An Optical Booster for Internet Routers

Joe Bannister, Joe Touch, Purushotham Kamath, and Aatash Patel

University of Southern California
Information Sciences Institute
Los Angeles, California, U.S.A.
{joseph, touch, pkamath}@isi.edu, aatashpa@usc.edu

Abstract. Although optical technologies have been effectively employed to increase the capacity of communication links, it has proven difficult to apply these technologies towards increasing the capacity of Internet routers, which implement the central forwarding and routing functions of the Internet. Motivated by the need for future routers that will forward packets among several high-speed links, this work considers the design of an Internet router that can forward packets from a terabit-per-second link without internal congestion or packet loss. The router consists of an optical booster integrated with a conventional (mostly electronic) Internet router. The optical booster processes Internet Protocol packets analogously to the hosting router, but it can avoid the time-consuming lookup function and keep packets in an entirely optical format. An optically boosted router is an inexpensive, straightforward upgrade that can be deployed readily in a backbone IP network, and provides optical processing throughput even when not deployed ubiquitously.

1 Introduction

Because of steady progress in the development of optical technology, the rate of data transfer over long distances has grown continuously and rapidly over the past several years. Data rates of 1 terabit/second (Tb/s) are attainable in the not-so-distant future. This paper describes the design of a future Internet router that forwards packets at the line rate of 1 Tb/s and can be built with optical and electronic components that will evolve easily from today's technology. The router is constructed by integrating an optical booster element with a conventional router.

With the advent of low-loss, dispersion-compensated optical fibers, high-quality lasers, optical amplifiers, and efficient modulation systems, link speeds have scaled rapidly upwards. A favored approach has been to apply wavelength-division multiplexing (WDM) on a single optical fiber to create multiple high-speed, independent channels that serve as virtual links between routers or other communication equipment. Often associated with WDM is the idea of an "all-optical network," which offers ingress-to-egress transport of data without intervening

Invited paper to appear in the 8th International Conference on High Performance Computing, Hyderabad, India, Dec. 17–20, 2001. This work was supported by the Defense Advanced Research Projects Agency.

optoelectronic conversions, promising reductions in end-to-end latency and data loss as well as improvements in throughput. Routers are sometimes enhanced by wavelength-selective optical crossconnect switches and label-switching software [1, 2, 3, 4], which maps flows of Internet Protocol (IP) packets to lightpaths of WDM channels created dynamically between specific packet-forwarding end-points. Label switching for optical networks is being developed by the Internet Engineering Task Force in its Multiprotocol Label/Lambda Switching (MPLS) [5] and (MPλS) [6] standards. Although MPLS/MPλS has proven valuable as a tool for traffic engineering, it has not made significant inroads as a method for dynamically managing label-switched paths. Label switching in a WDM optical network suffers from the disadvantage that it can be difficult to achieve acceptable channel utilization unless flows are sufficiently aggregated and can be mapped to the small set of WDM channels available in the network. Switching gains are achieved only for paths through contiguous sets of label-capable routers. Because WDM label-switched paths are scarce, they are used only for heavy flows. Detecting such a flow takes time, establishing a switched path requires several round-trip times, and deploying the flow on a label-switched path introduces further delay [7]. These combined latencies reduce the amount of time over which a flow can be switched.

To overcome the high latency of lightpath establishment in optical label switching, the alternative of optical burst switching [8, 9] has been considered as a way to exploit WDM. The first of a stream of packets destined for one address blazes a path through the network by setting up segments of the lightpath as it is being routed, the succeeding packets in the stream use the lightpath, and the final packet of the stream tears down the lightpath. Although this eliminates the wait for the signaling protocol to establish a lightpath, it also incurs the undesirable possibility that the burst will encounter a failure to complete the lightpath, because the search for a lightpath must be conducted without the benefit of backtracking. Streams of packets also must be buffered at the border routers so that the stream can be shaped by timing gaps between packet releases from the border router into the core network. Again, as with label switching, gains for burst switching are achieved only through contiguous paths of burst-capable routers.

Rather than continuing to force-fit connectionless IP onto these circuit-switching models, it is useful to consider routers that connect to other routers using a single high-speed channel on an optical fiber, i.e. each router pair may be connected by an unmultiplexed fiber. Router-to-router connections are largely provisioned on a link-by-link basis, e.g., SONET links and long-haul (gigabit) Ethernet links. This style of network deployment will likely persist into the foreseeable future, which implies that a network operator would derive greater benefit from the use of a conventional router that handles very-high-speed links than from a label- or burst-switching router that relies on dynamic mappings of IP flows to lower-speed WDM channels. The Internet performs best when the connectionless model at the network (IP) layer is replicated at the link layer; attempts to combine connectionless mode data transfer with connection-oriented data transfer have not been generally successful. A network of conventional (connectionless) routers is considerably easier to deploy, because these are the kind of routers the Internet already uses — the kind that network operators understand and appreciate.

This paper describes the architecture of an optically boosted Internet router that adds a fast, simple optical switch to the conventional router. The level of integration between these two components is moderate, so that the booster switches can be added to a network of conventional routers easily, modularly, and inexpensively. Minor redesign of router control interfaces would be necessary so that the boosting router and the base router can exchange control and routing information. A network may be upgraded router by router, rather than as a whole (as required by label and burst switching). A proper subset of router links may be boosted, at the discretion of the network operator. No new protocols are required, and interoperability is not affected in any way. The resulting enhanced router operates like the base router, except that the forwarding speed is increased substantially.

The remainder of this paper is organized into five sections. Section 2 discusses the challenges that router designers face in scaling up to Tb/s speeds. Section 3 presents the architecture of the optically boosted Internet router. Section 4 provides a simple analysis of the router's performance. Section 5 presents the results of a detailed simulation of the router, using real traffic traces taken from a production network. Section 6 offers conclusions to be drawn from this study.

2 Background

Consider a wide-area backbone network composed of IP routers connected by 1-Tb/s links. Clearly, the connecting links would be optical fibers. Although IP's soft-state properties and proven flexibility are strong attractions, IP forwarding is definitely expensive. Forwarding speed is key to the router's overall performance and is often its bottleneck. In a router with 1-Tb/s links, the worst-case scenario is to forward a continuous stream of minimum-size IPv4 packets, each consisting of 20 bytes (i.e., IP headers only). This scenario requires that a packet be forwarded every 160 picoseconds (ps), which equates to 6.25 billion packets per second per interface. Even if the interface were receiving a stream of back-to-back maximum-size packets of 1500 bytes each, then the forwarding rate would still be over 83 million packets per second per interface. These performance levels are considerably beyond the capabilities of today's routers, which today achieve about 10 million packets per second. The central task of forwarding is to look up the packet's destination IP address in the routing table and dispatch it to the correct next hop. Routers maintain a table of IP address prefixes, and the lookup of the address is actually a longest-prefix match among all routing-table entries and the packet's destination address. Given the rapid growth of routing tables in the core of the Internet, longest-prefix matching requires a significant — and growing — amount of time; to illustrate the magnitude of the problem, the size of a typical routing table in early 2001 was nearly 110,000 entries, up from about 70,000 entries at the start of 2000 [10].

Finding the longest matching prefix of an address every 160 ps is difficult, because it requires several memory accesses to be completed during the 160-ps interval. Routing-table memory is relatively large (needing to store 100,000 entries or more at core routers), and it is normally implemented with commodity, high-density random-access memory (RAM) parts, which have access times well above the subnanosecond times required by the worst-case routing scenario. Moreover, RAM access times have historically decreased at a slow pace; it is unlikely that high-density RAMs will be available with the required memory-access times in the near future.

The high cost of searching the routing table for the longest-prefix match with a destination IP address has motivated researchers to seek innovative algorithms and data structures to perform this search efficiently [11, 12, 13, 14]. These schemes have sought to exploit the fact that RAM cost and density are steadily improving. It is, however, difficult to extend these techniques to very high-speed links, because their performance is limited by RAM access times. Growth of the routing table, which is exacerbating lookup times, will undoubtedly continue. The joint growth in routing table size and link speed means that it will be difficult to keep lookup times low enough to handle packets at the line rate.

Concerns for the future of routing surround the planned transition from version 4 of IP (IPv4) to version 6 (IPv6). The 32-bit addresses of IPv4 are conceptually represented in the routing table by a binary tree of 2^{32} leaves, which must be matched to the lookup IP address and referred back to the smallest subtree that corresponds to an entry in the routing table (i.e. matched to the longest prefix of the address). The use of 128-bit addresses in IPv6 strains the lookup process even more, as there are, *prima facie*, 2^{128} distinct leaves in the lookup tree. Although the prefixes in the IPv6 routing tree are significantly shorter than 128 bits, and some schemes even route in two phases based on a 64-bit provider part initially, there is, nevertheless, still a heavy burden in searching the tree for prefixes that match the lookup address. If the Internet widely adopts IPv6 as the replacement for IPv4, then matters will become only grimmer for router design.

To summarize the current trends: (1) routing table sizes are growing rapidly because of the addition of new customers to the Internet; (2) for technological reasons the access times of the RAM in which routing tables are stored are decreasing very slowly; (3) as optical technology spreads, the link speeds and packet arrival rates are growing extremely rapidly; and (4) the gradual acceptance of longer IPv6 address formats is making routing tables larger and prefix matching more complicated. These trends make it more and more difficult to construct a cost-effective router that can operate at line speeds. The next section describes a router architecture designed to meet the challenges described above.

3 The Architecture of an Optically Boosted Router

The optically boosted Internet router is designed to enhance an existing conventional router by coupling with the base router to accelerate packet forwarding. If the base router has a number of high-speed optical (Tb/s) interfaces, then the boosting router is inserted into the path of these interfaces, as pictured in Fig. 1. Whereas the boosting router is an optical device that maintains incoming packets in an entirely optical format, the base router is normally a conventional router implemented in digital electronics. It is therefore possible to boost a router with heterogeneous links, because only a subset of the base router's links might need to be boosted. In this way the boosting router may be added to any router, enabling incremental deployment and upgrading of a population of routers — it is not necessary for all links of the router to have the same speed and signal formats.

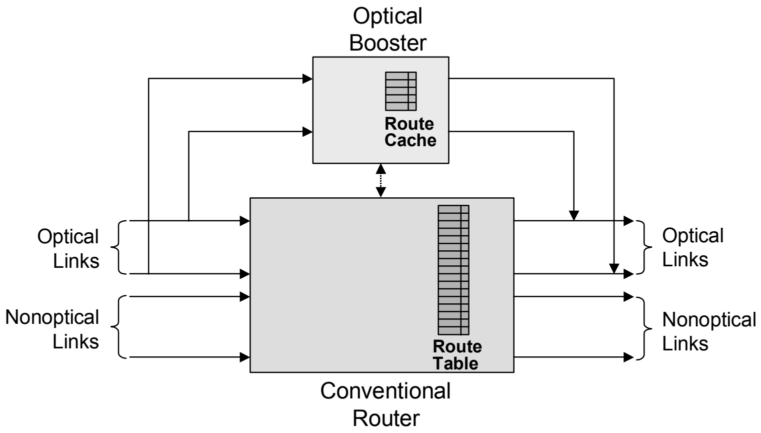

Fig. 1. A Base Router Enhanced by a Boosting Router

A central feature of the architecture is the use of routing caches in the boosting router. It is well established that destination IP addresses exhibit a high degree of temporal locality, often achieving hit rates above 90% with modest cache sizes and simple cache-replacement policies [15]. The routing cache reproduces entries of the full routing table. The cache is a commodity content-addressable memory (CAM), which looks up the destination address and returns the identity of the outgoing interface for that address, if it appears in the cache. If a packet's destination address does not appear in the cache, the boosting router ignores the packet and allows the base router to handle the packet. When addresses hit in the cache, their packets are passed to a totally nonblocking optical space switch that attempts delivery to the correct output link. The most-likely candidate for the optical switch is a $LiNbO_3$ device. Compared to RAMs, CAMs are relatively fast, but smaller and more expensive. Today's CAMs accommodate a few thousand entries and can achieve access times of about a nanosecond. Although it is impossible to predict the future reliably, CAMs and optical space switches of the required performance will likely be available in the next few years. Specifically, to route a stream of back-to-back 20-byte IPv4 packets, the boosting router requires CAMs with access times of less than 160 ps and a totally nonblocking optical space switch that can switch at the rate of more than 6 gigahertz (GHz).

Because it is implemented optically, the boosting router supports a much higher forwarding rate than the base router. The boosting router incorporates an optical space switch that switches packets among the inlets and outlets of the boosting router. If the routing-cache hit rate of the incoming packet stream is high, then a large fraction of the input packet stream is offered to the optical switch. A packet whose destination address is not found in the routing cache will be passed on to the base router, which will buffer and switch the packet to its next hop.

Whenever a set of packets is submitted simultaneously to the optical (or any) space switch, more than one packet might need to be switched to the same outlet. Such contention for outlets demands that all but one of the contending packets be "dropped" from the switch; in the boosting router these packets are then submitted for forwarding via the base router. This packet loss reduces the number of packets that pass through the boosting router. These packets, however, are not truly lost, since they get passed to the base router. Contention for outlets of the optical space switch causes packets to be forwarded through the slower base router. The reason that contending packets are diverted through the base router is that the optical boosting router cannot buffer packets, given that there are no true optical buffers in existence at this time.

A function diagram of the optically boosted Internet router is shown in Fig. 2. Note the absence of packet buffers in the optical subsystems; all packet buffers are in the electronic subsystems of the base router, and hold packets dropped by the boosting router after contention. Fig. 2 shows the base router functionality distributed among the booster's line cards and switching fabric.

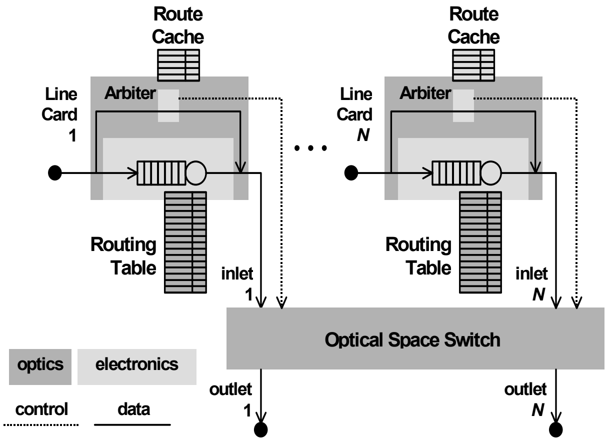

Fig. 2. The Architecture of the Optically Boosted Router

When a packet arrives on a link, it is replicated physically so that two identical copies are fed in parallel to the electronic and optical subsystems, until one of the copies is deleted or "killed" and the other is allowed to proceed and be forwarded to the next hop. First, the destination IP address is read by the booster, latched, and presented to the routing cache. This requires that the address be converted from its native optical format to an electrical format so that it can control the digital electronics of the booster; it means, also, that the initial segment of the packet header must be parsed at line rate to locate the destination IP address. (To enable line-rate parsing, parts of the packet header could be encoded at a lower bit rate than the main portion of the packet.) If the sought-after address is not in the cache, then the optical signal in the booster is deleted, and the copy of the packet in the base router is allowed to proceed through. If the address is found in the cache, then the booster's arbiter determines whether the desired output is in use or is sought by another input packet. If the arbiter finds no contention, the switch fabric is configured so that the packet's inlet and outlet are physically bridged; the packet is effectively forwarded to its next hop without optoelectronic conversion; the copy of the packet, however, must be deleted from the base router. If the booster's arbiter detects contention for the output, then all but one of the contending packets must be deleted from the booster and allowed to survive in the base router, where they will be forwarded to their next-hop routers.

The system incorporates some subtleties in how packets are handled and the switch controlled. To choreograph the movement of packets and to coordinate the control actions require very accurate and precise timing. Cache lookup and switch control must be completed in less than the critical minimum-size–packet time of 160 ps, so that a succeeding packet will be served without competing for these resources. Although the degree of integration between the base and boosting routers is low, there is a coupling so that the two subsystems can coordinate packet-deletion signals with each other as well as the exchange of routing-table information. Although this does not require a redesign of the base router, it would be necessary to introduce minimal modifications into the router that allow for routing-cache updates and packet deletions. Because packets from both subsystems merge onto the outgoing links of the router, the control of the base router is modified to defer to the boosting router when transmitting a packet, because — unlike the booster — the base router has buffering in which to defer packets.

It should be mentioned that not all addresses in the full routing table might be cached. For example, only destination addresses that use an outgoing link to which the booster is connected may appear in the cache (note in Fig. 1 that not all the base router's links need to be served by the booster, e.g. those that are not optical). It is difficult to cope with multicast addresses. Addresses that correspond to the router itself, such as an endpoint of an IP tunnel, should not appear either.

Internet routers modify packet headers, in addition to forwarding them. All Internet routers are required to decrement the IP time-to-live (TTL) field by at least one. IPv4 routers are further required to update the header checksum as well. Both of these operations present challenges to an optical booster, because they affect all packets processed by the router. Header option fields, though rarely used, require more complex processing, but those are likely to be processed outside the "fast-path"; even in conventional routers packets with options are relegated to slow-path processing in a separate processor. Fragmentation is to be avoided, because of the additional computational complexity it would require. The booster, however, should never face a fragmentation decision, because its input and output links are homogeneous and have the same maximum frame size.

The TTL can be decremented optically by zeroing the least-significant nonzero bit (the TTL needs to decrease by at least one). Alternatively, the TTL can be used as an index into a 256-entry, fast, read-only memory that stores the hard-wired ones' complement value of k–1 in location k, or copied to electronics and decremented by conventional means. The IPv4 checksum must be recomputed electronically; this is simple when only the modified TTL need be incorporated. IPv6 omits the header checksum, so it is simpler to process. All variations of these operations require that the electronics be pipelined and parallelized sufficiently to accommodate the header arrival rate, and also require optical bit, byte, or word replacement of header data. None of these is an insurmountable engineering challenge.

4 Performance Analysis

The optical booster increases throughput by sending some of the router's incoming packets through a low-latency optical booster capable of forwarding at very high speed. A certain portion of the incoming packets are not able to pass through the booster, because either their destination addresses are not in the fast-lookup routing cache or they experience contention at an outlet of the optical switch. Cache hit rate and output-port–contention probability are thus critical performance parameters. The higher the hit rate and the lower the contention probability, the better will be the boosted router's performance. The hit rate is determined by the traffic characteristics, the routing cache size, and the replacement policy for newly accessed addresses. The contention probability is influenced by the switch size and structure.

To gain insight into the performance of the booster, consider a simple model of the optical switch in the worst-case scenario mentioned above: each of the switch's N links has an average load of λ minimum-size packets per 160-ps timeslot (where $0 \le \lambda \le 1$). Equivalently, λ is the probability of a minimum-size packet arriving on an incoming link during a timeslot. Assume that packets are generated independently of each other, and that the next hop of a packet is chosen randomly from among the N outlets. For this analysis, suppose that all packet arrivals are synchronized to the timeslot. λ is the rate of arrival after the cache hit rate has been factored in, i.e., if γ is the arrival rate to an incoming link of the router and h is the hit rate, then $\lambda = h\gamma$ packets per timeslot.

Of the packets whose destination addresses hit in the cache, only a fraction of them will pass through the switch, on account of the contention that causes some of these packets to be deleted from the booster and sent back through the base router. Given the traffic load of λ packets per timeslot per booster inlet, let η be the switch throughput efficiency, i.e., the fraction of offered packets that pass successfully through the optical switch. The throughput efficiency of the booster switch fabric is [16]

$$\eta = \frac{1-(1-\lambda/N)^N}{\lambda}$$

which has a limiting value of $1-1/e \approx 0.632$ as N approaches ∞ and λ approaches 1. Thus, in a large, fully loaded router with a hit rate of 100%, about 37% of the packets presented to the switch will be turned away from the optical subsystem for handling by the base router. Another way of looking at this is to observe that 63% of a stream of back-to-back minimum-size IP packets presented to the router will take the all-optical booster path; this path has no loss and virtually no latency. The load on the base router is reduced dramatically.

Increasing port contention as the traffic load increases forces the value of η down to a fairly low level. A tactic to reduce contention is to allow the packets contending for the same outlet to choose a second, alternate outlet from the routing cache. These alternate outlets might indeed lead to longer routes than would the primary outlet, but the cost is often less than sending the packet through the base router. The risk of going through the base router, especially under high loads, is that congestion in the buffers will cause packets to be dropped, which carries a high penalty for many Internet end-to-end protocols.

Next assume that the base router is able to determine (and communicate to the cache) both a primary and an alternate route for every address in its routing table. Although routing protocols seldom compute alternate routes to a destination, it is straightforward to do so. The process of switching now becomes a two-stage event. In the first stage arriving packets are switched to their outlets. Some packets will have sole access to the outlet and some will contend with other packets for an outlet. Once the arbiter determines exactly which of the contending packets is granted access to the outlet and which of them is to be turned away, the second stage may begin. If the mean number of packets offered to the switch is k during the first stage, then the average number of packets granted access at the end of the first stage is $k_1 = \eta k$. Remaining packets must now be submitted to their alternate routes. After submission, some of the packets might actually find themselves tentatively switched to those outlets that already have been granted to the k_1 successful packets of stage 1; these packets are immediately deleted from the booster. In stage 2 the surviving k_4 packets are sent to their alternate, open outlets, and some of these packets might experience contention with each other. The contention is resolved just as it was in stage 1, but the new switch size is N' and the new traffic load is λ'. After resolving contention there are $k_5 = \eta' k_4$ packets that emerge unscathed from stage 2. The net throughput of k_6 packets consists of the successful packets from both stages, and the overall throughput efficiency is $\eta *$. The derivation outlined above is depicted in Fig. 3, and the equations for the two stages are given below.

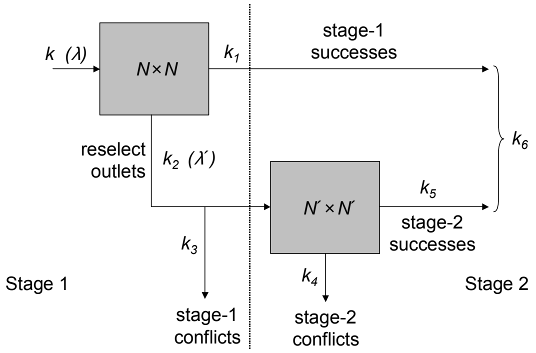

Fig. 3. Two-Stage Process Used in the Derivation of Throughput Efficiency for Dual-Entry Caches

$k = N\lambda =$ packets submitted to switch in stage 1
$\alpha \equiv (1 - \lambda/N)$
$\eta = \left(1 - \alpha^N\right)/\lambda$ = efficiency of stage-1 switch
$k_1 = \eta k =$ successful packets in stage 1
$k_2 = k - k_1 =$ packets resubmitted in stage
$k_3 = (k_1/N)k_2 =$ immediately failed reroutes
$k_4 = k_2 - k_3 =$ packets resubmitted in stage 2
$N' = N - k_1 =$ free outlets in stage 2
$\lambda' = k_4/N' =$ load on stage-2 switch
$\beta \equiv (1 - \lambda'/N')$
$\eta' = \left(1 - \beta^{N'}\right)/\lambda' =$ efficiency of stage-2 switch
$k_5 = \eta' k_4 =$ successful packets in stage 2
$k_6 = k_1 + k_5 =$ successful packets in stages 1 and 2
$\eta^* = k_6/k =$ overall switch efficiency

The equations above may be simplified to yield an expression for the overall switch efficiency in terms of the given parameters λ and *N*. Combining and simplifying the equations, one gets the result

$$\eta^* = \eta + \frac{\alpha^N}{\lambda}\left\{1 - \left[1 - \frac{\delta\lambda(1-\eta\lambda)}{N\alpha^{2N}}\right]^{N\alpha^N}\right\}$$

where $\delta \equiv 1 - \eta$. It is instructive to consider the limits of performance as the load and switch size grow. Taking the limit of η^* as λ approaches 1 and *N* approaches ∞, one gets

$$\begin{aligned}\eta^*_{\lim} &= \lim_{\lambda\to 1, N\to\infty} \eta^* \\ &= \lim_{\substack{\lambda\to 1 \\ N\to\infty}} \eta + \frac{\alpha^N}{\lambda}\left\{1 - \left[1 - \frac{\delta\lambda(1-\eta\lambda)}{N\alpha^{2N}}\right]^{N\alpha^N}\right\} \\ &= 1 - \frac{1}{e^{1+1/e}} \approx 0.745\end{aligned}$$

Shown in Figs. 4 through 6 are comparisons of the theoretical and simulated values of throughput efficiency η (and η^*) vs. traffic load λ for different switch sizes *N*. In each graph one set of curves represents the results obtained when the routing cache has only one next-hop interface per destination address, and the second set of curves represents the results obtained when the routing cache has both a primary and an alternate next-hop interface per destination address. We assume that the primary and alternate next-hop interfaces are chosen uniformly from among the available outlets. In each of the graphs the efficiency for dual-entry caches exceeds single-entry caches by a comfortable margin. The asymptotic efficiency for single-entry caches is 63%, while for dual-entry caches it is 74%. The real difference may be seen, however, in the degree of roll-off in the curves: for example, in Fig. 6 (16 interfaces) the efficiency drops below 95% at a load of 0.1 packets per timeslot per interface with a single-entry cache, but the efficiency with a dual-entry cache drops below 95% only after a load of 0.4 packets per timeslot per interface. Thus, for a given level of performance, a booster with a dual-entry cache can carry four times as much traffic as one with a single-entry cache. Given the modest cost of extending the cache to hold two entries, this is clearly a winning strategy to reduce outlet contention in the booster. One notes that the analytical model produces reasonably accurate results compared to the simulator, especially as the size of the switch grows.

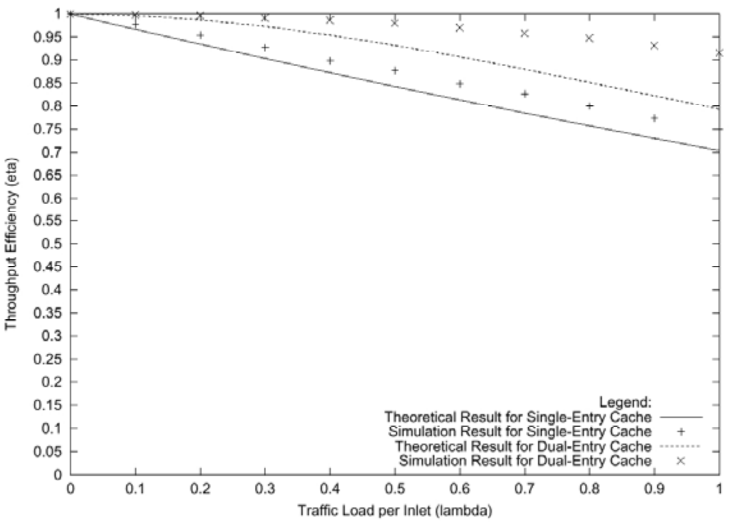

Fig. 4. Comparison of Theoretical and Simulation Results for a Booster with 3 Links.

5 Simulation Studies

Although the analytical results of the previous section are useful in characterizing the gross behavior of the optically boosted router, one cannot be sure how the system performs under real conditions unless a more-detailed and higher-fidelity model or prototype is developed and exercised. In this section the results from such a detailed simulation model driven by traffic traces collected from the Internet are presented.

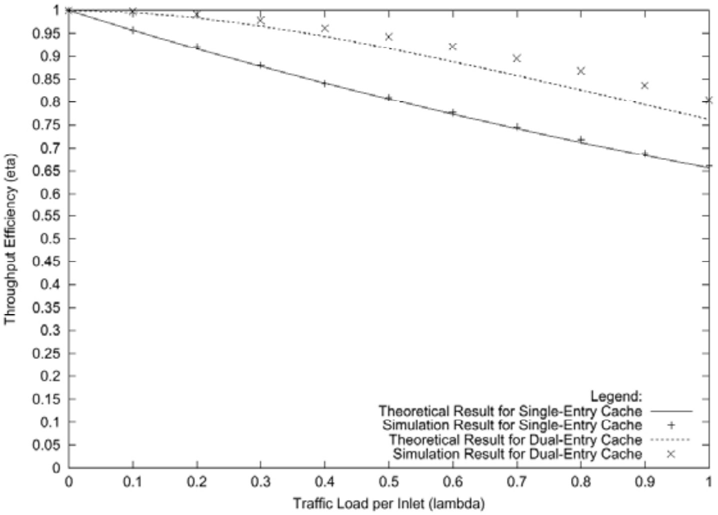

Fig. 5. Comparison of Theoretical and Simulation Results for a Booster with 8 Links

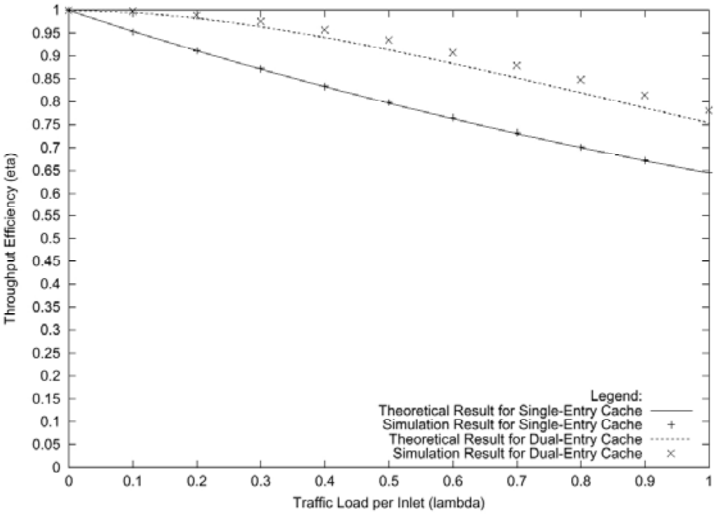

Fig. 6. Comparison of Theoretical and Simulation Results for a Booster with 16 Links

The simulator models the base and boosting routers at the packet level. The packets are copied to both the electronic and optical subsystems, and lookups of their destination address are performed. If the routing cache misses, then the packet is deleted from the optical subsystem and the packet proceeds only through the base router. The base router normally queues the incoming packet until lookup in the full routing table is completed. The base router has a 10-megabyte packet buffer per interface. The routing cache can hold 100 addresses. The simulator uses 160 ps as the lookup time in the routing cache and 200 ns for the lookup time in the full routing table. Given today's technology, a lookup time of 200 ns is conservative for the base router, corresponding to a packet-forwarding rate of 5 million packets per second. In principle, such lookup times could be achieved by using a fast RAM large enough to contain all IPv4 address prefixes of 24 bits or fewer (16 megabytes of RAM with better than 200-ns cycle time) [13]. After address lookup the base router switches the packet to the appropriate output link; the simulation assigns a negligible delay to packet processing and switching. It is assumed that no slow-path options processing occurs. As described above, packets emerging from a booster outlet are granted higher priority over packets emerging from the base router's corresponding output link.

Providing real traffic to drive the simulator is not a trivial matter, given that there are no links today that operate at Tb/s speeds. Thus, all traces collected have inherent bit rates far less than 1 Tb/s. It is therefore necessary to scale the empirical traffic in a way that the apparent bit rate approaches 1 Tb/s. This is accomplished by upscaling times in the router rather than downscaling times in the packet stream. In this way, the relative timings of packets and router are such that the link utilization is moderate to high. The simulator does not implement a routing protocol. Instead it is assumed that all traffic is uniformly destined to one of the router's output links. Two different schemes for updating the routing cache are evaluated, first-in first-out (FIFO) and random replacement. In the FIFO scheme a missed entry is read from the full routing table into the routing cache such that the oldest entry in the cache is replaced. In the random scheme, the missed entry is placed into a random location of the cache. The packet traces are from the University of Auckland's Internet uplink (between New Zealand and the United States) and are provided by NLANR MOAT and the WAND research group [17]. Several traces, taken from November 1999 to June 2000, are used to feed the individual input ports of the router, and routing is assumed to be uniform among the *N* output ports (actually *N*–1 ports, since packets are never looped back).

The averaged results of four groups of simulation runs are displayed in Table 1. For each value of *N* (3, 8, and 16) and cache-replacement scheme (FIFO and random), the mean values of switching gain (defined as the fraction of traffic that passes completely through the booster), hit rate, and latency are measured. The latency measurements are for the overall packet delay (from first bit in to last bit out) through the entire router and for the packet delay through just the booster; the astonishing variance between them covers four orders of magnitude. The results further validate the design by demonstrating that high gain and hit rates are achieved under realistic traffic conditions. The simulated switching gain is somewhat lower than the analytical results, because there is considerably more burstiness in the real traffic than in the artificial traffic of the theoretical model.

Table 1. Average Performance with Scaled Packet Traces

Number of Links	Cache Replacement Algorithm	Switching Gain	Cache Hit Rate	Booster Latency (ns)	Total Latency (us)
3	FIFO	0.89	0.95	3.23	41.02
3	Random	0.89	0.94	3.23	51.02
8	FIFO	0.87	0.95	3.34	39.53
8	Random	0.87	0.94	3.35	50.32
16	FIFO	0.87	0.95	3.39	44.96
16	Random	0.87	0.94	3.39	55.99

6 Conclusion

The optically boosted router is an approach to leveraging optics more effectively for packet switching. The simple ideas of a fast electronic cache and an all-optical path through the router come together to enable the implementation of an Internet router that could operate with high port counts and 1-Tb/s links, very low packet loss, and negligible latency. Analytical and detailed simulation models indicate that an optically boosted router under moderately heavy traffic can forward close to 90% of its packets through the all-optical switch path. The latency through the all-optical path is gauged in nanoseconds.

However, further studies of the optically boosted router need to be conducted, as important questions remain to be answered. The traffic model used is inadequate in two ways: (1) it must be scaled to mimic high data rates and (2) the routing is arbitrarily chosen to be uniform. The reason for the first shortcoming is that there appear to be no packet traces in the very high-speed regimes. The reason for the second shortcoming is that it is exceptionally difficult to collect packet traces along with associated routing-table information from commodity Internet providers. Our future approach will be to generate statistically aggregated packet traces that mimic real high-speed traffic and assign addresses that link back to actual routing tables.

The use of an electronic routing cache in the booster could eventually be replaced by an optical function to improve further the performance of the router. Although optical header recognition and switching are feasible [18], the capabilities of present implementations are limited. Similarly, dual-entry routing caches used for contention resolution might one day be supplanted by optical contention-resolution subsystems [19]. These technologies are maturing and might soon prove useful in the booster.

-

References

1. Y. Rekhter *et al.*, "Cisco Systems' Tag Switching Architecture Overview," IETF RFC 2105, Feb. 1997.
2. P. Newman *et al.*, "IP Switching — ATM Under IP," *IEEE/ACM Trans. Networking*, vol. 6, no. 2, pp. 117–129, Apr. 1998.
3. D. Blumenthal *et al.*, "WDM Optical IP Tag Switching with Packet-Rate Wavelength Conversion and Subcarrier Multiplexed Addressing*," Proc. OFC '99*, pp. 162–164, San Diego, Feb. 1999.
4. J. Bannister, J. Touch, A. Willner, and S. Suryaputra, "How Many Wavelengths Do We Really Need? A Study of the Performance Limits of Packet Over Wavelengths," *SPIE/Baltzer Optical Networks*, vol. 1, num. 2, pp. 1–12, Feb. 2000.
5. E. Rosen, A. Viswanathan, and R. Callon, "Multiprotocol Label Switching Architecture," IETF RFC 3031, Jan. 2001.
6. D. Awduche, Y. Rekhter, J. Drake, and R. Coltun, "Multi-Protocol Lambda Switching: Combining MPLS Traffic Engineering Control with Optical Crossconnects," IETF Internet Draft, Mar. 2001.
7. S. Suryaputra, J. Touch, and J. Bannister "Simple Wavelength Assignment Protocol," *Proc. SPIE Photonics East 2000*, vol. 4213, pp. 220–233, Boston, Nov. 2000.
8. J. Turner, "Terabit Burst Switching," Tech. Rep. WUCS-9817, Washington Univ., Dept. of Comp. Sci., July 1998.
9. C. Qiao and M. Yoo, "Optical Burst Switching — A New Paradigm for an Optical Internet," *J. High Speed Networks*, vol. 8, no. 1, pp. 69–84, 1999.
10. "AS1221 BGP Table Data," http://www.telstra.net/ops/bgp/bgp-active.html, Aug. 2000.
11. M. Degernark *et al.*, "Small Forwarding Tables for Fast Routing Lookups," *Proc. ACM Sigcomm '97*, pp. 3–14, Cannes, Sept. 1997.
12. M. Waldvogel *et al.*, "Scalable High Speed IP Lookups," *Proc. ACM Sigcomm '97*, pp. 25–36, Cannes, Sept. 1997.
13. P. Gupta, S. Lin, and N. McKeown, "Routing Lookups in Hardware at Memory Access Speeds," *Proc. IEEE INFOCOM '98*, pp. 1240–1247, San Francisco, Apr. 1998.
14. B. Lampson, V. Srinivasan, and G. Varghese, "IP Lookup Using Multiway and Multicolumn Binary Search," *Proc. IEEE INFOCOM '98*, pp. 1248–1256, San Francisco, Apr. 1998.
15. B. Talbot, T. Sherwood, and B. Lin, "IP Caching for Terabit Speed Routers," *Proc. IEEE GlobeCom '99*, pp. 1565–1569, Rio de Janeiro, Dec. 1999.
16. M. Karol, M. Hluchyj, and S. Morgan, "Input Versus Output Queueing on a Space-Division Packet Switch," *IEEE Trans. Commun.*, vol. COM-35, no. 12, pp. 1347–1356, Dec. 1987.
17. "Auckland Internet Packet Traces," http://moat.nlanr.net/Traces/Kiwitraces/auck2.html, Jan. 2001.
18. M. Cardakli *et al.*, "All-Optical Packet Header Recognition and Switching in a Reconfigurable Network Using Fiber Bragg Gratings for Time-to-Wavelength Mapping and Decoding," *Proc. OFC '99*, San Diego, Feb. 1999.
19. W. Shieh, E. Park, and A. Willner, "Demonstration of Output-Port Contention Resolution in a WDM Switching Node Based on All-Optical Wavelength Shifting and Subcarrier-Multiplexed Routing Control Headers," *IEEE Photonics Tech. Lett.*, vol. 9, pp. 1023–1025, 1997.