As appeared in *Protocols for High-Speed Networks VI*, ed. J. Touch, J. Sterbenz, Kluwer, 1999, pp. 43-60.

HOW MANY WAVELENGTHS DO WE REALLY NEED IN AN INTERNET OPTICAL BACKBONE?

Joe Bannister, Joe Touch, Alan Willner
University of Southern California/ISI and Dept. of EE-Systems
joseph@isi.edu, touch@isi.edu, willner@solar.usc.edu

Stephen Suryaputra
Nortel Networks
ssuryapu@nortelnetworks.com

ABSTRACT

Coupling Internet protocol (IP) routers with wavelength-selective optical crossconnects makes it possible to support existing Internet infrastructur in a wavelength-division multiplexing optical network. Because optical wavelength routing is transparent to IP, very high throughput and low delay can be achieved when packets are made to bypass the IP forwarding process by being switched directly through the optical cross-connect. One version of this approach is called packets over wavelengths (POW). This paper presents the POW architecture in detail and discusses its salient features. Realistic simulations of the POW that use actual packet traces in a well-known Internet backbone network reveal the level of performance that can be expected from POW under various options. Specifically, the fraction of packets that are switched through the crossconnect is evaluated as a function of the number wavelengths and the degree of flow aggregation that can be achieved. The resulting analysis, conducted in the context of the very-high bandwidt network service (vBNS) Internet backbone, suggests that as few as four wavelengths combined with a high degree of traffic aggregation can carry more than 98% of IP packets in the streamlined switched mode. In cases where it is not possible to aggregate traffic, the deployment of wavelength-merging technology would increase the fraction of IP packets carried in streamlined switched mode by up to 52%.

1. INTRODUCTION

The deployment of wavelength-division multiplexing (WDM) links has begun [7], and it is highly desirable to use these links to interconnect the routers that comprise the global Internet. Consider a network architecture — called packets over wavelengths (POW) and described in full below — in which packets can be forwarded by both Internet protocol (IP) routers and optical crossconnect switches. The goal of such an architecture is to switch as much traffic as possible directly by means of optical crossconnect switches, because IP forwarding is relatively expensive by comparison. Wavelength routing through an optical crossconnect switch is limited by the fact that only a few (four to 64) WDM channels per link are supported by today's commodity technology. This paper characterizes the expected performance of POW in such a sparse-WDM environment. Different options are examined for recognizing which packets should be switched through an optical crossconnect switch and which packets should be forwarded by an IP router. Simulations determine the level of WDM needed to carry a substantial fraction of packets in a switched (rather than a routed) mode.

POW shares features with IP switching [14], tag switching [16], and multiprotocol label switching [6], all of which are henceforth referred to by the vendor-neutral term "label switching." Label switching is used when an IP router includes a switching fabric that can be used to bypass IP forwarding. Because switching speeds are much greater than forwarding speeds (stimated by some [14, 12] to be 20 times greater for comparably priced hardware), the goal is to place as large a fraction of packets as possible on th streamlined switched path and to leave as small a fraction of packets as possible on the slower forwarded path. This feat requires some above-average intelligence in the switch–router. The router must have software that recognizes that a flow of packets can be passed through the switching fabric. A signaling protocol then notifies switches that the recognized flow should be carried over a switched path rather than a routed path. Eventually a hop-by-hop sequence of switches carries the flow of packets from one router to another. WDM equipment is on the verge of deployment in the Internet, and there are a number of projects to evaluate and implement label switching or burst switching in WDM networks [3, 15, 19], so it is crucial to fully understand fully their engineering tradeoffs.

This paper examines whether optical label switching is feasible and beneficial in the near-to-medium term. This necessitates investigating th behavior of real Internet traffic in an optical label-switching backbone with a limited number of wavelengths. It also requires the evaluation of the prformance improvement achieved by schemes that aggregate traffic to increase the utilization of WDM channels.

The remainder of this paper is organized into four sections. Section 2 describes the POW architecture and principles of operation. Section 3 prsents analytical results to characterize the overall gain that could be expected from the introduction of WDM. Section 4 provides the details of the simulation, the traffic model, and the experiments used to evaluate POW's performance. Section 5 presents the results of the evaluation. Section 6 offers conclusions to be drawn from the study.

2. POW ARCHITECTURE

A starting point for this work is to consider a wide-area backbone network that would be based upon advanced optical technology. In today's Internet a user's organization (business concern, educational institute, government agency, etc.) operates an enterprise network that attaches to an Internet service provider (ISP). A packet going from one customer to another then traverses the sending customer's enterprise network, one or more ISPs, and — finally — the receiving customer's enterprise network. More frequently the user's ISP provides wide-area transit of packets over its own backbone network; this ISP will typically hand off the packet to the receiving customer's ISP (also likely to be a wide-area backbone operator). A packe thus suffers a significant part of its IP-forwarding hops in the backbone network. It is not uncommon for a packet that travels coast-to-coast across North America to experience more than a dozen IP-forwarding hops. IP forwarding is expensive, because it is normally a software-controlled process. The dominant costs of forwarding come from matching the packet's destination IP address prefix to an entry in a routing table and accessing the packet's next hop from the table, which in a backbone today can exceed 60,000 entries. Although promising techniques for rapid lookup of addresses have been proposed [4, 11, 20] and are under consideration by router manufacturers, they have not been demonstrated widely in actual networks. Even if fast lookup is employed, there is still a significant store-and-forward delay associated with each hop when the forwarding path is used; this store-and-forward penalty is avoided in the switched mode, because switching is normally cut-through, allowing the head of the packet to exit the switch even before its tail has entered. One subsequent goal is to reduce the number of hops incurred by a packet while traveling through a large backbone network.

The introduction of WDM into the telecommunications network offers ISPs the opportunity to achieve greater performance and to scale their networks in speed and size. Consider an ISP-operated backbone that consists of routers connected by optical fibers that support WDM. Further assume that wavelength-selective optical crossconnect switches are available to channel wavelengths from incoming optical fibers to outgoing fibers [17]. A functional depiction of a wavelength-selective optical crossconnect switch (also known as a wavelength router) is shown in Fig. 1. This switch is a wavelength-selective optical crossconnect device that is capable of routing a specific wavelength of an incoming fiber to an outgoing fiber. The path is entirely optical and free from buffering or other delays. The wavelength routings are independent of each other, so that wavelength 1 arriving fro incoming fiber 1 may be switched to outgoing fiber 1, while wavelength 2 arriving from incoming fiber 1 may be switched independently and simultaneously to outgoing fiber 2. The optical crossconnect switch is not a rapidly switching device; it is configured on time scales of microseconds to milliseconds and typically is left in a specific configuration for an extended period of time (e.g. the lifetime of an IP flow, typically tens to hundreds of seconds).

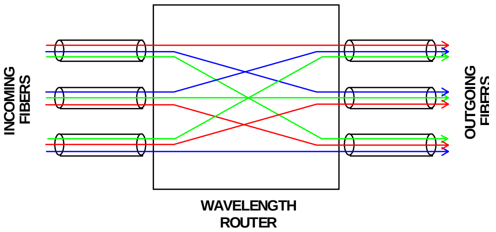


Figure 1 Optical Crossconnect Switch

The combination of an optical crossconnect switch and an IP router is employed in the POW switch–router to implement a node that is able to reassign an IP flow from the IP-forwarding process directly to a wavelength. An IP flow is can be defined as a sequence of packets that travel together along a subset of the same route in the network before exiting the network. This is a generalization of the more-common, narrow definition which identifies a flow as a sequence of packets with the same source and destination IP addresses and transport port numbers. POW's definition can focus on aggregated flows of greater intensity than narrowly defined flows.

By default, all packets flow initially through an IP router, which runs a process that detects and classifies flows of sufficient intensity and duration to merit fast-path switching. Each incoming fiber uses a special wavelength for the default traffic. When a flow is recognized, the router's control softwar attempts to shunt the flow straight through on its own wavelength. Shunting requires that the optical crossconnect switch be configured to support a wavelength that is routed from the flow's incoming fiber to its outgoing fiber. Suppose that a strong flow (call it flow 9) has been detected coming in on the default wavelength of fiber 1 and exiting on the default wavelength of fiber 3. The control software would seek to identify an unused wavelength on both incoming fiber 1 and outgoing fiber 3. If wavelength 2 is unused on both these fibers, then the router would signal the upstream router on the other end of incoming fiber 1 that it should bind all flow-9 packets to wavelength 2 going out on fiber 1. Similar actions are coordinated with the downstream router at the other of fiber 3 that flow-9 packets will be coming in over wavelength 2. In this way flow 9 will be carried from its ingress router to its egress router in the network. This sequence of steps is shown in Fig. 2.

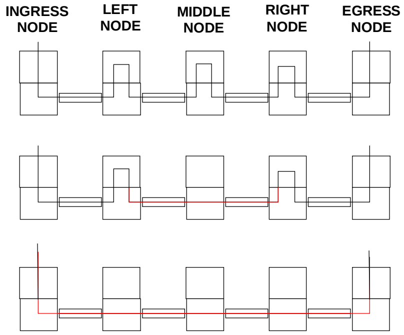


Figure 2 Assigning Wavelengths to Flows

Network architects have long recognized the desirability of assigning an IP flow to a wavelength so that the packets of the flow move along an all-optical path (sometimes called a lightpath or lightpipe) the network. The earliest attempts at this sought to create an overlay on top of a physical WDM-based network of a specific virtual topology optimized for the predicted traffic patterns [1, 13]. These attempts relied on a central controller that pro cessed the network's long-term traffic statistics and preformed an optimization to identify the wavelength assignment (virtual topology) that maximized a chosen performance metric under the network's prevailing traffic conditions. The process is essentially static. It is computationally challenging, attempting a large-scale global optimization. Finally, it is subject to a single point of failure. These approaches implicitly assume that whichever controller identified the best virtual topology would be responsible for reconfiguring the ne twork to realize the desired topology. It is questionable whether an ope ational network could implement this without imposing severe penalties on users. The assignment of flows to wavelengths in the backbone must be done dynamically, adapting to short-term traffic fluctuations and not dependent on a central point of control or requiring large-scale interruptions of service.

Signaling Protocol

The POW signaling protocol — called the flow-management protocol (FMP) — is built under the assumption of reliable message delivery, thus reducing the complexity of the signaling protocol. This assumption is nforced by running the protocol on top of a reliable transport protocol, e.g. the transmission control protocol (TCP). The POW flow analyzer recognizes three granularities of flows: fine-, medium-, and coarse-grain flows. A fine-grain flow is a sequence of packets with the same source and destination IP addresses, and the same source and destination TCP or UDP ports, i.e., a flow defined by a session between two applications. A medium-grain flow is an aggregation of fine-grain flows with the same source and destination IP addresses, i.e. a flow defined as the stream of information between two hosts. A coarse-grain flow is an aggregation of medium-grain flows that pass through the same ingress and egress nodes, i.e., a flow defined by the strea of packets that enter and exit the backbone at two given points of presence The three granularities of flows are illustrated in Fig. 3.

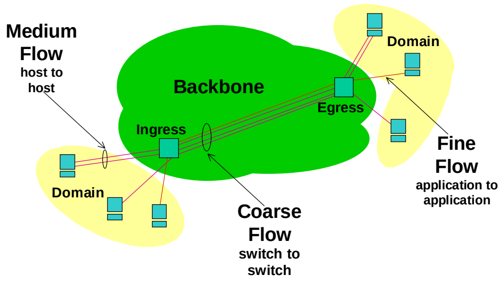


Figure 3 Flow Granularit

A flow is detected by means of the common *X/Y* flow classifier [12], in which a flow is declared eligible for switching whenever the switch–router observes *X* packets of a flow within a time period of *Y* seconds or less. Onc a node detects a flow of the targeted granularity, it uses FMP to bind the flo to a wavelength that traverses the n twork.

Immediately after the detection of a suitable flow, FMP initiates m s-sages to agree upon the existence of a set of free wavelengths along the route taken by the flow and to choose one wavelength common to each hop, thereby establishing a contiguous lightpath for the flow. FMP's strategy is to construct lightpaths from the egress node back towards the ingress node. Th lightpath lasts as long as there is sufficient momentum in the flow to justify its assignment to a dedicated wavelength. A weakened flow causes a hop to disengage and propagates teardown messages along the lightpath.

Routing Requirements

The POW architecture depends on the ability of nodes to monitor and classify flows of packets. Because packets transit an optical network at very high rates, it is essential to monitor the network in real time and with little or no interference. Given such a feat, it is necessary to identify a flow on the basis of its routing. Although a challenging performance problem, recognizing a fine- or medium-grain flow from source and destination IP addresses poses no fundamental difficulties, because these addresses part of the IP headers of the packets that comprise the flow.

More problematic is that coarse-grain flows are the aggregations of packets that might not have common IP addresses. Their commonality stems from sharing the same ingress or egress nodes of the backbone network. However, ingress and egress points are not usually expressed explicitly in packets, unless they happen to be source-routed (as is possible — but not widely supported — in IP). It is critical for POW to be able to deduce at least the ingress and egress nodes of a packet by examining only the header of the packet. Happily, this requirement is supported easily by the most-commonly encountered backbone routing protocols. For example, the IS–IS (intermediate system to intermediate system) routing protocol, which is used by many of the largest backbone operators, provides the entire path specification of all routes through its network [5]. Such information is easily incorporated into the routing table, and it can be henceforth assumed that the POW router nod software can lookup the next-hop, ingress, and egress nodes of a packet.

Routes used by the IP protocol may change in response to network conditions. Most commonly, a new route is computed whenever there is a failur in the network. Less commonly, a new route might be computed to optimiz a specific performance or cost metric. POW lives comfortably with route updates, which are typically on time scales of seconds. POW might not function well where routes changing dynamically and more frequently. Fortunately, routes in today's Internet backbones are extremely stable, with average route lifetimes lasting several days [9].

Node Design

A functional diagram of the POW node is shown in Fig. 4. The router is a general-purpose computing platform such as a PC used as a forwarding engine. It includes software for monitoring packet flows, FMP signaling software, as well as software to control the associated optical crossconnect switch. The router supports the backbone network's chosen interior routing protocol, which identifies the egress router of a packet in transit.

The POW node is connected to other POW nodes by high-bandwidth optical fibers that employ WDM to carry several channels of information. The link protocol should be transparent to the optical crossconnect switch, its implementation residing principally in the router. The exact link protocol is at the discretion of the router operator, and it might differ from node to nod (except where interoperability is needed). SONET, gigabit ethernet, or the point-to-point protocol (PPP) are likely candidates. This study does not assume the use any specific link protocol in the simulation model.

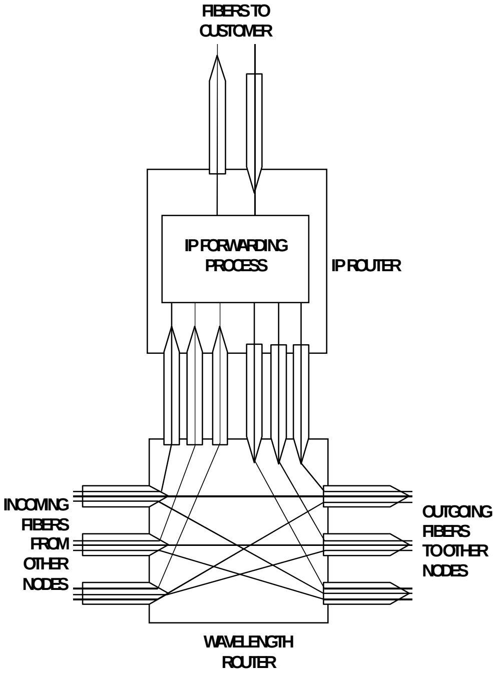


Figure 4 POW Node Architecture

The optical crossconnect switch is connected to the IP router by high-bandwidth optical fibers. These intranode fibers support only a single wavelength, which is the default channel over which all IP-forwarded traffic and signaling packets move. The IP router is the interface to the customer(s), with which it shares one or more links of a chosen technology (optical, elctronic, etc.). The IP router is thus a standard router with a specially designed interface to the optical crossconnect switch.

Wavelength Merging

The reuse of precious wavelengths is supported by aggregating tributary flows by merging packets from several streams. The optical crossconnect component of the POW node requires enhanced capabilities to perform this merging function. The design and implementation of a wavelength-selectiv optical cross-connect with merge capabilities are being pursued as part of th

POW project [2]. The device can route the same wavelength from different incoming fibers into a single outgoing fiber. It requires that contention between bits on the wavelength must be resolved before they are multiplexed into the common outgoing fiber.

Using the merge function for traffic grooming is not a new concept in the telecommunications arena [21]. It is possible to use spare capacity on an already allocated wavelength to compensate for the scarcity of flows. Th optical crossconnect switch can be integrated with a contention-resolution subsystem that time-multiplexes simultaneously arriving packets from a common wavelength but different input fibers onto the same wavelength on the same output fiber [18]. The contention resolver uses a combination of compression, subcarrier multiplexing, and time-shifting.

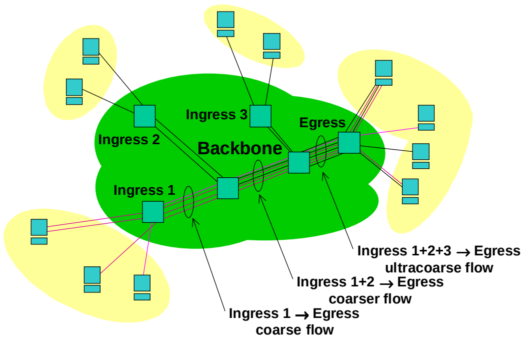


Figure 5 Wavelength Merging

Wavelength merges allow several ingress nodes to feed their flows to a single egress node, as depicted in Fig. 5. The signaling protocol must be modified a bit to allow for the allocation of wavelengths to “light-trees” rather than lightpaths, and it is also possible to merge wavelengths after they have been assigned.

3. THEORETICAL LIMITS OF WDM

As an abstract representation of a WDM backbone, consider a network of N nodes and the links that interconnect them. Suppose that the links can carry information on separate channels, one may ask how many channels are required to create a virtual overlay on the physical network that interconnects all nodes by exactly one hop. The goal of using WDM in an IP backbone is to put each pair of routers in the backbone within a single hop of each other, so that switching is favored over forwarding. It is therefore instructive to xplore how many wavelengths are needed to realize a fully connected virtual topology in an arbitrary graph.

Although it is difficult to answer this question for all graphs, it may be answered for specific graphs that represents extremes of physical connectivity. Consider first the graph *K* in which each pair of nodes is connected by two links; *K* represents the idealized physical topology with maximal connectivity. Next consider the graph *R* in which all nodes are arranged in a ring, the links of which are all unidirectional; *R* is the idealized physical topology with the poorest connectivity (subject to the constraint that all nodes are connected by at least one path). How many wavelengths are needed in *K* and *R* to connect every pair of nodes by one hop?

It is clear that only one wavelength is needed in *K* to realize a single-hop topology, since the underlying physical topology is already single-hop. The number of wavelengths required to create a single-hop virtual topology in the ring *R* is much larger and depends on *N*.

Let f_N be the number of wavelengths required to overlay a single-hop virtual topology on top of the ring physical topology *R*. f_{N+1} can be computed inductively by observing that a new (N+1)-node ring can be created by inserting a node between two specific neighboring nodes of *R*. Using the original f_N wavelengths in addition to *N* new wavelengths to connect the new nod to the original *N* nodes, full connectivity is achieved in the (N+1)-node ring. This yields a simple recurrence relation

$$f_{N+1} = f_N + N$$

It is clear that f_1=0, since a single-node degenerate network requires no wavelengths. Take the z-transform of the recurrence relation to obtain

$$\sum_{k=0}^{\infty} f_{k+1} z^{-k} = \sum_{k=0}^{\infty} f_k z^{-k} + \sum_{k=0}^{\infty} k z^{-k}$$

After algebraic manipulation of this last equation, the z-transform $F(z)$ of f_N is seen to b

$$F(z) = \frac{1}{(z-1)^3}$$

That this function is the z-transform of the sequence

$$f_N = N(N-1)/2$$

may be verified by consulting a table of common z-transform pairs [8].

To summarize, in a richly connected physical topology (*K*, the bidirectional complete graph) one wavelength per link is required to create a single-hop virtual topology, whereas in a poorly connected physical topology (*R*, the *N*-node unidirectional ring) $N(N–1)/2$ wavelengths per link are required to create a single-hop virtual topology.

If the volume of traffic between each pair of nodes is uniformly γ, then the throughput per node in the fully connected network K is

$$T_K = 2(N-1)\gamma$$

where the factor of 2 accounts for traffic both originating from and destined to the node. On the other hand, the throughput of a router in the ring under uniform traffic is

$$T_R = (N-1)(N+2)\gamma/2$$

since $(N–1)\gamma$, units of traffic are sourced by the router, $(N–1)\gamma$ units of traffic are sinked by the router, and $\sum_{k=1}^{N-2} k\gamma = (N-2)(N-1)\gamma/2$ units of traffic are transited by the router. If the N-node ring is provisioned wit $N(N–1)/2$ wavelengths, then the amount of traffic that flows through a node in packet-forwarding mode can be reduced by as much as

$$T_R - T_K = (N-1)(N-2)\gamma/2$$

which is a substantial fraction of the total load offered to the ring.

The discussion above bounds the limits of performance that can b achieved by employing WDM in the network. In a poorly connected physica topology, routers can be unburdened of a large portion of their load (up to a factor that grows quadratically in the number of nodes N). The price paid for this is an increase in the number of wavelengths required per link (up to a factor that varies as the square of N). When dealing with real networks that have arbitrary physical topologies and nonuniform traffic demands, fewer than $O(N^2)$ wavelengths are expected to be used. In the next section simulations of actual networks under realistic traffic conditions will expose th practical tradeoffs between performance improvements and the number of usable wavelengths.

4. SIMULATION AND TRAFFIC MODELS

To evaluate POW a detailed simulation has been constructed for th purpose of running experiments. The goal of these experiments is to estimate the fraction of packets that could be switched (vs. forwarded) in a realistic network of POW nodes. To this end an actual topology and real traffic traces were used to drive a model built in the virtual Internet testbed network simulator (VINT/ns).

While earlier simulations focused on assessing performance in a singl switch [14, 12], this study focuses in overall performance in a wavelength-limited environment. Such performance is presumably influenced by the competition for wavelengths by different nodes. It is imperative to simulate an entire multinode network rather than a single node.

VINT/ns Simulation Model

The VINT/ns tool is a popular simulation package used for evaluating protocols in large-scale networks [10]. VINT/ns performs packet-level simulation according to specified set of protocols. It has extensive facilities for the purposes of gathering data and testing functionality, and a large library of existing protocols. Most importantly for this work, it accepts as inputs log files of real packet traces.

Essential components of the simulation model include the flow classifier, which is constructed as an *X/Y* classifier with *X* set to 10 packets and *Y* set to 20 seconds, the forwarding functions, and the high-speed transmission links. The model implements the FMP signaling system (described above) for establishing lightpipes upon recognition of candidate flows. FMP is implemented on a hop-by-hop basis above TCP, which VINT/ns provides as a library protocol. The internode WDM links operate at OC-48 speeds (2.5 Gb/s), while the intranode links operate at OC-12 speeds (622 Mb/s). The node model does not use a routing protocol, but instead relies upon static routes that are preloaded in the nodes.

The nodes are interconnected in VINT/ns according to the vBNS (very high bandwidth network service) backbone topology, which is shown in Fig. 6. The vBNS network matches well the type of environment that POW would be used in: vBNS provides IP service on top of an asynchronous transfer mode network. However, the vBNS establishes a complete mesh of peranent virtual circuits among all nodes; POW would establish “circuits” (or wavelengths) dynamically in accordance with the amount of flow to be carried from one node to another.

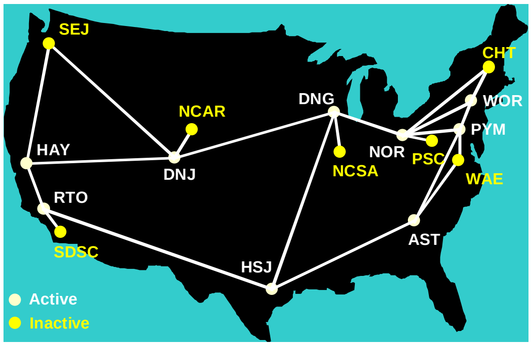


Figure 6 vBNS Backbone Topology

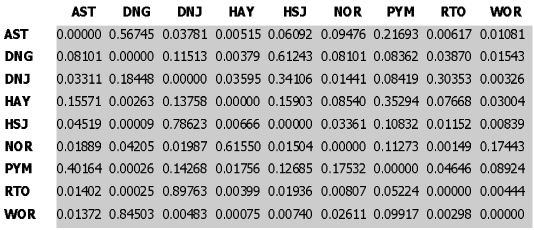
Each POW node is connected to its neighbors by an optical fiber that carries *W* WDM channels. In addition to these *W* channels, there is always one WDM channel reserved exclusively for routed traffic and signaling between any pair of neighboring nodes. The model of POW simulates its wavelength-management functions as well as the interactions of nodes through the FMP signaling protocol.

The simulation model is instrumented to measure several quantities. Th principal metrics computed are the number of packets switched vs. the number of packets routed, the number of FMP packets exchanged as overhead, the transit delay of packets, and the number of wavelengths utilized.

Traffic Model

The simulation is based on an actual topology and real traffic traces. The vBNS backbone consists of 16 nodes, of which nine were passing traffic on September 18, 1998, when our traffic measurements were taken. These measurements are collected by the National Laboratory for Advanced Network Research and represent the average of five-minute samples taken hourly over the entire day. This data was used to compute a traffic matrix, an entry of which is the probability that a node's packet would exit the vBNS via another specified node. Thus, entry (*i*, *j*) of the traffic matrix represents the probability that a packet from node *i* is destined for nod *j*. Traffic on the vBNS is relatively light, loading none of its links by more than 10% of capacity. However, it is the traffic pattern that interests us, rather than the actual loading. The matrix is displayed in Table 1.

Table 1 vBNS Traffic Matrix



	AST	DNG	DNJ	HAY	HSJ	NOR	PYM	RTO	WOR
AST	0.00000	0.56745	0.03781	0.00515	0.06092	0.09476	0.21693	0.00617	0.01081
DNG	0.08101	0.00000	0.11513	0.00379	0.61243	0.08101	0.08362	0.03870	0.01543
DNJ	0.03311	0.18448	0.00000	0.03595	0.34106	0.01441	0.08419	0.30353	0.00326
HAY	0.15571	0.00263	0.13758	0.00000	0.15903	0.08540	0.35294	0.07668	0.03004
HSJ	0.04519	0.00009	0.78623	0.00666	0.00000	0.03361	0.10832	0.01152	0.00839
NOR	0.01889	0.04205	0.01987	0.61550	0.01504	0.00000	0.11273	0.00149	0.17443
PYM	0.40164	0.00026	0.14268	0.01756	0.12685	0.17532	0.00000	0.04646	0.08924
RTO	0.01402	0.00025	0.89763	0.00399	0.01936	0.00807	0.05224	0.00000	0.00444
WOR	0.01372	0.84503	0.00483	0.00075	0.00740	0.02611	0.09917	0.00298	0.00000

The traffic matrix represents the averages of millions of packets. It is not feasible either to collect or simulate such a large sampling of traffic. real trace of about one hour's worth of traffic was thus used. The packets were collected in tcpdump format from a router at the Lawrence Berkeley National Laboratory in 1994. The packet trace — known as LBL-PKT-5 — has "sanitized" IP addresses to protect the privacy of users whose packets were traced. The IP addresses in the trace are nonsense (and do not correspond to real hosts), except that a single value is used to replace a real traced address. Because the IP addresses refer to hosts that reside on customer n t-works that might or might not be attached to the backbone, it was necessary to devise a rule to assign a sanitized address to an egress router. When a packet is injected into the VINT/ns model, its address is read and randomly assigned an egress node in accordance with the probabilities in the traffic matrix. This indicates that the address is found on a network on the "other side" of the egress point. The assignment of an egress node to an IP address is consistent across a single node, but one IP address injected at different points will not nec ssarily exit the network from the same point.

This simulation model does not completely capture the dynamics of th end-to-end protocols. Because actual traces are used as inputs to the routers at POW points of presence, TCP behavior may not be modeled with total accuracy. The tcpdump traces that make up LBL-PKT-5 already reflect time-dependent end-to-end behavior that is governed by TCP's congestion-avoidance and flow-control mechanisms. As traffic loads fluctuate, TCP end-to-end throughput is expected to change adaptively. The packet timestamps should also change. Short of developing a complete simulation that includes thousands of hosts, it is not feasible to model the detailed behavior of TCP flows, so static traces are used to approximate the steady-state network flow.

5. PERFORMANCE EVALUATION

In partial response to the question "how many wavelengths are really needed," POW was simulated over a range of from four to 64 wavelengths (in addition to the default wavelength) deployed in the vBNS physical topo l-ogy, using the traffic traces described above to drive the model. The principal performance metric is the switching gain, measured as the ratio of the nu m-ber of packets that travel along an allocated lightpath (as opposed to a default wavelength) to the total number of packets submitted to the network. Also of interest is the signaling overhead, measured as the ratio of the number of FMP packets to the total number of packets submitted to the ne twork.

As a performance metric, switching gain is an indirect reflection of throughput and delay. It relates directly to the goals of label switching, to maximize traffic over switching paths rather than routing paths.

The graph of Fig. 7 shows how much traffic can be switched as th wavelength count is increased, progressively aggregating flows. As expected, the switching gain grows steadily when the wavelength count is increased from four to 64. Less intuitive is the dramatic rise in switching gain as traffic is aggregated: when POW defines a flow according to the coarsest granular-ity, it can carry more that 98.59% of its traffic over dedicated lightpaths u s-ing a few as four wavelengths. When aggregation is weak (as in variants of POW that use medium- and fine-grain flows), switching gain is low, reaching about 84.56% and 65.94%, respectively, for medium- and fine-grain flows when 64 wavelengths are available. These latter figures hint at the high cost of operating POW without sufficient aggregation of traffic.

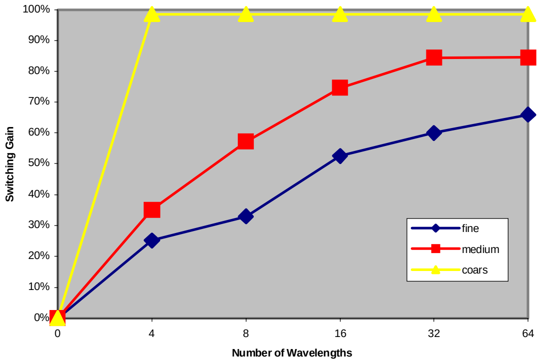


Figure 7 Switching Gain vs. Number of Wavelengths for Different Flow Granularities

Being software-based, the aggregation of traffic in a POW node does not come for free. Packet addresses must be matched, looked up, and tallied according to affinities with other addresses. Some of this dovetails easily with packet forwarding, but there is extra effort in aggregation. The wavelength-merging technology being developed for POW is a natural way to aggregate traffic by joining flows at intermediate nodes. When the fine-grain flows in POW are simulated with and without wavelength merging, an improvement in the switching gain is seen for all wavelength counts. As shown in Fig. 8, the improvement ranges from 21% to over 52%. At low wavelengt counts the amount of switched traffic remains low. Given the excellent switching gain achieved with coarse-grain flows, it is unclear whether introducing wavelength-merging technology in POW will ultimately pay off.

The signaling protocol FMP imposes a penalty on the network by introducing overhead traffic that competes with user traffic for bandwidth and processing cycles. Although FMP’s traffic is restricted to the default wavelength, its presence on that link still deprives other unswitched traffic of bandwidth. If the overhead of FMP is kept low as a percentage of overall traffic, then improvements in switching gain are clearly achieved. If FMP overhead is high, then one must weigh any improvements in switching gain against the cost of this ov rhead.

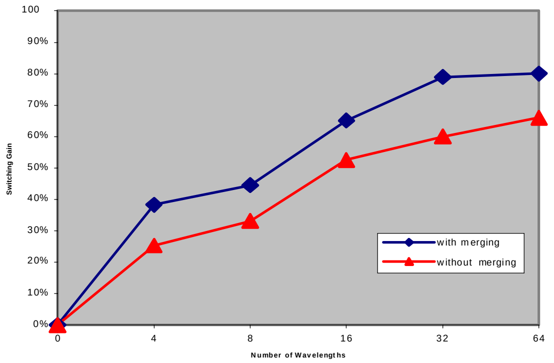


Figure 8 Switching Gain vs. Number of Wavelengths with and without Wavelength Merging for Fine-Grain Flows

Table 2 shows an account of the signaling overhead as a function of flow granularity and wavelength count for POW networks without wavelength merging. Signalling overhead remains low over most of the operating regimes of POW. The exception is when the granularity is fine and th wavelength count is high, where FMP is obliged to search the wavelengt space for an available WDM channel. This search is exacerbated by temporarily locking wavelengths while FMP probes the entire path. FMP thus places significant additional traffic in the network in some circumstances. Note that signaling overhead is calculated as a fraction of all offered traffic. In the 64-wavelength, fine-grain POW configuration, the 8.60% of the offered traffic that is associated with FMP is comparable to the 34.06% of offered traffic that is not switched (see Fig. 7). In the case of coarse flows, th overhead is steady regardless of the number of wavelengths available, because four wavelengths are adequate to switch most eligible traffic, and th addition of wavelengths has no impact at all on the operation of FMP.

Table 2 Signaling Overhead

Granularity	Wavelength Count				
	4	8	16	32	64
Fine	0.18%	0.39%	0.94%	5.00%	8.60%
Medium	0.14%	0.37%	0.88%	1.41%	1.46%
Coarse	0.02%	0.02%	0.02%	0.02%	0.02%

6. CONCLUSION

This study considered the question of how many wavelengths ar needed to achieve good performance for a packet-over-wavelength architecture for providing high-speed Internet service in an optical backbone network. Focusing on an existing backbone topology (vBNS) and using real traffic traces, simulation was used to evaluate the switching gain achievabl as a function of wavelength count and traffic aggregation. The central conclusion is that very high switching gain (on the order of 98% of offered traffic is switched) can be achieved when traffic is coarsely aggregated according to its ingress and egress nodes, even for low wavelength counts. It is rasonable to expect that a network comparable to vBNS could benefit fro POW with as few as four wavelengths.

Introducing wavelength-merging technology into POW improves th switching gain with fine-grain flows by as much as 52%. However, overall switching gain is low, even when wavelength merging is employed with fine-grain flows, reaching only 80.14% with 64 wavelengths.

The signaling overhead that is imposed by POW is generally low, except when traffic is finely aggregated. In the case of fine-grain flows with a high wavelength count, overhead amounts to more that 8.60% of the offered traffic. Because the overhead is carried entirely on links shared by unswitched traffic, it can negatively impact the network's performance.

In summary, four wavelengths are sufficient to achieve very high prformance when traffic is aggregated according to its ingress and egress nodes. The conclusions drawn from this study apply to one relatively small backbone, using a traffic model based on data collected from older packet traces and traffic patterns on a lightly loaded backbone. Evaluating POW more effectively would require larger, topologically diverse backbones and packet traces that are more representative of a high-performance network. Completely ignored in this study are the issues of stability and transient response when traffic patterns change abruptly and new wavelength assignments are effected; however, such concerns would probably be among the most critical in the view of a backbone operator and its customers.

REFERENCES

[1] J. Bannister, L. Fratta, and M. Gerla, "Topological Design of the Wavelength-Division Optical Network," *Proc. IEEE INFOCOM '90*," San Francisco, pp. 1005–1013, Apr. 1990.

[2] J. Bannister *et al*., "How Many Wavelengths Do We Really Need? A Study of Packets Over Wavelengths," presented at GBN '99, New York, Mar. 1999.

[3] D. Blumenthal *et al*., "WDM Optical IP Tag Switching with Packet-Rate Wavelength Conversion and Subcarrier Multiplexed Addressing," *Proc. OFC '99*, San Diego, Feb. 1999.

[4] A. Brodnik *et al*., "Small Forwarding Tables for Fast Routing Lookups," *Proc. ACM Sigcomm '97*, Cannes, pp. 3–14, Sept. 1997.

[5] R. Callon, "Use of OSI IS–IS for Routing in TCP/IP and Dual Environments," IETF RFC 1195, Dec. 1990.

[6] R. Callon *et al.*, "A Framework for Multiprotocol Label Switching (MPLS)," (work in progress), Nov. 1997.

[7] "WDM Equipment Buyer's Guide," *Data Communications Magazine*, http://www.data.com, Apr. 1999.

[8] R. Dorf, ed., *The Electrical Engineering Handbook*, CRC Press, Boca Raton, Fla., 1993.

[9] R. Govindan and A. Reddy, "An Analysis of Inter-Domain Topology and Route Stability," *Proc. IEEE INFOCOM '97*, Kobe, pp. 850–857 Apr. 1997.

[10] P. Huang, D. Estrin, and J. Heidemann, "Enabling Large-scale Simulations: Selective Abstraction Approach to the Study of Multicast Protocols," *Proc. IEEE MASCOTS '98*, Montreal, pp. 241–248, Jul. 1998.

[11] B. Lampson, V. Srinivasan, and G. Varghese, "IP Lookup Using Multiway and Multicolumn Binary Search," *Proc. IEEE INFOCOM '98*, San Francisco, pp. 1248–1256, Apr. 1998.

[12] S. Lin and N. McKeown, "A Simulation Study of IP Switching," *Proc. ACM Sigcomm '97*, Cannes, pp. 15–24, Sept. 1997.

[13] B. Mukherjee *et al.*, "Some Principles for Designing a Wide-Area Optical Network," *Proc. IEEE INFOCOM '94*, Toronto, pp. 110–119, Jun. 1994.

[14] P. Newman *et al.*, "IP Switching — ATM Under IP," *IEEE/ACM Trans. Networking*, vol. 6, no. 2, pp. 117–129, Apr. 1998.

[15] C. Qiao and M. Yoo, "Optical Burst Switching — A New Paradigm," *J. High Speed Networks*, to appear.

[16] Y. Rekhter *et al.*, "Cisco Systems' Tag Switching Architecture Overview," IETF RFC 2105, Feb. 1997.

[17] R. Schmidt and R. Alferness, "Directional Coupler Switches, Modulators, and Filters Using Alternating $\delta\beta$ Techniques," in *Photonic Switching*, H. Hinton and J. Midwinter, eds., IEEE Press, New York, 1990.

[18] 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.

[19] J. Turner, "Terabit Burst Switching," Tech. Rep. WUCS-9817, Washington Univ., Dept. of Comp. Sci., Jul. 1998.

[20] M. Waldvogel *et al.*, "Scalable High Speed IP Lookups," *Proc. ACM Sigcomm '97*, Cannes, pp. 25–36, Sept. 1997.

[21] X. Zhang and C. Qiao, "An Effective and Comprehensive Solution to Traffic Grooming and Wavelength Assignment in SONET/WDM Rings," *Proc. SPIE Conf. on All-Optical Networking*, vol. 3531, pp. 221–223, Nov. 1998.