From Network Design to Dynamic Provisioning and Restoration in Optical Cross-Connect Mesh Networks: An Architectural and Algorithmic Overview

Sudipta Sengupta and Ramu Ramamurthy, Tellium, Inc.

Abstract

This article presents a broad overview of the architectural and algorithmic aspects involved in deploying an optical cross-connect mesh network, starting from the network design and capacity planning phase to the real-time network operation phase involving dynamic provisioning and restoration of lightpaths and online algorithms for route computation. Frameworks for offline design and capacity planning of optical networks based on projected future lightpath demands are discussed. The essential components of an IP-centric control architecture for dynamic provisioning and restoration of lightpaths in optical networks are outlined. These include neighbor discovery, topology discovery, route computation, lightpath establishment, and lightpath restoration. Online algorithms for route computation of unprotected, 1+1 protected, and mesh-restored lightpaths are discussed in both the centralized and distributed scenarios.

To accommodate the exponential growth of the Internet, transport networks based on wavelength-division multiplexing (WDM) technology [1] are increasingly being deployed in carrier networks. WDM technology harnesses the enormous bandwidth of the fiber (potentially tens of terabits per second) by multiplexing hundreds of optical channels, each of which operates at speeds of several gigabits per second. Point-to-point WDM links that multiplex several tens of optical channels are currently being deployed in the carrier networks.

Figure 1 illustrates the architecture of an optical network where multiple optical cross-connects/switches (OXCs) are interconnected via WDM links in a general topology (referred to as an *optical mesh network* [1]). An optical mesh network allows for flexible use of wavelength channels, enabling capacity-efficient provisioning and restoration. An OXC has multiple ports and is capable of switching an optical wavelength channel (e.g., at OC-48, OC-192 rates) from an input port to an output port. In general, optical switches can be purely optical, electronic, or a combination of optical and electronic, depending on the degree to which signals remain in the optical or electronic domains within the switch. In this article we assume that the optical network allows for full wavelength conversion (i.e., any wavelength channel on an input port can be switched to any wavelength channels on output ports). These switches typically have an electronic switch fabric and perform opto-electronic conversion of the incoming light signal for intelligent processing. They are referred to as *optical-electronic-optical* (O-E-O) switches. Wavelength conversion is enabled by the O-E-O switch and/or transponders at the WDM system.

An optical network allows dynamic provisioning of *optical layer connections*, called *lightpaths*, between clients connected to the network. The provisioning activity consists of establishing suitable cross-connects in each OXC in the connection path such that end-to-end connectivity is realized. Lightpaths can be protected against failure of network components such as lasers, fiber links, and nodes. The interaction between OXCs in a network is over a well-defined signaling and routing interface. This interface includes both IP and multiprotocol label switching (MPLS)-based protocols for provisioning of optical layer connections as well as proprietary protocols for restoration of such connections after failures. A consensus is emerging in the industry on utilizing an IP-centric control plane within optical networks to support dynamic provisioning and restoration of optical layer connections. Figure 1 also depicts the interfaces to client networks (the user–network interface, UNI [2]) and to other external optical (sub)networks (the network–network interface, NNI [2]). Details of these interfaces are not considered in this article.

A *lightpath* in an optical network is an alternating sequence of OXCs and ports/channels traversed by the connection, starting with the source OXC and ending at the destination OXC. A lightpath in an optical network is typically bidirectional. A port in an OXC contains both a transmitter and a receiver (transceiver card), thus allowing two unidirectional paths (traversing the same set of OXC ports) to be set up in either direction between the source and destination. The following three types of lightpaths are considered in this article:

- **Unprotected lightpaths**: An unprotected lightpath is not protected upon the failure of any resources (fiber links,

0890-8044/01/$10.00 © 2001 IEEE

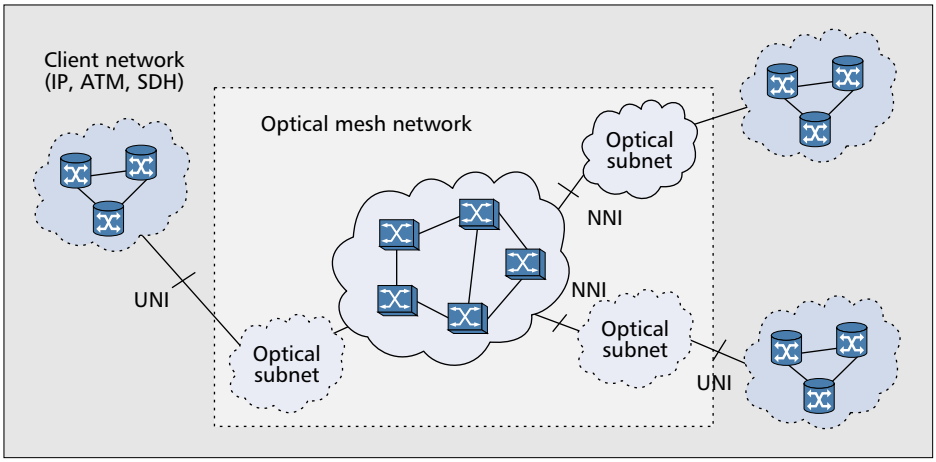


Figure 1. *An optical mesh network (consisting of subnetworks) connected externally to client networks.*

OXCs, transceivers, etc.) along the lightpath route.

- **1+1 protected lightpaths**: A 1+1 protected lightpath has a primary route and a diversely routed *dedicated* backup route. The source node transmits (bridges the signal) simultaneously on the primary and backup paths, and the receiver switches from primary to backup when the former fails.[1] A 1+1 protected lightpath can recover from any failure (of fiber links, transceivers, etc.) on its working route.
- **Mesh-restored lightpaths**: A mesh-restored lightpath has a primary route and a diversely routed *shared* backup route. The wavelength channels on the backup route for the mesh-restored lightpath are shared among different mesh-restored lightpaths. Wavelength channels are shared in such a way as to ensure that any single failure on the primary route of any mesh-restored lightpath can be restored. Two mesh-restored lightpaths with primary routes P1 and P2 and backup routes B1 and B2, respectively, are illustrated in Fig. 2. Since P1 and P2 are link route diverse, their backup routes B1 and B2 can share a channel on link C1–C2.

The primary and backup route for a lightpath must be diverse so that no single event failure can affect both these paths. Diversity can be defined at the link or node level, thus protecting against single link or node failures. However, note that multiple links in a network could pass through the same physical conduit or share the same "right-of-way." Damage to a conduit, for example, can cause the failure of all fiber links passing through that conduit. This motivates the concept of a *shared risk link group* (SRLG), which identifies a point of failure (a fiber, cable, conduit, or WDM system) that could affect all channels belonging to the SRLG. Thus, even if two paths are link disjoint, they could have an SRLG in common, the failure of which will cause both these paths to fail. Hence, in this article we consider diverse routing of lightpaths whose primary and backup routes are SRLG-disjoint.

The objective of this article is to present a broad overview of the issues involved in architecting an optical cross-connect mesh network, starting from the network design and planning phase to the real-time network operation phase involving dynamic provisioning and restoration of lightpaths and online algorithms for route computation. We discuss frameworks for offline design and planning of optical networks based on projected future lightpath demands. We outline the essential components of an IP-centric control architecture for dynamic provisioning and restoration of lightpaths in optical networks. Online algorithms for route computation of unprotected, 1+1 protected, and mesh-restored lightpaths are discussed. Finally, we conclude.

Network Design and Capacity Planning

During network design and capacity planning, the network operator has a demand forecast for a future time period, and decides how to add capacity to the network in an optimal manner to support the demand. Network design and planning may involve addition of capacity to existing links, or the addition of new links and nodes to the topology. Such network design and planning activities are performed offline or non-real-time. Offline routing algorithms are used to optimally design networks based on forecast lightpath demands. Thus, they help in predeployment capacity planning. Because offline algorithms have knowledge of the entire set of demands (as opposed to online algorithms that route without knowledge of future demands), they make more efficient use of network capacity and project a lower capacity requirement. To prevent blocking of lightpath demands during network operation (the online scenario), planned capacities must be increased by an adjustment factor for deployment in the actual network. This factor can be determined through experimental comparison of the capacity required to route all the given lightpath demands in the offline and online cases separately.

Design Approaches

The following network design problems are considered:

- *Greenfield network design*: Under this, a network is designed from scratch with no previously provisioned lightpath demands. Primary and backup routes need to be determined for each lightpath so that the total link capacity is minimized.
- *Incremental network design*: Under this, new lightpath requests arrive in batches over multiple time periods. Each

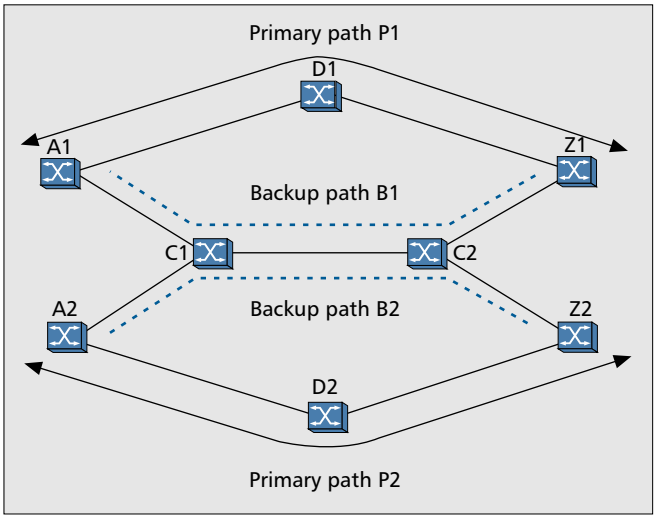


Figure 2. *Mesh-restored lightpaths sharing backup paths.*

[1] *In 1:1 protection, transmission occurs on the primary path only, while the backup path may be used for carrying low-priority traffic which must be preempted to carry restored traffic when the primary path fails.*

batch of lightpaths is routed incrementally starting from an existing base network. Primary and backup routes need to be determined for each lightpath so that minimum additional capacity is required.

- *Backup path reoptimization*: Under this, primary and backup routes for lightpaths already exist in an operational network. An attempt is made to recover restoration capacity by rerouting the backup routes of existing lightpaths (carrying traffic on their primary path). Since past demands were routed in an online fashion, it is conceivable that this offline reoptimization will reduce restoration capacity overhead.

Each of the above models allows the following optimization scenarios:

- *Restricted link optimization (RLO)*: In this case, capacity is added to existing links only.
- *Unrestricted link optimization (ULO)*: In this case, capacity is added to existing links as well as a subset of new links.

Both the above scenarios can be incorporated into the network design algorithms described in this section.

Solution Approaches

We discuss two frameworks for offline routing algorithms. These are broad frameworks in the sense that they can be adapted to fit each of the three types of lightpath demands: unprotected, 1+1 protected, and mesh-restored. The first approach is based on integer linear programming (ILP) [3]; the second involves combinatorial techniques. In general, ILPs are more computation-intensive than combinatorial algorithms. They also provide a mechanism for achieving tradeoffs between running time and optimality of the solution.

Integer Linear Programming — Under this framework, the problem is modeled as an ILP [3], which can be solved using techniques from the linear programming literature such as Lagrangian relaxation [3] or through standard ILP solvers like CPLEX.

We discuss briefly an ILP formulation given in [4]. Given a network in terms of nodes and links and a set of lightpath demands, the aim is to find the primary and backup paths for each demand so that total network capacity is minimized. Each link can have an associated cost (capacity weight), so the objective function is the weighted sum of the capacity usage on each link. The paths from source to destination pairs are modeled as network flow [5] (i.e., a path from source s to destination d is associated with one unit of flow from s to d). Constraints corresponding to the following are imposed on the ILP, and define the space of feasible solutions:

- Flow conservation constraints for the primary path of each demand.
- Flow conservation constraints for the backup path of each demand.
- Primary and backup paths for the same demand are link disjoint.
- Two primary paths can share a link (channel) on their backup path only if they are link disjoint.
- Backup path of a demand consumes capacity if and only if a link on the primary path fails.

Note that the only way the allocations of lightpaths for two demands interact is through the sharing constraint. This corresponds to a bundling constraint in the ILP given in [4]. Without this constraint, the problem would decompose into a number of independent subproblems, one for each demand. This decomposition would significantly reduce the computational complexity of the problem, originally formulated as an ILP. A common technique used to relax such constraints is the Lagrangian relaxation method [3]. The application of this technique to this problem and the corresponding algorithm are given in [4].

Combinatorial Algorithms — Under this framework, combinatorial algorithms are used to solve the problem. Combinatorial approaches have the advantage of running much faster than ILPs and often give useful insight into the problem and its solution. We explore a framework in which the space of all feasible solutions can be generated by varying two dimensions:

- Sequence of routing the individual lightpath demands
- Method of routing a lightpath demand independent of the others

Some of these approaches require the initial capacity of links in the network to be known. This can be estimated by computing a feasible solution that routes all demands, either through ILP-based techniques or by a simple shortest-path-based routing approach. In the latter case, the diversity of primary and backup paths can be ensured by avoiding the primary path links when computing the backup path for a lightpath.

The following approaches can be used in ordering the computation of routes for each lightpath demand:

- *Minimum max flow first*: The maximum flow [5] that can be sent from a source to a destination is a measure of the number of lightpaths that can be set up between that source-destination pair. Hence, it makes sense to route lightpaths between those source-destination pairs first. Otherwise, when these source-destination lightpath demands are routed at a later stage, all paths between the source and destination might be blocked (zero max flow), since demands routed at an earlier stage might have used up the links on the few remaining paths between this source-destination pair. This approach tries to maximize the number of lightpath demands that can be satisfied under given link bandwidth availabilities.
- *Longest shortest path first*: The length of a routed path is a measure of the bandwidth used on the path. Hence, from a minimum bandwidth usage point of view, it makes sense to route lightpath demands along shortest paths. Now, consider source-destination pairs for which the shortest path (or the average of the first few shortest paths) is highest. Deferring the routing of demands between these source-destination pairs might increase the length of the shortest paths between these pairs even further, leading to very high bandwidth usage when these demands are routed at a later stage; hence, the rationale behind this heuristic. This approach tries to minimize the bandwidth usage over all routed lightpaths.
- *Highest critical links first*: A link is said to be critical [5] with respect to a source-destination pair if reducing the bandwidth on that link reduces the maximum flow between that source-destination pair. The criticality of a link is very sensitive to the topology of the network, since a critical link appears in some minimum cost cut between the source-destination pair. Thus, a source-destination pair with a high number of critical links is very vulnerable to loss of connectivity due to routing of demands between other pairs (the chance of the critical links being used in routing between other pairs is high). This heuristic routes lightpath demands first between those source-destination pairs that have the highest number of critical links.

In addition, genetic algorithms [6] can be used to select certain permutations of lightpath demands from a large population. The algorithms used for routing each individual lightpath demand depend on the nature of the lightpath being routed (i.e., unprotected, 1+1 protected, or mesh-restored). These are discussed in detail later. Note that these algorithms can also be used for routing individual demands in the Lagrangian relaxed ILP discussed in the previous section. Thus, a combinatorial algorithm for offline routing can be obtained by choosing a permutation of the lightpath demand requests and

then using an appropriate algorithm for routing each individual lightpath demand.

Dynamic Provisioning and Restoration: IP-Centric Control Architecture

In this section we briefly describe an IP-centric control architecture for optical networks, and outline the dynamic lightpath provisioning and restoration processes. This sets the context for the online route computation algorithms discussed in the next section. A consensus is emerging in the industry on utilizing an IP-centric control plane within optical networks to support dynamic provisioning and restoration of optical layer connections [7]. The rationale behind this is as follows. First, IP-based routing protocols and MPLS signaling protocols are readily available through third-party implementations (the adoption of these protocols for optical network applications requires some modifications). Second, an IP-based control plane is expected to ease end-to-end control and facilitate unified traffic engineering in an environment where IP (router) networks are interconnected via an optical core network. Recent advances in control plane technology for MPLS traffic engineering can be adapted for use in OXC networks by generalizing the concept of a label in traditional MPLS. Generalized MPLS (GMPLS) provides a framework for dynamic provisioning of lightpaths in optical networks and allows the use of uniform semantics for network management and operations control in hybrid networks consisting of OXCs and label switching routers. For an overview of the GMPLS architecture (work in progress at the Internet Engineering Task Force, IETF), see [8, 9].

Provisioning and restoration of lightpaths in an optical network requires protocol and signaling support. The topology of an optical network is illustrated in Fig. 1. Here, neighboring OXCs may have multiple links between them. Each OXC shown in the figure is capable of switching a data stream from a given input port to a given output port. This switching function is controlled by appropriately configuring a cross-connect table. A lightpath from an ingress port in an OXC to an egress port in a remote OXC is established by setting up suitable cross-connects in the ingress, intermediate, and egress OXCs such that a continuous physical path exists from the ingress to the egress port. Lightpaths are assumed to be bidirectional — the return path from the egress port to the ingress port follows the same route as the forward path.

It is assumed that each OXC in the optical network has a unique IP address which serves to identify the OXC and as a basis for creating an IP-centric control plane. Provisioning a lightpath requires identification of the OXC ports that originate and terminate the path. Such ports may be referred to by locally unique indices. In provisioning requests, endpoints are therefore referred to by the pair <*OXC address, port index*>. Additional addressing information pertaining to channels may be present.

For implementing an IP-centric plane, it is necessary to have a bidirectional point-to-point control channel between adjacent OXCs. For instance, with synchronous optical network (SONET)-based optical links, unused SONET overhead bytes may be used to define this control channel. For our discussion, it is sufficient to abstract away the specific implementation and assume that there is exactly one IP control channel between adjacent OXCs and that the control channel will be available as long as there is one functioning bidirectional logical link between the corresponding OXCs.

The following mechanisms are required to support automated provisioning and restoration of lightpaths in optical networks; we discuss each of these in detail in the following subsections:

- **Neighbor discovery**: Automatic detection of links (including port-to-port connectivity and SRLG information) between neighboring OXCs and keeping track of their status (e.g., up/down, bandwidth availability, etc.).
- **Topology discovery**: Disseminating the link state information from each OXC so that every OXC has knowledge of current topology and link state characteristics of the entire network.
- **Route computation**: Computation of a primary and backup route (the latter for protected lightpaths) for the demand being serviced, taking into account the bandwidth needs and other constraints specified for the path, and the state of the network.
- **Lightpath establishment**: Signaling to set up the cross-connects in each OXC for the primary (and backup for 1+1 protected) path and to "soft-reserve" the cross-connects and channels in each OXC for the shared backup path (for mesh-restored).
- **Lightpath restoration**: Fault detection and signaling to set up the backup path when any link on the primary path fails. Current restoration latency (time taken to restore service after failure) requirements are on the order of a couple of 100 ms.

For an overview and comparative discussion of routing and restoration architectures in optical cross-connect mesh networks, see [10].

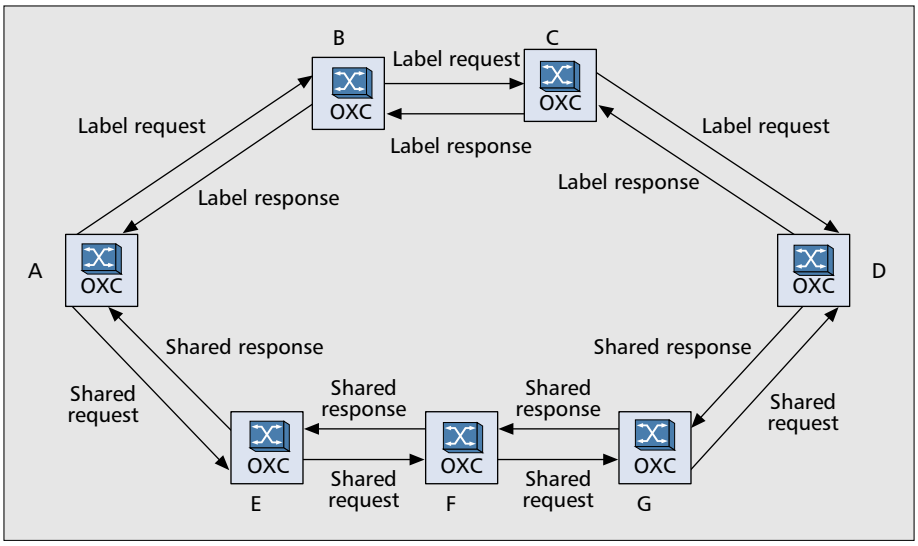
Neighbor Discovery

Neighbor discovery is the procedure by which each OXC determines the status of each optical link, the bandwidth and other parameters of the link, the remote node and port identities, and the consistency of link parameters with the information available at the other end of the link. Neighbor discovery is based on a combination of manual configuration and a protocol running between adjacent switches. One such protocol is *Link Management Protocol* (LMP) [11]. In an O-E-O switch network, LMP messages can be exchanged in an in-band fashion using, for example, the SONET overhead bytes. LMP also contains specifications for out-of-band control in an all-optical (O-O-O) switch network.

LMP runs directly over IP with a distinct protocol ID. Under LMP, an OXC sends a Hello message periodically over the control channel of each link to which it is connected. This message is encapsulated in an IP packet and sent to a designated IP multicast address. The content of the Hello message includes the IP address of the sending OXC, the port number of the link over which the packet is sent, and other parameters (e.g., SRLG information) whose consistency must be verified. This packet is received and processed by the neighbor, which repeats the received information along with the corresponding information from its side. The received information is also used to populate the *port state database* which has both configured information (local attributes) and information discovered using LMP (neighbor attributes). LMP monitors the health of the control channel between two adjacent OXCs and is also responsible for detection of any misconfiguration of port-to-port fiber connectivity, SRLG information, and so on.

Topology Discovery

Topology discovery allows each OXC in a network to build a database representing the network topology and resource availability. Topology discovery is accomplished by running a modified version of a distributed IP routing protocol such as Open Shortest-Path First (OSPF) or Intermediate System to Intermediate System (IS-IS) in each OXC. The following discussion uses OSPF, but the essential ideas can also be applied



■ Figure 3. *Lightpath establishment using CR-LDP.*

to IS-IS. The topology (or *link state*) information essentially consists of the representation of links and nodes in the network along with certain associated parameters (link cost, resource information, etc). Under the modified OSPF scheme, each OXC broadcasts the local link state information to other OXCs periodically, as well as when a change occurs in topology or resource availability. Using the link state information, an OXC could compute the path for an optical layer connection that originates from itself and terminates in a remote OXC. OSPF also allows hierarchical routing, whereby a large network may be treated as a collection of smaller areas with limited information exchange between areas.

Most of the OSPF functionality is maintained when applied to optical networks. The following new features are introduced in the modified version of OSPF:

- **Link bundling**: Under standard OSPF, each physical link between a pair of switches would result in a routing adjacency. This means that routing protocol messages would be exchanged over each such link, and each such link would be advertised to other OXCs in broadcast messages. Since the number of physical links between a pair of nodes could be large in optical networks, this would result in increased message broadcasting and processing overhead. To eliminate this, *all the links* between a pair of neighbors could be treated as a single logical routing adjacency. This procedure is called *link bundling* [12]. With link bundling, routing protocol messages are sent over exactly one link even if there are multiple links connecting a pair of OXCs. Furthermore, the entire set of links is advertised in a single message. In this message, individual links are represented as members of a *link group* with identical resource parameters.
- **Resource parameters**: In an optical network, link state and resource information is used to compute diverse primary and backup paths for a lightpath connection. This information consists of the representation of links and nodes in the network along with certain associated resource parameters (e.g., link cost, resource type and availability, SRLG information) that are critical to optimal and diverse routing of lightpaths. Standard OSPF does not provide the mechanism to disseminate such information through link state messages; hence, the need to introduce traffic engineering extensions to OSPF [12, 13].
- **Link state advertisement thresholds**: Because link state advertisements capture resource availability, care must be taken to ensure that this information is not generated too frequently with minor changes in resource status. A configurable thresholding scheme needs to be introduced whereby an OXC would generate a link state update only if the resource information changes "significantly." This modification reduces the number of link state advertisements.
- **Source routing methodology**: Standard OSPF is designed for routing IP datagrams. Hence, under standard OSPF each participating node would use an identical algorithm to compute a *forwarding table* that allows packets to be routed based on the destination address. Routing of an optical layer connection, on the other hand, requires that the entire path for the connection be computed at the source OXC and signaled to other OXCs in the path.

The new link representation and resource parameters are incorporated into OSPF through traffic engineering extensions [13]. Extensions to OSPF for supporting GMPLS are described in [14].

Route Computation

Route computation is based on specialized algorithms that take into account both the requirements for the primary path as well as the backup path (if any). In the *centralized scenario*, route computation takes place at the network management system (NMS) or a route server, using information about the entire network. In the *distributed scenario*, the ingress OXC for the lightpath computes the route for the lightpath, using *only* the information that is contained in its local databases.

Each OXC has a *connection* database that contains an entry for each connection that traverses the OXC. The connection database is updated when a connection is provisioned, and when its attributes change. Note that the connection database in an OXC does not contain information about lightpaths that do not traverse that OXC.

The topology database obtained via OSPF is converted into a graph representation as follows. Nodes and links in this graph correspond to OXCs and links in the optical network, respectively. Labels on the edges of the graph indicate the cost and resource information. This graph representation is used to find paths in the network. For example, the shortest-cost path between two switches can be computed using a version of the Bellman-Ford algorithm [15]. Edge disjoint paths in the network are found by determining link/node disjoint routes in the graph between the given pair of nodes. We discuss online path computation algorithms in a later section.

For mesh-restored lightpaths, each OXC also maintains an additional *sharing database*, which contains, for each link adjacent to that OXC and for each shared channel on that link, the set of lightpaths whose backup routes use that channel. The significance of this sharing database will be pointed out in the context of routing mesh-restored lightpaths in a later section.

The centralized NMS in the optical network contains the snapshot of the entire network. Specifically, the NMS contains the topology database and the connection and sharing databases at each OXC in the network.

Lightpath Establishment

The MPLS architecture for IP networks defines protocols for establishing label switched paths (LSPs) [8]. LSPs allow for provisioning of traffic-engineered virtual circuits in an IP network, and the signaling protocols to establish LSPs may be

adapted for provisioning lightpaths in optical networks. There are two choices for MPLS-based signaling protocols: Resource Reservation Protocol with Traffic Engineering (RSVP-TE) or Constraint Routed Label Distribution Protocol (CR-LDP). Both these protocols allow hop-by-hop signaling from a source to a destination node to establish unidirectional LSPs. Certain new features must be introduced in these protocols to accommodate the peculiarities of lightpath provisioning in optical networks, including support for establishing bidirectional paths, support for OXCs without wavelength conversion (O-O-O switches), support for establishing shared backup paths, and fault tolerance. Extensions for some of these requirements have already been proposed and are described for RSVP-TE in [16] and for CR-LDP in [17].

We illustrate in Fig. 3 the process of establishment of a primary and shared backup path using CR-LDP signaling. The primary path A-B-C-D is established by a CR-LDP Label Request message which propagates hop by hop from A to D. This message has the path identifier, destination OXC address and port information, path route, path parameters, local port selected at the previous hop, and so on. The destination OXC establishes the local cross-connect for the path, and returns a Label Response message to the previous hop. Each OXC receiving the message establishes the local cross-connects for the path and forwards the message to the previous hop. The establishment of a shared backup path, A-E-F-G-D, is also shown in the figure. The shared request and response messages are treated in a manner similar to the label request and response messages. Cross-connects on the shared backup paths, however, are just "soft-reserved" and not set up until the restoration process is invoked.

Lightpath Restoration

Protected paths within an optical network can be restored at two levels:

- The local span or optical link level
- The end-to-end or path level

Whenever a failure is detected, the OXC closest to the failure first attempts to restore the connections by performing local span switching. If span switching fails, end-to-end restoration is attempted by the OXCs at the endpoints of the path. The span switching protocol is referred to as the LR (local restoration) protocol, and the end-to-end path restoration protocol as the EER protocol. In the following discussion, we consider a restricted version of span switching where the LR protocol attempts to find a replacement channel from the set of all available channels between the OXCs adjacent to the failed span. A more general case of span switching would allow multihop (instead of single-hop) paths for local restoration of a failed span.

Every bidirectional optical link is associated with an instance of an LR protocol state machine. In local span switching involving two protocol instances running on the peer ports of the neighboring OXCs, the LR instance on the OXC with the higher node ID (IP address) is considered the master. The master is responsible for selecting the spare optical link pair (if available) that will be used to restore the failed optical link pair. Once the protection optical link pair is selected, the LR protocol instances running on the two ends of the link coordinate to restore the failed optical link over the protection optical link.

The endpoints of every path pair (primary and backup) are controlled by EER protocol instances associated with the drop ports at the end OXCs. If LR fails to restore the failed connection, EER is invoked by one or both OXC endpoints of the path. Depending on the type of path restoration scheme (1+1 or mesh-restored), different actions are taken by the EER protocol. If both LR and EER protocols fail, a new provisioning event is invoked by the network management system, and a new path is set up by the ingress OXC. This newly provisioned path is used to restore the failed connection.

Both the LR and EER protocols allow for reversion of traffic onto the primary span or path, respectively, when the failure triggering the restoration process is repaired.

In the following discussion, the LR and EER restoration protocols are illustrated using a particular connection shown in Fig. 4. The primary path is from OXC A to OXC E through switches B, C, and D. It is precisely represented by the sequence <(1:1,A,5:1), (4:1,B,7:1), (3:1,C,10:1), (7:1,D,5:1), (7:1,E,9:1)>. Each tuple in this sequence identifies the ingress port:channel, the node, and egress port:channel. The backup path is from OXC A to OXC E through switches F, G, and H. More specifically, the backup path is represented by the sequence <(1;1,A,7:1), (5:1,F,8:1), (7:1,G,9:1), (4:1,H,8:1)>. During provisioning, the primary path is signaled (using RSVP or CR-LDP) and the cross-connects are set up. The backup path is also signaled, and depending on the type of backup path, different actions are taken. For dedicated backup paths (for 1+1 protection), the cross-connects are established on all the OXCs, and the signal is bridged at the head-ends. If the backup path is shared, signaling does not establish the cross-connects. Rather, *soft reservations* are made to mark the channels on the backup path as allocated for shared backup paths. In this case, cross-connects are established during restoration using a separate signaling process.

LR Protocol Overview — In this section, the LR protocol is described briefly using the example depicted in Fig. 4. Suppose that the optical link from port 7 on switch D to port 10 on switch C has failed. The nearest switch, C in this case, detects the failure condition and inserts an alarm indication in the downstream direction. It then initiates the local restoration process using the LR protocol. Also suppose that C assumes the role of the master and starts the restoration link selection pro-

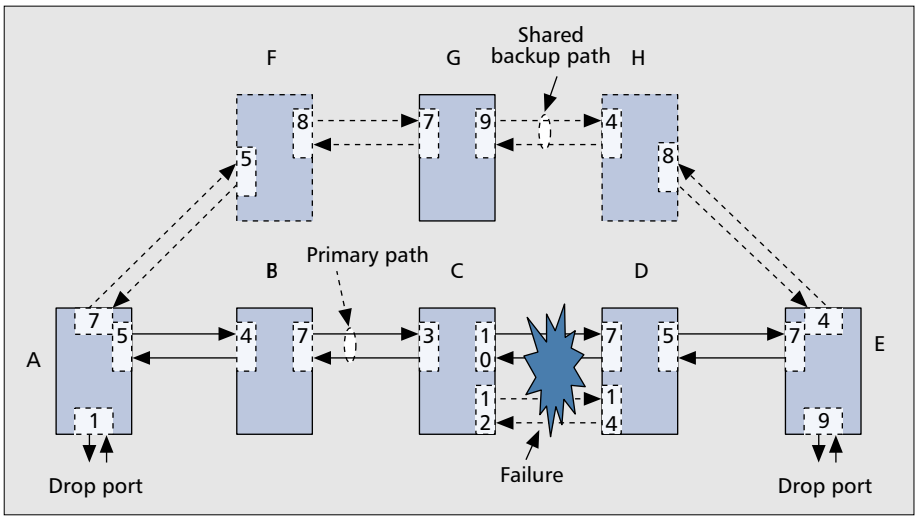


Figure 4. *LR and EER restoration protocols.*

cess. As a part of the selection process, it first checks if a similar optical link (OC-48, OC-192, etc.) is available on the same link. Let us assume that an appropriate optical link is available (link between port 12 on C and port 14 on D). The LR protocol instance at port 12 on C then engages the corresponding instance on its peer port 14 on switch D. When the protocol is successfully completed, port 3 is bridged (bidirectional) to port 12 in switch C, and port 5 on D is similarly bridged to port 14. When the failure is repaired, port 10 in C receives a valid signal and initiates the switchback process.

If the LR protocol fails, C triggers the EER protocol. A signal to trigger EER is dispatched to one or both endpoints of the connection. The connection endpoint (either one or both of them) then begins the end-to-end restoration.

EER Protocol Overview — In the above example, if local restoration fails, EER is invoked at OXC A by the trigger from OXC C. Since the backup path is soft-reserved during the provisioning process, explicit signaling is required to establish the cross-connects along the backup path. Restoration signaling is initiated by A and carried hop-by-hop in the SONET overhead bytes. This signal identifies the backup path to be activated. Each OXC en route retrieves the path-related information (established during backup path provisioning) from a database. This information indicates the cross-connect entry to be set up, and the previous and next OXCs in the path. If the ports associated with the backup path are available, the cross-connect is set up and the signaling message forwarded to the next hop. If the ports are not available, an error message is propagated backward. If the backup path can be successfully set up, switches A and E move the signal from the failed primary path to the backup path.

Restoration Latency — For 1+1 protected paths, since the cross-connects on the backup path are set up in advance and the signal is bridged at the source, restoration involves a switch at the receiver side and takes a few milliseconds. Since the LR protocol involves local channel replacement, restoration latency for LR is in the range of a few tens of milliseconds and is dominated by the bridging/switching time and the time to find a replacement channel. Restoration latency concerns crop-up for the EER protocol for mesh-restored lightpaths where the cross-connects on the backup path are only soft-reserved in advance and need to be set up after the primary path fails. This involves two-way handshaking between the end nodes of the lightpath being restored and incurs round-trip propagation time, cross-connect setup time at the intermediate nodes on the backup path, bridging/switching time at the end nodes, and associated protocol processing time. Hence, some of the factors that affect restoration latency for mesh-restored lightpaths are hop count of backup path, OXC cross-connect setup and bridging/switching time, OXC processor speed for protocol processing, and processor load (number of backup paths being restored that pass through an OXC). The second and third factors depend heavily on the OXC switch fabric and processor architecture. Restoration latency constraints can potentially be incorporated during route computation. Factors like backup path hop constraint can easily be handled in route computation algorithms, but some of the other factors such as OXC processing time and load seem more difficult to model and incorporate. Given the flexibility of a mesh architecture, carriers may be willing to relax the 50 ms restoration latency typical of SONET rings and accept mesh restoration times in the ballpark of a couple of hundreds of milliseconds (this also depends on the type of application). Restoration latency targets of 200 ms seem to be achievable for mesh networks, as reported through delay modeling and simulation based on a real OXC architecture and EER protocol [18].

Online Algorithms for Route Computation

In this section we discuss algorithms for routing unprotected, 1+1 protected, and mesh-restored lightpaths. For protected lightpaths, we consider end-to-end restoration where traffic switches from the primary to a route diverse backup path after the former fails. The algorithms are suited to an opaque O-E-O switch network with full wavelength conversion, where wavelength assignment to enforce the wavelength continuity constraint for a lightpath is not required. For the more difficult problem of combined routing and wavelength assignment (RWA) in transparent all-optical (O-O-O) networks, see [1].

In the centralized scenario, route computation occurs at the NMS and the route may be specified up to the port/channel level. In the distributed scenario, the ingress OXC for the lightpath computes the route. Since an OXC does not have information about the local port/channel state at other OXCs, the route is only specified at the link level in the distributed scenario. Port/channel selection is done during signaling of the path, using MPLS-based signaling protocols like RSVP-TE [16] or CR-LDP [17], as discussed earlier. For unprotected and 1+1 protected lightpaths, algorithms for route computation are identical in the centralized and distributed scenarios, since the algorithm needs only information about links with available channels. This is available from the OSPF topology database. For mesh-restored lightpaths, complete knowledge of the sharing databases at all OXCs in the centralized scenario allows more efficient computation of shared backup paths (from a capacity utilization point of view). Hence, for mesh-restored lightpaths, we discuss algorithms for centralized and distributed scenarios separately. For a discussion of the capacity performance for provisioning different types of lightpaths, see [19].

In our model, there is a cost associated with each link in the optical network. This is the estimated cost of using a new channel on that link. This cost is user-defined and advertised through OSPF LSAs. The cost is computed taking several factors into consideration such as configuration parameters and equipment cost (fiber, dense WDM common equipment, amplifier, transponder, regeneration transponder, OXC common equipment, OXC transceiver, floor space, power, etc.). The route computation algorithms use this cost at the link level to route lightpath demands through primary and backup paths of "least cost."

Unprotected Lightpaths

The path cost for an unprotected lightpath is defined to be the sum of the costs of the links on the path. The Bellman-Ford or Dijkstra algorithm [15] can be used to compute the minimum cost path for an unprotected lightpath, using the graph representation of the topology and connection attributes.

1+1 Protected Lightpaths

A 1+1 protected lightpath from ingress OXC A to egress OXC Z is allocated a pair of dedicated SRLG-disjoint paths in the network. One of these paths is the primary path, the other the backup path. The path cost for a 1+1 protected connection is the sum of the costs of the links of the primary and backup paths. Hence, the objective is to pick the minimum combined cost SRLG-disjoint path pair from A to Z.

A pair of minimum cost link-disjoint paths can be found by applying Suurballe's algorithm [15] on the graph representation

of the topology. In network flow terminology [5], Suurballe's algorithm essentially finds a minimum cost flow of value 2 from A to Z, where each link in the network has capacity 1. The latter constraint ensures the link disjointness of the paths. The shorter-hop path can be chosen as the primary path and the other as the backup path. In the general case of SRLGs, the problem of computing even a feasible (not minimum cost) pair of SRLG-disjoint paths becomes NP-complete. In this case, heuristic approaches like enumeration can be used to consider primary paths and then compute the best backup path for each. The primary-backup pair with the least cost is chosen.

Mesh-Restored Lightpaths

A mesh-restored lightpath from ingress OXC A to egress OXC Z is allocated a pair of SRLG-disjoint paths in the network, where one of the paths is the primary path, and the other the backup path. Each link in the working path has dedicated capacity allocated to the connection (and carries traffic under normal conditions). The capacity allocated on the backup path, however, is shared with backup paths for other mesh-restored connections. For any two mesh-restored lightpaths L1 (primary path = P1, backup path = B1) and L2 (primary path = P2, backup path = B2), B1 and B2 can share channels on common links only if P1 and P2 are SRLG-disjoint, as illustrated in Fig. 2. This sharing condition ensures that all mesh-restored lightpaths can be restored after any single SRLG failure. For a discussion of backup path sharing scenarios for both local and end-to-end restoration under single link, single SRLG, and single node failures, see [20].

Each OXC maintains a sharing database as described in an earlier section. This enables the path computation algorithm to decide if a given connection can share wavelength links with other existing connections. For a given shared channel c at an OXC, a lightpath with working path p can share c with other mesh-restored lightpaths if p is SRLG-disjoint with the primary paths of all lightpaths that currently use c on their backup paths. The cost of a mesh-restored lightpath whose primary path is p and backup path is b is defined as the sum of the costs of the links in p and costs of the links *not shared* in b. This definition sums up the costs of "new" capacity allocated to the lightpath. Additionally, it may be a requirement that the number of hops on the backup path be bounded. This constraint avoids long backup paths that will violate the restoration time guarantee. This constraint is absent in the 1+1 protected case, since cross-connects are set up on the backup path during lightpath establishment, thus requiring only minimal signaling during path restoration. The problem of computing a minimum cost mesh-restored lightpath (even without the backup hop constraint) is NP-complete.

Centralized Scenario — Since the NMS has complete knowledge of the sharing database at each OXC, it can determine the backup path sharability of a channel on any link of the network for a given primary path. Hence, the algorithm used is similar to the one for 1+1 protected lightpath computation with a modification of the cost tructure on the backup path to reflect the fact that shared links on the backup path do not incur any additional cost (as defined above).

Distributed Scenario — In this scenario the route computation occurs at the ingress OXC, which, in the absence of knowledge of sharing databases of other OXCs, is unable to determine backup sharability of links for any given primary path. The ingress OXC has only aggregated information about the number of available and (shared) backup channels on a link, disseminated by traffic engineering extensions to a link state protocol like OSPF. The approach for routing 1+1 protected lightpaths suggests a heuristic scheme for routing mesh-restored lightpaths, described below. The aim is to increase sharing of backup paths and reduce restoration capacity overhead in the network by using sharability information that is available at the intermediate OXCs on the backup path, but not at the ingress OXC where the path is computed. This approach [20] involves distributed determination of the sharability of a link on the backup path during path signaling using the sharing database at each OXC on the backup path.

A pair of SRLG-disjoint routes is determined using the 1+1 scheme outlined above. Sharability of any link on the backup path is determined during signaling of the backup path. Each OXC on the backup path, when it receives the signaling message, independently makes a decision, using its local sharing database, about whether it can share a backup channel with the given lightpath request. If so, it updates its sharing database appropriately. Otherwise, it allocates an unallocated channel for the backup path, which is added to the sharing database and can then be shared with future lightpaths. In either case, the sharing database is updated to reflect the new backup path going through it and the corresponding primary path.

Note that 1+1 path computation uses only available channel information about a link. Consider a scenario where the network is heavily loaded so that no 1+1 primary-backup path pair is available. However, even in such a case, a primary-backup pair where some channels on the backup path are shared may be available. Can the path computation algorithm at the ingress OXC use information about the number of (shared) backup channels per link to choose links on the backup path (during 1+1 computation) that have a higher chance of containing a sharable channel? The answer lies in the fact that *the probability that a link contains a sharable channel increases as the number of shared channels in it increases*.

This suggests the following modification [21]. If a 1+1 link-disjoint primary-backup path pair cannot be found, links with no available channels will also have to be included in the graph as a second attempt for path computation. In this case, higher priority must be given to links with higher numbers of (shared) backup channels for backup path computation, since these have a higher probability of being sharable. However, this is an optimistic heuristic. Backup path provisioning can fail in the unlikely event of a link on the computed backup path having no available capacity but a (possibly large) number of shared backup channels, none of which are sharable with the given primary path (as determined at the local OXC during signaling). In this case, information about the link can be included in an error message to the ingress OXC, and the lightpath computation can be repeated with the knowledge that this link should be avoided. Such a *retry* scheme is facilitated by *crankback routing* extensions to RSVP and CR-LDP [22]. This reduces blocking of lightpaths in the network. A configurable upper limit can be placed on the number of retries allowed before failure is declared for provisioning the lightpath under consideration.

Conclusion

This article considers the architectural and algorithmic aspects of optical cross-connect mesh networks, starting from network design and planning to real-time provisioning and restoration of lightpaths. Frameworks for offline design and planning of optical networks based on projected future lightpath demands are discussed. Currently, optical cross-connects with full wavelength conversion capability are being deployed in core optical networks to interconnect IP (router) networks, and it is becoming increasingly important to have an architecture that

facilitates easy control and management of such networks. A consensus is emerging in the industry on utilizing an IP-centric control plane within optical networks to support dynamic provisioning and restoration of lightpaths. The essential components of such an IP-centric control architecture for optical networks are outlined. These include neighbor discovery, topology and resource discovery, route computation, lightpath establishment, and lightpath restoration. Online algorithms for route computation of unprotected, 1+1 protected, and mesh-restored lightpaths are discussed in both centralized and distributed scenarios. Some important criteria in the performance evaluation of the architectural and algorithmic approaches discussed in this article include standardization, interoperability and scalability of routing, topology and resource discovery, and signaling protocols, realistic network design and capacity planning based on forecasted lightpath demands, capacity efficiency of online routing algorithms, and restoration latency. As optical networks are deployed to meet the explosive growth of traffic on the Internet, the merits and demerits of the proposed architectural and algorithmic approaches will become clearer, and new challenges will emerge as requirements for intelligent and automatic control of such networks continue to increase.

Acknowledgments

Thanks to Debanjan Saha and Bala Rajagopalan for useful discussions.

References

[1] T. E. Stern and K. Bala, *Multiwavelength Optical Networks: A Layered Approach*, Prentice Hall, May 1999.
[2] D. Pendarakis, B. Rajagopalan, and D. Saha, "Routing Information Exchange over the UNI and the NNI," OIF cont. 2000.083, Apr. 2000.
[3] L. Nemhauser and L. A. Wolsey, *Integer and Combinatorial Optimization*, Wiley, Nov. 1999.
[4] B. T. Doshi *et al.*, "Optical Network Design and Restoration," *Bell Labs Technical Journal*, vol. 4, no. 1, Jan.–Mar. 1999.
[5] R. K. Ahuja, T. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*, Prentice Hall, 1993.
[6] L. A. Cox, Jr., J. R. Sanchez, and L. Lu, "Cost Savings from Optimized Packing and Grooming of Optical Circuits: Mesh vs. Ring Comparisons," *Opt. Networks*, May–June 2001, pp. 72–90.
[7] B. Rajagopalan *et al.*, "IP over Optical Networks: Architectural Aspects," *IEEE Commun. Mag.*, Sept. 2000.
[8] D. Awduche and Y. Rekhter, "Multiprotocol Lambda Switching: Combining MPLS Traffic Engineering Control with Optical Crossconnects," *IEEE Commun. Mag.*, vol. 39, no. 3, Mar. 2001.
[9] P. Ashwood-Smith *et al.*, "Generalized Multi-Protocol Label Switching (GMPLS) Architecture," Internet draft, <draft-many-gmpls-architecture-00.txt>, Feb. 2001.
[10] E. Bouillet *et al.*, "Routing and Restoration Architectures in Mesh DWDM Networks," submitted for publication.
[11] J. P. Lang *et al.*, "Link Management Protocol (LMP)," Internet draft, <draft-lang-mpls-lmp-02.txt>, July 2000.
[12] K. Kompella *et al.*, "Link Bundling in MPLS Traffic Engineering," Internet draft, <draft-kompella-mpls-bundle-05.txt>, Feb. 2001.
[13] D. Katz *et al.*, "Traffic Engineering Extensions to OSPF," Internet draft, <draft-katz-yeung-ospf-traffic-04.txt>, Feb. 2001.
[14] K. Kompella *et al.*, "OSPF Extensions in Support of Generalized MPLS," Internet draft, <draft-kompella-ospf-gmpls-extensions-01.txt>, Feb. 2001.
[15] R. Bhandari, *Survivable Networks: Algorithms for Diverse Routing*, Kluwer, Jan. 1999.
[16] P. Ashwood-Smith *et al.*, "Generalized MPLS Signaling - RSVP-TE Extensions," Internet draft, <draft-ietf-mpls-generalized-rsvp-te-03.txt>, May 2001.
[17] P. Ashwood-Smith *et al.*, "Generalized MPLS Signaling — CR-LDP Extensions," Internet draft, <draft-ietf-mpls-generalized-cr-ldp-03.txt>, May 2001.
[18] S. Biswas, S. Datta, and S. Sengupta, "Performance of Restoration Protocols in Optical Mesh Networks," *NFOEC*, July 2001.
[19] R. Ramamurthy *et al.*, "Capacity Performance of Dynamic Provisioning in Optical Networks," *J. Lightwave Tech.*, vol. 19, no. 1, Jan. 2001, pp. 40–48.
[20] S. Kini *et al.*, "Shared Backup Label Switched Path Restoration," Internet draft, <draft-kini-restoration-shared-backup-01.txt>, May 2001.
[21] S. Sengupta and R. Ramamurthy, "Capacity Efficient Distributed Routing of Mesh-Restored Lightpaths in Optical Networks," submitted for publication.
[22] A. Iwata *et al.*, "Crankback Routing Extensions for MPLS Signaling," Internet draft, <draft-fujita-mpls-crldp-crankback-01.txt>, Nov. 2000.

Biographies

SUDIPTA SENGUPTA (sudipta@tellium.com) is currently a senior architect at Tellium, where he works on control and management protocols and routing algorithms for intelligent optical networking. Prior to this he was at Oracle Corp., where he was involved in the design and development of the wireless networking platform for Oracle Mobile Applications. He was also at Bell Labs Research, Lucent Technologies, where he worked on QoS routing and traffic engineering under the MPLS framework. He holds an M.S. degree from Massachusetts Institute of Technology, Cambridge, and a B.Tech. degree from Indian Institute of Technology, Kanpur, both in computer science. He received the President of India Gold Medal at IIT-Kanpur for academic excellence.

RAMU RAMAMURTHY (ramu@tellium.com) is currently a senior architect at Tellium, where he works on the design of algorithms and protocols for dynamic provisioning and restoration in optical networks. Prior to this he was a research scientist at Telecordia Technologies, where he worked on network control and management of IP/WDM optical networks. He holds a B.Tech. degree from the Indian Institute of Technology, Madras, and M.S. and Ph.D. degrees from the University of California, Davis.