. Owing to the synchronous nature of the fabric, the duplicated chunk information passes through each fabric simultaneously. A bad chunk is detected at the destination line shelf by use of a cyclic redundancy check (CRC) code that is appended on the chunk's payload at the source line shelf. When a bad chunk is detected, it is logged. The CRC check of both received chunks is used to select which chunk to keep. Chunk selection based on CRC occurs each chunk period. Having duplicate fabrics allows a fabric module or cable to be pulled or fail without warning and not lose any traffic. In addition, duplicate fabrics allow system maintenance and upgrades to occur without affecting traffic. One fabric can be taken offline and upgraded or repaired without affecting the other fabric. However, associated packets may be lost if a corrupted chunk is received while the fabric is in simplex operation. The two fabrics are located in different racks to allow physical separation. A sprinkler system going off over one fabric rack would take down one of the fabrics but not both.

6. Overall Performance Measurements and Analysis

Numerous measurements of the performance have been made. One important measure is the latency of the router. Whereas average and maximum latency are dependent on the scheduling and traffic characteristics, the minimum latency is dependent on the switch fabric, pipeline delays, and synchronization. For an OC-192 interface, and small packet sizes (where the packetization delay is small), the latency at 100% loading was measured at 27.75 μs. This delay includes ingress to the router, removing packets from the OC-192 POS frame, IPv4 address lookup and next-hop resolution, scheduling and arbitration of the 315-port switch fabric, transit delay across the optical fabric and interconnecting fiber, reconversion back into a POS frame, and egress from the router at OC-192. The equivalent minimum latency for the same test, except when run with a gigabit Ethernet (GbE) interface was 23.18 μs. This is comparable with the minimum latency through an all-electronic IPv4 router of less than 16 ports total size, with OC-192 and GbE interfaces, respectively.

Figure 5 shows the minimum latency over a sweep of packet sizes from one input to one output. There are four packetization steps involved: one at the test equipment GbE source, one at the router GbE ingress receiver, one at the router GbE egress transmitter, and the fourth at the test equipment GbE receiver. At a packet size of 1500 Bytes, these four packetization delays account for 48 μs, accounting for the slope of the line in Fig. 5.

The router was also tested for single-faults in the optical switch plane. In this case faults were inserted into one of the two switch planes by various techniques, including removal of power to one of the planes, and pulling the interconnecting fiber to one of the switch planes. In both cases the router did not experience loss of packets. This is because of the

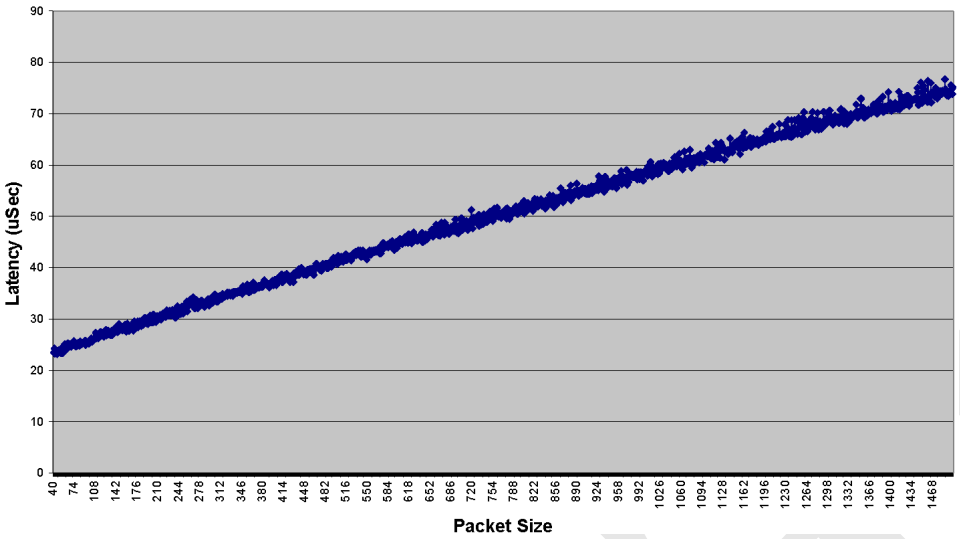

Fig. 5. Minimum latency versus packet size for optical router with 1-Gbit Ethernet interface, with four GbE packetize–depacketize steps.

$1+1$ redundancy employed: the ability of the receiving line shelf to listen to both switch fabrics and select, on a chunk-by-chunk basis, the best-quality data chunk. Failure of one path corrupted only the chunks in that data path.

The automatic chunk synchronization method was tested over a range of fiber lengths up to 100 m of intrasystem interconnect. The synchronization system was able to maintain approximately ± 2 ns of differential chunk arrival time at the inlet of the optical fabric, which allows the system to be deployed without the need to worry about cutting interconnecting cables to specific lengths.

7. Conclusion

We have demonstrated an optically switched IP router of very large scale that is fully compatible with existing packet networks and interfaces and that operates with line rate performance over scale. Tests of the router indicate that it meets the set performance objectives and that the optical fabric provides both the optical performance, architecture reliability, and large scale needed to build next-generation IP/MPLS switching and routing platforms for the core of data networks.

Lessons learned. It proved difficult to communicate the need for dc coupling of the burst-mode electronics components to suppliers. Many problems in the use of transmitters and receivers are simply solved with ac coupling. However, ac coupling can cause large voltage offsets when used in a burst-mode system, and readily available dc stable components were sometimes hard to come by; we had to design our own. Fortunately most test equipment was flexible enough to serve both purposes.

The optical switch has reasonable uniformity of path-to-path loss. However, we found that the optical fiber connectors used sometimes exhibited a wide variation in loss across a switch. The net combination of connector and switch variations required enhancing the ability of the burst-mode receiver to accommodate larger chunk-to-chunk power variations.

In addition, designing such a large router involved widely different engineering disciplines, including optics, device physics, system timing, digital ASICs, and complex embedded and system software. It took us several months before the design team was able to

effectively merge acronyms and communicate with a common engineering language across these disciplines.

Future research should be done in the area of higher-speed burst mode optics (40, 160 Gbit/s), reduction of the optical chunk-to-chunk power variation, and increases in the scale of the switch element itself.

References and Links

[1] E. Shekel, A. Feingold, Z. Fradkin, A. Geron, J. Levy, G. Matmon, D. Majer, E. Rafaely, M. Rudman, G. Tidhar, J. Vecht, and S. Ruschin, "64x64 fast optical switching module," *Optical Fiber Communication Conference (OFC 2002),* Vol. 70 of OSA Trends in Optics and Photonics Series (Optical Society of America, Washington, D.C., 2002), paper TuF3, pp. 27–29.

[2] E. Shekel, D. Majer, G. Matmon, A. Krauss, S. Ruschin, T. McDermott, M. Birk, and M. Boroditsky, "Broadband testing of a 64x64 nanosecond optical switch," presented at the 15th Annual Meeting of the IEEE Lasers and Electro-Optics Society (LEOS), Glasgow, Scotland, 10–14 November 2002, paper WA2.

[3] E. L. Goldstein, L. Eskildsen, and A. F. Elrefaie, "Performance implications of component crosstalk in transparent lightwave networks," IEEE Photon. Technol. Lett.**6,** 657–660 (1994).

[4] K. P. Ho, C. K. Chan, F. Tong, and L. K. Chen, "Exact analysis of homodyne crosstalk induced penalty in WDM networks," IEEE Photon. Technol. Lett. **10,** 457–458 (1998).

[5] H. J. Chao, "Next generation routers," Proc. IEEE **90**(9), 1518–1558 (2002).

[6] S. Yamashita, S. Ide, K. Mori, A. Hayakawa, N. Ueno, K. and Tanaka, "Novel cell-AGC technique for burst-mode CMOS preamplifier with wide dynamic range and high sensitivity for ATM-PON system," presented at the European Solid-State Circuits Conference, Villach, Austria, 18–22 September 2001, session 2.2, paper 3.

[7] N. McKeown, "Scheduling algorithms for input-queued cell switches," Ph.D. dissertation (University of California at Berkeley, Berkeley, Calif., 1995).

[8] M. McKeown and T. Anderson, "A quantitative comparison of scheduling algorithms for input-queued switches," Comput. Netw. ISDN Syst. **30,** 2309–2326 (1998).