How Many Wavelengths Do We Really Need? A Study of the Performance Limits of Packet Over Wavelengths

JOE BANNISTER
JOE TOUCH
ALAN WILLNER
University of Southern California

STEPHEN SURYAPUTRA
Nortel Networks

ABSTRACT

Coupling Internet protocol (IP) routers with wavelength-selective optical crossconnects makes it possible to extend the existing Internet infrastructure to a wavelength-division–multiplexing optical network. Because optical wavelength routing is transparent to IP, one can achieve very high throughput and low delay when packets are made to bypass the IP forwarding process by being switched directly through the optical crossconnect. We study the performance of a specific instantiation of this approach,which we call packet over wavelengths (POW). We present the POW architecture in detail and discuss 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, we evaluate the fraction of packets that are switched through the crossconnect as a function of the number of wavelengths and the degree of flow aggregation that can be achieved.Our study, conducted in the context of the very-high bandwidth 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 [Dat99], and it is highly desirable to use these links to interconnect the routers that comprise the Internet. We consider a network architecture - called packet 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 this architecture is to switch as much traffic as possible directly by means of optical crossconnect switches, because IP forwarding is relatively expensive by comparison. However, wavelength routing through an optical crossconnect switch is constrained by the fact that only a few (four to 64) WDM channels per link are supported by today's commodity technology. Our intent is to study and characterize the expected performance of POW in such a sparse-WDM environment. To this end we examine different options for recognizing which packets should be switched through an optical crossconnect switch and which packets should be forwarded by an IP router. We conduct simulations to 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 [New96], tag switching [Rek97], and multiprotocol label switching [Cal97], all of which we shall henceforth refer 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. Since switching speeds are much greater than forwarding speeds (estimated by some [New96, Lin97] to be 20 times greater for comparably priced hardware), one attempts to place as large a fraction of the packets as possible on the streamlined switched path, leaving as small a fraction of the packets as possible on the slower forwarded path. To accomplish this requires above-average intelligence in the switch–router. The router must have software that recognizes that a flow of packets can be labeled and passed through the switching fabric. A signaling protocol then assists in notifying 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 [Blu98, Qia99, Tur98], so it is crucial to understand fully the engineering tradeoffs that one encounters.

The goals of our work are to determine whether optical label switching is feasible and beneficial in the near-to-medium term. We therefore investigate the behavior of real Internet traffic in an optical label-switching backbone with a limited number of wavelengths. We also evaluate the performance 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 outlines the functions of the signaling protocol used by POW. In Section 4 we present analytical results to characterize the overall gain that one could expect from the introduction of WDM.Section 5 provides the details of the simulation, the traffic model, and the experiments used to evaluate POW's performance. Section 6 presents the results of the evaluation.Section 7 offers conclusions to be drawn from the study.

2 - POW ARCHITECTURE

A starting point for our 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 and 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). Thus,a packet 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, in that it is normally a software-controlled process. The dominant costs of forwarding come from the act of 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 [Bro97, Lam98, Wal97] 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 cut-through switching allows the head of the packet to exit the switch even before its tail has entered.The effect can be significant when packets are large. Furthermore, routing and forwarding paths are more prone to loss than are optical-switching paths because of the reliance on memory to buffer packets. Thus, one of the objectives of our work is to reduce the number of hops suffered 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. We consider an ISP-operated backbone that consists of routers connected by optical fibers that support WDM. We further assume that wavelength-selective optical crossconnect switches are available to channel wavelengths from incoming optical fibers to outgoing fibers [Sch90]. A functional depiction of a wavelength-selective optical crossconnect switch (also known as a wavelength router) is shown in Fig. 1.This switch is an optical 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 from 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).

As shown in Fig. 2 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. For our purposes we define an IP flow as a sequence of packets that travel together along a subset of the same route in the network before exiting the network.This definition 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. Our definition permits us to 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 suffi-

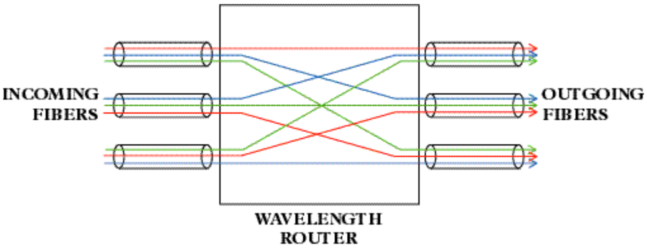


FIGURE 1: *Wavelength-Selective Optical Crossconnect Switch.*

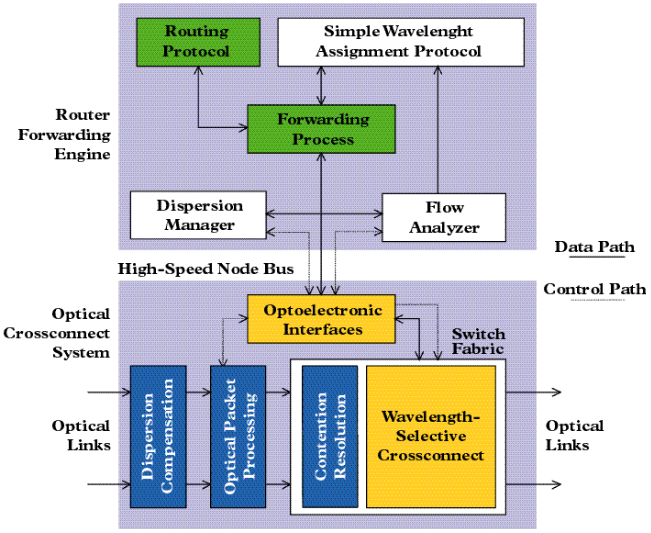


FIGURE 2: *POW Node Architecture.*

cient intensity and duration to merit fast-path switching.Each incoming fiber uses a special wavelength (which we designate as l0) for the default traffic. When a flow is recognized, however, the router's control software attempts to shunt the flow straight through on its own wavelength. To do so 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, say, 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. Likewise, similar actions are coordinated with the downstream router at the other end 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.The process is illustrated in Fig. 3.

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) of the network [Chl92]. The earliest attempts at this sought to overlay on top of a physical WDM-based network a specific virtual topology optimized for the predicted traffic patterns [Ban90, Muk94]. These attempts relied exclusively on a central controller that processed the network's long-term traffic statistics and performed an optimization that sought 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.And it is subject to a single point of failure. Implicit in these approaches is the assumption that the controller that identified the best virtual topology would be responsible for reconfiguring the network to realize the desired topology. It is questionable whether this in fact could be implemented in an operational network without imposing severe penalties on users. The assignment of flows to wavelengths in the backbone must therefore be done in a dynamic manner that adapts to short-term traffic fluctuations and does not depend on a central point of control or require large-scale interruptions of service.

2.1 - ROUTING REQUIREMENTS

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 (transmission control protocol) or UDP (user datagram protocol) 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 stream of packets that enter and exit the backbone at two given points of presence (but might originate and terminate at many different hosts). The three granularities of flows are illustrated in Fig. 4.

A flow is detected by means of the common *X/Y* flow classifier [Lin97], 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. Immediately after the detection of a suitable flow, a signaling protocol - called the simple wavelength assignment protocol (SWAP) - initiates messages to choose one wavelength common to each hop of the flow, thereby establishing a continuous lightpath for the flow.

The POW architecture depends on the ability of nodes to monitor and classify flows of packets. Because one expects packets to 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 that such a feat can be accomplished, it is still necessary to identify a flow on the basis of its routing. Although challenging from the viewpoint of performance,recognizing a fine- or medium-grain flow from source and destination IP addresses poses no fundamental difficulties, since these addresses are carried explicitly in the IP headers of the packets that comprise the flow.

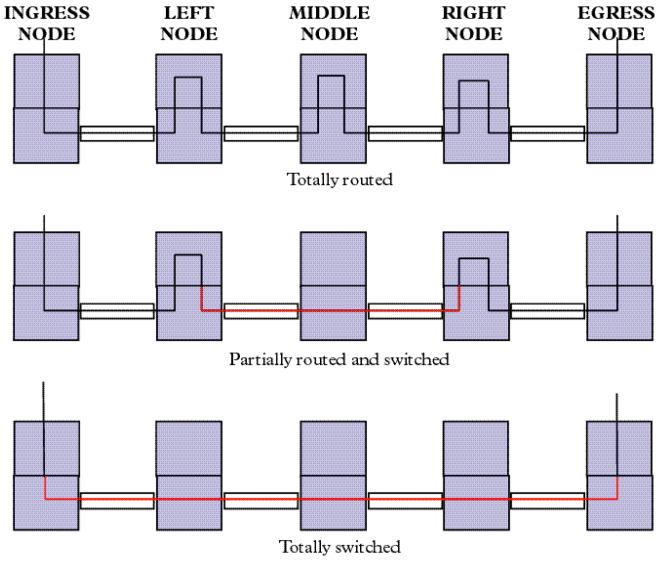


FIGURE 3: *Assigning Wavelengths to Flows.*

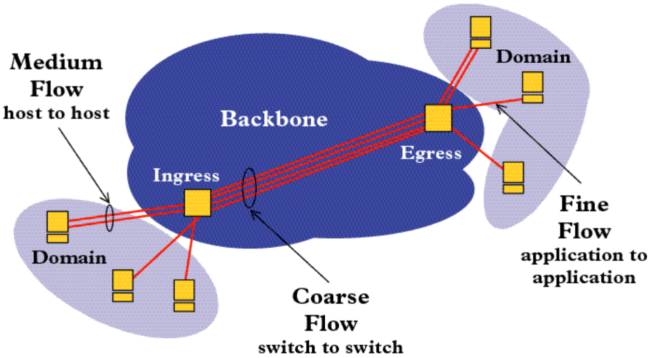


FIGURE 4: *Flow Granularity.*

More problematic is that coarse-grain flows are the aggregations of packets that might not have common IP addresses. Instead, their commonality stems from their 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). So a fundamental requirement of POW is the ability 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 vector of the routes through its network [Cal90]. Such information is easily incorporated into the routing table employed by the forwarding process, and we henceforth assume that the software in the router component of the POW node can look up the next-hop, ingress, and egress nodes of any packet that it processes.

Routes used by the IP protocol may change in response to network conditions. Most commonly, a new route is computed whenever there is a failure in the network. Less commonly, a new route might be computed to optimize 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 were routes changing dynamically and frequently. Fortunately, routes in today's Internet backbones are extremely stable, with average route lifetimes lasting several days [Gov97].

2.2 - Node Design

A functional diagram of the POW node is shown in Fig. 2. The IP router is a general-purpose computing platform, such as a personal computer,used as a forwarding engine.The use of a special-purpose IP router would also be possible, but this would require implementers to have access to the internals of the router; a general-purpose processor allows the use of open networking software, such as the GateD Consortium routing package [Gat99].The IP router includes software for monitoring packet flows. Also included in the IP router is the SWAP signaling software, as well as software to control the associated optical crossconnect switch. Of course, the router supports the backbone network's chosen interior routing protocol, which is assumed to identify the egress router of any packet in transit.The node implements a signaling protocol, which is described in Section 3.

The POW node is connected to other POW nodes by high-bandwidth optical fibers that employ WDM to carry several independent 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 node (except where interoperability is needed). SONET, gigabit ethernet, or the point-to-point protocol (PPP) are likely candidates. For the purposes of this study, we assume in our simulation model the use of PPP.

The optical crossconnect switch is connected to the IP router by high-speed bus. This bus can be implemented electronically or optically, but an optical implementation that uses the capabilities of the crossconnect device would consist of multiple fibers and wavelengths that could be fed by the internode links. The bus 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, electronic, etc.).Thus,the IP router is a standard router enhanced with a specially designed interface to the optical crossconnect switch.

2.3 - Wavelength Merging

A strategy for reusing precious wavelengths allows tributary flows to be aggregated 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-selective optical crossconnect with merge capabilities are being pursued as part of the POW project [Ban99]. The device would be able to route the same wavelength from different incoming fibers into a single outgoing fiber. The key property of the device is 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 [Zha98]. It is possible to use spare capacity on an already-allocated wavelength so as to compensate for the scarcity of available flows.The 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 [Shi97a, Shi97b]. The contention resolver uses a combination of compression, subcarrier multiplexing, and time shifting [Hav99].

The purpose of the wavelength merger is to 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 ”lighttrees” rather than lightpaths, and it is also possible to merge wavelengths after they have been individually assigned.

3 - THE SIMPLE WAVELENGTH ASSIGNMENT PROTOCOL

The strategy of the SWAP signaling protocol is to construct lightpaths between the ingress and egress nodes (we assume that coarse flows are used).The lightpath lasts as long as there is sufficient momentum in the flow to justify its assignment to a dedicated wavelength. A weakened flow will cause a hop to disengage and propagate a teardown message to other hops along the lightpath.

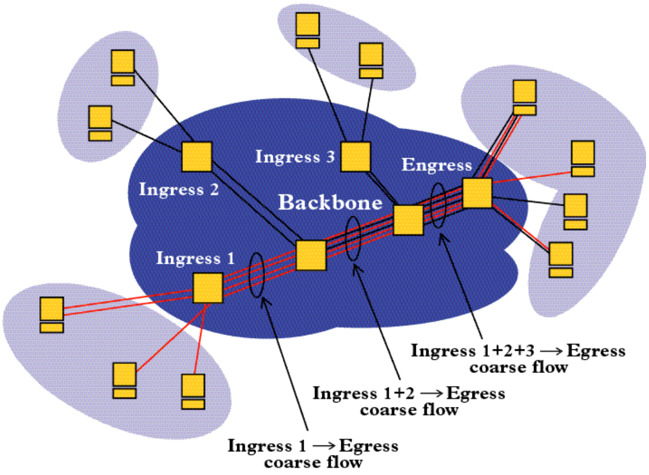

Ingress 3
Engress
Ingress 2
Backbone
Ingress 1
Ingress 1+2+3 → Egress coarse flow
Ingress 1+2 → Egress coarse flow
Ingress 1 → Egress coarse flow

FIGURE 5: *Wavelength Merging.*

There are several high-level requirements for the signaling protocol. First, it must be kept as simple as possible (using few, brief messages). Second, the signaling should construct the continuous lightpath as far as possible. Discontinuous lightpaths force the flow to be routed or demand wavelength conversion, which requires expensive hardware[1]. Third, the protocol must support flow merging (grooming).

To conform to the requirement of simplicity, SWAP is implemented on top of a reliable transport layer; this decouples the protocol from issues of reliable transmission and reuses protocols designed for this purpose. SWAP components establish per-neighbor TCP connections, over which the signaling messages are sent. Neighbor connections enable the node to determine whether it is the first, the last, or an intermediate hop for a particular flow (i.e.,closest to the flow's origin, destination, or in between, respectively), which is used to simplify the signaling, failure detection, and recovery. Neighbor relationships are maintained for each link of the node (one connection per link).The neighbor connections are not optimized because they are trivial and similar to many of the common routing protocols.

Constructing the continuous lightpath as far as possible requires SWAP to pick a common free wavelength along the flow path; therefore SWAP must collect the list of free wavelengths for each hop. If there is one free wavelength common to all the hops, it will be picked; if not, SWAP may choose to construct a discontinuous lightpath (if so configured) if there are sufficient wavelength converters available. Compared to generic label-swapping techniques, this feature is unique to SWAP, because SWAP's decision to pick a "label" (a wavelength here functions as a label) is a global decision. SWAP tries to minimize or eliminate the swapping of "labels". Therefore, it incurs a round-trip time to select a wavelength. The resource (set of free wavelengths) must be also locked during signaling to prevent different flows from trying to acquire the same resource.

Next SWAP decides where to initiate the signaling.Either end of the path is appropriate, being natural places where SWAP can efficiently gather complete path information.The first hop is good because a source can propagate its free-wavelength set when it detects an active flow (SETUP).When the next hop receives that set, it intersects it with its own free-wavelength set, and forwards the result to the next hop. If the final set is not empty, the last hop picks one free wavelength from the resulting set, configures its local node, and sends an acknowledgement (COMMIT) back to the previous hop with the chosen wavelength. Upon receiving an acknowledgement, the previous hop configures its local node and passes the acknowledgement to its previous hop, until the packet is received by the first hop.

However, it is often better to initiate the signaling from the last hop, because wavelength merging is better served. When wavelength merging is permitted, there is a single last hop but multiple first hops,which would complicate a source-initiated protocol.The last hop will also notice the flow earlier, as the traffic merges there and therefore has higher intensity. Second, it simplifies the protocol because there could be more than one outstanding setup request from upstream, and the protocol must keep track of the upstream status so that it can selectively send the COMMIT messages back downstream.Third, last-hop-initiated signaling will ease the interaction with other signaling mechanisms, such as the reservation protocol (RSVP) [Zha93] which also uses received-initiated setup. ARIS (aggregate route-based IP switching) [Dav98] takes the same approach to enable route aggregation. For these reasons SWAP employs last-hop-initiated signaling.

The drawback of last-hop-initiated signaling is that it will take more time to complete.The first phase is similar to first-hop-initiated signaling, except that the previous hop should not start sending packets using the new wavelength unless the next hop has already set up the node (to avoid losses). SWAP could do something similar to IP Switching, so that the node would send the packets using the slow path (through the IP router) while waiting for a response from the next hop. However, because this technique requires temporary path termination and optical switches require nontrivial setup times, it is undesirable to do so. Instead, SWAP forces the first hop to wait one round-trip time for signal propagation to complete. The process is shown in Fig. 6. Flow aggregation (grooming) also affects where to initiate the teardown mechanism.The last hop is undesirable, because drops in an aggregated flow are noticed at the sources first. Merging hops is also not desired because it requires the hop to monitor the

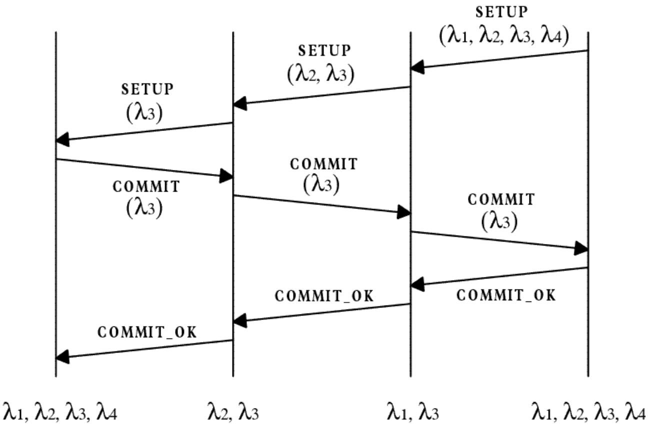


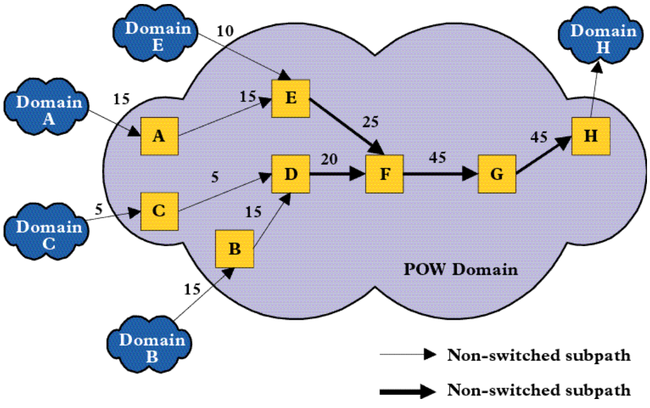
FIGURE 6: *Time-Sequence Diagram of Last-Hop-Initiated Signaling.*

[1] *We do not consider wavelength conversion in this paper, but SWAP is designed to support wavelength conversion.*

optical signal. Therefore, it is the responsibility for the first hops to initiate the teardown. If the first-hop node of a switched flow detects a drop in the throughput, it will send a TEARDOWN message to the next hop and the next hop will pass it to further hops if there are no switched incoming branches. Fig. 7 illustrates this effect on an aggregated flow.

Referring to Fig. 7, suppose SWAP were set to regard 20 packets per second (pps) as the threshold to switch a flow. Nodes D, E, F, G, and H all see an aggregate outbound throughput of 20 pps or higher for flow F_1 (i.e., all traffic going to domain H), and because node H knows it is the last hop, it locks the free-wavelength resource λ_{GH} and sends SETUP(F_1, λ_{GH}) to G. Upon receiving the SETUP message, G forms the wavelength-set intersection $\lambda_x = \lambda_{GH} \cap \lambda_{FG} = \{\lambda_2, \lambda_3, \lambda_4\}$,locks λ_x, and sends SETUP(F_1, λ_x) to F. F forms the intersections $\lambda_y = \lambda_x \cap \lambda_{DF} = \{\lambda_2\}$ and $\lambda_z = \lambda_x \cap \lambda_{EF} = \{\lambda_3, \lambda_4\}$, locking both λ_y and λ_z. Then it sends SETUP(F_1, λ_z) to E because E contributes more to the aggregate throughput than D. E knows that it is the first hop for that path[2] and arbitrarily picks any element of λ_z, e.g. λ_3. E then sends COMMIT(F_1, λ_3) to F. Meanwhile, F waits for a response from E. Upon receiving the response, F removes λ_3 from the free-wavelength set λ_{EF}, unlocks the set λ_z, and sends COMMIT(F_1, λ_3) to G. F then tries to add D: it computes $\lambda_w = \{\lambda_3\} \cap \lambda_{DF}$ and if $\lambda_w \neq$, then it sends COMMIT(F_1, λ_3) to D (if the intersection is empty, then the flow cannot be assigned without a wavelength conversion) and unlocks λ_w. G removes λ_3 from the set λ_{FG}, unlocks λ_x, and forwards COMMIT(F_1, λ_3) to H. Then H removes λ_3 from the set λ_{GH}, unlocks it, configuring its optical switch to pass data on λ_3 back to the IP router, and sends COMMIT_OK(F_1) to G. Upon receiving COMMIT_OK(F_1) from H, G sends COMMIT_OK(F_1) to F, and node F sends COMMIT_OK(F_1) to E. E configures its switch and IP router so that flow F1 is bound to wavelength λ_3.

Branch E is selected because the more a branch contributes to the aggregate throughput, the more likely it is to stay significant or become even more significant. If other



Node	Link	Free Wavelength Set
H	GH	$\lambda_{GH} = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$
G	FG	$\lambda_{FG} = \{\lambda_2, \lambda_3, \lambda_4\}$
F	EF	$\lambda_{EF} = \{\lambda_3, \lambda_4\}$
	DF	$\lambda_{DF} = \{\lambda_1, \lambda_2\}$
D	CD	$\lambda_{CD} = \{\lambda_1, \lambda_2, \lambda_4\}$
	BD	$\lambda_{BD} = \{\lambda_3, \lambda_4\}$

FIGURE 7: *Flow Aggregation Scheme.*

branches become inactive, they will likely be demoted to subpaths. The wavelengths of low-flow branches are picked arbitrarily and merged to the target wavelength. For instance, although the F_1 flow from node D to F was rejected for assignment to a wavelength, if λ_3 were to become free on this hop, the flow would be merged to the lightpath on wavelength λ_3 from node E to H.This step would require node F to reconfigure its switch so that links EF and DF merge their wavelength λ_3 on the outgoing link FG.

Now suppose there is an increase in the F_1 throughput from B to D. D realizes that the flow has been switched, so it sends SETUP(F_1, λ_{BD}) to B. B determines that it is the first hop and sends COMMIT(F_1, λ_3). Upon receiving from B, D configures its optical switch and sends COMMIT_OK(F_1) to B, and B sends the flow using λ_3 when it receives the message. If E subsequently detects a drop in F_1 throughput, it sends TEARDOWN(F_1) to F. Node F will not forward the message further, because it still has a switched branch, i.e., DF. As a result, F frees the wavelength, removes the switched path from its optical switch and rebinds it to the default wavelength λ_0. Since the flow intensity near the root of the tree forming the lightpath is always greater than the intensity near the leaves,the switched path is torn down from the leaves towards the root.

Finally, SWAP requires that neighboring protocol entities emit periodic keep-alive messages so the POW node can detect neighbor failures. The keep-alive message should incorporate a mechanism to detect the case of neighbor failure and subsequent recovery prior to the timeout of its neighbor entry. This is handled by the routing protocol.If the failed neighbor is the previous hop, then the node will behave as if it received TEARDOWN(F_1),TEARDOWN(F_2),...,TEARDOWN(F_n), where F_i are the switched flows coming from the neighbor. If the failed neighbor is the next hop, and there is no previous hop switched, then the node just sends the flow using the default wavelength λ_0. However,if there is a switched previous hop, the switch takes the last-hop role, i.e., it converts the switched flow back to λ_0.

We summarize below the major design points of SWAP:

1 SWAP is implemented on top of a reliable transport protocol, such as TCP.

2 A first-hop node is defined as the point where there is no upstream neighbor, or there are upstream neighbors but the incoming throughputs are not high enough to constitute a bonafide flow. A last-hop node is defined as the point where there is no downstream neighbor.

3 SWAP uses last-hop-initiated setup if there is no switched path and aggregation-point-initiated setup if the aggregation point already has the flow switched (i.e., the aggregation point is the last hop of the augmentation to the existing switched path).

4 SWAP uses first-hop-initiated teardown. Teardown messages are terminated at the merging point if there is still a switched incoming branch.

5 Resources (free wavelengths) are maintained and locked independently for each incoming link.

6 Grooming points should maintain the flow status for each incoming branch and the flow packet count for each unswitched incoming branch.

[2] *Nodes use two criteria to determine whether they are the first hop: (1) there is no upstream neighbor for the path or (2) there are no incoming branches that carry high-throughput flows.*

7. Whether the node is the first, an intermediate, or the last hop is determined using the neighbor protocol.
8. The neighbor protocol is assumed to be available or be provided by routing protocols.
9. If the neighbor protocol reports to a node that there are no upstream and downstream neighbors, SWAP will be disabled, because the SWAP implementation assumes there are at least two POW nodes in sequence (actually, it is useful only if there are at least three switches in a row).
10. SWAP makes use of six simple messages: SETUP, SETUP_CONFLICT, SETUP_FAIL, COMMIT, COMMIT_OK, and TEARDOWN. The semantics of the messages and behavior of protocol are specified in greater detail in [Sur99].

4 - THE LIMITS OF WDM

As an abstract representation of a WDM backbone, we consider a network of N nodes and the links that interconnect them. If we suppose that the links can carry information on separate channels, we may certainly 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 explore how many wavelengths are needed to realize a fully connected virtual topology in an arbitrary graph.

Although it is difficult to answer the foregoing question for all graphs, it may be answered for specific graphs that represent 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. We may compute f_{N+1} 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 node to the original N nodes, we achieve full connectivity 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. We 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 be

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

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 [Dor93].

To summarize, in a richly connected physical topology (K, the N-node bidirectional complete graph) we require 1 wavelength per link to create a single-hop virtual topology, whereas in a poorly connected physical topology (R, the N-node unidirectional ring) we require $N(N-1)/2$ wavelengths per link 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 R under uniform traffic is

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

If we provision the N-node ring with $N(N-1)/2$ 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.

To gauge the extent to which flow aggregation and merging can improve performance, we consider the N-node ring topology R under uniform traffic loading. Specifically, we are interested in maximizing the switching gain, defined as the fraction of traffic that is carried from the source node to the destination node on a single wavelength and without intervening router hops. For a specific switching gain σ, how many wavelengths W per fiber link are needed in R when the traffic is uniformly distributed?

We assume first that coarse-grain flows are identified in the ring, i.e. all traffic between any pair of nodes should be carried on a wavelength that starts at the first node and ends at the final node without intervening router hops (we cannot always realize this, as we might not have enough wavelengths at our disposal). We would like to calculate the number of wavelengths that are needed to achieve a given switching gain. Since node i in the uniformly loaded ring R communicates with N–1 other nodes, there can be at most N–1 coarse-grain flows emanating from node i, and each of these flows requires a different wavelength. If k flows of node i are switched, then the switching gain will be $\sigma = k/(N-1)$. It is clear that to minimize the number of wavelengths W needed to achieve the specified switching gain σ, it will always be the case that the wavelengths from node i to the k destinations must connect i to its k nearest neighbors, because all nodes look similar (as the network is homogeneous and traffic is uniform) and wavelength minimization demands that a wavelength traverse as few physical links as possible. In R we can support this switching gain with $W = k(k+1)/2$ distinct wavelengths; to see this we observe that the physical link from node i to i+1 will multiplex k wavelengths that terminate at i+1, k–1 wavelengths that terminate at i+2, …, and 1 wavelength that terminates at i+k. Therefore

$$W = \sum_{j=k}^{1} j = k(k+1)/2$$

is the number of wavelengths that are multiplexed onto the physical link from i to i+1 when k flows are switched from each node. This means that a switching gain of $k/(N–1)$ can be achieved in R with W wavelengths per fiber link. Solving the previous equation for k in terms of W by applying the quadratic formula, we find that

$$k = \left\lfloor \frac{(\sqrt{8W+1})-1}{2} \right\rfloor$$

The switching gain may be expressed in terms of the given network parameters W and N:

$$\sigma = \frac{(\sqrt{8W+1})-1}{2(N-1)}$$

We see that the switching gain grows no faster than the root of W in a sparsely connected topology under uniform traffic loading. Because R may be viewed as a worst-case topology, this result should be considered to be a lower bound on the benefits of adding additional wavelengths to realize greater switching gains. Nonetheless, the marginal improvement in switching gain derived from adding extra wavelengths is expected to be small, unless the physical topology is very richly connected.

Now we consider the same ring network R under uniform traffic loading when wavelength merging is used. In this case a single wavelength is used to deliver the aggregated flows from all other nodes to node i. At most N–1 wavelengths are sufficient to achieve 100% switching gain, but if only W ($W < N–1$) wavelengths are available per fiber, then the achievable switching gain is

$$\sigma_{\text{merge}} = \frac{W}{N-1}$$

Compared to the results above, the switching gain in R without wavelength merging is $O(\sqrt{W}/N)$, while the gain with wavelength merging is $O(W/N)$. The advantages of using wavelength merging are clear: performance scales linearly with the number of wavelengths, whereas performance without merging scales slowly with the square root of the number of wavelengths.

The discussion above frames the limits of performance that we can achieve by employing WDM, flow aggregation, and wavelength merging in the network. Clearly, in a poorly connected physical topology, we could unburden a node's router of a large portion of its traffic load (up to a factor that grows quadratically in the number of nodes N) by passing the traffic directly through the node's crossconnect switch. 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, we expect to use fewer than $O(N^2)$ wavelengths. We also see that in the idealized case the switching gain grows very slowly as we add wavelengths, unless some form of merging is applied. In the next section our simulations of actual networks under realistic traffic conditions will expose the practical tradeoffs between performance improvements and the number of usable wavelengths.

5 - SIMULATION AND TRAFFIC MODELS

To evaluate POW a detailed simulation has been constructed for the 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 single switch [New96, Lin97], we are interested in overall performance in a wavelength-limited environment. This performance is presumably influenced by the competition for wavelengths by different nodes. Therefore, it is imperative to simulate an entire multinode network rather than a single node.

5.1 - VINT/ns Simulation Model

The VINT/ns tool is a popular simulation package used for evaluating protocols in large-scale networks [Hua98]. 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 it is a valuable tool of many protocol designers. Most importantly for our work, it accepts as inputs log files of real packet traces. It has a large library of existing protocols.

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 SWAP signaling system (described above) for establishing lightpipes upon recognition of candidate flows. SWAP 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.

	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

TABLE 1: *vBNS Traffic Matrix*

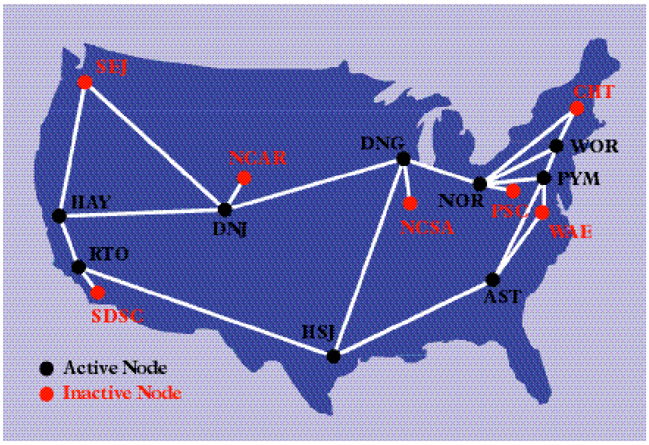


FIGURE 8: *vBNS Backbone Topology.*

The nodes are interconnected in VINT/ns according to the vBNS (very high bandwidth network service) backbone topology, which is shown in Fig.8. 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 permanent 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.

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[3]. The model of POW simulates its wavelength-management functions as well as the interactions of nodes through the SWAP signaling protocol.

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

5.2 - 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. From this data we computed 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 node 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.

	Wavelength Count				
Granularity	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%

TABLE 2: *Signaling Overhead*

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. We therefore used a real trace of about one hour’s worth of traffic. 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. Essentially, the IP addresses in the trace are nonsense (and probably do not correspond to real hosts), except that a single value is used to replace a real traced address. Since the IP addresses refer to hosts that reside on customer networks 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. Basically, this says 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 the same IP address injected at two different points will not necessarily exit the network from the same point.

It is important to note that the simulation model does not capture completely all the dynamics of the end-to-end protocols. For instance, because we use actual traces as inputs to the routers at POW points of presence, we cannot be assured that TCP behavior is being 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, we expect TCP end-to-end throughput to change adaptively. Thus, the timestamps of packets of the traces should change. Short of developing a complete simulation that includes thousands of hosts, it is not feasible to model the detailed behavior of TCP flows, and we must use the static traces as an approximation of steady flow through the network.

6 - PERFORMANCE EVALUATION

In partial response to the question ”how many wavelengths are really needed”, we simulated POW over a range of from four to 64 wavelengths (in addition to the default wavelength) deployed in the vBNS physical topology, using the traffic traces described above to drive the model. The principal performance metric that we evaluated is the switching gain, measured as the ratio of the number 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 SWAP packets to the total number of packets submitted to the network.

As a performance metric, switching gain is an indirect reflection of throughput and delay. However, it relates directly to the goals of label switching, which is to carry as much

[3] *In reporting our results we always refer to W as the number of wavelengths, assuming that the default wavelength is implicit in the architecture.*

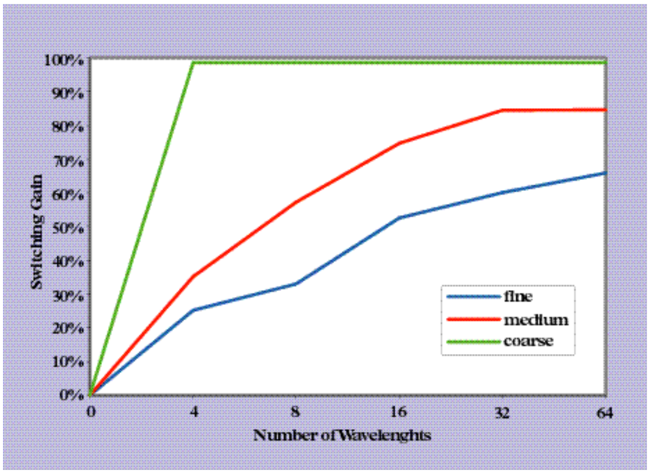


FIGURE 9: *Switching Gain vs. Number of Wavelengths for Different Flow Granularities.*

traffic as possible over the network's switching paths rather than its routing paths.

The graph of Fig. 9 shows how much traffic can be switched as we increase the wavelength count and progressively aggregate flows. As expected, the switching gain grows steadily when we increase the wavelength count 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 granularity, it can carry more that 98.59% of its traffic over dedicated lightpaths using as 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.

Being a software-based function, 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 always extra effort in the aggregation process. The wavelength-merging technology being developed for POW is a natural way to aggregate traffic by joining flows at intermediate nodes. When we simulated fine-grain flows in POW with and without wavelength merging, we saw improvement in the switching gain for all wavelength counts. As may be seen in Fig. 10, the improvement ranges from about 21% to over 52%. However, at low wavelength 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 SWAP imposes a penalty on the network by introducing overhead traffic that competes with user traffic for bandwidth and processing cycles. Although SWAP's traffic is restricted to the default wavelength,its presence on that link still deprives other unswitched traffic of bandwidth. If the overhead of SWAP is kept low as a percentage of overall traffic, then improvements in switching gain are clearly achieved. If SWAP overhead is high, then one must weigh any improvements in switching gain against the cost of this overhead.

We see in Table 2 an account of the signaling overhead as a function of flow granularity and wavelength count for POW networks without wavelength merging. On the whole, signaling overhead remains low over most of the operating regimes of POW. The exception is when the granularity is fine and the wavelength count is high. In this case SWAP is obliged to search the wavelength space for an available WDM channel. This search is exacerbated by the fact that wavelengths must be locked down temporarily while SWAP probes the entire route.Thus, much additional traffic can be placed in the network by SWAP in some circumstances. It is important to note that signaling overhead is calculated as a fraction of all offered traffic. Thus, in the 64-wavelength, fine-grain POW configuration, the 8.60% of the offered traffic that is associated with SWAP is comparable to the 34.06% of offered traffic that is not switched (see Fig.9).In the case of coarse flows,the overhead is steady regardless of the number of wavelengths available. This is because four wavelengths are completely adequate to switch essentially all eligible traffic, and the addition of wavelengths has no impact at all on the operation of SWAP.

7 - CONCLUSION

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

We also evaluated the effect of introducing wavelength-merging technology into POW, finding that the switching gain with fine-grain flows could be improved 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.

We presented the design of the simple wavelength assignment protocol and outlined its behavior. The signaling overhead that is imposed by SWAP is generally low, except when

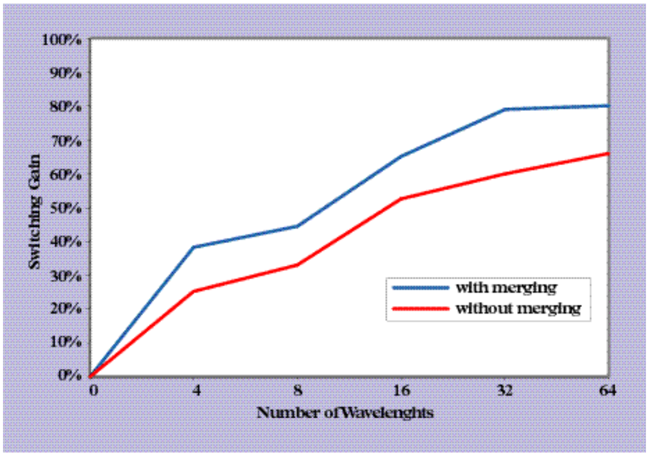


FIGURE 10: *Upper bound of wavelength without failures.*

traffic is finely aggregated. The complexity of allocating and deallocating wavelengths to flows increases with both the number of flows present and the number of wavelengths that need to be searched.In the case of fine-grain flows with a high wavelength count, overhead amounts to more that 8.60% of the offered traffic. Since the overhead is carried entirely on links shared by unswitched traffic, it can negatively impact the network's performance.

In summary, we conclude that four wavelengths are entirely sufficient to achieve very high performance when traffic is aggregated according to its ingress and egress nodes. The conclusions drawn from this study apply to one relatively small backbone. Our traffic model is based on data collected from older packet traces and traffic patterns on a lightly loaded backbone. To evaluate POW more effectively, we will need to model 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

[Ban90] J. Bannister, L. Fratta, and M. Gerla, ”Topological Design of the Wavelength-Division Optical Network”, Proc. IEEE INFOCOM ’90, pp. 11005–1013, San Francisco, Apr. 1990.

[Ban99] 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.

[Blu99] 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.

[Bro97] A. Brodnik et al., ”Small Forwarding Tables for Fast Routing Lookups”, Proc. ACM Sigcomm ’97, pp. 3–14, Cannes, Sept. 1997.

[Cal90] R. Callon, ”Use of OSI IS–IS for Routing in TCP/IP and Dual Environments”, IETF RFC 1195, Dec. 1990.

[Cal97] R. Callon et al., ”A Framework for Multiprotocol Label Switching (MPLS)”, IETF Internet Draft, Nov. 1997.

[Chl92] I. Chlamtac, A. Ganz and G. Karmi, ”Lightpath Communications: An Approach to High Bandwidth Optical WANs”, IEEE Trans. Commun., vol. 40, no. 7, pp. 1171–1182, 1992.

[Dat99] Data Communications Magazine,WDM Equipment Buyer’s Guide, http://www.data.com, Apr. 1999.

[Dav98] B. Davie, P. Doolan and Y. Rekh-ter,Switching in IP Networks, Morgan Kaufmann, San Francisco, 1998.

[Dor93] R. Dorf, ed.,The Electrical Engineering Handbook, CRC Press,Boca Raton, Fla., 1993.

[Gat99] GateD Consortium, www.gated.org.

[Gov97] R. Govindan and A. Reddy, ”An Analysis of Inter-Domain Topology and Route Stability”, Proc. IEEE INFOCOM ’97, Kobe, pp. 850–857,Apr. 1997.

[Hav99] S. Havstad et al., ”Dynamic fiber-loop-mirror-filter (LMF) based on pump-induced saturable gain or absorber gratings”,Proc. OFC ’99, San Diego, Jan.1999.

[Hua98] P. Huang, D. Estrin and J. Heidemann, ”Enabling Large-scale Simulations: Selective Abstraction Approach to the Study of Multicast Protocols”, Proc. IEEE MASCOTS ’98, pp. 241–248, Montreal, Jul. 1998.

[Lam98] 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.

[Lin97] S. Lin and N. McKeown, ”A Simulation Study of IP Switching”, Proc. ACM Sigcomm ’97, pp. 15–24, Cannes, Sept. 1997.

[Muk94] B. Mukherjee et al.,”Some Principles for Designing a Wide-Area Optical Network”, Proc. IEEE INFOCOM ’94, pp. 117– 129, Toronto, Jun. 1994.

[New96] P. Newman et al.,”IP Switching - ATM Under IP”, IEEE/ACM Trans. Networking, vol. 6, no. 2, Apr. 1998.

[Qia99] C.Qiao and M.Yoo, ”Optical Burst Switching (OBS) - A New Paradigm for an Optical Internet”, J. High Speed Networks, vol. 8, no. 1, pp. 69–84, 1999.

[Rek97] Y. Rekhter et al., ”Cisco Systems’ Tag Switching Architecture Overview”, IETF RFC 2105, Feb. 1997.

[Sch90] R. Schmidt and R. Alferness, ”Directional Coupler Switches, Modulators, and Filters Using Alternating db Techniques”, in: Photonic Switching, H. Hinton and J. Midwinter, eds., IEEE Press, New York, 1990.

[Shi97a] 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.

[Shi97b] W. Shieh and A. Willner, ”A Wavelength-Routing Node Using Multifunctional Semiconductor Optical Amplifiers and Multiple-Pilot-Tone-Coded Subcarrier Control Headers”, IEEE Photonics Tech. Lett., vol. 9, pp. 1268–1270, 1997.

[Sur99] S. Suryaputra, J. Touch and J. Bannister, ”Simple Wavelength Assignment Protocol”, USC/ISI Technical Report, Sept. 1999.

[Tur98] J. Turner, ”Terabit Burst Switching”, Tech. Rep. WUCS-9817, Washington Univ., Dept. of Comp. Sci., Jul. 1998.

[Wal97] M. Waldvogel et al., ”Scalable High Speed IP Lookups”,Proc. ACM Sigcomm ’97,pp. 25–36,Cannes, Sept. 1997.

[Zha93] L. Zhang et al., ”RSVP: A New Resource ReSerVation Protocol”, IEEE Network, vol. 7, no. 5, pp. 8–18, Sept. 1993.

[Zha98] 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, Boston, Nov. 1998.

Joe Bannister
joseph@isi.edu

Joe Bannister is the Director of the Computer Networks Division at the Information Sciences Institute and Research Assistant Professor in the Electrical Engineering-Systems Department of the University of Southern California.He received his B.A.with High Distinction in Mathematics from the University of Virginia in 1977. From UCLA he received his M.S.in Electrical Engineering, his M.S.in Computer Science, and his Ph.D. in Computer Science in 1980, 1984,and 1989,respectively. He has over 50 publications in high-speed networking, distributed computing, and network management. His research is or has been supported by the Defense Advanced Research Projects Agency and the National Science Foundation in the areas of high-speed networks, performance evaluation,wireless networks, and network management.Mr. Bannister actively participates in the research community, having served on the program committees or as chair of INFOCOM,Interop,ICNP, ICCCN, the IEEE LAN/MAN Workshop, and the IEEE TCCC Computer Communications Workshop. He is on the editorial boards of Optical Networks Magazine and Computer Networks. He has edited special issues of Computer Networks and Wireless Networks. Before joining USC/ISI he held positions at the Aerospace Corporation, System Development Corporation, Sytek, Research Triangle Institute, and Xerox. He is a member of IEEE, ACM SIGCOMM,AAAS, and Sigma Xi.

Joe Touch
touch@isi.edu

Joe Touch received a B.S.in Biophysics and CS from the Univ. of Scranton in 1985,a M.S.in CS from Cornell Univ. in 1988,and his Ph.D. in CIS from the University of Pennsylvania in 1992. Since then he has been at USC/ISI, where he is currently a Project Leader in the Computer Networks Division and a Research Assistant Professor in the Department of Computer Science. Joe has led projects ranging from gigabit LANs (ATOMIC2),NIC design (PC-ATOMIC), multicast web caching (LSAM), to his current project in the automated deployment and management of overlay networks (XBONE).He is also developing projects in optical internet WANs and LANs, fault tolerant networks, and Smart Space devices for user presence. His primary research interests include multicast variants of network management, high-speed protocols, empirical protocol performance analysis, Internet architecture, and protocols for latency reduction. Joe is a member of Sigma Xi (since 1984), IEEE, and ACM, and is the Technical Activities and Vice Chair of the IEEE TCGN (gigabit networking).He also serves on the editorial boards of IEEE Network, and Elsevier's Computer Networks, and is a member of several program committees, including IEEE Infocom (since 1994), and was co-chair of the IFIP/IEEE Protocols for High Speed Networks Workshop '99.

Stephen Suryaputra
ssuryapu@nortelnetworks.com

Stephen Suryaputra received an S.T. (B.S.E.E equivalent) degree from Sekolah Tinggi Teknik,Surabaya,Indonesia,in 1996,and an M.S.E.E degree from the University of Southern California in 1998.He is currently a senior software engineer at Nortel Networks where he is responsible for the design and implementation of Versalar 25K OC-12 ATM and OC-48 POS Linecard Hardware Simulator. His current interests are in optical networks, packet-switch and network-interface design, device drivers, protocol software implementation,and hardware/software codesign.

Alan Willner
willner@solar.usc.edu

Alan Willner received his Ph.D. in Electrical Engineering from Columbia University in 1988. He was a Postdoctoral Member of the Technical Staff at AT&T Bell Laboratories (Crawford Hill) and a Member of Technical Staff at Bellcore. He is currently a professor in the Department of Electrical EngineeringSystems at the University of Southern California.He is the Associate Director for the USC Center for Photonics Technology and is an Associate Director for Student Affairs for the NSF Engineering research Center in Multimedia.He has over 200 publications, including one book.His research is in the area of optical fiber communication systems, wavelength division multiplexing, optical networks, optical switching, optical amplification, and optical interconnections. Dr. Willner has received the following national awards:the NSF-sponsored Presidential Faculty Fellows Award from the White House,the David and Lucile Packard Foundation Fellowship in Science and Engineering, and the NSF National Young Investigator Award (formerly known as the Presidential Young Investigator Award). He is a Fellow of the Optical Society of America, a Fellow of the Semiconductor Research Corporation, and an IEEE Senior Member. He has received the Fulbright Foundation Senior Scholar Fellowship, the USC/Northrop Outstanding Junior Engineering Faculty Research Award (given annually to the highest-ranked assistant professor in engineering), the USC Outstanding Engineering Teacher Award, and the Armstrong Foundation Memorial Prize for the highest-ranked Electrical Engineering graduate student at Columbia University. Dr. Willner has served in many professional capacities, including:Vice-President for Technical Affairs for the IEEE Lasers and Electro-Optics Society (LEOS), Elected Member of the Board of Governors for IEEE LEOS, Program Chair of the IEEE LEOS Annual Meeting, Chair of the Optical Communications and Optical Networks IEEE LEOS Technical Committees, the Photonics Division Chair and Optical Communications Vice-Chair for the Technical Council of the Optical Society of America (OSA),General Co-Chair of the IEEE LEOS Topical Meeting on Technologies for a Global Information Infrastructure, Program Co-Chair for the OSA Optical Amplifier Conference, Member of the Scientific Advisory Board for the OIDA/OP, Steering Committee Member for OFC, and a Committee Member for OFC and CLEO. Prof. Willner is Editor-in-Chief of the IEEE Journal of Selected Topics in Quantum Electronics, an Associate Editor of the IEEE/OSA Journal of Lightwave Technology (JLT), a Guest Editor for the Special Issue of JLT on Multiple-Wavelength Technologies and Networks, and a Guest Editor for the Journal of Quantum Electronics Focus Issue on Fundamental Challenges in Ultra-High-Capacity Optical Communication Systems. Dr. Willner's research has been supported by many agencies including:(i) DARPA under the Ultra Photonics Program, the NGI Supernet Program, and an Optoelectronics Center, and (ii) the NSF under the Presidential Faculty Fellows Award, the National Young Investigator Award, the Interdisciplinary All-Optical Networks Initiative, and the Integrated Media Systems Engineering Research Center.

An early version of this paper was presented at the Sixth International Workshop on Protocols for High-Speed Networks. This work was partially supported by DARPA under contract MDA 972-99-C-0022.