Switched Optical Backbone for Cost-Effective Scalable Core IP Networks

Sudipta Sengupta and Vijay Kumar, Bell Laboratories, Lucent Technologies
Debanjan Saha, IBM T. J. Watson Research Center

ABSTRACT

With the advent of WDM technology, IP backbone carriers are now connecting core routers directly over point-to-point WDM links (IP over WDM). Recent advances and standardization in optical control plane technologies like GMPLS have substantially increased the intelligence of the optical layer and shown promise toward making dynamic provisioning and restoration of optical layer circuits a basic capability to be leveraged by upper network layers. In light of this, an architecture where a reconfigurable optical backbone (IP over OTN) consisting of SONET/SDH crossconnects/switches interconnected via DWDM links provides connectivity among IP routers is an emerging alternative. As carriers evolve their networks to meet the continued growth of data traffic in the Internet, they have to make a fundamental choice between the above architectural alternatives. In the current business environment, this decision is likely to be guided by network cost and scalability concerns. A reconfigurable optical backbone provides a flexible transport infrastructure that eases many operational hurdles, such as fast provisioning, robust restoration, and disaster recovery. It can also be shared with other service networks such as ATM, frame relay, and SONET/SDH. From that perspective, an agile transport infrastructure is definitely the architecture of choice. The IP-over-OTN solution is also more scalable since the core of the network in this architecture is based on more scalable optical switches rather than IP routers. But what about cost? Since the IP-over-OTN solution introduces a new network element, the optical switch, is it more expensive? In this article we address that question by comparing IP-over-WDM and IP-over-OTN architectures from an economic standpoint using real-life network data. We show that contrary to common wisdom, IP over OTN can lead to substantial reduction in capital expenditure through reduction of expensive transit IP router ports. The savings increases rapidly with the number of nodes in the network and traffic demand between nodes. The economies of scale for the IP-over-OTN backbone increase substantially when we move traffic restoration from the IP layer to the optical layer. We also compare the two architectures from the perspective of scalability, flexibility, and robustness. Our observations make a strong case in favor of a switched optical backbone for building scalable IP networks.

INTRODUCTION

With IP traffic continuing to grow at a healthy rate [1, 2], scalability of IP backbones is one important problem, if not the most important, facing service providers today. Historically, IP backbones have consisted of core routers interconnected in a mesh topology over asynchronous transfer mode (ATM) or synchronous optical network/synchronous digital hierarchy (SONET/SDH) links. With the advent of wavelength-division multiplexing (WDM) technology, service providers are now connecting core routers directly over point-to-point WDM links. This architecture, referred to as IP over WDM, is illustrated in Fig. 1a. In this figure we show an IP traffic flow from point of presence (PoP) 1 to PoP 4 passing through PoP 2 as an intermediate PoP. Note that transit traffic at PoP 2 (for this IP flow) uses IP router ports. In IP over WDM, traditional transport functions such as switching, grooming, configuration, and restoration are eliminated from the SONET/SDH layer. These functions are moved to the IP layer and accomplished by protocols like multiprotocol label switching (MPLS) [3].

In an alternative approach, referred to as IP over optical transport network (OTN), routers are connected through a reconfigurable optical backbone, or OTN, consisting of SONET/SDH optical cross-connects (OXCs) interconnected in a mesh topology using WDM links. The core optical backbone consisting of such OXCs takes over the functions of switching, grooming, and restoration at the optical layer. IP over OTN is illustrated in Fig. 1b. The IP traffic flow (as shown for IP over WDM) from PoP 1 to PoP 4 is carried on an optical layer circuit from PoP 1

0163-6804/03/$17.00 © 2003 IEEE

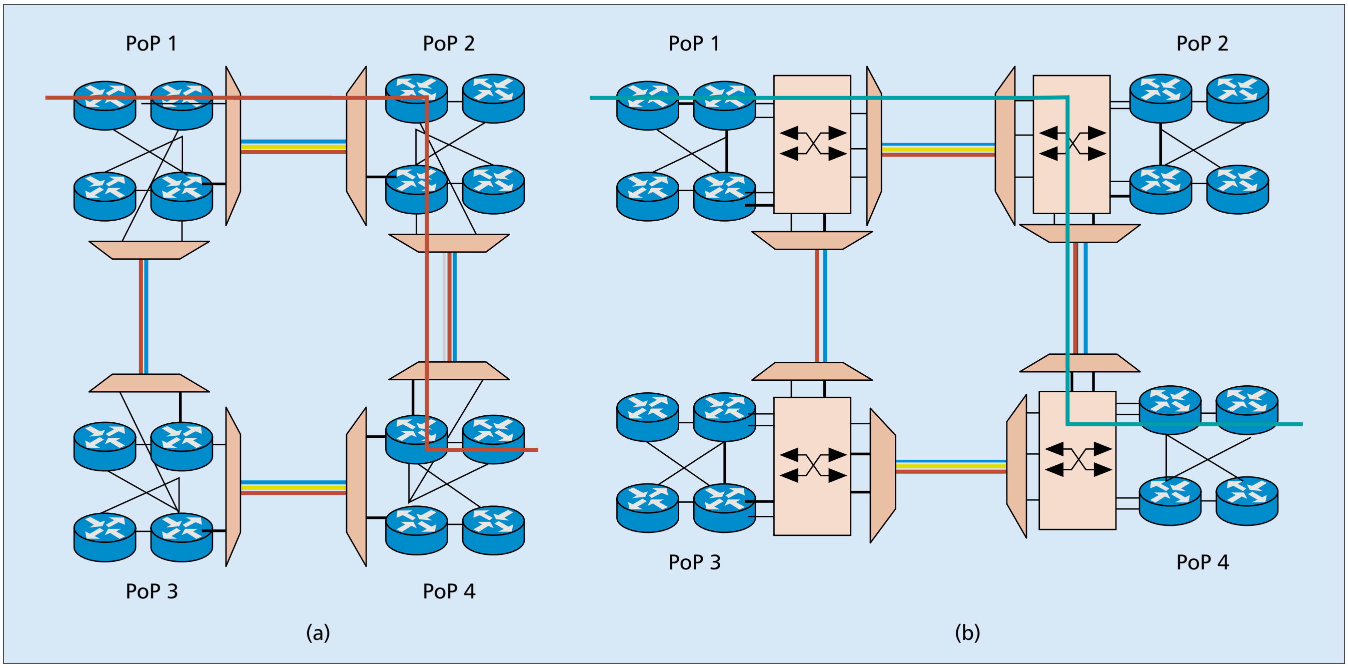


Figure 1. *Alternative architectures for interconnecting IP routers: a) IP over WDM; b) IP over OTN.*

to PoP 4. Note that in this case, the transit traffic at PoP 2 (for this IP flow) uses OXC ports that are typically a third as expensive as IP router ports. This *bypass of router ports for transit traffic* is the basis for the huge economies of scale reaped by interconnecting IP routers over an optical backbone in IP over OTN. We often use the term *lightpath* to refer to an optical layer circuit in IP over OTN.

While IP over WDM is very popular with service providers, it raises a number of issues about scalability and economic feasibility. Specifically, the ability of router technology to scale to port counts consistent with multiterabit capacities without compromising performance, reliability, restoration speed, and software stability is questionable [4]. Also, according to [5], IP routers are 200 times less reliable than traditional carrier-grade switches and average 1219 min of downtime per year. In the following paragraphs, we discuss some of the shortcomings of IP-over-WDM architecture and present the alternatives offered by an IP-over-OTN solution.

Scalability: IP routers are difficult to scale. The largest routers commercially available have 16-32 OC-192 (10 Gb/s) ports. Compare that with OXCs, which can easily support 128–256 10 Gb/s ports. The scalability of a backbone that consists of IP routers connected directly over WDM links depends directly on the scalability of the IP routers. An alternative architecture where OXCs interconnected via WDM links forms the core with IP routers feeding into the optical switches is clearly a more scalable solution.

Resiliency: In traditional IP backbones, core routers were connected over SONET/SDH links. SONET/SDH provides fast restoration, which masks failures at the transport layer from the IP layer. In IP over WDM, failures at the physical and transport layers are handled at the IP layer. For example, if there is a fiber cut or an optical amplifier failure, a number of router-to-router links may be affected at the same time, triggering restoration at the IP layer. Traditional IP layer restoration is performed through IP rerouting, which is slow and can cause instability in the network. MPLS-based restoration, a relatively new addition to IP, can be fast, but has its own scalability issues. In IP over OTN, the transport layer can provide the restoration services, making the IP backbone much more resilient.

Flexibility: One of the problems with IP-over-WDM architecture is that the transport layer is very static. Given that IP traffic is difficult to measure and traffic patterns can change often and significantly, this lack of flexibility forces network planners to be conservative and provision based on peak IP traffic assumptions. Consequently, IP backbones are underutilized and often cost more than they should. Lack of flexibility at the transport layer is also an impediment to disaster recovery after a large failure. IP over OTN alleviates this problem and provides fast and easy provisioning at the transport layer. This obviates worst case network engineering based on peak IP traffic assumptions and allows variations in traffic patterns to be handled effectively through just-in-time reconfiguration of the switched optical backbone.

Degree of Connectivity: An OXC or IP router in a typical central office (CO)/PoP has a small adjacency; it is connected to two, sometimes three, and rarely four other COs/PoPs. Because of this, it is not possible to connect IP routers with a high degree of connectivity in IP over WDM. On the other hand, because of the reconfigurable optical backbone in IP over OTN, a router can set up a logical adjacency with any other router by establishing a lightpath between them through the optical backbone. Hence, it is possible to interconnect routers in an arbitrary (logical) mesh topology in IP over OTN.

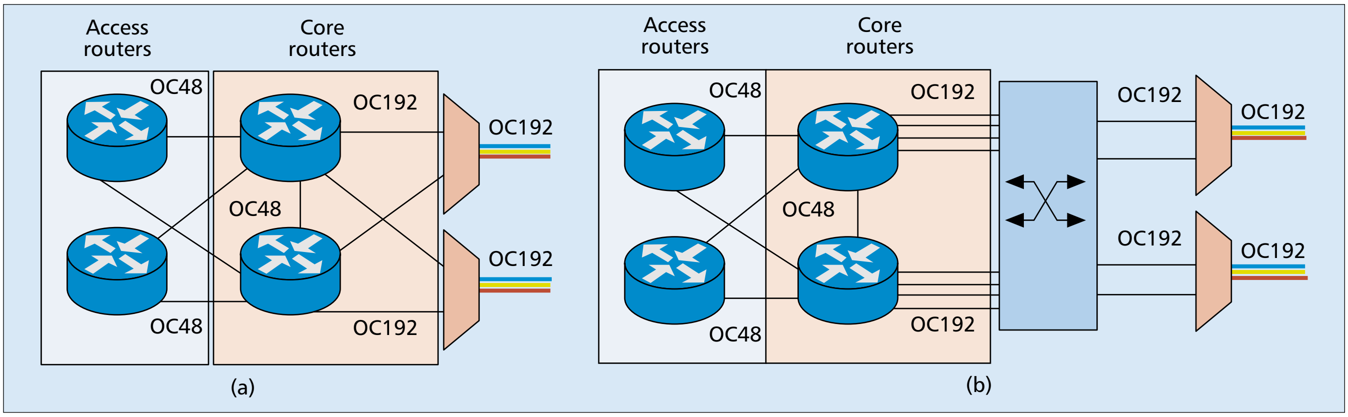


Figure 2. *PoP architectures for a) IP over WDM and b) IP over OTN.*

The arguments presented above highlight the advantages of IP-over-OTN architecture in terms of scalability, resiliency, flexibility, and degree of connectivity. The lingering question, however, is cost. IP over OTN introduces a new network element into the equation: the OXC. Does the cost of deploying the OXC into the network outweigh the potential benefits it brings? In the rest of the article we address this question using real-life network data representative of IP backbones operated by leading service providers. We show that contrary to the common wisdom, IP-over-OTN architecture can lead to a significant decrease in network cost through reduction of expensive transit IP router ports. The savings increase rapidly with the number of nodes in the network and traffic demand between nodes. The economies of scale for the IP-over-OTN backbone increase substantially when we move the restoration function from the IP layer to the optical layer.

Note that in IP-over-OTN architecture, the OXC backbone could have different switching granularity (e.g., STS-1, STS-3, or STS-48). Given that the current level of traffic in IP carrier backbones is at sub-STS-48 (less than 2.5 Gb/s) levels between PoP pairs, a lower-granularity switch provides the flexibility of grooming at the optical layer (vs. at the IP layer) and increases utilization of the OXC backbone. For the results presented in this article, we have assumed an STS-48 switched optical backbone for IP over OTN; this requires efficient packing of IP flows onto 2.5 Gb/s optical layer circuits (as discussed later). Our assumption of a wavelength switched backbone leads to conservative estimates of network cost savings with IP over OTN. The savings will increase when sub-STS-48 grooming functionality is provided by the optical layer (e.g., STS-1 switched backbone).

The rest of this article is organized as follows. We provide an overview of the network architecture, including PoP configuration and restoration options at IP and optical layers. We briefly describe the routing methodology (algorithms) used for the results reported here. We outline the network topology/traffic and equipment pricing model used. We then discuss the network cost savings with IP over OTN. We also investigate the impact of a high degree of IP layer adjacency in IP over OTN on routing protocols such as Open Shortest Path First (OSPF). Finally, we conclude the article.

Network Architecture

As mentioned before, an IP backbone consists of core routers interconnected in a mesh topology. Typically, a router is connected to its immediate neighbors. Sometimes express links are established between routers that are not physical neighbors, but exchange large volumes of traffic. For an express link, WDM terminals at each intermediate node are connected in a glass-through fashion without using IP router ports. An architecture where all IP layer links are express links was considered in [6]. In this section we discuss how the routers are interconnected in IP-over-WDM and IP-over-OTN architectures. We also present different alternatives for restoration in the two architectures.

PoP Configuration

Figure 2 shows the PoP configuration in the two different architectures. Notice that in both architectures routers are configured in a similar fashion. The routers to the left, called *access routers*, connect to the client devices, and the routers to the right, called *core routers*, connect to the transport systems. There may be more than two access routers in a PoP depending on traffic volume, traffic mix, and capacity of the routers. Most PoPs use two core routers to protect against router failures. It may be necessary to add more routers as traffic volume increases. In IP over WDM, the core routers are connected directly to the WDM systems, which connect them to neighboring PoPs. In IP over OTN, the core routers are connected to the OXCs, which in turn are connected to the WDM systems.

A client device attached to this PoP sends (and receives) 50 percent of its traffic to (from) one access router and 50 percent to (from) the other, in a load-balanced fashion. Also, the intra-PoP links connecting the access and core routers are at most 50 percent utilized. This allows either of the access routers to carry the entire traffic when the other goes down. A similar load-balancing strategy could be applied to all transit and add/drop traffic that flows through the core routers. When the core or access routers run out of port capacity, the entire quad configuration at a PoP needs to be replicated in order for the PoP to handle additional traffic.

Traffic Restoration

Restoration of service after a failure is an important consideration in carrier networks. In this section, we outline the various restoration options available in the two architectures. In IP over WDM, restoration occurs in IP layer. IP over OTN allows the flexibility of optical layer and/or IP layer restoration.

Restoration in IP over WDM — IP-over-WDM architecture allows two different restoration options: vanilla IP rerouting and MPLS-based restoration. IP rerouting is the typical mode of operation in most carrier networks today. Some service providers are exploring MPLS-based restoration in order to address some of the problems with IP rerouting.

Vanilla IP restoration: In the event of a link or node failure, routing tables change automatically to reroute around the failure. Under normal circumstances, traffic is sent along shortest paths through next-hop forwarding tables at each router. In order to accommodate restoration traffic on a link, bandwidth is overprovisioned on every link with link (router interface) utilization typically between 30 and 50 percent. One of the problems with restoration using IP rerouting is that it takes a long time (sometimes 15 minutes [5]) for the network to reach stability after a major failure. Also, network utilization has to be kept at a low level in order to accommodate rerouted traffic after a failure.

MPLS-based restoration: Each IP flow is routed over diverse primary and backup MPLS label switched paths (LSPs) for end-to-end path-based restoration. Backup paths may also protect individual links for local span-based restoration (MPLS fast reroute). We discuss these below.

Fast reroute: Fast reroute is a form of span protection. In this mode, segments of an MPLS path are protected, segment by segment, by different backup paths. Fast reroute is typically used for fast restoration around failed routers and links.

End-to-end path protection: In this mode an MPLS path is protected end to end by a backup path between the same source and destination routers. An MPLS path can be 1:1 protected, where bandwidth on the backup path is dedicated to the associated LSP. Alternatively, a shared backup path can protect it. In that case, bandwidth between different backup paths could be shared in a way that guarantees restoration for any single event failure.

For MPLS-based restoration, label mappings at routers on the backup paths are set up during LSP provisioning, so the restoration process involves just a switch at either of the end nodes of the LSP. MPLS restoration alleviates some of the problems of vanilla IP rerouting. Services are restored much faster, and sophisticated traffic engineering can improve network utilization. However, failures still affect underlying IP routing infrastructure leading to instability in the network for a prolonged period of time. Also, scalability of MPLS-based networks is still unproven, to say the least.

Restoration in IP over OTN — IP-over-OTN architecture allows multiple restoration options. IP backbones can be protected using optical layer restoration. It can also be protected at the IP layer using MPLS or IP rerouting.

IP layer restoration: This is analogous to the restoration options in IP over WDM. Lightpaths in the optical layer (which appear as express links at the IP layer) are unprotected, so failures are restored at the IP layer. For vanilla IP restoration, optical layer lightpaths (express links) are provisioned with typically at most 50 percent utilization to accommodate restoration traffic (as in IP over WDM).

Optical shared mesh restoration: Traffic is restored at the optical layer through diverse primary and backup lightpaths. Backup paths share channels in a way that guarantees complete restoration against single event failures. Thus, two backup paths can share a channel only if their corresponding primary paths are diverse (i.e., a single failure cannot affect both of them). IP layer restoration would kick in if optical layer restoration fails, say, due to multiple concurrent failures. However, since the latter is a rare event, IP layer provisioning may utilize shared mesh restoration to a higher degree.

One of the major advantages of optical layer restoration is that it masks optical layer failures from the IP layer. Consequently, IP routing is not affected even after major failures such as a fiber cut or WDM failures.

One of the major advantages of optical layer restoration is that it masks optical layer failures from the IP layer. Consequently, IP routing is not affected even after major failures such as a fiber cut or WDM failures.

Routing Methodology

In this section, we discuss how IP traffic is routed in the two architectures. Routing in IP over WDM is straightforward. For vanilla IP routing, the Dijkstra or Bellman-Ford shortest path algorithm [7] can be used. For routing MPLS LSPs, we use an enumeration-based algorithm to generate a set of candidate primary paths. For each primary path, the least-cost backup path is computed taking into account backup bandwidth sharing. Finally, the least-cost primary-backup path pair is chosen. Routing of protected MPLS LSPs is similar to routing of mesh-restored optical layer lightpaths. The latter is discussed in more detail later where the cost model for backup path bandwidth sharing is outlined.

Routing in IP-over-OTN architecture is more complex. In this case, the optical layer is flexible, allowing us to create different topologies for the IP layer. Integrated routing involving both IP and optical layers is a hard algorithmic problem and difficult to handle. Consequently, we separate the overall problem into two subproblems:

- Packing IP flows into lightpaths at the optical layer
- Routing of primary and backup lightpaths at the optical layer

Both of these problems are NP-complete [7]

Determination of backup path bandwidth sharability is based on the following rule: Two demands can share bandwidth on any common link on their backup paths only if their primary paths are link disjoint. This guarantees complete recovery from single-link failures.

and hence do not allow polynomial time exact algorithms. Before discussing algorithmic approaches to each problem, let us first try to understand why packing of IP flows is important. Typical IP flows between PoPs are currently well below WDM channel capacity (2.5–10 Gb/s). For example, in the traffic scenario considered later, the average IP traffic between any pair of nodes is about 1.7 Gb/s, which is a fraction of the bandwidth available on a single wavelength. Example 1 illustrates how intelligent packing of IP flows (beyond simple aggregation at the ingress router) can lead to increased utilization of the optical backbone.

Example 1: Consider 1.25 Gb/s of IP traffic demand between each pair of PoPs A, B, and C in a network. Simple aggregation of IP traffic at the ingress router requires one 2.5 Gb/s lightpath to be provisioned between each pair of these nodes. This creates three 2.5 Gb/s lightpaths, each 50 percent utilized. In a more efficient flow packing scenario, the IP router at node B can be used to reduce the number of lightpaths in the optical backbone as follows: provision one lightpath, L1, from A to B and another lightpath, L2, from B to C. Lightpaths L1 and L2 can carry the IP traffic between their corresponding PoP pairs. Also, the IP flow from A to C can ride on these two lightpaths with packet grooming at intermediate PoP B. This creates two 2.5 Gb/s lightpaths, each 100 percent utilized.

An integer linear programming (ILP) formulation for the problem of routing primary and shared backup paths is given in [8]. The problem of packing IP flows into 2.5 Gb/s circuits can also be formulated as an ILP. Depending on network size and the number of demands, both these ILP formulations may take a few minutes to several hours to run to completion on industry grade ILP solvers like `cplex`. Since the packing ILP for the second subproblem operates on a complete graph (there can be an optical layer connection between potentially every pair of nodes), its running time increases much more rapidly with increasing network size. For these reasons, we discuss here iterative combinatorial heuristics for both problems that perform fairly close (within 5–10 percent) to the ILP solutions.

Packing of IP Flows onto Optical Layer Circuits

In this section we discuss the packing algorithm for routing IP flows onto 2.5 Gb/s lightpaths at the optical layer. We start with the physical topology and transform it to a fully connected logical graph. Since the underlying physical network can be assumed to be biconnected — a diverse primary and backup path exists between every pair of nodes — the graph on which the packing algorithm operates is a complete graph. Each link of the graph corresponds to a protected 2.5 Gb/s lightpath. In other words, link (i, j) is representative of a 2.5 Gb/s lightpath between nodes i and j that is protected using shared mesh restoration. Each link in the logical graph is marked with a cost figure estimated to be the cost of the protected lightpath between the node pairs. Since backup paths are shared, the exact cost of the protected lightpaths cannot be determined without knowledge of the entire set of lightpaths. However, we can use an estimate of the cost of such a circuit by computing a 1 + 1 (dedicated backup) circuit and reducing the cost of the backup path by a certain factor. This factor is indicative of the savings in restoration capacity of shared backup paths over dedicated backup paths and is typically in the range of 30–50 percent, as reported in [9].

The demands to be routed are considered in some arbitrary sequence. Each IP flow is routed one by one on the logical graph using the Dijkstra or Bellman-Ford shortest path algorithm. In order to understand how routing is performed, let us consider an IP flow of bandwidth b. For each link (i, j) in the logical graph, let T_{ij} be the remaining bandwidth on the last (partially utilized) lightpath from i to j. If $b \leq T_{ij}$, routing the flow on logical link (i, j) does not require provisioning a new lightpath from i to j, and we set the cost of this link as $c_{ij}^{\text{pack}} = c_{ij}^{\text{orig}} * K_1$, where c_{ij}^{orig} is the original cost of the link in the logical graph (as discussed above). Otherwise, routing the flow on logical link (i, j) requires provisioning of a new lightpath from i to j, and we set the cost of this link as $c_{ij}^{\text{pack}} = c_{ij}^{\text{orig}} * K_2$. Here, K_1 and K_2 are constants, and the ratio K_2/K_1 should be large enough to bring out the difference of provisioning a new lightpath at the optical layer vs. utilizing capacity on an existing one. For example, one could choose $K_1 = 0.1$ and $K_2 = 1$.

Finally, since this is an offline planning scenario where all the demands are available at once, multiple passes can be made on the demand sequence and during each such pass, the packing of each IP flow can be recomputed. From our implementation experience, most of the benefit of further optimization is obtained over the second and third passes and further iterations are not required.

Routing of Primary and Backup Paths on Physical Topology

In this section we discuss routing of primary and backup paths. The same algorithm is used to route lightpaths in the optical layer in IP-over-OTN architecture and MPLS LSPs in IP-over-WDM architecture. The optimization problem involves finding the primary and shared backup path for each demand so as to minimize total network cost.

Consider the demands to be routed in some arbitrary sequence. For a given demand, a list of candidate primary paths is enumerated using Yen's K-shortest path algorithm [10]. For each choice of primary path, a link disjoint back path is computed as follows. First, links that belong to the primary path are removed from the network graph. This ensures that the backup path corresponding to this primary path is link disjoint from the primary path. Second, the cost of each remaining link is set to 0 (or small value) if the link contains sharable backup channel bandwidth. Otherwise, the cost is set to the original cost. This transformation helps encourage sharing bandwidth on the backup path. A shortest cost path is then computed between the source and destination, and set as the backup path for the current primary path. Finally, the primary-backup path pair with the least cost is chosen. Determination of backup path bandwidth sharability is based on the following rule: Two demands can share bandwidth on any common link on their backup paths only if their primary paths are link disjoint. This guarantees complete recovery from single-link failures.

Since this is an offline planning scenario where all demands are available at once, multiple passes can be made on the demand sequence, and during each such pass, the primary and backup path of each demand can be rerouted. As before, we observed that most of the benefit of further optimization is obtained over the second and third passes, and further iterations are not required.

Network and Cost Models

In the network studies discussed below, we used representative but desensitized topological and traffic data from service providers in the United States. Interestingly, most service providers' networks have similar topologies and traffic distribution. This is not surprising given that demographics drive traffic patterns, which in turn affect network buildouts. We also used realistic industry pricing for different equipment (IP routers and optical switches) in the study.

Network Topology and Traffic

We consider a representative carrier backbone topology for comparing the two architectures from a carrier economics standpoint. The 12-node U.S. backbone is shown in Fig. 3. The network has 17 links, and hence a small average node degree of 2.8. Carrier backbone physical topologies are characterized by small node degrees due to limitations in acquiring right of way and laying out fibers. In fact, a typical CO/PoP in a carrier network has two, in some cases three, and on rare occasions four conduits connecting it to neighboring central offices.

The traffic matrix for this topology specifies 66 bidirectional IP demands, one between each pair of nodes. The average traffic between a node pair is 1.7 Gb/s, the maximum is 8 Gb/s, and the minimum is 80 Mb/s. We also consider higher traffic scenarios by multiplying the traffic matrix by a factor of 2, 4, 8, and 16. In the results presented below, traffic matrices are represented as x1, x2, x4, x8, and x16. Traffic mix and volume (x1) is representative of major backbone service providers in the United States. Although Internet traffic growth has slowed, most service providers reported 100 percent year-over-year growth in 2002 [1]. Another way of looking at the traffic matrices is to consider x1–x16 as the estimated traffic matrices over a five-year planning horizon with x1 as year 1 traffic and x16 as year 5 traffic at 100 percent year-over-year growth rate.

Equipment	Common equipment (chassis)	2.5G interface	10G interface
Router (64 × 64 2.5G port equivalent)	\$50,000	\$37,000	\$119,000
Optical switch (256 × 256 2.5G port equivalent)	\$500,000	\$13,000	\$41,000

Table 1. *Industry pricing for IP routers and optical switches.*

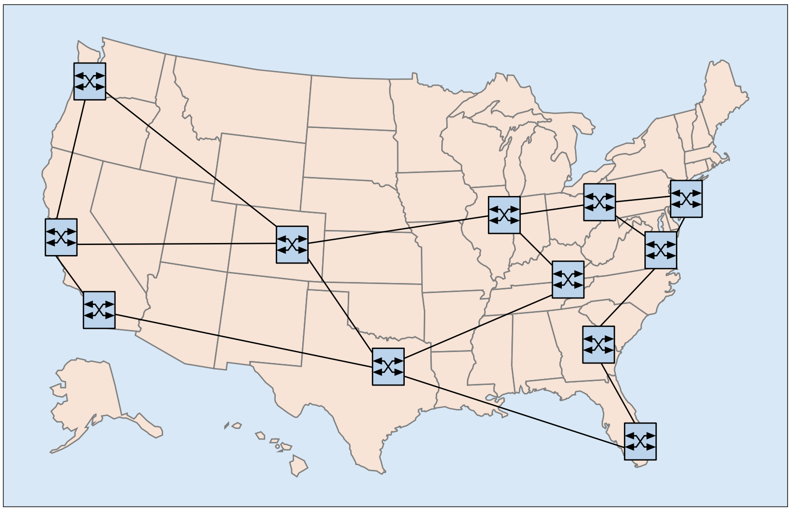


Figure 3. *The 12-node U.S. backbone topology.*

Equipment Pricing Model

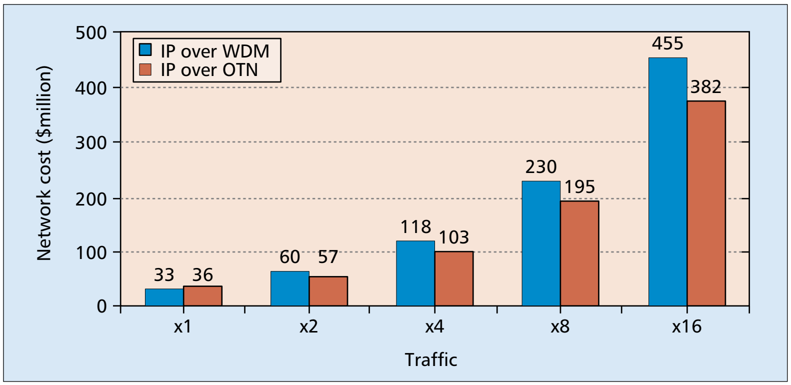
Each IP router and OXC has a common equipment cost and a per-port cost for its 2.5 Gb/s and 10 Gb/s interfaces. An IP router or OXC at a PoP may not be fully loaded, so the per-port cost is incurred only for the required number of routers and OXC ports at each PoP. The overall cost of routing a given traffic demand is the sum of the common equipment and port costs at each PoP. We ignore the cost of WDM systems since it turns out to be about the same in both architectures and does not affect the relative cost analysis.

The pricing structure is outlined in Table 1. This is representative of industry pricing of IP routers [11] and OXCs [12, 13] in the current market (in 2003). We use a chassis port count consistent with capacity of currently shipped products: IP routers can have up to 64 and OXCs up to 256 2.5 Gb/s equivalent ports.

Network Cost Computation

We have developed a sophisticated tool for routing traffic in the IP-over-WDM and IP-over-OTN architectures. For the purpose of the study, we used this tool to find the lowest cost solution in both architectures given the traffic demand and the network topology. The output of the routing algorithm consists of the following:

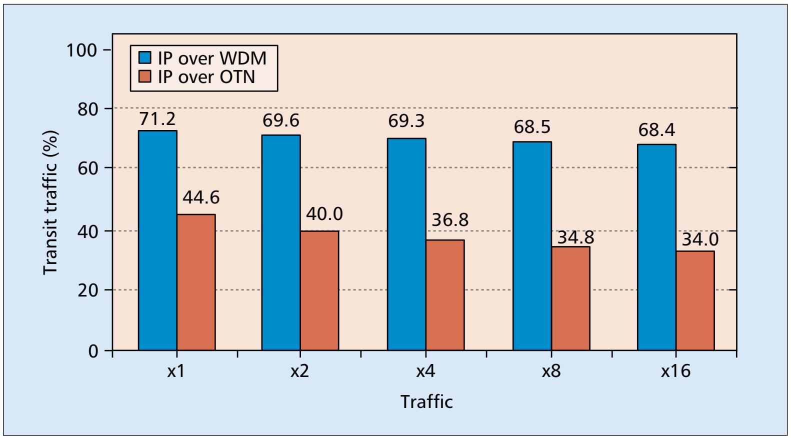
- Number of router drop and line ports at each PoP in IP over WDM and IP over OTN, and number of OXC drop and line ports at each PoP in IP over OTN
- Number of IP routers at each PoP in IP over WDM and IP over OTN, and number of OXCs at each PoP in IP over OTN



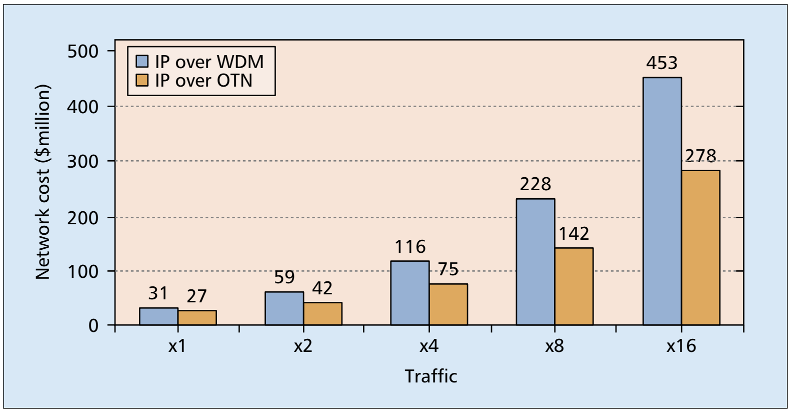
■ **Figure 4.** *Network cost (excluding WDM lambda) with IP layer restoration for both IP over WDM and IP over OTN; $73 million (16 percent) savings with IP over OTN for x16 traffic.*

- Number of WDM channels on each link (total number about the same for both architectures)

In IP over WDM, each path traverses two or four router ports at each intermediate PoP it traverses, depending on whether it enters and exits out of the same or different router(s), respectively. On the average transit traffic passes through three ports in IP over WDM. In IP over OTN, a lightpath traverses two OXC ports at each of the intermediate PoPs it passes through. An IP layer path that traverses multiple express links (optical layer lightpaths) in IP over OTN consumes router ports at intermediate PoPs in accordance with the same model outlined for IP over WDM. Since access routers carry only add/drop traffic, which is the same for both architectures, they have not been included in the network cost analysis.



■ **Figure 5.** *Percentage of router transit traffic (averaged over all PoPs) with IP layer restoration for both IP over WDM and IP over OTN.*



■ **Figure 6.** *Network cost (excluding WDM lambda) with MPLS restoration for IP over WDM and optical mesh restoration for IP over OTN; $175 million (39 percent) savings with IP over OTN for x16 traffic.*

Case Study Results

In this section we present results from our case studies comparing the costs of IP-over-WDM and IP-over-OTN solutions. In order to understand the impact of different architectural choices, we have considered two different network scenarios. First, we analyze the impact of transit router bypass. We then quantify the impact of optical layer restoration on network cost. Finally, we discuss the benefits of dynamic optical layer provisioning at a qualitative level.

The Impact of Transit Router Bypass

Figure 4 plots the network cost with IP layer restoration for both IP-over-WDM and IP-over-OTN. In order to accommodate restoration traffic on a link in both architectures, bandwidth is over-provisioned on every link with router interface utilization set to at most 50 percent (a typical value used by carriers). The analysis is over a range of up to 16 times (x16) growth from initial traffic (x1). Since *WDM lambda costs are about the same in both architectures* (as produced by network planning), they are left out of the network cost. IP over OTN gives a network cost savings of $73 million (16 percent) for x16 traffic and is a less expensive architectural choice from x2 traffic onward. This comparative scenario shows the effect of *just moving transit traffic from routers to OXC ports*.

IP over OTN reaps huge economies of scale by moving transit traffic from routers to optical ports. In Fig. 5 we plot the router transit traffic as a percentage of the total traffic (drop and transit) at a router PoP and averaged over all PoPs for the scenarios considered in Fig. 4. The router transit traffic percentage for IP over OTN is about 68–70 percent. IP over OTN clearly decreases the router transit traffic percentage. Note, however, that for IP layer restoration, restoration traffic still flows in the router (IP) domain for IP over OTN.

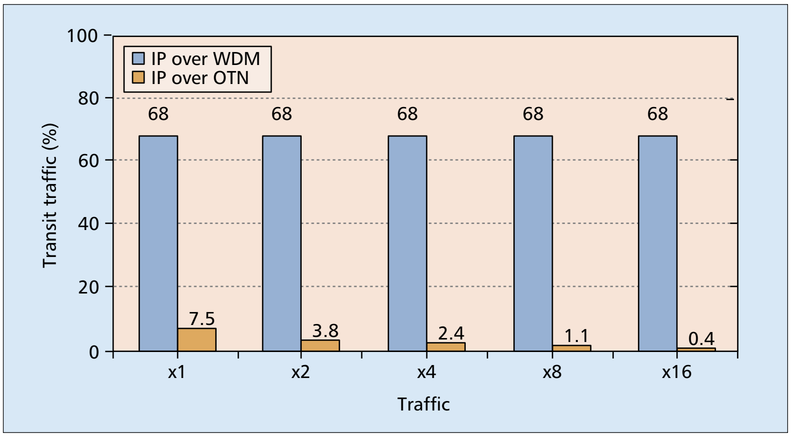
The Impact of Restoration at the Optical Layer

Figure 6 plots the network cost with MPLS shared backup path restoration for IP over WDM and optical shared mesh restoration for IP over OTN. The analysis is over a range of up to 16 times (x16) growth from initial traffic (x1). As before, WDM lambda costs are left out of the network cost. IP over OTN gives network cost savings of $175 million (39 percent) for x16 traffic. IP over OTN is a less expensive architectural choice as early as x1 traffic onward.

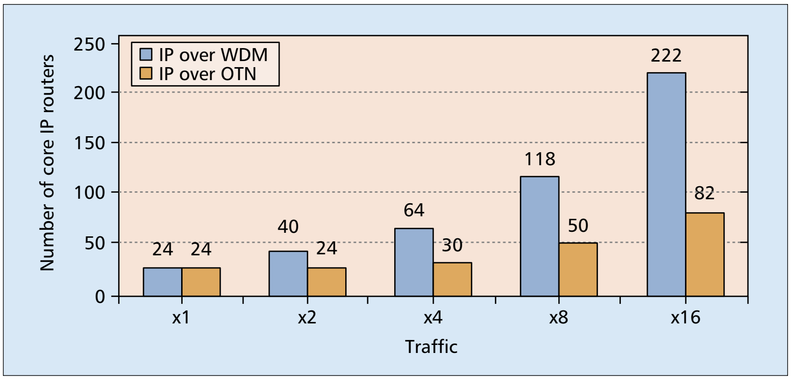
The full benefit of an optical backbone is realized when restoration occurs at the optical layer for IP over OTN through shared mesh restoration (the scenario in Fig. 6). Apart from the network cost savings, this is also evident when we consider router transit traffic decrease when moving from IP over WDM to IP over OTN. In Fig. 7 we plot the router transit traffic as a percentage of the total traffic (drop and transit) at a router PoP and averaged over all PoPs for the scenario considered in Fig. 6. The router transit traffic for IP over OTN is much lower (less than 8 percent for x1 traffic and dropping down to 0.4 percent for x16 traffic) when restoration occurs at the optical layer than when it occurs at the IP layer (34–44 percent).

The decrease in router transit traffic in IP over OTN translates directly to a reduction in the number of core IP routers and associated network cost savings (note that the number of access routers is the same in both architectures since it is determined by add/drop traffic only). In Fig. 8 we plot the number of core IP routers required for routing traffic in the two architectures. Note that since the routers are deployed in a dual redundant configuration, we require at least two core routers at each PoP for a total of 24 in the entire network. For IP over WDM, the number of core IP routers grows from 24 for x1 traffic to 222 for x16 traffic. For IP over OTN, the number of core IP routers grows to just 82 for x16 traffic, a 63 percent reduction in IP routers. This not only saves network cost but also eases IP traffic management; router quad configuration at a PoP scales for a longer time with traffic growth, and router upgrades become less frequent.

At present, the common equipment cost of an OXC is several times that of a router, and this is responsible for offsetting some of the savings obtained by bypassing traffic through OXC interfaces in IP over OTN as opposed to carrying traffic on IP router interfaces in IP over WDM. In the future, as OXC technologies become cheaper and more scalable in terms of port count, common equipment and per-port OXC costs will go down further, increasing the economic attractiveness of IP over OTN over IP over WDM.



■ **Figure 7.** *Percentage of router transit traffic (average over all PoPs) with MPLS restoration for IP over WDM and optical mesh restoration for IP over OTN.*



■ **Figure 8.** *Number of core IP routers with MPLS restoration for IP over WDM and optical mesh restoration for IP over OTN; 63 percent decrease in core IP routers with for IP over OTN for x16 traffic.*

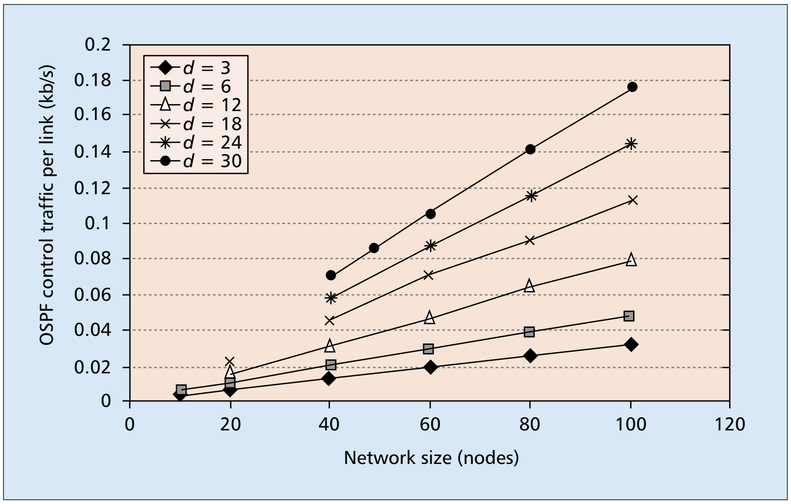
The Impact of Optical Layer Reconfiguration

One of the problems in designing IP networks is estimating the traffic between PoPs accurately. Even if traffic can be estimated, it can change often and by large amounts due to various factors. In order to alleviate this problem, IP network designers often provision networks based on peak IP traffic assumptions. This leads to underutilization of the network and bloated cost. Fast provisioning of the transport network offered by IP over OTN architecture makes it possible to dynamically reconfigure the backbone as traffic demand changes. Since an optical layer connection corresponds to an IP layer (logical) link, fast reconfiguration at the optical layer allows dynamic changes to the IP topology in response to changing traffic patterns. Thus, the carrier need not overprovision the network in advance to accommodate peak IP traffic. Rather, periodic variations (time of day, week, month, etc.) in IP traffic patterns can be handled effectively through just-in-time reconfiguration of the switched optical backbone.

This dynamic optical layer reconfiguration, enabled by a distributed optical control plane, requires interoperability between the IP (client) network and the optical backbone through a mechanism for IP routers to request bandwidth from the optical layer. The user-to-network interface (UNI) [14] is one such protocol that has been proposed by the Optical Internetworking Forum (OIF) for this purpose. Over the longer term, tighter integration between the IP and optical layers is expected to be achieved through the generalized MPLS (GMPLS) architecture [15] being developed by the Internet Engineering Task Force (IETF).

The Impact on Routing Protocols

In the previous section we discussed how IP over OTN leads to cost savings by bypassing IP traffic through the optical layer. In other words, IP over OTN increases the degree of connectivity



■ **Figure 9.** *OSPF control traffic per link for network size of up to 100 nodes and node degree* d *up to 30.*

of the underlying transport network. Degree of connectivity has a direct impact on the control traffic exchanged between routers. Specifically, an increase in the degree of connectivity increases the volume of control traffic and processing power required at the routers to process them. In this section we discuss the impact of degree of connectivity on routing protocols with OSPF [16] as an example.

OSPF uses a reliable flooding mechanism [16] to disseminate topology information. Link state advertisements (LSAs) are generated by each router describing its local piece of the routing topology (i.e., adjacent links and neighbors). Five different types of LSAs are defined in OSPF; they are:

Router LSA: The router LSA includes state and cost of all point-to-point links that terminate on a router. There is only one router LSA associated with a router. (In optical networks, all links are point to point and hence can be captured in router LSAs).

Network LSA: This includes a representation for each broadcast network (e.g., Ethernet, token ring). (Network LSAs are not necessary in optical networks)

Summary LSA: For reasons of scalability, OSPF networks are often divided into multiple areas. Router and network LSAs pertaining to nodes and links in an area are contained within the area. Information about reachable destinations within an area is sent in summary LSAs to nodes outside the area. (Summary LSAs may be used in optical networks with multi-area routing).

Other LSAs: Besides router, network, and summary LSAs, OSPF uses other LSAs. Routes learned form other autonomous systems are distributed using external and autonomous system border router (ASBR) LSAs. Opaque LSAs provide a standard way to extend OSPF. GMPLS optical extensions [17] to OSPF use this mechanism to disseminate resource information.

For the following analysis, we make the assumption that there is a single OSPF area. This is conservative since it leads to overestimation of the control traffic compared to multi-area OSPF implementations. For single-area OSPF deployments, we need to consider only router LSAs. Reliable flooding of a particular LSA is either routine (every 30 min, also called *routine refresh*) or triggered by a change in the content of the LSA (e.g., in link status). We use routine refresh traffic as an estimate of the total OSPF control traffic. This is realistic since change triggered LSA updates due to link failures are rare and would constitute a negligible fraction of the total control traffic.

The average size of a router LSA is (24 + 12 * *d*) bytes, where *d* is the average node degree (router adjacency). In calculating the average OSPF control traffic due to router LSAs, it is helpful to note that when an LSA is flooded, either the LSA or its acknowledgment (ACK) is sent on every link, but not both [16]. Let *n* be the number of nodes in the network. Thus, routine refresh of router LSAs generates *n**(24 + 12**d*)/1800 bytes/s bandwidth per link on the average. Using the above formula, we see that the link bandwidth is in the sub-kilobit-per-second range even for large networks with high IP layer connectivity. In Fig. 9 we consider networks of up to 100 nodes having node degrees up to 30.

We now investigate the scalability of OSPF control traffic as we move from IP over WDM to IP over OTN for the topology and traffic considered earlier. Concerns about OSPF scalability for IP over OTN spring from the fact that the degree of connectivity of the IP layer topology increases on deploying a switched optical mesh backbone. The overhead of reliable flooding could be expected to increase due to the above. Can we estimate this increase? Is it within an acceptable range?

Traffic	#IP routers (*n*)	Total number of (OC-192) links (*m*)	Average router adjacency (*d*)	Average OSPF control traffic bandwidth/ link (kb/s)	Control traffic per link per refresh interval (kbytes)	Average #LSAs received per node after single link status change
X1	24	67	5.58	0.0097	2.18	11.2
X2	40	123	6.15	0.0174	3.92	12.3
X4	64	242	7.56	0.0326	7.34	14.1
X8	118	474	8.03	0.0631	14.2	16.1
X16	222	943	8.5	0.1243	27.97	17

■ **Table 2.** *OSPF control traffic for IP over WDM (12-node U.S. network).*

Traffic	#IP routers (n)	Total number of (OC-48) links (m)	Average router adjacency (d)	Average OSPF control traffic bandwidth/ link (kb/s)	Control traffic per link per refresh interval (kbytes)	Average #LSAs received per node after single link status change
X1	24	59	4.92	0.0089	1.99	9.58
X2	24	101	8.42	0.0133	3.00	16.8
X4	30	191	12.73	0.0236	5.30	25.5
X8	50	369	14.76	0.0447	10.06	29.5
X16	82	722	17.61	0.0858	19.30	35.2

■ **Table 3.** *OSPF control traffic for IP over OTN (12-node U.S. network).*

OSPF Control Traffic for IP over WDM

For IP-over-WDM architecture, the number of nodes n is equal to the number of IP routers and is given in column 2 of Table 2. An IP layer adjacency corresponds a WDM OC-192 link. Hence, the total number of links m in the network is given by column 3. Using this, we compute the average node degree in column 4 as $d = 2*m/n$, and the average OSPF control traffic bandwidth per link in column 5 as $n*(24 + 12 * d)/1800$ bytes/s (shown in kilobits per second).

The average control traffic reflects the fact that the refreshes originated by different routers do not occur at the same time. In the very unlikely event that they do, the total amount of traffic (in kilobytes) passing over a link (over a small time interval) is given in column 6. Finally, notice that in the case of a change triggered update for a link, the two routers adjacent to the link refresh their router LSAs. Thus, any router in the network processes at most both these LSAs received on each of its incoming links, the average being 2 * d. This is shown in column 7. Given that an OSPF adjacency in this case corresponds to an OC-192 (10 Gb/s) link, the control traffic shown in Table 2 for IP over WDM is negligibly small.

OSPF Control Traffic for IP over OTN

For IP-over-OTN architecture, the number of nodes n is equal to the number of IP routers (as before) and is given in column 2 of Table 3. However, in this case, an IP layer adjacency corresponds to an optical layer OC-48 lightpath. Hence, the total number of links m in the network is given by column 3. Using this, we compute the average node degree in column 4 as $d = 2 * m/n$, and the average OSPF control traffic bandwidth per link as $n*(24 + 12 * d)/1800$ bytes/s (shown in kilobits per second).

As mentioned earlier, we use routine refresh traffic as an estimate of the total OSPF control traffic. Failures of IP layer (logical) links in IP over OTN are even less frequent than in IP over WDM, since they are protected by optical layer restoration. Note that we might have addition and deletion of links corresponding to provisioning and deletion of lightpaths. However, lightpath hold times would be on the order of at least hours (e.g., for short-term time-of-day variations in IP traffic). Hence, the additional control traffic due to such changes would be much less than that for routine refresh.

Comparing Tables 2 and 3, we see that the average OSPF control traffic bandwidth per link is *actually smaller* for IP over OTN. This can be explained by the fact that in moving from IP over WDM to IP over OTN, an increase in the average router adjacency (IP layer meshiness) is more than offset by a decrease in the number of IP routers attributable to optical layer bypass of transit traffic.

Conclusion

The network analysis presented in this article leads to a number of insightful observations. We observe that for any given physical transport topology, the volume of transit traffic and number of transit interfaces grow rapidly with traffic. Hence, as traffic increases, IP-over-OTN architecture drives the network cost down by moving transit traffic from the IP layer to the optical layer. We also observe that reduction in transit traffic is much higher when restoration occurs at the optical layer than when restoration occurs at the IP layer. Consequently, restoration at the optical layer further reduces network cost. Although not presented here, cost savings from IP-over-OTN architecture increase as the network grows in terms of the number of backbone PoPs.

As mentioned before, IP-over-OTN architecture is also more scalable, flexible, and robust than IP-over-WDM architecture. We have also investigated the effect of increased degree of adjacency (logical meshiness) at the IP layer in IP over OTN on IP layer routing (i.e., control traffic and processing overhead) in the context of a link state routing protocol like OSPF. Our analysis shows that OSPF protocol overheads remain within acceptable levels in IP over OTN, and hence, an increased degree of connectivity at the IP layer does not impose significant overheads on IP layer routing in IP over OTN. We would like to conclude by pointing out that a switched optical backbone can also be used as a shared common infrastructure for other services like ATM, frame relay, and voice traffic.

References

[1] "Internet Traffic Soars, But Revenues Glide," *RHK Inc. Industry Report*, May 2002.

[2] "IP Markets Will Enjoy Shelter from the Telecom Storm," *RHK Inc. Industry Report*, Dec. 2002.

Restoration at the optical layer further reduces network cost. Although not presented here, cost savings from IP-over-OTN architecture increases as the network grows in terms of the number of backbone PoPs.

[3] B. Davie and Y. Rekhter, *MPLS: Technology and Applications*, Morgan Kaufmann, Jan. 2000.

[4] M. Reardon and S. Saunders, "Terabit Trouble," *Data Communications*, Aug. 1999, pp. 11–16.

[5] C. Labovitz, A. Ahuja, and F. Jahanian, "Experimental Study of Internet Stability and Wide-Area Backbone Failures," Univ. of MI tech. rep. CSE-TR-382-98.

[6] S. Phillips, N. Reingold, and R. Doverspike, "Network Studies in IP/Optical Layer Restoration," *OFC 2002*, Mar. 2002.

[7] T. Cormen, C. Leiserson, and R. Rivest, *Introduction to Algorithms*, McGraw-Hill, June 1990.

[8] B. T. Doshi *et al.*, "Optical Network Design and Restoration," *Bell Labs Tech. J.*, vol. 4, no. 1, Jan.–Mar. 1999.

[9] R. Ramamurthy *et al.*, "Capacity Performance of Dynamic Provisioning in Optical Networks," *IEEE/OSA J. Lightwave Tech.*, vol. 19, no. 1, Jan. 2001.

[10] J. Y. Yen, "Finding the K Shortest Loopless Paths in a Network," *Mgmt. Sci.*, vol. 17, no. 11, July 1971.

[11] "Worldwide Router Market Forecast," *Dell 'Oro Group Industry Report*, Jan. 2003.

[12] R. Kline, A. Dwivedi, and D. Cooperson, "Optical Networks: North America Market Forecast," *RHK Inc. Industry Report*, Oct. 2002.

[13] "Worldwide Optical Switch Market Forecast," *Dell 'Oro Group Industry Report*, Jan. 2003.

[14] OIF, "User Network Interface (UNI) v1.0 Signaling Specification," Oct. 2001.

[15] E. Mannie *et al.*, "Generalized Multi-Protocol Label Switching (GMPLS) Architecture," Internet draft, <draft-ietf-ccamp-gmpls-architecture-05.txt>, Mar. 2003.

[16] J. Moy, *OSPF: Anatomy of an Internet Routing Protocol*, Addison-Wesley, Jan. 1998.

[17] K. Kompella *et al.*, "OSPF Extensions in Support of Generalized MPLS," Internet draft, <draft-ietf-ccamp-ospf-gmpls-extensions-09.txt>, Dec. 2002.

Biographies

Sudipta Sengupta (sudipta@research.bell-labs.com) is currently in the Optical Networks Research Department at Bell Laboratories, Lucent Technologies. In addition to pursuing research in network protocols and routing algorithms, he is also a lead architect for distributed control plane architecture for Lucent's optical networking product portfolio. Prior to this, he was a senior architect at Tellium and worked on the dynamic provisioning and restoration architecture and routing algorithms for one of the industry's earliest mesh optical core switch. Previously, he was a senior project leader for the design and development of the wireless networking platform for Oracle Corporation's Mobile Applications Division. He holds an M.S. degree from the Massachusetts Institute of Technology (MIT), Cambridge, and a B.Tech. degree from the Indian Institute of Technology (IIT), Kanpur, both in computer science. He received the President of India Gold Medal at IIT-Kanpur for academic excellence. He has authored numerous publications for conferences, journals, and technical magazines, and has filed U.S. patents in the area of computer networking. He has also taught advanced courses on optical networking at academic/research and industry conferences.

Debanjan Saha is currently with the Network Services and Software Department at IBM T. J. Watson Research Center. He is one of the primary authors of GMPLS standards in IETF. He also serves as editor of international journals and magazines, and is a technical committee member of workshops and conferences. He is a notable author of numerous technical articles on networking, and is a frequent speaker at academic and industry events. He holds a B.Tech. degree from IIT, and M.S. and Ph.D. degrees from the University of Maryland at College Park, all in computer science.

Vijay P. Kumar is director of Intelligent Optical Networks Research at Bell Laboratories, Lucent Technologies, where he leads research in architectures, algorithms, and protocols for the optical network control plane. Prior to undertaking this position in 2001, he was director of High Speed Networks Research in Bell Laboratories and led the teams that created the ATLANTA ATM chip set (the industry-leading silicon solution for ATM switches) and the PacketStar IP router (the industry's first gigabit router with per-flow QoS). He received his B.Eng. (electronics and communication engineering) from Osmania University, and M.S. and Ph.D. degrees (electrical and communication engineering) from the University of Iowa. He is a recipient of the 1998 Ellersick Prize Paper Award, and a Fellow of Bell Laboratories.