

Chapter 2

Structural Overview of ISP Networks

Robert D. Doverspike, K.K. Ramakrishnan, and Chris Chase

2.1 Introduction

An *Internet Service Provider (ISP)* is a telecommunications company that offers its customers access to the Internet. This chapter specifically covers the design of a large Tier 1 ISP that provides services to both residential and enterprise customers. Our primary focus is on a large *IP backbone* network in the continental USA, though similarities arise in smaller networks operated by telecommunication providers in other parts of the world. This chapter is principally motivated by the observation that in large carrier networks, the IP backbone is not a self-contained entity; it co-exists with numerous access and transport networks operated by the same or other service providers. In fact, how the IP backbone interacts with its neighboring networks and the transport layers is fundamental to understanding its structure, operation, and planning. This chapter is a hands-on description of the practical structure and implementation of IP backbone networks. Our goal is complicated by the complexity of the different network layers, each of which has its own nomenclature and concepts. Therefore, one of our first tasks is to define the nomenclature we will use, classifying the network into *layers* and *segments*. Once this partitioning is accomplished, we identify where the IP backbone fits and describe its key surrounding layers and networks.

This chapter is motivated by three aspects of the design of large IP networks. The first aspect is that the design of an IP backbone is strongly influenced by the details of the underlying network layers. We will illustrate how the evolution

R.D. Doverspike (✉)
Executive Director, Network Evolution Research, AT&T Labs Research,
200 S. Laurel Ave, Middletown, NJ 07748, USA
e-mail: rdd@research.att.com

K.K. Ramakrishnan
Distinguished Member of Technical Staff, Networking Research, AT&T Labs Research,
Shannon Labs, 180 Park Avenue, Florham Park, NJ 07932, USA

C. Chase
AT&T Labs, 9505 Arboretum Blvd, Austin, TX 78759, USA
e-mail: chase@labs.att.com

of customer access through the metro network has influenced the design of the backbone. We also show how the evolution of the *Dense Wavelength-Division Multiplexing* (DWDM) layer has influenced core backbone design.

The second aspect presents the use of *Multiprotocol Label Switching* (MPLS) in large ISP networks. The separation of routing and forwarding provided by MPLS allows carriers to support *Virtual Private Networks* (VPNs) and *Traffic Engineering* (TE) on their backbones much more simply than with traditional IP forwarding.

The third aspect is how network outages manifest in multiple network layers and how the network layers are designed to respond to such disruptions, usually through a set of processes called *network restoration*. This is of prime importance because a major objective of large ISPs is to provide a known level of quality of service to its customers through *Service Level Agreements* (SLAs). Network disruptions occur from two major sources: failure of network components and maintenance activity. Network restoration is accomplished through preplanned network design processes and real-time network control processes, as provided by an *Interior Gateway Protocol* (IGP) such as *Open Shortest Path First* (OSPF). We present an overview of OSPF reconvergence and the factors that affect its performance. As customers and applications place more stringent requirements on restoration performance in large ISPs, the assessment of OSPF reconvergence motivates the use of MPLS Fast Reroute (FRR).

Beyond the motivations described above, the concepts defined in this chapter lay useful groundwork for the succeeding chapters. Section 2.2 provides a structural basis by providing a high-level picture of the network layers and segments of a typical, large nationwide terrestrial carrier. It also provides nomenclature and technical background about the equipment and network structure of some of the layers that have the largest impact on the IP backbone. Section 2.3 provides more details about the architecture, network topology, and operation of the IP backbone (the IP layer) and how it interacts with the key network layers identified in Section 2.2. Section 2.4 discusses routing and control protocols and their application in the IP backbone, such as MPLS. The background and concepts introduced in Sections 2.2–2.4 are utilized in Section 2.5, where we describe network restoration and planning. Finally, Section 2.6 describes a “case study” of an IPTV backbone. This section unifies many of the concepts presented in the earlier sections and how they come together to allow network operators to meet their network performance objectives. Section 2.7 provides a summary, followed by a reference list, and a glossary of acronyms and key terms.

2.2 The IP Backbone Network in Its Broader Network Context

2.2.1 Background and Nomenclature

From the standpoint of large telecommunication carriers, the USA and most large countries are organized into metropolitan areas, which are colloquially referred to as *metros*. Large intrametro carriers place their transmission and switching equipment

in buildings called *Central Offices (COs)*. Business and residential customers typically obtain telecommunication services by connecting to a designated first CO called a *serving central office*. This connection occurs over a *feeder network* that extends from the CO toward the customer plus a *local loop* (or *last mile*) segment that connects from the last equipment node of the feeder network to the customer premise. Equipment in the feeder network is usually housed in above-ground huts, on poles, or in vaults. The feeder and last-mile segments usually consist of copper, optical fiber, coaxial cable, or some combination thereof. Coaxial cable is typical to a cable company, also called a *Multiple System Operator (MSO)*. While we will not discuss metro networks in detail in this chapter, it is important to discuss their aspects that affect the IP backbone. However, the metro networks we describe coincide mostly with those carriers whose origins are from large telephone companies (sometimes called “Telcos”).

Almost all central offices today are interconnected by optical fiber. Once a customer’s data or voice enters the serving central office, if it is destined outside that serving central office, it is routed to other central offices in the same metro area. If the service is bound for another metro, it is routed to one or more gateway COs. If it is bound for another country, it eventually routes to an international gateway. A metro gateway CO is often called a *Point of Presence (POP)*. While POPs were originally defined for telephone service, they have evolved to serve as intermetro gateways for almost all telecommunication services. Large intermetro carriers have one or more POPs in every large city.

Given this background, we now employ some visualization aids. Networks are organized into *network layers*, which we depict vertically with two network graphs vertically stacked on top of one another in Fig. 2.1. Each of the network layers can be considered to be an *overlay network* with respect to the network below.

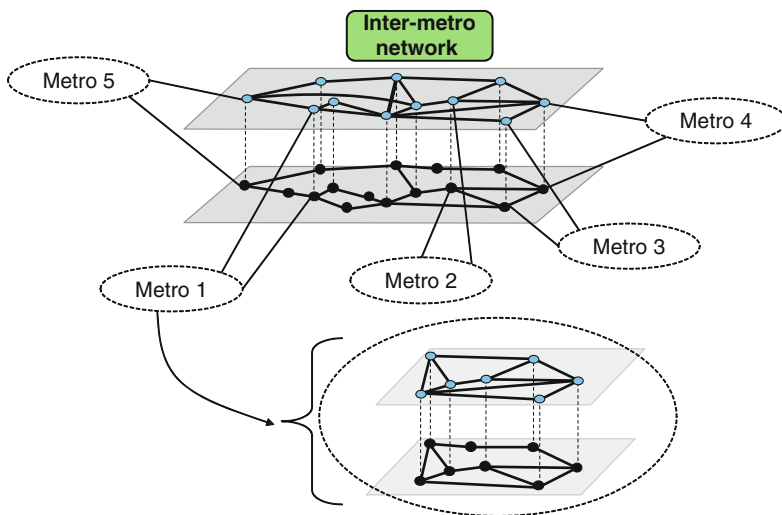


Fig. 2.1 Conceptual network layers and segmentation

We can further organize these layers into *access*, *metro*, and *core* network segments. Figure 2.1 shows the core segment connected to multiple metro segments. Each metro segment represents the network layers of the equipment located in the central offices of a given metropolitan area. The access segment represents the feeder network and loop network associated with a given metro segment. The core segment represents the equipment in the POPs and network structures that connect them for intermetro transport and switching.

In this chapter, we focus on the ISP backbone network, which is primarily associated with the core segment. We refer only briefly to access architectures and will discuss portions of the metro segment to the extent to which they interact and connect to the core segment. Also, in this chapter we will not discuss broader telecommunication contexts, such as international networks (including undersea links), satellite, and wireless networks. More detail on the various network segments and their network layers and a historical description of how they arose can be found in [11].

Unfortunately, there is a wide variety of terminology used in the industry, which presents a challenge for this chapter because of our broad scope. Some of the terminology is local to an organization, application, or network layer and, thus, when used in a broader context can be confused with other applications or layers. Within the context of network-layering descriptions, we will use the term *IP layer*. However, we use the term “IP backbone” interchangeably with “IP layer” in the context of the core network segment. The terms *Local Area Network (LAN)*, *Metropolitan Area Network (MAN)*, and *Wide Area Network (WAN)* are also sometimes used and correlate roughly with the access, metro, and core segments defined earlier; however, LAN, MAN, and WAN are usually applied only in the context of packet-based networks. Therefore, in this chapter, we will use the terms access, metro, and core, since they apply to a broader context of different network technologies and layers. Other common terms for the various layers within the core segment are *long-distance* and *long-haul* networks.

2.2.2 Simple Graphical Model of Network Layers

The following simple graph-oriented model is helpful when modeling routing and network design algorithms, to understand how network layers interact and, in particular, how to classify and analyze the impact of potential network disruptions. This model applies to most *connection-oriented* networks and, thus, will apply to some higher-layer protocols that sit on top of the IP layer. The IP layer itself is *connectionless* and does not fit exactly in this model. However, this model is particularly helpful to understand how lower network layers and neighboring network layers interact.

In the layered model, a network layer consists of *nodes*, *links* (also called *edges*), and *connections*. The nodes represent types of switches or cross-connect equipment that exchange data in either digital or analog form via the links that connect

them. Note that at the lowest layer (such as fiber) nodes represent equipment, such as fiber-optic patch panels, in which connections are switched manually by cross-connecting fiber patch cords from one interface to another. Links can be modeled as *directed* (unidirectional) or *undirected* (bidirectional). Connections are cross-connected (or switched) by the nodes onto the links, and thus form paths over the nodes and links of the graph. Note that the term *connection* often has different names at different layers and segments. For example, in most telecommunication carriers, a connection (or portions thereof) is called a *circuit* in many of the lower network layers, often referred to as *transport* layers. Connections can be *point-to-point* (unidirectional or bidirectional), *point-to-multipoint* or, more rarely, *multipoint-to-multipoint*. Generally, connections arise from two sources. First, telecommunication services can arise “horizontally” (relative to our conceptual picture of Fig. 2.1) from a neighboring network segment. Second, connections in a given layer can originate from edges of a higher-layer network layer. In this way, each layer provides a connection “service” for the layer immediately above it to provide connectivity. Sometimes, a “client/server” model is referenced, such as the *User-Network Interface (UNI)* model [29] of the *Optical Internetworking Forum (OIF)*, wherein the links of higher-layer networks are “clients” and the connections of lower-layer networks are “servers”. For example, see G.7713.2 [19] for more discussion of connection management in lower-layer transport networks.

Recall that the technology layers we define are differentiated by the nodes, which represent actual switching or cross-connect equipment, rather than more abstract entities, such as protocols within each of these technology layers that can create multiple protocol *sublayers*. An early manifestation of protocol layering is the OSI model developed by the ISO standards organization [37] and the resulting classification of packet layering, such as *Layer 1*, *Layer 2*, *Layer 3*, which subsequently emerged in the industry. Although these layering definitions can be somewhat strained in usage, the industry generally associates IP with Layer 3 and MPLS or Ethernet VLANs with Layer 2 (which will be described later in the chapter). Layer 1, or the Physical Layer (PHY layer) of the OSI stack, covers multiple technology layers that we will cover in the next section.

We illustrate this graphical network-layering model in Fig. 2.2, which depicts two layers. Note that for simplicity, we depict the edges in Fig. 2.2 as undirected. The cross-connect equipment represented by the nodes of Layer *U* (“upper layer”) connect to their counterpart nodes in Layer *L* (“lower layer”) by interlayer links, depicted as lightly dashed vertical lines. While this model has no specific geographical correlation, we note that the switching or cross-connect equipment represented in Layer *U* usually are colocated in the same buildings/locations (central offices in carrier networks) as their lower-layer counterparts in Layer *L*. In such representations, the interlayer links are called *intra-office* links. The links of Layer *U* are transported as connections in lower Layer *L*. For example, Fig. 2.2 highlights a link between nodes 1 and 6 of layer *U*. This link is transported via a connection between nodes 1 and 6 of Layer *L*. The path of this connection is shown through nodes (1, 2, 3, 4, 5, 6) at Layer *L*.

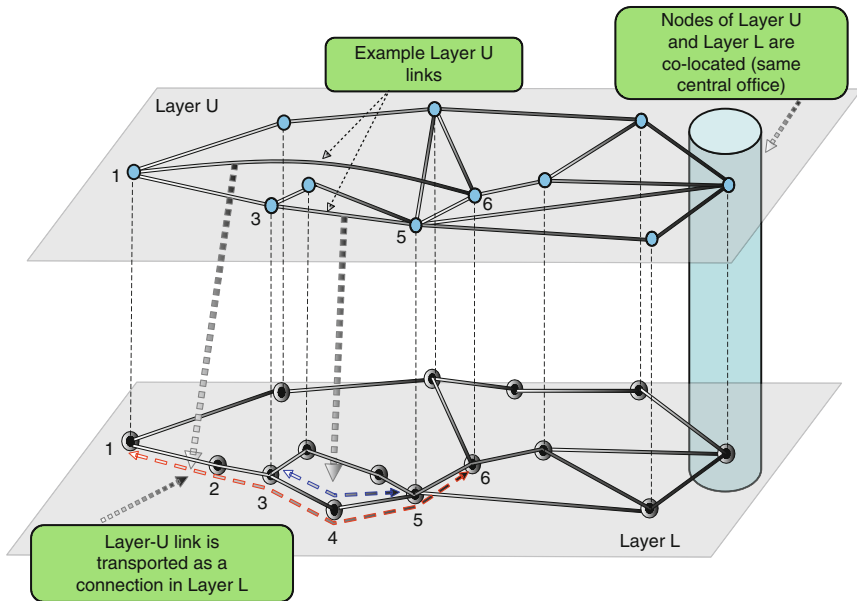


Fig. 2.2 Example of network layering

Another example is given by the link between nodes 3 and 5 of Layer U . This routes over nodes (3, 4, 5) in Layer L . As this layered model illustrates, the concept of a “link” is a logical construct, even in lower “physical layer(s)”. Along these lines, we identify some interesting observations in Fig. 2.2:

1. There are more nodes in Layer L than in Layer U .
2. When viewed as separate abstract graphs, the degree of logical connectivity in Layer L is less than that for Layer U . For example, there are at the most three edge-diverse paths between nodes 1 and 6 in layer U . However, there are at the most, only two edge-diverse paths between the corresponding pair of nodes in Layer L .
3. When we project the links of Layer U onto their connection paths in Layer L , we see some overlap. For example, the two logical links highlighted in Layer U overlap on links (3, 4) and (4, 5) of Layer L .

These observations generalize to the network layers associated with the IP backbone and affect how network layers are designed and how network failures at various layers affect higher-layer networks. The second observation says that while the logical topology of an upper-layer network, such as the IP layer, looks like it has many alternate paths to accommodate network disruptions, this can be deceiving unless one incorporates the lower-layer dependencies. For example, if link 3–4 of Layer L fails, then both links 1–6 and 3–5 of Layer U fail. Put more generally, failures of links of lower-layer networks usually cause multiple link failures in higher-layer networks. Specific examples will be described in Section 2.3.2.

2.2.3 Snapshot of Today’s Core Network Layers

Figure 2.3 provides a representation of the set of services that might be provided by a large US-based carrier, and how these services map onto different network layers in the core segment. This figure is borrowed from [11] and depicts a mixture of legacy network layers (i.e., older technologies slowly being phased out) and current or emerging network layers. For a connection-oriented network layer (call it layer L), demand for connections comes from two sources: (1) links of higher network layers that route over layer L and (2) demand for telecommunications services provided by layer L but which originate outside layer L ’s network segment. The second source of demand is depicted by rounded rectangles in Fig. 2.3. Note that Fig. 2.3 is a significant simplification of reality; however, it does capture most predominant layers and principal interlayer relationships relevant to our objectives. Note that an important observation in Fig. 2.3 is that links of a given layer can be spread over multiple lower layers including “skipping” over intermediate lower layers.

Before we describe these layers, we provide some preliminary background on Time Division Multiplexing (TDM), whose signals are often used to transport links of the IP layer. Table 2.1 summarizes the most common TDM transmission rates. The Synchronous Optical Network (SONET) digital-signal standard [35], pioneered

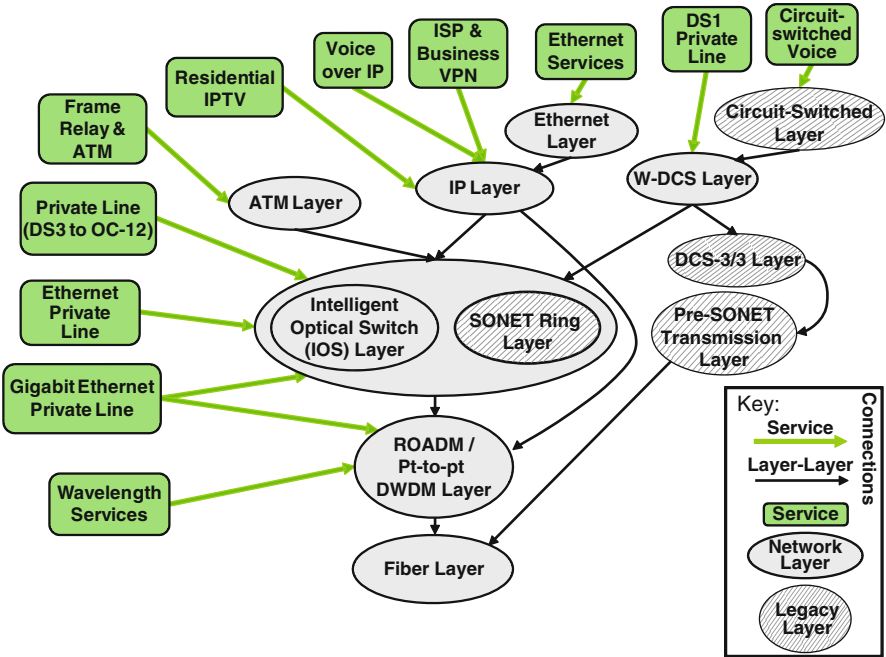


Fig. 2.3 Example of core-segment network layers

Table 2.1 Time division multiplexing (TDM) digital hierarchy (partial list)

Approximate rate	DS- <i>n</i>	Plesiosynchronous	SONET	SDH	OTN wrapper
64 Kb/s	DS-0	E0			
1.5 Mb/s	DS-1				
2.0 Mb/s		E-1			
34 Mb/s		E-3			
45 Mb/s	DS-3				
51.84 Mb/s			STS-1	VC-3	
155.5 Mb/s			OC-3	STM-1	
622 Mb/s			OC-12	STM-3	
2.5 Gb/s			OC-48	STM-16	ODU-1
10 Gb/s			OC-192	STM-48	ODU-2
40 Gb/s			OC-768	STM-192	ODU-3
100 Gb/s					ODU-4

Kb/s = kilobits per second; Mb/s = megabits per second; Gb/s = gigabits per second.
OTN line rates are higher than payload. ODU-2 includes 10 GigE and ODU-3 includes 40 GigE (under development). ODU-4 only includes 100 GigE

by Bellcore (now Telcordia) in the early 1990s, is shown in the fourth column of Table 2.1. SONET is the existing higher-rate digital-signal hierarchy of North America. Synchronous Digital Hierarchy (SDH) is a similar digital-signal standard later pioneered by the International Telecommunication Union (ITU-T) and adopted by most of the rest of the world. The *DS-*n** column represents the North American pre-SONET digital-signal rates, most of which originated in the Bell System. The Plesiosynchronous column represents the pre-SDH rates used mostly in Europe. However, after nearly 30 years, both *DS-*n** and Plesiosynchronous are still quite abundant and their related private-line services are still sold actively. Finally, in the last column, we show the more recent *Optical Transport Network (OTN)* signals, also standardized by the ITU-T [18]. Development of the OTN signal standards were originally motivated by the need for a more robust standard to achieve very high bit rates in *DWDM* technologies; for example, it was needed to incorporate and standardize various bit-error recovery techniques, such as *Forward Error Correction (FEC)*. As such, the OTN rates were originally termed “digital wrappers” to contain high rate SONET, SDH, or Ethernet signals, plus provide the extra fault notification information needed to reliably transport the high rates. Although there are many protocol layers in OTN, we just show the *Optical channel Data Unit (ODU)* rates in Table 2.1. To minimize confusion, in the rest of this chapter, we will mostly give examples in terms of *DS-*n** and SONET rates.

Referring back to the layered network model of the previous section, Table 2.2 gives some examples of the nodes, links, and connections in Fig. 2.3. We only list those layers that have relevance to the IP layer. We will briefly describe these layers in the following sections.

Table 2.2 Examples of nodes, links, and connections for network layers of Fig. 2.3

Core layer	Typical node	Typical link	Typical connection
IP	Router	SONET OC- n , 1/10 gigabit Ethernet, ODU- n	IP is connection-less
Ethernet	Ethernet switch or router with Ethernet functionality	1/10 Gigabit Ethernet or rate-limited Ethernet private line	Ethernet can refer to both connection-less and connection-oriented services
Asynchronous transfer mode (ATM)	ATM switch	SONET OC-12/48	Permanent virtual circuit (PVC), Switched virtual circuit (SVC)
W-DCS	Wideband digital cross-connect system (DCS)	SONET STS-1 (channelized)	DS1
SONET Ring	SONET add-drop multiplexer (ADM)	SONET OC-48/192	SONET STS- n , DS-3
IOS	Intelligent optical switch (IOS) or broadband digital cross-connect system (DCS)	SONET OC-48/192	SONET STS- n
DWDM	Point-to-point DWDM terminal or reconfigurable optical add-drop multiplexer (ROADM)	DWDM signal	SONET, SDN, or 1/10/100 gigabit Ethernet
Fiber	Fiber patch panel or cross-connect	Fiber optic strand	DWDM signal or SONET, SDH, or Ethernet signal

2.2.4 Fiber Layer

The commercial intercity fiber layer of the USA is privately owned by multiple carriers. In addition to owning fiber, carriers lease bundles of fiber from one another using various long-term *Indefeasible Right of Use (IROU)* contracts to cover needed connectivity in their networks. Fiber networks differ significantly between metro and rural areas. In particular, in carrier metro networks, optical fiber cables are usually placed inside PVC pipes, which are in turn placed inside concrete conduits. Additionally, fiber for core networks is often corouted in conduit or along rights-of-way with metro fiber. Generally, in metro areas, optical cables are routed and spliced between central offices. In the central office, most carriers prefer to connect the fibers to a fiber patch panel. Equipment that use (or will eventually use) the interoffice fibers are also cross-connected into the patch panels. This gives the carrier flexibility to connect equipment by simply connecting fiber patch cords on the patch panels. Rural areas differ in that there are often long distances between central offices and, as such, intermediate huts are used to splice fibers and place equipment, such as optical amplifiers.

2.2.5 DWDM Layer

Although many varieties of DWDM systems exist, we show a simplified view of a (one-way) point-to-point DWDM system in Fig. 2.4. Here, *Optical Transponders (OTs)* are *Optical-Electrical-to-Optical (O-E-O)* converters that input optical digital signals from routers, switches, or other transmission equipment using a receive device, such as a photodiode, on the *add/drop* side of the OT. The input signal has a standard intra-office wavelength, denoted by λ_0 . The OT converts the signal to electrical form. Various other physical layer protocols may be applied at this point, such as incorporating various handshaking called *Link Management Protocols (LMPs)* between the transmitting equipment and the receiving OT. A transponder is in *clear channel* mode if it does not change the transport protocols of the signal that it receives and essentially remains invisible to the equipment connecting to it. For example, *Gigabit Ethernet (GigE)* protocols from some routers or switches sometimes incorporate signaling messages to the far-end switch in the interframe gaps. If clear channel transmission is employed by the OT, such messages will be preserved as they are routed over the DWDM layer.

After conversion to electrical form, the signal is retransmitted using a laser on the *network* or *line-side* of the OT. However, typical of traditional point-to-point systems, the wavelength of the laser is fixed to correspond to the wavelength assigned to a specific channel of the DWDM system, λ_k . The output light pulses from multiple OTs at different wavelengths are then multiplexed into a single fiber by sending them through an optical multiplexer, such as an *Arrayed Waveguide Grating*

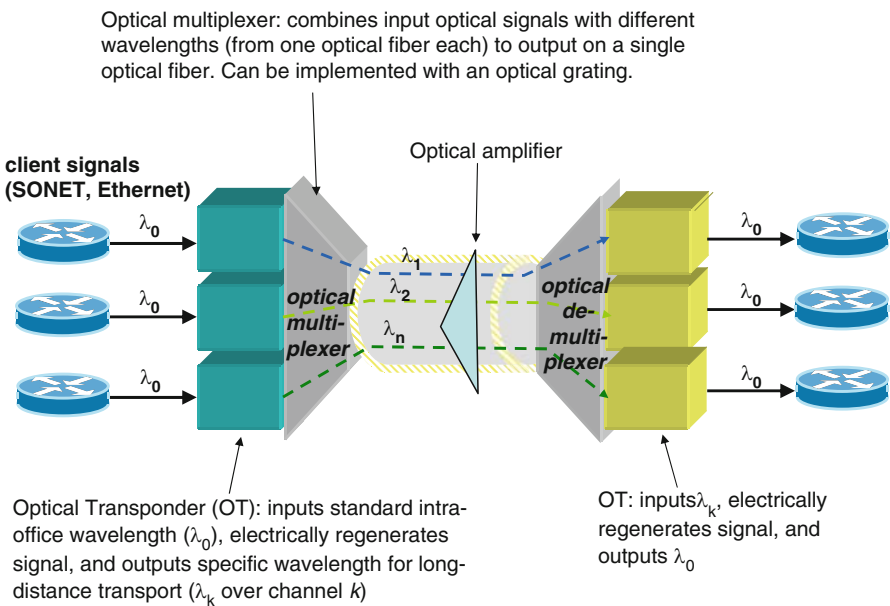


Fig. 2.4 Simplified view of point-to-point DWDM system

(AWG) or similar device. If the distance between the DWDM terminals is sufficiently long, optical amplifiers are used to boost the power of the signal. However, power balancing among the DWDM channels is a major concern of the design of the DWDM system, as are other potential optical impairments. These topics are beyond the scope of this chapter. On the right side of Fig. 2.4, typically, the same (or similar) optical multiplexer is used in reverse, in which case, it becomes an optical *demultiplexer*. The OTs on the right side (the receive direction of the DWDM system) basically work in reverse to the transmit direction described above, by receiving the specific interoffice wavelength, λ_k , converting to electrical, and then using a laser to generate the intra-office wavelength, λ_0 .

Carrier-based DWDM systems are usually deployed in bidirectional configurations. To see this, the reader can visually reproduce the entire system in Fig. 2.4 and then flip it (mirror it) right to left. The multiplexed DWDM signal in the opposite direction is transmitted over a separate fiber. Therefore, even though the electronics and lasers of the one-way DWDM system in the reverse direction operate separately from the shown direction, they are coupled operationally. For example, the two fiber ports (receive and transmit) of the OT are usually deployed on the same line card and arranged next to one another.

Optical amplification is used to extend the distance between terminals of a DWDM system. However, multiple systems are required to traverse the continental USA. Connections can be established between different point-to-point DWDM systems in an intermediate CO via an *intermediate-regenerator* OT (not pictured in Fig. 2.4). An intermediate-regenerator OT has the same effect on a signal as back-to-back OTs. Since the signal does not have to be cross-connected elsewhere in the intermediate central office, cost savings can be achieved by omitting the intermediate lasers and receivers of back-to-back OTs. However, we note that most core DWDM networks have many vintages of point-to-point systems from different equipment suppliers. Typically, an intermediate-regenerator OT can only be used to connect between DWDM systems of the same equipment supplier.

A difficulty with deploying point-to-point DWDM systems is that in central offices that interface multiple fiber spans (i.e., the node in the fiber layer has degree >2), all connections demultiplex in that office and pass through OTs. OTs are typically expensive and it is advantageous to avoid their deployment where possible. A better solution is the Reconfigurable Optical Add-Drop Multiplexer (ROADM). We show a simplified diagram of a ROADM in Fig. 2.5. The ROADM allows for multiple interoffice fibers to connect to the DWDM system. Appropriately, it is often called a *multidegree ROADM* or *n-degree ROADM*. As Fig. 2.5 illustrates, the ROADM is able to optically (i.e., without use of OTs) cross-connect channel k (transmitting at wavelength λ_k) arriving on one fiber to channel k (wavelength λ_k) outgoing on another fiber. Note that the same wavelength must be used on the two fibers. This is called the *wavelength continuity* constraint. The ROADM can also be configured to terminate (or “drop”) a connection at that location, in which case it is cross-connected to an OT to connect to routers, switches, or transmission equipment. A “dropped” connection is illustrated by λ_2 on the second fiber from the top on the left in Fig. 2.5 and an “added” connection is illustrated by λ_n on the bottom

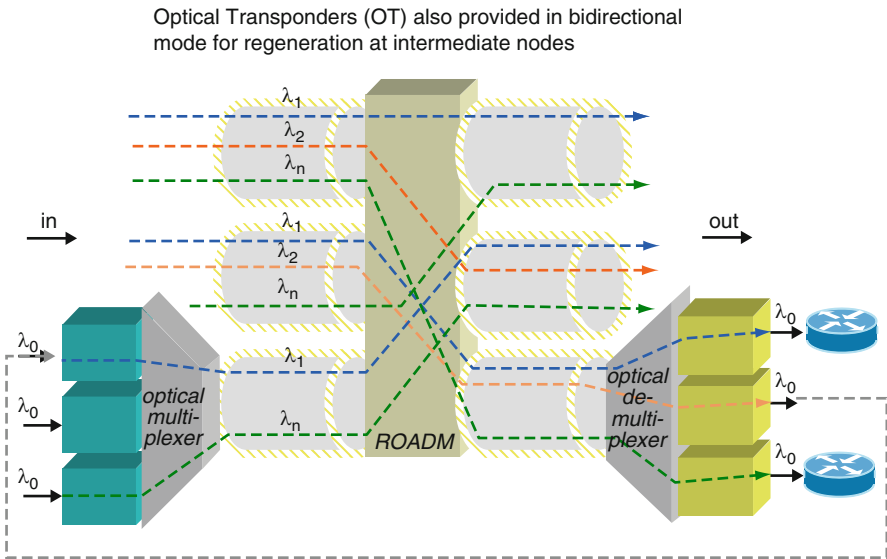


Fig. 2.5 Simplified view of Reconfigurable Optical Add-Drop Multiplexer (ROADM)

fiber on the left. As with the point-to-point DWDM system, optical properties of the system impose distance (also called *reach*) constraints.

Many transmission technologies, including optical amplification, are used to extend the distance between the optical add/drop points of a DWDM system. Today, this separation is designed to be about 1,500 km for a long-distance DWDM system, as a trade-off between cost and the all-optical distance for a US-wide network. Longer connections have to regenerate their signals, usually with an intermediate-regenerator OT. As with point-to-point DWDM systems, connections crossing ROADMS from different equipment suppliers usually must add/drop and connect through OTs.

We illustrate a representative ROADM layer for the continental USA in Fig. 2.6. The links represent fiber spans between ROADMS. As described above, to route a connection over the network of Fig. 2.6 may require points of regeneration. We also note, though, that today's core transport carriers usually have many vintages of DWDM technology and, thus, there may be several ROADM networks from different equipment suppliers, plus several point-to-point DWDM networks. All this complexity must be managed when routing higher-layer links, such as those of the IP backbone, over the DWDM layer.

We finish this introduction of the DWDM layer with a few observations. While most large carriers have DWDM technology covering their core networks, this is not generally true in the metro segment. The metro segment typically consists of a mixture of DWDM spans and fiber spans (i.e., spans with no DWDM). In fact, in metro areas usually only a fraction of central office fiber spans have DWDM technology routed over them. This affects how customers interface to the IP backbone network for higher-rate interfaces. Finally, we note that while most

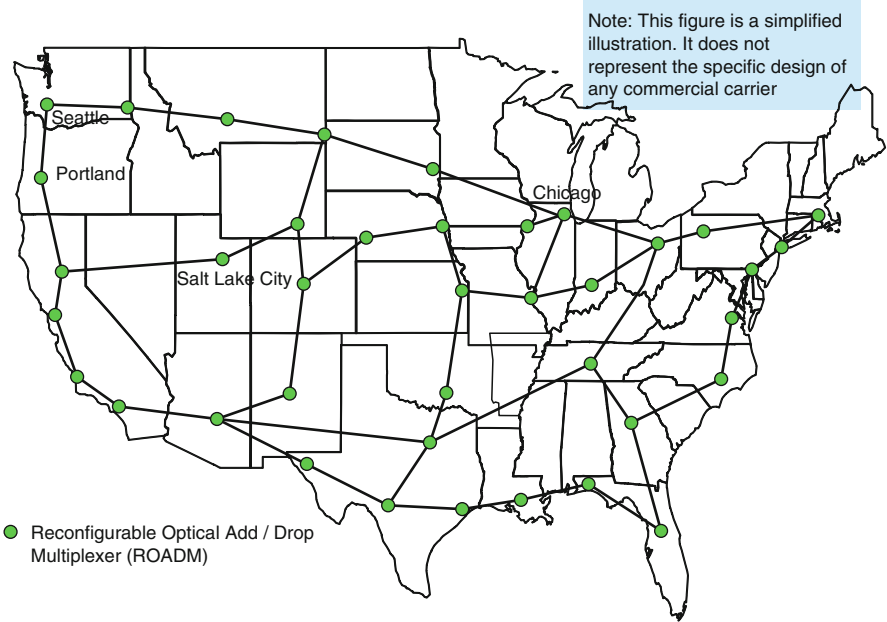


Fig. 2.6 Example of ROADM Layer topology

of the connections for the core DWDM layer arise from links of the IP layer, many of the connections come from what many colloquially call “wavelength services” (denoted by the rounded rectangle in Fig. 2.3). These come from high-rate private-line connections emanating from outside the core DWDM layer. Examples are links between switches of large enterprise customers that are connected by leased-line services.

2.2.6 TDM Cross-Connect Layers

In this section, we will briefly describe the TDM cross-connect layers. TDM cross-connect equipment can be basically categorized into two common types: a SONET/SDH *Add-Drop Multiplexer (ADM)* or a *Digital Cross-Connect System (DCS)*. Consistent with our earlier remark about the use of terminology, the latter often goes by a variety of colloquial or outmoded model names of equipment suppliers, such as DCS-3/1, DCS-3/3, DACS, and DSX. A TDM cross-connect device interfaces multiple high-rate digital signals, each of which uses time division multiplexing to break the signal into lower-rate channels. These channels carry lower-rate TDM connections and the TDM cross-connect device cross-connects the lower-rate signals among the channels of the different high-rate signals. Typically, an ADM only interfaces two high-rate signals, while a DCS interfaces many. However, over time these distinctions have blurred. Telcordia classified DCSs into three layers:

a narrowband DCS (N-DCS) cross-connects at the DS-0 rate, a wideband-DCS (W-DCS) cross-connects at the DS-1 rate, and a broadband-DCS (B-DCS) cross-connects at the DS-3 rate or higher. ADMs are usually deployed in SONET/SDH *self-healing rings*. The IOS and SONET Ring layers are shown in Fig. 2.3, encircled by the (broader) ellipse that represents the TDM cross-connect devices. More details on these technologies can be found in [11]. Self-healing rings and DCSs will be relevant when we illustrate how services access the wide-area ISP network layer later in this chapter.

Despite the word “optical” in its name, an *Intelligent Optical Switch (IOS)* is a type of B-DCS. Examples can be found in [6, 34]. The major differentiator of the IOS over older B-DCS models is its advanced control plane. An IOS network can route connection requests under distributed control, usually instigated by the source node. This requires mechanisms for distributing topology updates and internodal messaging to set up connections. Furthermore, an IOS usually can restore failed connections by automatically rerouting them around failed links. More detail is given when we discuss restoration methods.

Many of the connections for the core TDM-cross-connect layers (ring layers, DCS layers, IOS layer) come from higher layers of the core network. For example, many connections of the IOS layer are links between W-DCSs, ATM networks, or lower-rate portions of IP layer networks. However, much of their demand for connections comes from subwavelength private-line services, shown by the rounded rectangle in Fig. 2.3. A portion of this private-line demand is in the form of *Ethernet Private Line (EPL)* services. These services usually represent links between Ethernet switches or routers of large enterprise customers. For example, the Gigabit Ethernet signal from an enterprise customer’s switch is transported over the metro network and then interfaces an Ethernet card either residing on the IOS itself or on an ADM that interfaces directly onto the IOS. The Ethernet card encapsulates the Ethernet frames inside concatenated $n \times$ STS-1 signals that are transported over the IOS layer. The customer can choose the rate of transport, and hence the value of n he/she wishes to purchase. The ADM Ethernet card polices the incoming Ethernet frames to the transport rate of $n \times$ STS-1.

2.2.7 IP Layer

The nodes of the IP layer shown in Fig. 2.3 represent routers that transport packets among metro area segments. IP generally define pairwise *adjacencies* between ports of the routers. In the IP backbone, these adjacencies are typically configured over SONET, SDH, or Ethernet, or OTN interfaces on the routers. As described above, these links are then transported as connections over the interoffice lower-layer networks shown in Fig. 2.3. Note that different links can be carried in different lower-layer networks. For example, lower-rate links may be carried over the TDM cross-connect layers (IOS or SONET Ring), while higher-rate links may be carried directly over the DWDM layer, thus “skipping” the TDM cross-connect layers. We will describe the IP layer in more detail in subsequent sections.

2.2.8 Ethernet Layer

The *Ethernet layer* in Fig. 2.3 refers to several applications of Ethernet technology. For example, Ethernet supports a number of physical layer standards that can be used for Layer 1 transport. Ethernet also refers to connection-oriented Layer 2 *pseudowire* services [16] and connection-less *transparent LAN* services. For example, intra-office links between routers often use an Ethernet physical layer riding on optical fiber.

An important application of Ethernet today is providing wide-area Layer 2 Virtual Private Network (VPN) services for enterprise customers. Although many variations exist, these services generally support enterprise customers that have Ethernet LANs at multiple locations and need to interconnect their LANs within a metro area or across the wide area. Most large carriers provide these services as an overlay on their IP layer, and hence, why we show the layered design in Fig. 2.3. Prior to the ability to provide such services over the IP layer, Ethernet private lines were supported by TDM cross-connect layers (i.e., Ethernet frames encapsulated over Layer 1 TDM private lines as described in Section 2.2.6). However, analogous to why wide-area Frame Relay displaced wide-area DS-0 private lines in the 1990s, wide-area packet networks are often more efficient than private lines to connect LANs of enterprise customers.

The principal approach that intermetro carriers use to provide wide-area Ethernet private network services is *Virtual Private LAN Service (VPLS)* [24, 25]. In this approach, carriers provide such Ethernet services with routers augmented with appropriate Ethernet capabilities. The reason for this approach is to provide the robust carrier-grade network capabilities provided by routers. With wide-area VPLS, the enterprise customer is connected via the metro network to the edge routers on the edge of the core IP layer. We describe how the metro network connects to the core IP layer network in the next section. The VPLS architecture is described in more detail in Section 2.4.2 when we describe MPLS.

We conclude this section with the comment that standards organizations and industry forums (e.g., IEEE, IETF, and Metro Ethernet Forum) have explored the use of Ethernet switches with upgraded carrier-grade network control protocols rather than using routers as nodes in the IP layer. For example, see *Provider Backbone Transport (PBT)* [27] and *Provider Backbone Bridge – Traffic Engineering (PBB-TE)* [15]. However, most large ISPs are deploying MPLS-based solutions. Therefore, we concentrate on the layering architecture shown in Fig. 2.3 in the remainder of this chapter.

2.2.9 Miscellaneous/Legacy Layers

For completeness, we depict other “legacy” network layers with dashed ovals in Fig. 2.3. These technologies have been around for decades in most carrier-based core networks. They include network layers whose nodes represent ATM

switches, Frame-Relay switches, DCS-3/3s (a B-DCS that cross-connects DS3s), Voice-switches (DS-0 circuit switches), and pre-SONET ADMs. Most of these layers are not material to the spirit of this chapter and we do not discuss them here.

2.3 Structure of Today’s Core IP Layer

2.3.1 Hierarchical Structure and Topology

In this chapter, we further break the IP layer into *Access Routers (ARs)* and *Backbone Routers (BRs)*. Customer equipment homes to access routers, which in turn home onto backbone routers. An AR is either colocated with its backbone routers or not; the latter is called a *Remote Access Router (RAR)*. Of course, there are alternate terminologies. For example, the IETF defines similar concepts to customer equipment, access routers, and backbone routers with its definitions, respectively, of *Customer-Edge (CE)* equipment, *Provider-Edge (PE)* routers, and *Provider (P)* routers. A simplified picture of a typical central office containing both ARs and BRs is shown in Fig. 2.7. Access routers are *dual-homed* to two backbone routers to enable higher levels of service availability. The links between routers in the same office are typically Ethernet links over intra-office fiber. While we show only two ARs in

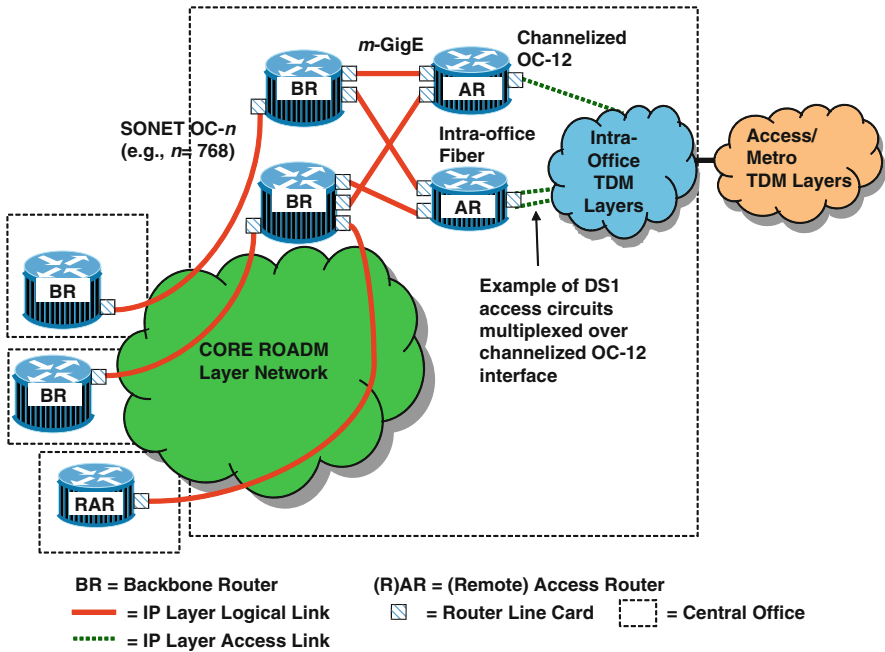


Fig. 2.7 Legacy central office interconnection diagram (Layer 3)

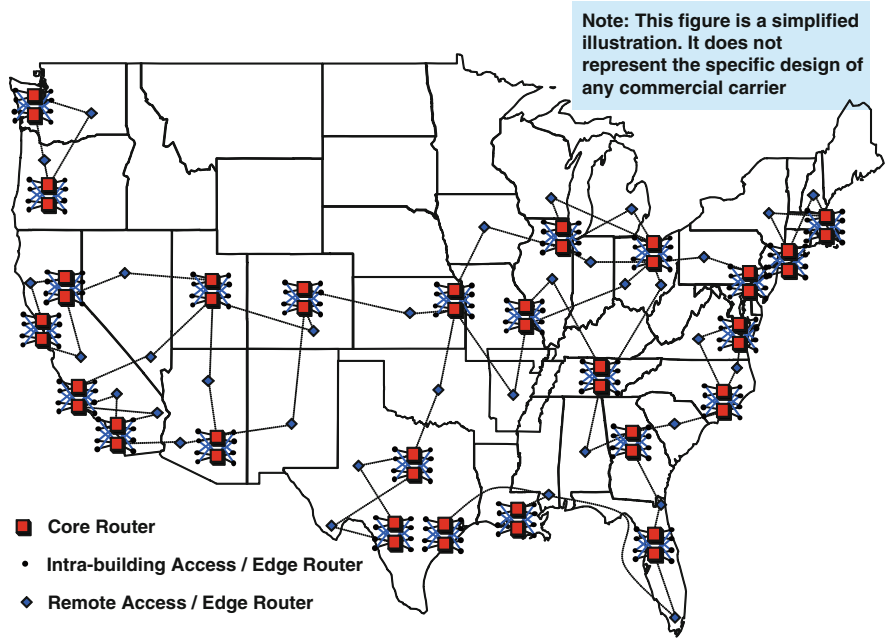


Fig. 2.8 Example of IP layer switching hierarchy

Fig. 2.7, note that typically there are many ARs in large offices. Also, due to scaling and sizing limitations, there may be more than two backbone routers or switches per central office used to further aggregate AR traffic before it enters the BRs.

Moreover, we show a remote access router that homes to one of the BRs. Figure 2.8 illustrates this homing arrangement in a broader network example, where small circles represent ARs, diamonds represent RARs, and large squares represent BRs. Note that remote ARs are homed to BRs in different offices. Homing remote ARs to BRs in different central offices raises network availability. However, a stronger motivation for doing this is that RAR–BR links are usually routed over the DWDM layer, which generally does not offer automatic restoration, and so the dual-homing serves two purposes: (1) protect against BR failure or maintenance activity and (2) protect against failure or maintenance of a RAR–BR link.

While the homing scheme described here is typical of large ISPs, other variations exist. For example, there are dual-homing architectures where (nonremote) ARs are homed to a BR colocated in the same central office and then a second BR in a different central office. While this latter architecture provides a slightly higher level of network availability against broader central office failure, it can be more costly owing to the need to transport the second AR–BR link. However, the latter architecture allows more load balancing across BRs because of the extra flexibility in homing ARs.

Improved load balancing can offer other advantages, including lower BR costs. Also, for ISPs with many scattered locations, but less total traffic, this latter architecture may be more cost-effective than colocating two BRs in each BR-office.

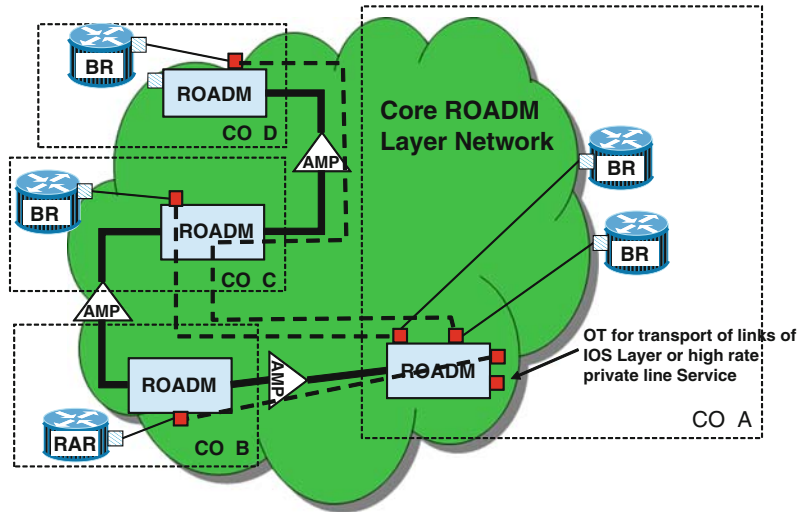
The right side of Fig. 2.7 also shows the metro/access network-layer clouds to connect customer equipment to the ARs. In particular, we illustrate DS1 customer interfaces. The left side of Fig. 2.7 also shows the lower-layer DWDM clouds to connect the interoffice links between BRs. We will expand these clouds in the next sections.

The reasons for segregating the IP topology into access and backbone routers are manifold:

- Access routers aggregate lower-rate interfaces from various customers or other carriers. This function requires significant equipment footprint and processor resources for customer-related protocols. As a result, major central offices consist of many access routers to accommodate the low-rate customer interfaces. Without the aggregation function of the backbone router, each such office would be a myriad of tie links between access routers and interoffice links.
- Access routers are often segregated by different services or functions. For example, general residential ISP service can be segregated from high-priority enterprise private VPN service. As another example, some access routers are sometimes segregated to be peering points with other carriers.
- Backbone routers are primarily designed to be IP-transport switches equipped only with the highest speed interfaces. This segregation allows the backbone routers to be optimally configured for interoffice IP forwarding and transport.

2.3.2 Interoffice Topology

Figure 2.9 expands the core lower ROADM Layer cloud of Fig. 2.7. It shows ports of interoffice links between BRs connecting to ports on ROADMs. These links are transported as connections in the ROADM network. For example, today these links go up to 40 gigabits per second (Gb/s) or SONET OC-768. These connections are routed optically through intermediate ROADMs and regenerated where needed, as described in Section 2.2.5. Also, we note that the link between the remote ARs and BRs route over the same ROADM network, although the rate of this RAR-BR link may be at lower rate, such as 10 Gb/s. Figure 2.10 shows a network-wide example of the IP layer interoffice topology. There are some network-layering principles illustrated in Fig. 2.10 that we will describe. First, if we compare the IP layer topology of Fig. 2.8 with that of the DWDM layer (ROADM layer) of Fig. 2.10, we note that there is more connectivity in the IP layer graph than the DWDM layer. The reason for this is the existence of what many IP layer planners call *express* links. If we examine the link labeled “direct link” between Seattle and Portland, we find that when we route this link over the DWDM layer topology, there are no intermediate ROADMs. In fact, there are two types of direct links. The first type connects through



ROADM = Reconfigurable Optical Add-Drop Multiplexer (R)AR = (Remote) Access Router
BR = Backbone Router □ = Central Office (CO)
■ = ROADM Optical Transponder (OT) □ = Router Line Card
-- = ROADM Layer connection transporting IP layer link

Fig. 2.9 Core ROADM Layer diagram

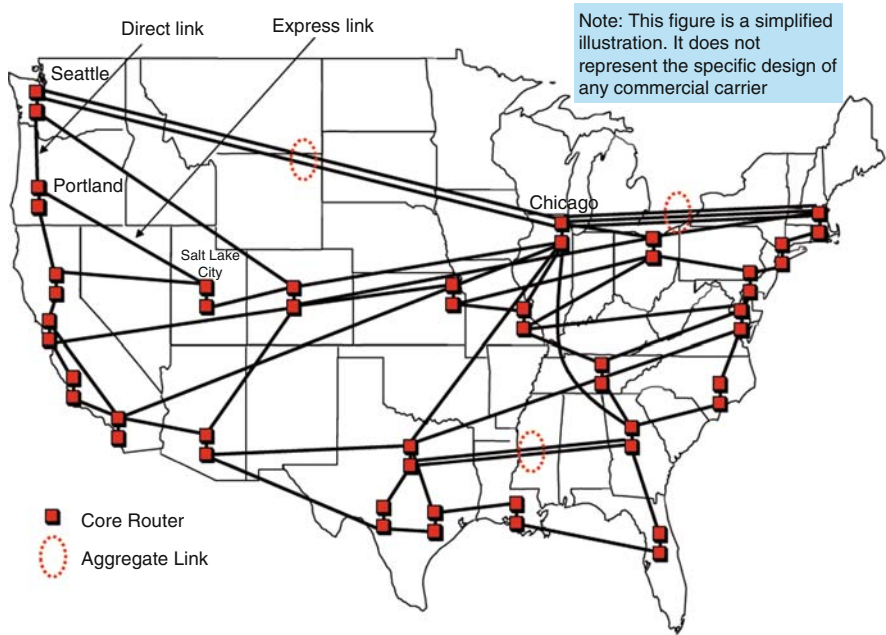


Fig. 2.10 Example of IP layer interbackbone topology

no intermediate ROADMs, as illustrated by the Seattle–Portland link. The second type connects through intermediate ROADMS, but encounters no BRs in those intermediate central offices, as illustrated by the Seattle–Chicago link.

In contrast, if we examine the express link between Portland and Salt Lake City, we find that any path in the DWDM layer connecting the routers in that city pair bypasses routers in at least one of its intermediate central offices. Express links are primarily placed to minimize network costs. For example, it is more efficient to place express links between well-chosen router pairs with high network traffic (enough to raise the link utilization above a threshold level); otherwise the traffic will traverse through multiple routers. Router interfaces can be the most-expensive single component in a multilayered ISP network; therefore, costs can usually be minimized by optimal placement of express links.

It is also important to consider the impact of network layering on network reliability. Referring to the generic layering example of Fig. 2.2, we note that the placement of express links can cause a single DWDM link to be shared by different IP layer links. This gives rise to complex network disruption scenarios, which must be modeled using sophisticated network survivability modeling tools. This is covered in more detail in Section 2.5.3.

Returning to Fig. 2.10, we also note the use of *aggregate links*. Aggregate links also go by other names, such as *bundled links* and *composite links*. An aggregate link bundles multiple physical links between a pair of routers into a single virtual link from the point of view of the routers. For example, an aggregate link could be composed of five OC-192 (or 10 GigE) links. Such an aggregate link would appear as one link with 50 Gb/s of capacity between the two routers. Generally, aggregate links are implemented by a load-balancing algorithm that transparently switches packets among the individual links. Usually, to reduce jitter or packet reordering, packets of a given IP flow are routed over the same component link. The main advantage of aggregate links is that as IP networks grow large, they tend to contain many lower-speed links between a pair of routers. It simplifies routing and topology protocols to aggregate all these links into one. If one of the component links of an aggregate link fails, the aggregate link remains up; consequently, the number of topology updates due to failure is reduced and network rerouting (called *reconvergence*) is less frequent. Network operators seek to achieve network stability, and therefore shy away from many network reconvergence events; aggregate links result in less network reconvergence events.

On the downside, if only one link of a (multiple link) aggregate link fails, the aggregate link remains “up”, but with reduced capacity. Since many network routing protocols are capacity in-sensitive, packet congestion could occur over the aggregate link. To avoid this situation, router software is designed with capacity thresholds for aggregate links that the network operator can set. If the aggregate capacity falls below the threshold, the entire aggregate link is taken out of service. While the network “loses” the capacity of the surviving links in the bundle when the aggregate link is taken out of service, the alternative is potentially significant packet loss due to congestion on the remaining links.

2.3.3 Interface with Metro Network Segment

Figure 2.11 is a blowup of the clouds on the right side of Fig. 2.7. It provides a simplified example of how three business ISP customers gain access to the IP backbone. These could be enterprise customers with multiple branches who subscribe to a VPN service. Each access method consists of a DS1 link encapsulating IP packets that is transported across the metro segment. In carrier vernacular, using packet/TDM links to access the IP backbone is often called *TDM backhaul*. We do not show the inner details of the metro network here. Detailed examples can be found in [11]. Even suppressing the details of the complex metro network, the TDM backhaul is clearly a complicated architecture. To aid his/her understanding, we suggest the reader to refer back to the TDM hierarchy shown in Table 2.1.

The customer's DS-1 (which carries encapsulated IP packets) interfaces to a low-speed multiplexer located in the customer building, such as a small SONET ADM. This ADM typically serves as one node of a SONET ring (usually a 2-node ring). Each link of the ring is routed over diverse fiber, usually at OC-3 or OC-12 rate. Eventually, the DS-1 is routed to a SONET OC-48 or OC-192 ring that has one of its ADMs in the POP. The DS-1 is transported inside an STS-1 signal that is divided into 28 time slots called *channels* (a *channelized STS-1*), as specified by the SONET standard. The ADM routes all the SONET STS-1s carrying DS-1 traffic bound for the core carrier to a metro W-DCS. Note that there are often multiple

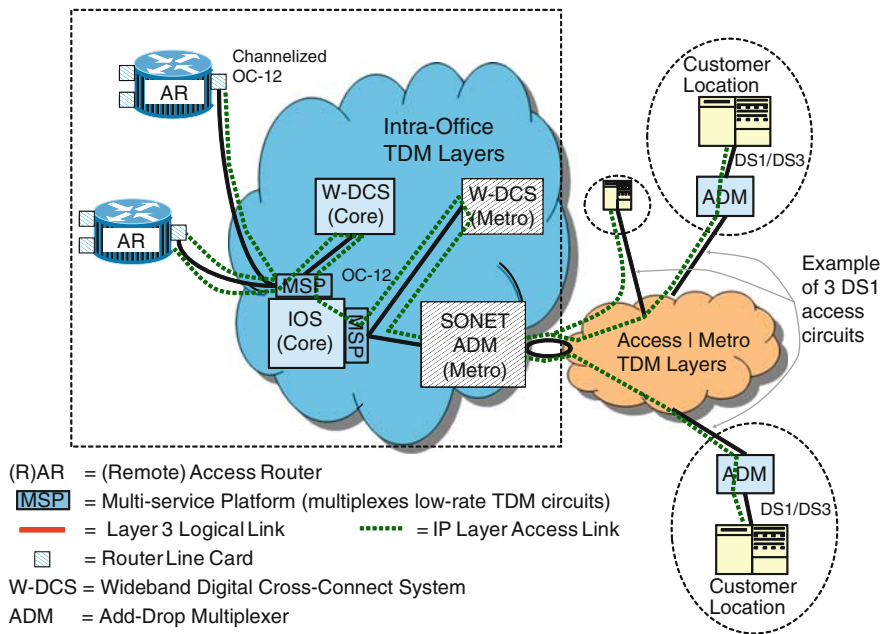


Fig. 2.11 Legacy central office interconnection diagram (intra-office TDM layers)

core carriers in a POP, and hence, the metro W-DCS cross-connects all the DS-1s destined for a given core carrier into channelized STS-1s and hands them off to the core W-DCS(s) of that core carrier. However, note that this handoff does not occur directly between the two W-DCSs, but rather passes through a higher-rate B-DCS, in this case the Intelligent Optical Switch (IOS) introduced in Section 2.2.6. The IOS cross-connects most of the STS-1s (multiplexed into OC- n interfaces) in a central office. Also, notice that the IOS is fronted with *Multi-Service Platforms (MSPs)*. An MSP is basically an advanced form of SONET ADM that gathers many types of lower-speed TDM interfaces and multiplexes them up to OC-48 or OC-192 for the IOS. It usually also has Ethernet interfaces that encapsulate IP packets into TDM signals (e.g., for Ethernet private line discussed earlier). The purpose of such a configuration is to minimize the cost and scale of the IOS by avoiding using its interface bay capacity for low-speed interfaces.

Finally, the core W-DCS cross-connects the DS1s destined for the access routers in the central office onto channelized STS-1s. Again, these STS-1s are routed to the AR via the IOS and its MSPs. The DS-1s finally reach a channelized SONET card on the AR (typically OC-12). This card on the AR de-multiplexes the DS-1s from the STS-1, de-encapsulates the packets, and creates a virtual interface for each of our three example customer access links in Fig. 2.11. The channelized SONET card is colloquially called a *CHOC* card (CHannelized OC- n).

Note that the core and metro carriers depicted in Fig. 2.11 may be parts of the same corporation. However, this complex architecture arose from the decomposition of long-distance and local carriers that was dictated by US courts and the *Federal Communications Commission (FCC)* at the breakup of the Bell System in 1984. It persists to this day.

If we reexamine the above TDM metro access descriptions, we find that there are many restoration mechanisms, such as dual homing of the ARs to the BRs and SONET rings in the metro network. However, there is one salient point of potential failure. If an AR customer-facing line card or entire AR fails or is taken out of service for maintenance in Fig. 2.11, then the customer's service is also down. Carriers offer service options to protect against this. The most common provide two TDM backhaul connections to the customer's equipment, often called *Customer Premise Equipment (CPE)*, each of which terminates on a different access router. This architecture significantly raises the availability of the service, but does incur additional cost. An example of such a service is given in [1].

To retain accuracy, we make a final technical comment on the example of Fig. 2.11. Although we show direct fiber connections between the various TDM and packet equipment, in fact, most of these usually occur via a fiber patch panel. This enables a craftsperson to connect the equipment via a simple (and well-organized) patch chord or cross-connect. This minimizes expense, simplifies complex wiring, and expedites provisioning work orders in the CO.

Figure 2.12 depicts how customers access the AR via emerging metro packet network layers instead of TDM. Here, instead of the traditional TDM network, the customer accesses the packet core via Ethernet. The most salient difference is the substantially simplified architecture. Although many different types of services

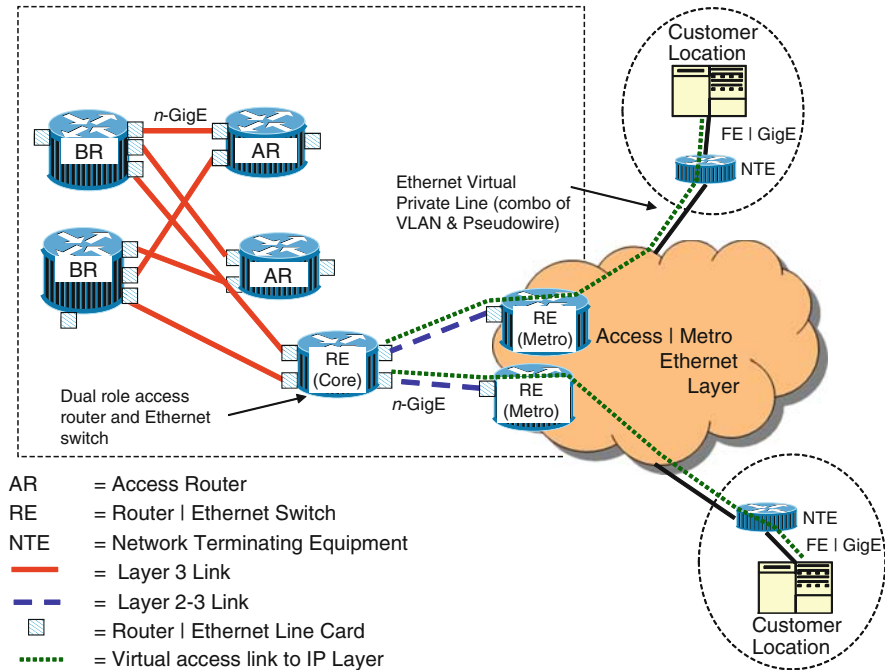


Fig. 2.12 Central office interconnection diagram (metro Ethernet interface)

are possible, we describe two fundamental types of Ethernet service: Ethernet virtual circuits and Ethernet VPLS. Most enterprise customers will use both types of services.

There are three basic types of connectivity for Ethernet virtual circuits: (1) intrametro, (2) ISP access via establishment of Ethernet virtual circuits between the customer location and IP backbone, and (3) intermetro. Since our main focus is the core IP backbone, we discuss the latter two varieties. For ISP access, in the example of Fig. 2.12, the customer's CPE interfaces the metro network via *Fast Ethernet (FE)* or GigE into a small Ethernet switch placed by the metro carrier called *Network Terminating Equipment (NTE)*. The NTE is the packet analog of the small ADM in the TDM access model in Fig. 2.11. For most metro Ethernet services, the customer can usually choose which policed access rate he/she wishes to purchase in increments of 1 Mb/s or similar. For example, he/she may wish 100 Mb/s for his/her *Committed Information Rate (CIR)* and various options for his/her *Excess Information Rate (EIR)*. The EIR options control how his bandwidth bursts are handled/shared when they exceed his CIR. The metro packet networks uses *Virtual Local Area Network (VLAN)* identifiers [14] and pseudowires or MPLS LSPs to route the customer's Ethernet virtual circuit to the metro Ethernet switch/router in the POP, as shown in Fig. 2.12. VLANs can also be used to segregate a particular customer's services, such as the two fundamental services (VPLS vs Internet access) described here. The *metro* Ethernet switch/router has high-speed links

(such as 10 Gb/s) to the *core* Ethernet switch/router. However, the core Ethernet switch/router is fundamentally an access router, but with the needed features and configurations needed to provide Ethernet and VPLS, and thus homes to backbone routers as any other access router. Thus, the customer's virtual circuit is mapped to a virtual port on the core AR/Ethernet-Switch and from that point onward is treated similarly as the TDM DS-1 virtual port in Fig. 2.11. If an intermetro Ethernet virtual circuit is needed, then an appropriate pseudowire or tunnel can be created between the ARs in different metros. Such a service can eventually substitute for traditional private-line service as metro packet networks are deployed.

The second basic type of Ethernet service type is generally provided through the VPLS model described in Section 2.2.8. For example, the customer might have two LANs in metro-1, one LAN in metro-2 and another LAN in metro-3. Wide-area VPLS interconnects these LANs into a large transparent LAN. This is achieved using pseudowires (tunnels) between the ARs in metros-1, 2, and 3. Since the core access router has a dual role as access router and Ethernet VPLS switch, it has the abilities to route customer Ethernet frames among pseudowires among the remote access routers.

Besides enterprise Ethernet services, connection of cellular base stations to the IP backbone network is another important application of Ethernet metro access. Until recently, this was achieved by installing DS-1s from cell sites to circuit switches in *Mobile Telephone Switching Offices (MTSOs)* to provide voice service. However, with the advent and rapid growth of cellular services based on 3G or 4G technology, there is a growing need for high-speed packet-based transport from cell sites to the IP backbone. The metro Ethernet structure for this is similar to that of the enterprise customer access shown in Fig. 2.12. The major differences occur in the equipment at the cell site, the equipment at the MTSO, and then how this equipment connects to the access router/Ethernet switch of the IP backbone.

2.4 Routing and Control in ISP Networks

2.4.1 IP Network Routing

The IP/MPLS routing protocols are an essential part of the architecture of the IP backbone, and are key to achieving network reliability. This section introduces these control protocols.

An Interior Gateway Protocol (IGP) disseminates routing and topology information within an *Autonomous System (AS)*. A large ISP will typically segment its IP network into multiple autonomous systems. In addition, an ISP's network interconnects with its customers and with other ISPs. The *Border Gateway Protocol (BGP)* is used to exchange global reachability information with ASs operated by the same ISP, by different ISPs, and by customers. In addition, IP multicast is becoming more widely deployed in ISP networks, using one of several variants of the *Protocol-Independent Multicast (PIM)* routing protocol.

2.4.1.1 Routing with Interior Gateway Protocols

As described earlier, Interior Gateway Protocols are used to disseminate routing and topology information within an AS. Since IGPs disseminate information about topology changes, they play a critical role in network restoration after a link or node failure. Because of the importance of restoration to the theme of this chapter, we discuss this further in Section 2.5.2.

The two types of IGPs are distance vector and link-state protocols. In link-state routing [32], each router in the AS maintains a view of the entire AS topology using a *Shortest Path First (SPF)* algorithm. Since link-state routing protocols such as *Open Shortest Path First (OSPF)* [26] and *Intermediate System–Intermediate System (IS–IS)* [30] are the most commonly used IGPs among large ISPs, we will not discuss distance vector protocols further. For the purposes of this chapter, which focuses on network restoration, the functionality of OSPF and IS–IS are similar. We will use OSPF to illustrate how IGPs handle failure detection and recovery.

The view of network topology maintained by OSPF is conceptually a directed graph. Each router represents a vertex in the topology graph and each link between neighboring routers represents a unidirectional edge. Each link also has an associated weight (also called *cost*) that is administratively assigned in the configuration file of the router. Using the weighted topology graph, each router computes a shortest path tree (SPT) with itself as the root, and applies the results to build its forwarding table. This assures that packets are forwarded along the shortest paths in terms of link weights to their destinations [26]. We will refer to the computation of the shortest path tree as an *SPF computation*, and the resultant tree as an *SPF tree*.

As illustrated in Fig. 2.13, the OSPF topology may be divided into areas, typically resulting in a two-level hierarchy. Area 0, known as the “backbone area”, resides at the top level of the hierarchy and provides connectivity to the nonbackbone areas (numbered 1, 2, etc.). OSPF typically assigns a link to exactly one area. Links may be in multiple areas, and multi-area links are addressed in more detail in Chapter 11 (Measurements of Control Plane Reliability and Performance by Aman Shaikh and Lee Breslau). Routers that have links to multiple areas are called *border routers*. For example, routers E, F and I are border routers in Fig. 2.13. Every router maintains its own copy of the topology graph for each area to which it is connected. The router performs an SPF computation on the topology graph for each area and thereby knows how to reach nodes in all the areas to which it connects. To improve scalability, OSPF was designed so that routers do not need to learn the entire topology of remote areas. Instead, routers only need to learn the total weight of the path from one or more area border routers to each node in the remote area. Thus, after computing the SPF tree for the area it is in, the router knows which border router to use as an intermediate node for reaching each remote node.

Every router running OSPF is responsible for describing its local connectivity in a *Link-State Advertisement (LSA)*. These LSAs are flooded reliably to other routers in the network, which allows them to build their local view of the topology. The flooding is made reliable by each router acknowledging the receipt of every LSA it receives from its neighbors. The flooding is hop-by-hop and hence does not depend

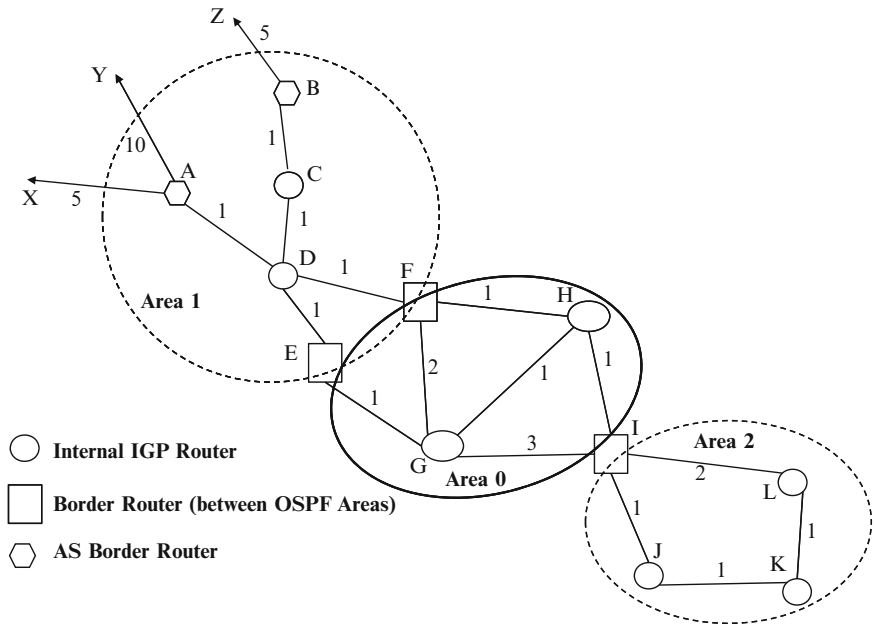


Fig. 2.13 OSPF topology: areas and hierarchy

on routing. The set of LSAs in a router’s memory is called a *Link-State Database (LSDB)* and conceptually forms the topology graph for the router.

OSPF uses several types of LSAs for describing different parts of topology. Every router describes links to all its neighbor routers in a given area in a *Router LSA*. Router LSAs are flooded only within an area and thus are said to have an area-level flooding scope. Thus, a border router originates a separate Router LSA for every area to which it is connected. Border routers summarize information about one area and distribute this information to adjacent areas by originating *Summary LSAs*. It is through Summary LSAs that other routers learn about nodes in the remote areas. Summary LSAs have an area-level flooding scope like Router LSAs. OSPF also allows routing information to be imported from other routing protocols, such as BGP. The router that imports routing information from other protocols into OSPF is called an *AS Border Router (ASBR)*. Routers A and B are ASBRs in Fig. 2.13. An ASBR originates External LSAs to describe the external routing information. The External LSAs are flooded in the entire AS irrespective of area boundaries, and hence have an AS-level flooding scope. While the capability exists to import external routing information from protocols such as BGP, the number of such routes that may be imported may be very large. As a result, this can lead to overheads both in communication (flooding the external LSAs) as well as computation (SPF computation scales with the number of routes). As a consequence of the scalability problems they pose, the importing of external routes is rarely utilized.

Two routers that are neighbor routers have link-level connectivity between each other. Neighbor routers form an *adjacency* so that they can exchange routing

information with each other. OSPF allows a link between the neighbor routers to be used for forwarding only if these routers have the same view of the topology, i.e., the same link-state database. This ensures that forwarding data packets over the link does not create loops. Thus, two neighbors have to make sure that their link-state databases are synchronized, and they do so by exchanging parts of their link-state databases when they establish an adjacency. The adjacency between a pair of routers is said to be “full” once they have synchronized their link-state databases. While sending LSAs to a neighbor, a router bundles them together into a *Link-State Update* packet. We will re-examine the OSPF convergence process in more detail when we discuss network disruptions in Section 2.5.2.1.

Although elegant and simple, basic OSPF is insensitive to network capacity and routes packets hop-by-hop along the SPF tree. As mentioned in Section 2.3.2, this has some potential shortcomings when applied to aggregate links. While aggregate-link capacity thresholds can be tuned to minimize this potentially negative effect, a better approach may be to use capacity-sensitive routing protocols, often called *Traffic Engineering (TE)* protocols, such as *OSPF-TE* [21]. Alternatively, one may use routing protocols with a greater degree of routing control, such as MPLS-based protocols. Traffic Engineering and MPLS are discussed later in this chapter.

2.4.1.2 Border Gateway Protocol

The Border Gateway Protocol is used to exchange routing information between autonomous systems, for example, between ISPs or between an ISP and its large enterprise customers. When BGP is used between ASs, it is referred to as *Exterior BGP (eBGP)*. When BGP is used within an AS to distribute external reachability information, it is referred to as *Interior BGP (iBGP)*. This section provides a brief summary of BGP. It is covered in much greater detail in Chapters 6 and 11.

BGP is a connection-oriented protocol that uses TCP for reliable delivery. A router advertises *Network Layer Reachability Information (NLRI)* consisting of an IP address prefix, a prefix length, a BGP next hop, along with path attributes, to its BGP peer. Packets matching the route will be forwarded toward the BGP next hop. Each route announcement can also have various attributes that can affect how the peer will prioritize its selection of the best route to use in its routing table. One example is the *AS_PATH* attribute which is a list of ASes through which the route has been relayed.

Withdrawal messages are sent to remove NLRI that are no longer valid. For example in Fig. 2.14, $A|z$ denotes an advertisement of NLRI for IP prefix z , and $W|s, r$ denotes that routes s and r are being withdrawn and should be removed from the routing table. If an attribute of the route changes, the originating router announces it again, replacing the previous announcement. Because BGP is connection-oriented, there are no refreshes or reflooding of routes during the lifetime of the BGP connection, which makes BGP simpler than a protocol like OSPF. However, like OSPF, BGP has various timers affecting behavior like hold-offs on route installation and route advertisement.

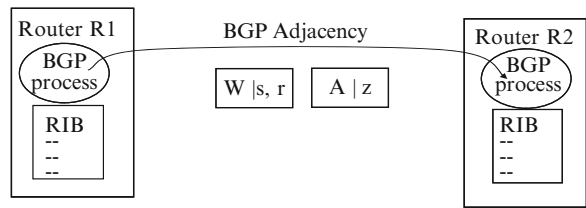


Fig. 2.14 BGP message exchange

BGP maintains tables referred to as *Routing Information Bases* (RIBs) containing BGP routes and their attributes. The Loc-RIB table contains the router’s definitive view of external routing information. Besides routes that enter the RIB from BGP itself, routes enter the RIB via distribution from other sources, such as static or directly connected routes or routing protocols such as OSPF. While the notion of a “route” in BGP originally meant an IPv4 prefix, with the standardization of Multi-protocol BGP (MP-BGP) it can represent other kinds of reachability information, referred to as address families. For example, a BGP route can be an IPv6 prefix or an IPv4 prefix within a VPN.

External routes advertised in BGP must be distributed to every router in an AS. The hop-by-hop forwarding nature of IP requires that a packet address be looked up and matched against a route at each router hop. Because the address information may match external networks that are only known in BGP, every router must have the BGP information. However, we describe later how MPLS removes the need for every interior router to have external BGP route state.

Within an AS, the BGP next hop will be the IP address of the exit router or exit link from the AS through which the packet must route and BGP is used by the exit router to distribute the routes throughout the AS. To avoid creating a full mesh of iBGP sessions among the edge and interior routers, BGP can use a hierarchy of *Route Reflectors* (RR). Figure 2.15 illustrates how BGP connections are constructed using a Route Reflector.

BGP routes may have their attributes manipulated when received and before sending to peers, according to policy design decisions of the operator. Of the BGP routes received by a BGP router, BGP first determines the validity of a route (e.g., is the BGP next hop reachable) and then chooses the best route among valid duplicates with different paths. The best route is decided by a hierarchy of tiebreakers among route attributes such as IGP metric to the next hop and BGP path attributes such as AS_PATH length. The best route is then relayed to all peers except the originating one. One variation of this relay behavior is that any route received from an iBGP peer on a nonroute reflector is not relayed to any other iBGP peer.

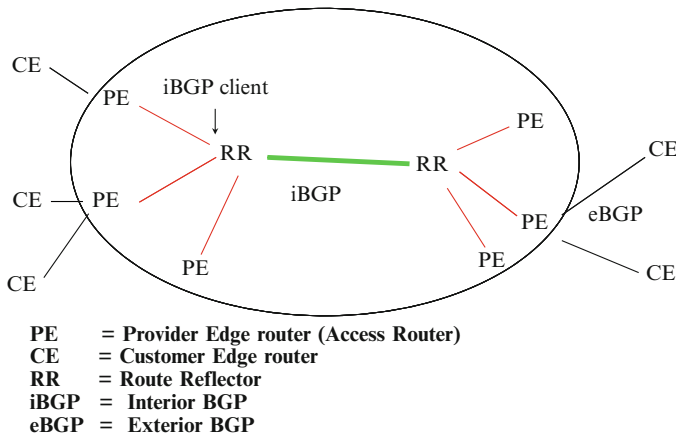


Fig. 2.15 BGP connections in an ISP with Route Reflectors (RR)

2.4.1.3 Protocol-Independent Multicast

IP Multicast is very efficient when a source sends data to multiple receivers. By using multicast at the network layer, a packet traverses a link only once, and therefore the network bandwidth is utilized optimally. In addition, the processing at routers (forwarding load) as well as at the end-hosts (discarding unwanted packets) is reduced. Multicast applications generally use UDP as the underlying transport protocol, since there is no unique context for the feedback received from the various receivers for congestion control purposes. We provide a brief overview of IP Multicast in this section. It is covered in greater detail in Chapter 11.

IP Multicast uses group addresses from the Class “D” address space (in the context of IPv4). The range of IP addresses that are used for IP Multicast group addresses is 224.0.0.0 to 239.255.255.255. When a source sends a packet to an IP Multicast group, all the receivers that have joined that group receive it. The typical protocol used between the end-hosts and routers is *Internet Group Management Protocol (IGMP)*. Receivers (end-hosts) announce their presence (join a multicast group) by sending an IGMP report to join a group. From the first router, the indication of the intent of an end-host to join the multicast group is forwarded through routers upwards along the shortest path to the root of the multicast tree. The root for an IP Multicast tree can be a source in a source-based distribution tree, or it may be a “rendezvous point” when the tree is a shared distribution tree. The routing protocol used in conjunction with IP multicast is called *Protocol-Independent Multicast (PIM)*. PIM has variants of the routing protocol used to form the multicast tree to forward traffic from a source (or sources) to the receivers. A router forwards a multicast packet only if it was received on the upstream interface to the source or to a rendezvous point (in a shared tree). Thus, a packet sent by a source follows the distribution tree. To avoid loops, if a packet arrives on an interface that is not on the shortest path toward the source of rendezvous point, the packet is discarded

(and thus not forwarded). This is called *Reverse Path Forwarding (RPF)*, a critical aspect of multicast routing. RPF avoids loops by not forwarding duplicate packets. PIM relies on the SPT created by the traditional routing protocols such as OSPF to find the path back to the multicast source using RPF.

IP Multicast uses soft-state to keep the multicast forwarding state at the routers in the network. There are two broad approaches for maintaining multicast state. The first is termed *PIM-Dense Mode*, wherein traffic is first flooded throughout the network, and the tree is “pruned” back along branches where the traffic is not wanted. The underlying assumption is that there are multicast receivers for this group at most locations, and hence flooding is appropriate. The flood and prune behavior is repeated, in principle, once every 3 min. However, this results in considerable overhead (as the traffic would be flooded until it is pruned back) each time. Every router also ends up keeping state for the multicast group. To avoid this, the router downstream of a source periodically sends a “state refresh” message that is propagated hop-by-hop down the tree. When a router receives the state refresh message on the RPF interface, it refreshes the prune state, so that it does not forward traffic received subsequently, until a receiver joins downstream on an interface.

While PIM-Dense Mode is desirable in certain situations (e.g., when receivers are likely to exist downstream of each of the routers – densely populated groups – hence the name), *PIM-Sparse Mode (PIM-SM)* is more appropriate for wide-scale deployment of IP multicast for both densely and sparsely populated groups. With PIM-SM, traffic is sent only where it is requested, and receivers are required to explicitly join a multicast group to receive traffic. While PIM-SM uses both a shared tree (with a rendezvous point, to allow for multiple senders) as well as a per-source tree, we describe a particular mode, *PIM-Source Specific Multicast (PIM-SSM)*, which is more commonly used for IPTV distribution. More details regarding PIM-SM, including PIM using a shared tree, is described in Chapter 11. PIM-SSM is adopted when the end-hosts know exactly which source and group, typically denoted (S, G) , to join to receive the multicast transmissions from that source. In fact, by requiring that receivers signal the combination of source and group to join, different sources could share the same group address and not interfere with each other. Using PIM-SSM, a receiver transmits an IGMP join message for the (S, G) and the first hop router sends a (S, G) join message directly along the shortest path toward the source. The shortest path tree is rooted at the source.

One of the key properties of IP Multicast is that the multicast routing operates somewhat independently of the IGP routing. Changes to the network topology are reflected in the unicast routing using updates that operate on short-time scales (e.g., transmission of LSAs in OSPF reflect a link or node failure immediately). However, IP Multicast routing reflects the changed topology only when the multicast state is refreshed. For example, with PIM-SSM, the updated topology is reflected only when the join is issued periodically (which can be up to a minute or more) by the receiver to refresh the state. We will examine the consequence of this for wide-area IPTV distribution later in this chapter.

2.4.2 Multiprotocol Label Switching

2.4.2.1 Overview of MPLS

Multiprotocol Label Switching (MPLS) is a technology developed in the late 1990s that added new capabilities and services to IP networks. It was the culmination of various IP switching technology efforts such as multiprotocol over ATM, Ipsilon's IP Switching, and Cisco's tag switching [7,20]. The key benefits provided by MPLS to an ISP network are:

1. Separation of routing (the selection of paths through the network) from forwarding/switching via IP address header lookup
2. An abstract hierarchy of aggregation

To understand these concepts, we first consider how normal IP routing in an ISP network functions. In an IP network without MPLS, there is a topology hierarchy with edge and backbone routers. There is also a routing hierarchy with BGP carrying external reachability information and an IGP like OSPF carrying internal reachability information. BGP carries the information about which exit router (BGP next hop) is used to reach external address space. OSPF picks the paths across the network between the edges (see Fig. 2.16). It is important to note that every OSPF router knows the complete path to reach all the edges. The internal paths that OSPF picks and the exit routers from BGP are determined before the first packet is forwarded. The connection-less and hop-by-hop forwarding behavior of IP routing requires that every router have this internal and external routing information present.

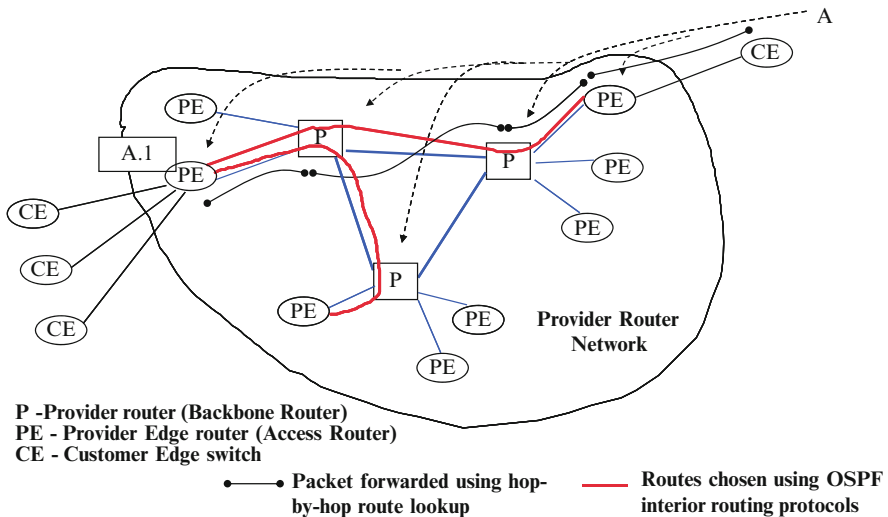


Fig. 2.16 Traditional IP routing with external routes distributed throughout backbone

Consider the example in Fig.2.16, where a packet enters on the left with address A.1 destined to the external network A on the upper right. When the first packet arrives, the receiving provider edge router (PE) looks up the destination IP address. From BGP, it learns that the exit router for that address is the upper right PE. From OSPF, the path to reach that exit PE is determined. Even though the ingress PE knows the complete path to reach the exit PE, it simply forwards the packet to the next-hop backbone router, labeled as a P-router (P) in the figure. The backbone router then repeats the process: using the packet IP address, it determines the exit from BGP and the path to the exit from OSPF to forward the packet to the next-hop BR. The process repeats again until the packet reaches the exit PE.

The repeated lookup of the packet destination to find the external exit and internal path appears to be unnecessary. The lookup operation itself is not expensive, but the issue is the unnecessary state and binding information that must be carried inside the network. The ingress router knows the path to reach the exit. If the packet could somehow be bound to the path itself, then the successive next-hop routers would only need to know the path for the packet and not its actual destination. This is what MPLS accomplishes.

Consider Fig. 2.17 where MPLS sets up an end-to-end *Label Switched Path (LSP)* by assigning labels to the interior paths to reach exits in the network. The LSP might look like the one shown in Fig. 2.18. The backbone routers are now called *Label Switch Routers (LSR)*. Via MPLS signaling protocols, the LSR knows how to forward a packet carrying an incoming label for an LSP to an outgoing interface and outgoing label; this is called a “swap” operation. The PE router also acts as an LSR, but is usually at the head (start) or end (tail) of the LSP where, respectively, the initial label is “pushed” onto the data or “popped” (removed) from the data.

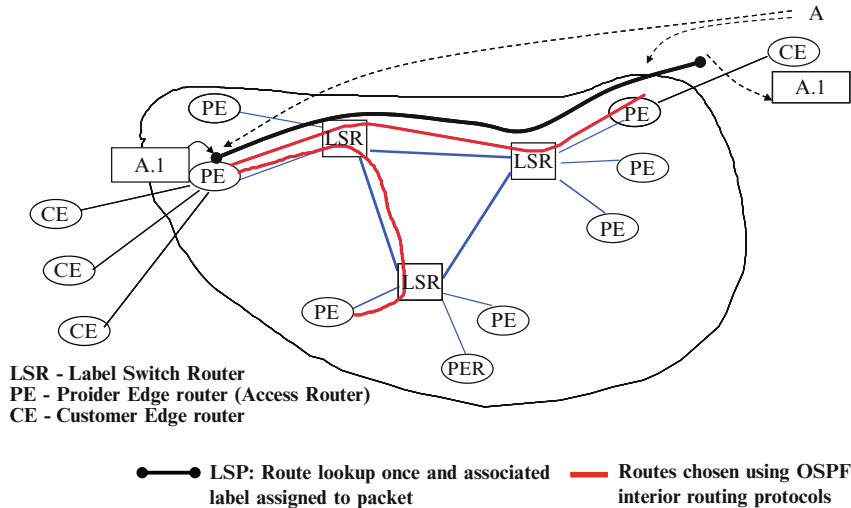


Fig. 2.17 Routing with MPLS creates Label Switched Paths (LSP) for routes across the network

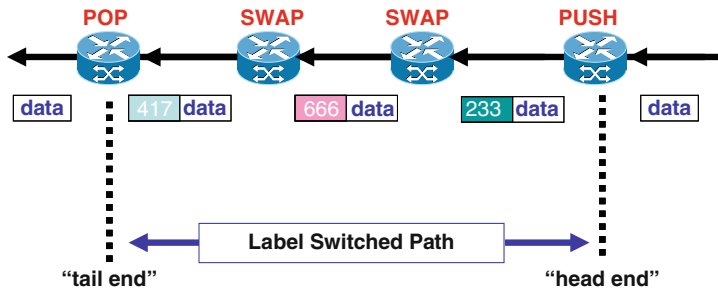


Fig. 2.18 Within an LSP, labels are assigned at each hop by the downstream router

In the example of Fig. 2.17, external BGP routing information such as routes to network A is only needed in the edges of the network. The interior LSRs only need to know the interior path among the edges as determined by OSPF. When the packet with address A.1 arrives at the ingress PE, the same lookup operation is done as previously: the egress PE is determined from BGP and the interior path to reach the egress is found from OSPF. But this time the packet is given a label for the LSP matching the OSPF path to the egress. The internal LSRs now forward the packet hop-by-hop based on the labels alone. At the exit PE, the label is removed and the packet is forwarded toward its external destination.

In this example, the binding of a packet to paths through the network is only done once – at the entrance to the network. The assignment of a packet to a path through the network is separated from the actual forwarding of the packet through the network (this is the first benefit that was identified above). Further, a hierarchy of forwarding information is created: the external routes are only kept at the edge of the network while the interior routers only know about interior paths. At the ingress router all received packets needing to exit the same point of the network receive the same label and follow the same LSP.

MPLS takes these concepts and generalizes them further. For example, the LSP to the exit router could be chosen differently from the IGP shortest path. IPv4 provides a method for explicit path forwarding in the IP header, but it is very inefficient. With MPLS, explicit routing becomes very efficient and is the primary tool for traffic engineering in IP backbones. In the previous example, if an interior link was heavily utilized, the operator may desire to divert some traffic around that link by taking a longer path as shown in Fig. 2.19. Normal IP shortest path forwarding does not allow for this kind of traffic placement.

The forwarding hierarchy can be used to create *provider-based VPNs*. This is illustrated in Fig. 2.20. Virtual private routing contexts are created at the PEs, one per customer VPN. The core of the network does not need to maintain state information about individual VPN routes. The same LSPs for reaching the exits of the network are used, but there are additional labels assigned for separating the different VPN states.

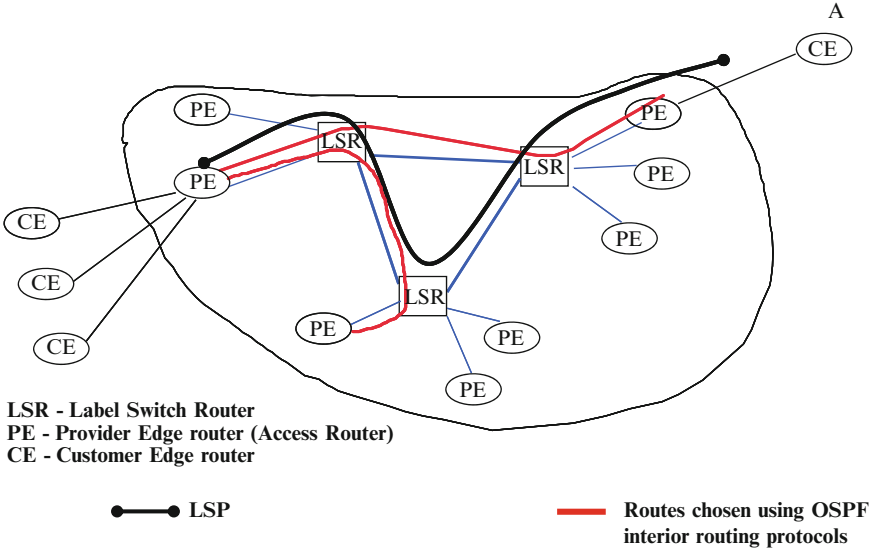


Fig. 2.19 MPLS with Traffic Engineering can use alternative to the IGP shortest path

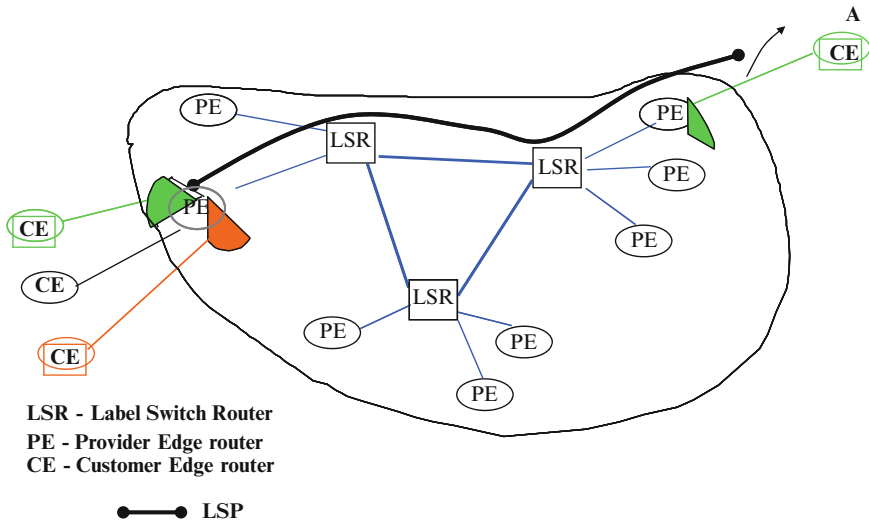


Fig. 2.20 MPLS VPNs support separated virtual routing contexts in PEs interconnected via LSPs

In summary, the advantages to the IP backbone of decoupling of routing and forwarding are:

- It achieves efficient explicit routing.
- Interior routers do not need any external reachability information.

- Packet header information is only processed at head of LSP (e.g., edges of the network).
- It is easy to implement nested or hierarchical identification (such as with VPNs).

2.4.2.2 Internet Route Free Core

The ability of MPLS to remove the external BGP information plus Layer 3 address lookup from the interior of the IP backbone is sometimes referred to as an *Internet Route Free Core*. The “interior” of the IP backbone starts at the left-side (BR-side) port of the access routers in Fig. 2.7. Some of the advantages of Internet Route Free Core include:

- Traffic engineering using BGP is much easier.
- Route reflectors no longer need to be in the forwarding plane, and thus can be dedicated to IP layer control plane functions or even placed on a server separate from the routers.
- *Denial of Service (DoS)* attacks and security holes are better controlled because BGP routing decisions only occur at the edges of the IP backbone.
- Enterprise VPN and other priority services can be better isolated from the “Public Internet”.

We provide more clarification for the last advantage. Many enterprise customers, such as financial companies or government agencies, are concerned about mixing their priority traffic with that of the public Internet. Of course, all packets are mixed on links between backbone routers; however, VPN traffic can be functionally segregated via LSPs. In particular, since denial of service attacks from the compromised hosts on the public Internet rely on reachability from the Internet, the private MPLS VPN address space isolates VPN customers from this threat. Further, enterprise premium VPN customers are sometimes clustered onto access routers dedicated to the VPN service. Furthermore, higher performance (such as packet loss or latency) for premium VPN services can be provided by implementing priority queueing or providing them bandwidth-sensitive LSPs (discussed later). A similar approach can be used to provide other performance-sensitive services, such as Voice-over-IP (VoIP).

2.4.2.3 Protocol Basics

MPLS encapsulates IP packets in an MPLS header consisting of one or more MPLS labels, known as a label stack. Figure 2.21 shows the most commonly used MPLS encapsulation type. The first 20 bits are the actual numerical label. There are three bits for inband signaling of class of service type, followed by an End-of-Stack bit (described later) and a time-to-live field, which serves the same function as an IP packet time-to-live field.

MPLS encapsulation does not define a framing mechanism to determine the beginning and end of packets; it relies on existing underlying link-layer technologies.

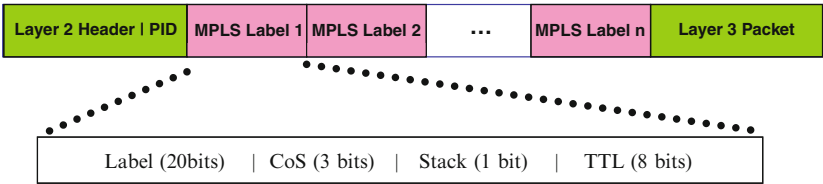


Fig. 2.21 Generic MPLS encapsulation and header fields

Existing protocols such as Ethernet, *Point-to-Point Protocol (PPP)*, ATM, and Frame Relay have been given new protocol IDs or new link-layer control fields to allow them to directly encapsulate MPLS-labeled packets.

Also, MPLS does not have a protocol ID field to indicate the type of packet encapsulated, such as IPv4, IPv6, Ethernet, etc. Instead, the protocol type of the encapsulated packet is implied by the label and communicated by the signaling protocol when the label is allocated.

MPLS defines the notion of a *Forwarding Equivalence Class (FEC)* (not to be confused with Forward Error Correction (FEC) in lower network layers defined earlier). All packets with the same forwarding requirements, such as path and priority queuing treatment, can belong to the same FEC. Each FEC is assigned a label. Many FEC types have been defined by the MPLS standards: IPv4 unicast route, VPN IPv4 unicast route, IPv6 unicast route, Frame Relay permanent virtual circuit, ATM virtual circuit, Ethernet VLAN, etc.

Labels can be stacked, with the number of stacked labels indicated by the end-of-stack bit. This allows hierarchical nesting of FECs, which permits VPNs, traffic engineering, and hierarchical routing to be created simultaneously in the same network. Consider the previous VPN example where a label may represent the interior path to reach an exit and an inner label may represent a VPN context.

MPLS is entitled “multiprotocol” because it can be carried over almost any transport as mentioned above, ironically even IP itself, and because it can carry the payload for many different packet types – all the FEC types mentioned above.

Signaling of MPLS FECs and their associated label among routers and switches can be done using many different protocols. A new protocol, the *Label Distribution Protocol (LDP)*, was defined specifically for MPLS signaling. However, existing protocols have also been extended to signal FECs and labels: *Resource Reservation Protocol (RSVP)* [3] and BGP, for example.

2.4.2.4 IP Traffic Engineering and MPLS

The purpose of IP traffic engineering is to enable efficient use of backbone capacity. That is, both to ensure that links and routers in the network are not congested and that they are not underutilized. Traffic engineering may also mean ensuring that certain performance parameters such as latency or minimum bandwidth are met.

To understand how MPLS traffic engineering plays a role in ISP networks, we first explain the generic problem to be solved – the multicommodity flow problem – and how it was traditionally solved in IP networks versus how MPLS can solve the problem.

Consider an abstract network topology with traffic demands among nodes. There are:

Demands $d(i, j)$ from node i to j

Constraints – link capacity $b(i, j)$ between nodes

Link costs $C(i, j)$

Path $p(k)$ or route for each demand

The traffic engineering problem is to find paths for the demands that fit the link constraints. The problem can be specified at different levels of difficulty:

1. Find any feasible solution, regardless of the path costs.
2. Find a solution that minimizes the costs for the paths.
3. Find a feasible or a minimum cost solution after deleting one or more nodes and/or links.

Traffic Engineering an IP Network

In an IP network, the capacities represent link bandwidths between routers and the costs might represent delay across the links. Sometimes, we only want to find a feasible solution, such as in a multicast IPTV service. Sometimes, we want to minimize the maximum path delay, such as in a Voice-over-IP service. And sometimes, we want to ensure a design that is survivable (meaning it is still feasible to carry the traffic) for any single- or dual-link failure.

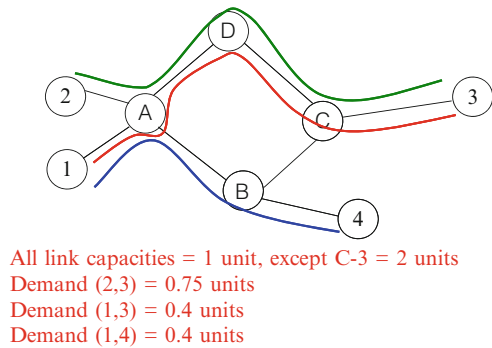
Consider how a normal ISP without traffic engineering might try to solve the problem. The tools available on a normal IP network are:

- Metric manipulation, i.e., pick OSPF weights to create a feasible solution.
- Simple topology or link augmentation: this tends to overengineer the network and restrict the possible topology.
- Source or policy route using the IPv4 header option or router-based source routes. Source routes are very inefficient resulting in tremendously lower router capacity and they are not robust, making the network very difficult to operate.

Figure 2.22 illustrates a network with a set of demands and an example of the way that particular demands might be routed using OSPF. Although the network has sufficient total capacity to carry the demands, it is not possible to find a feasible solution (with no congested links) by only setting OSPF weights. A small ISP facing this situation without technology like MPLS would probably resort to installing more link capacity on the A-D-C node path.

The generic solution to an arbitrary traffic engineering problem requires specifying the explicit route (path) for each demand. This is a complex problem that can take an indeterminate time to solve. But there are other approaches that can solve a large subset of problems. One suboptimal approach is *Constraint-based Shortest*

Fig. 2.22 IP routing is limited in its ability to meet resource demands. It cannot successfully route the demands within the link bandwidths in this example



Path First (CSPF). CSPF has been implemented in networks with ATM *Private Network-to-Network Interface (P-NNI)* and IP MPLS. For currently defined MPLS protocols, the constraints can be bandwidths per class of service for each link. Also, links can be assigned a set of binary values, which can be used to include or exclude the links from routing a given demand.

CSPF is implemented in a distributed fashion where all nodes have a full knowledge of network resource allocation. Then, each node routes its demands independently by:

1. Pruning the network to only feasible paths
2. Pick the shortest of the feasible paths on the pruned network

Although CSPF routing is suboptimal when compared with a theoretical multi-commodity flow solution, it is a reasonable compromise to solving many traffic engineering problems in which the nodes route their demands independently of each other. For more complex situations where CSPF is inadequate, network planners must use explicit paths computed by an offline system. The next section discusses explicit routing in more detail.

Traffic Engineering Using MPLS

The main problems with traffic engineering an IP backbone with only a Layer 3 IGP routing protocol (such as OSPF) are (1) lack of knowledge of resource allocation and (2) no efficient explicit routing. The previous example of Fig. 2.22 shows how OSPF would route all demands onto a link that does not have the necessary capacity. Another example problem is when a direct link is needed for a small demand between nodes to meet certain delay requirements. But OSPF cannot prevent other traffic demands from routing over this smaller link and causing congestion. MPLS solves this with extensions to OSPF (OSPF-TE) [21] to provide resource allocation knowledge and RSVP-TE [2] for efficient signaling of explicit routes to use those resources.

See Fig. 2.23 for a simple example of how an explicit path is created. RSVP-TE can create an explicit hop-by-hop path in the PATH message downstream. The PATH

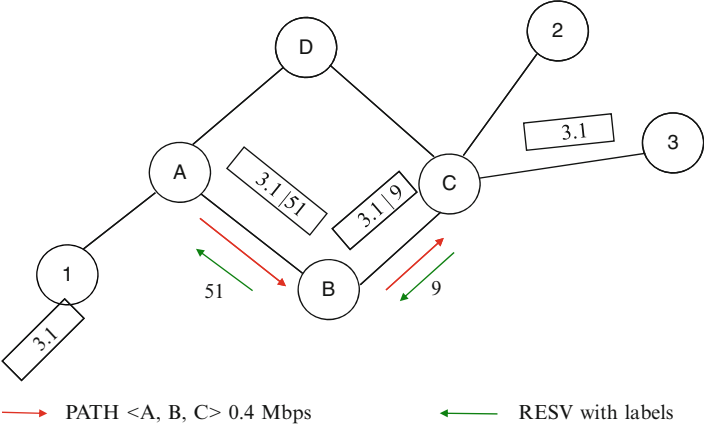
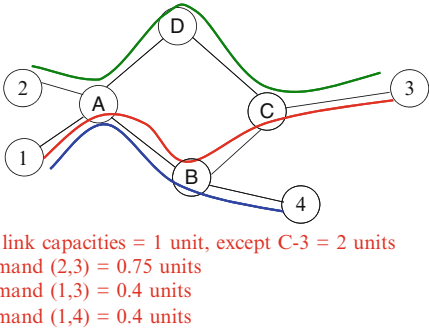


Fig. 2.23 RSVP messaging to set up explicit paths

Fig. 2.24 MPLS-TE enables efficient capacity usage through traffic engineering to solve the example in Fig. 2.22



message can request resources such as bandwidth. The return message is an RESV, which contains the label that the upstream node should use at each link hop. In this example, a traffic-engineered LSP is created along path A-B-C for 0.4 Mb/s. These LSPs are referred to as traffic engineering tunnels. Tunnels can be created and differentiated for many purposes (including restoration to be defined in later sections). But in general, primary (service route) tunnels can be considered as a routing mechanism for all packets of a given FEC between a given pair of routers or router interfaces. Using this machinery, Fig. 2.24 illustrates how MPLS-TE can be used to solve the capacity overload problem in the network shown in Fig. 2.22.

The explicit path used in RSVP-TE signaling can be computed by an offline system and automatically configured in the edge routers or the routers themselves can compute the path. In the latter case, the edge routers must be configured with the IP prefixes and their associated bandwidth reservations that are to be traffic-engineered to other edges of the network. Because the routers do this without knowledge of other demands being routed in the network, the routers must receive periodic updates about bandwidth allocations in the network.

OSPF-TE provides a set of extensions to OSPF to advertise traffic engineering resources in the network. For example, bandwidth resources per class of service can be allocated to a link. Also, a link can be assigned binary attributes, which can be used for excluding or including a link for routing an LSP. These resources are advertised in an opaque LSA via OSPF link-state flooding and are updated dynamically as allocations change. Given the knowledge of link attributes in the topology and the set of demands, the router performs an online CSPF to calculate the explicit paths. The path outputs of the CSPF are given to RSVP-TE to signal in the network. As TE tunnels are created in the network, the link resources change, i.e., available bandwidth is reduced on a link after a tunnel is allocated using RSVP-TE. Periodically, OSPF-TE will advertise the changes to the link attributes so that all routers can have an updated view of the network.

2.4.2.5 VPNs with MPLS

Figure 2.20 illustrates the key concept in how MPLS is used to create VPN services. VPN services here refer to carrier-based VPN services, specifically the ability of the service provider to create private network services on top of a shared infrastructure. For the purposes of this text, VPNs are of two basic types: a Layer 3 IP routed VPN or a Layer 2 switched VPN. *Generalized MPLS (GMPLS)* [19] can also be used for creating Layer 1 VPNs, which will not be discussed here.

A Layer 3 IP VPN service looks to customers of the VPN as if the provider built a router backbone for their own use – like having their own private ISP. VPN standards define the PE routers, CE routers, and backbone P-routers interconnecting the PEs. Although the packets share (are mixed over) the ISP's IP layer links, routing information and packets from different VPNs are virtually isolated from each other.

A Layer 2 VPN provides either point-to-point connection services or multi-point Ethernet switching services. Point-to-point connections can be used to support end-to-end services such as Frame Relay permanent virtual circuits, ATM virtual circuits, point-to-point Ethernet circuits (i.e., with no *Media Access Control (MAC)* learning or broadcasting) and even a circuit emulation over packet service. Interworking between connection-oriented services, such as Frame Relay to ATM interworking, is also defined. This kind of service is sometimes called a Virtual Private Wire Service (VPWS).

Layer 2 VPN multipoint Ethernet switching services support a traditional Transparent LAN over a wide-area network called Virtual Private LAN Service (VPLS) [24, 25].

Layer 3 VPNs over MPLS

As mentioned previously, Layer 3 VPNs maintain a separate virtual routing context for each VPN on the PE routers at the edge of the network. External CEs connect to the virtual routing context on a PE that belongs to a customer's VPN.

Layer 3 VPNs implemented using MPLS are often referred to as BGP MPLS VPNs because of the important role BGP has in the implementation. BGP is used to carry VPN routes between the edges of the network. BGP keeps the potentially overlapping VPN address spaces unique by prepending onto the routes a *route distinguisher (RD)* that is unique to each VPN. The RD + VPN IPv4 prefix combination creates a new unique address space carried by BGP, sometimes called the VPNv4 address space.

VPN routes flow from one virtual routing instance into other virtual routing instances on PEs in the network using a BGP attribute called a *Route Target (RT)*. An RT is an address configured by the ISP to identify all virtual routing instances that belong to a VPN. RTs constrain the distribution of VPN routes among the edges of the network so that the VPN routes are only received by the virtual routing instances belonging to the intended (targeted) VPN.

We note that RDs and RTs are only used in the BGP control plane – they are not values that are somehow applied to user packets themselves. Rather, for every advertised VPNv4 route, BGP also carries a label assignment that is unique to a particular virtual router on the advertising PE.

Every VPN packet that is forwarded across the network receives two labels at the ingress PE: an inner label associated with the advertised VPNv4 route and an outer label associated with the LSP to reach the egress advertising PE (dictated by the BGP next-hop address). See Fig. 2.25 for a simplified example. In this example,

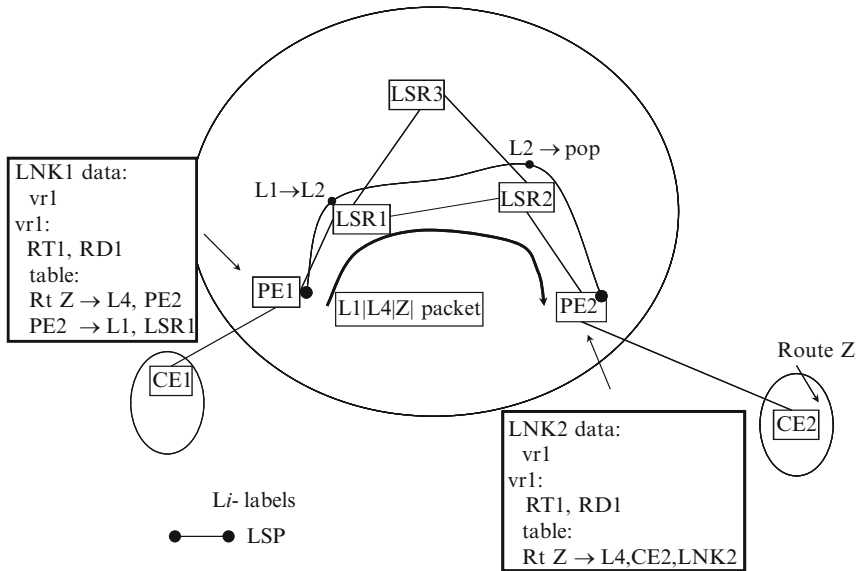


Fig. 2.25 In this VPN example, a virtual routing context (vr1) in the PEs contains the VPN label and routing information such as route target (RT1) and route distinguisher (RD1), attached CE interfaces, and next-hop lookup and label binding. VPN traffic is transported using a label stack of VPN label and interior route label

there is a VPN advertising a route Z, which enters the receiving virtual router (vr1) and is distributed by BGP to other PE virtual routers using RTs. A packet entering the VPN destined toward Z is looked up in the virtual routing instance, where the two labels are found – the outer label to reach the egress PE and the inner label for the egress virtual routing instance.

Layer 2 VPNs over MPLS

The implementation of Layer 2 VPNs over MPLS is similar to Layer 3 VPNs. Because there is no IP routing in the VPN service, there is instead a virtual switching context created on the edge PEs to isolate different VPNs. These virtual switching contexts keep the address spaces of the edge services from conflicting with each other across different VPNs.

Layer 2 VPNs use a two-label stack approach that is similar to Layer 3 VPNs. Reaching an egress PE from an ingress PE is done using the same network interior LSPs that the Layer 3 VPN service would use. And then, there is an inner label associated with either the VPWS or VPLS context at the egress PE. This inner label can be signaled using either LDP or BGP. The inner label and the packet encapsulation comprise a pseudowire, as defined in the PWE3 standards [16]. The pseudowire connects an ingress PE to an egress PE switching context and is identified by the inner label. The VPWS service represents a single point-to-point connection, so there will only be a single pseudowire setup in each direction. For VPLS however, carriers typically set up a full mesh of pseudowires/LSPs among all PEs belonging to that VPLS.

Forwarding for a VPWS is straightforward: the CE connection is associated with the appropriate pseudowires in each direction when provisioned. For VPLS, forwarding is determined by the VPLS forwarding table entry for the destination Ethernet MAC address. Populating the forwarding table is based on source MAC address learning. The forwarding table records the inbound interface on which a source MAC was seen. If the destination MAC is not in the table, then the packet is flooded to all interfaces attached to the VPLS. Flooding of unknown destination MACs and broadcast MACs follows some special rules within a VPLS. All PEs within a backbone are assumed to be full mesh connected with pseudowires. So, packets received from the backbone are not flooded again into the backbone, but are only flooded onto CE interfaces. On the other hand, packets from a CE to be flooded are sent to all attached CE interfaces and all pseudowire interfaces toward the other backbone PEs.

There is also a VPLS variation called Hierarchical VPLS to constrain the potential explosion of mesh point-to-point LSPs needed among the PE routers. This might happen with a PE that acts like a spoke with a single pseudowire attached to a core of meshed PEs. In this model, a flooding packet received at a mesh connected PE from a spoke PE pseudowire is sent to all attached CEs and pseudowires. In such a model, the PE interconnectivity must be guaranteed to be loop-free or a spanning tree protocol may be run among the PEs for that VPLS.

2.5 Network Restoration and Planning

The design of an IP backbone is driven by the traffic demands that need to be supported, and network availability objectives. The network design tools model the traffic carried over the backbone links not only in a normal “sunny day” scenario, but also in the presence of network disruptions.

Many carriers offer *Service Level Agreements (SLAs)*. SLAs will vary across different types of services. For example, SLAs for private-line services are quite different from those for packet services. SLAs also usually differ among different types of packet services. The SLAs for general Internet, VPN, and IPTV services will generally differ. A packet-based SLA might be expressed in terms of *Quality of Service (QoS)* metrics. For example, the SLA for a premium IP service may cover up to three QoS metrics: latency, jitter, and packet loss. An example of the latter is “averaged over time period Y , the customer will receive at least $X\%$ of his/her packets transmitted.” Some of these packet services may be further differentiated by offering different levels of service, also called *Class of Service (CoS)*.

To provide its needed SLAs, an ISP establishes internal network objectives. Network availability is a key internal metric used to control packet loss. Furthermore, network availability is also sometimes used as the key QoS metric for private-line services. Network availability is often stated colloquially in “9s”. For example, “four nines” of availability means the service is available at least 0.9999 of the time. Stated in the contra-positive, the service should not be down more than 0.0001 of the time (approximately 50 min per year). Given its prime importance, we will concentrate on network availability in the remainder of this section.

The single largest factors in designing and operating the IP backbone such that it achieves its target network availability are modeling its potential network disruptions and the response of the network to those disruptions. Network disruptions most typically are caused by network failures and maintenance activities. Maintenance activities include upgrading of equipment software, replacement of equipment, and reconfiguration of network topologies or line cards. Because of the complex layering and segmentation of networks surrounding the IP backbone and because of the variety and vintage of equipment that accumulates over the years, network planners, architects, network operators, and engineers spend considerable effort to maintain network availability. In this section, we will briefly describe the types of restoration methods we find at the various network layers. Then, we will describe how network disruptions affect the IP backbone, the types of restoration methods used to handle them, and finally how the network is designed to meet the needed availability.

Table 2.3 summarizes typical restoration methods used in some of today’s network core layers that are most relevant to the IP backbone. See [11] for descriptions of restoration methods used in other layers shown in Fig. 2.3. In the next sections, we will describe the rows of this table. Note that the table is approximate and does not apply universally to all telecommunication carriers.

Table 2.3 Example of core-segment restoration methods

Network layer	Restoration method(s) against network failures that originate at that layer or lower layers	Exemplary restoration time scale
Fiber	No automatic rerouting	Hours (manual)
DWDM	1) Manual 2) 1 + 1 restoration (also called <i>dedicated protection</i>)	1) Hours (manual) 2) 3–20 ms
SONET Ring	Bidirectional Line-Switched Rings (BLSR)	50–100 ms
IOS (DCS)	Distributed path-based mesh restoration	Sub-second to seconds
W-DCS	No automatic rerouting	Hours
IP backbone	1) IGP reconfiguration 2) MPLS Fast Reroute (FRR)	1) 10–60 s 2) 50–100 ms

2.5.1 Restoration in Non-IP Layers

2.5.1.1 Fiber Layer

As we described earlier, in most central offices today, optical interfaces on switching or transport equipment connect to fiber patch panels. Some carriers have installed an automated fiber patch panel, also called a *Fiber Cross-Connect (FXC)*, which has the ability for an operator to remotely control the cross-connects. Some of the enabling technologies include physical crossbars using optical collometers and *Micro-Electro-Mechanical Systems (MEMS)*. A good overview of these technologies can be found in [12]. When disruptions occur to the fiber layer, most commonly from construction activity, network operators can reroute around the failed fiber by using a patch panel to cross-connect the equipment onto undamaged fibers. This may require coordination of cross-connects at intermediate central offices to patch a path through alternate COs if an entire cable is damaged. Of course, this typically is a slow manual process, as reflected in Table 2.3 and so higher-layer restoration is usually utilized for disruptions to the fiber layer.

2.5.1.2 DWDM Layer

Some readers may be surprised to learn that carriers have deployed few (if any) automatic restoration methods in their DWDM layers (neither metro nor core segment). The one type of restoration occasionally deployed is *one-by-one (1:1)* or *one-plus-one (1 + 1)* tail-end protection switching, which switches at the end-points of the DWDM layer connection. With 1 + 1 switching, the signal is duplicated and transmitted across two (usually) diversely routed connections. The path of the connection during the nonfailure state is usually called the *working path* (also called the *primary* or *service path*); the path of the connection during the failure state is called the *restoration path* (also called *protection path* or *backup path*). The receiver

consists of a simple detector and switch that detects failure of the signal on the working path (more technically, detects performance errors such as average BER threshold crossings) and switches to the restoration path upon alarm. Once adequate signal performance is again achieved on the signal along the working path (including a time-out threshold to avoid link “flapping”), it switches back to the working path. In 1:1 protection switching, there is no duplication of signal, and thus the restoration connection can be used for other transport in nonfailure states. The transmitted signal is switched to the restoration path upon detection of failure of the service path and/or notification from the far end.

Technically speaking, in ROADM or Point-to-point DWDM systems, 1 + 1 or 1:1 protection switching is usually implemented electronically via the optical transponders. Consequently, these methods can be implemented at other transport layers, such as DCS, IOS, and SONET. The major advantage of 1 + 1 or 1:1 methods is that they can trigger in as little as 3–20 ms. However, because these methods require restoration paths that are dedicated (one-for-one) for each working connection, the resulting restoration capacity cannot be shared among other working connections for potential failures. Furthermore, the restoration paths are diversely routed and are often much longer than their working paths. Consequently, 1 + 1 and 1:1 protection switching tend to be the costliest forms of restoration.

2.5.1.3 SONET Ring Layer

The two most common types of deployed SONET or SDH self-healing ring technology are *Unidirectional Path Switched Ring (UPSR-2F)* and *Bidirectional Line-Switched Ring (BLSR-2F)*. The “2F” stands for “2-Fibers”. For simplicity, we will limit our discussion to SONET rings, but there is a very direct analogy for SDH rings. However, note that ADM-ADM ring links are sometimes transported over a lower DWDM layer, thus forming a “connection” that is routed over channels of DWDM systems, instead of direct fiber. Although there is no inherent topographical orientation in a ring, many people conceptually visualize each node of a SONET self-healing ring as an ADM with an *east* bidirectional OC- n interface (i.e., a transmit port and a receive port) and a *west* OC- n interface. Typically, $n = 48$ or 192. An STS- k SONET-Layer connection enters at an add/drop port of an ADM, routes around the ring on k STS-1 channels of the ADM–ADM links and exits the ring at an add/drop port of another ADM. The UPSR is the simplest of the devices and works similarly to the 1 + 1 tail-end switch described in Section 2.5.1.2, except that each direction of transmission of a connection routes counterclockwise on the “outer” fiber around the ring (west direction) and therefore an STS- k connection used the same k STS-1 channels on all links around the ring. At each add/drop transmit port, the signal is duplicated in the opposite direction on the “inner” fiber. The selector responds to a failure as described above.

The BLSR-2F partitions the bidirectional channels of its East and West high-speed links in half. The first half is used for working (nonfailure) state, and the second half is reserved for restoration. When a failure to a link occurs,

the surrounding ADMs loop back that portion of the connection paths onto the restoration channels around the opposite direction of the ring. The UPSR has very rapid restoration, but suffers the dedicated-capacity condition described in Section 2.5.1.2; as a consequence, today UPSRs are now confined mostly to the metro network, in particular to the portion closest to the customer, often extending into the feeder network. Because BLSR signaling is used to advertise failures among ADMs and real-time intermediate cross-connections have to be made, a BLSR restores more slowly than a UPSR. However, the BLSR is capable of having multiple connections share restoration channels over nonsimultaneous potential network failures, and is thus almost always deployed in the middle of the metro network or parts of the core network. Rings are described in more detail in [11].

2.5.1.4 IOS Layer

The typical equipment that comprise today's IOS layer use distributed control to provision (set-up) connections. Here, links of the IOS network (SONET bidirectional OC- n interfaces) are assigned routing weights. When a connection is provisioned over the STS-1 channels of an IOS network, its source node (IOS) computes its working path (usually along a minimum-weight path) plus also computes its restoration path that is diversely routed from the working path. After the connection is set up along its working path, the restoration path is stored for future use. The nodes communicate the state of the network connectivity via topology update messages transmitted over the SONET overhead on the links between the nodes. When a failure occurs, the nodes flood advertisement messages to all nodes indicating the topology change. The source node for each affected connection then instigates the restoration process for its failed connections by sending connection request messages along the links of the (precalculated) restoration path, seeking spare STS-1 channels to reroute its connections. Various handshaking among nodes of the restoration paths are implemented to complete the rerouting of the connections. Note that in contrast to the dedicated and ring methods, the restoration channels are not prededicated to specific connections and, therefore, connections from a varied set of source/destination pairs can potentially use them. Such a method is called *shared* restoration because a given spare channel can be used by different connections across nonsimultaneous failures. Shared mesh restoration is generally more capacity-efficient than SONET rings in mesh networks (i.e., networks with average connectivity greater than 2).

We now delve a little more into IOS restoration to make a key point that will become relevant to the IP backbone, as well. The example in Fig. 2.2 shows two higher-layer connections routing over the same lower-layer link. In light of the discussion above about the restoration path being diverse from the working path in the IOS layer, the astute reader may ask "diverse relative to what?" The answer is that, in general, the path should be diverse all the way down through the DWDM and Fiber Layers. This requires that the IOS links contain information about how they share these lower-layer links. Often, this is accomplished via a mechanism called

“bundle groups”. That is, a bundle group is created for each lower-layer link, but is expressed as a group of IOS links that share (i.e., route over) that link. Diverse restoration paths can be discovered by avoiding IOS links that belong to the same bundle group of a link on the working path. Of course, the equipment in the IOS-Layer cannot “see” its lower layers, and consequently has no idea how to define and create the bundle groups. Therefore, bundle groups are provisioned in the IOSs using an *Operations Support System (OSS)* that contains a database describing the mapping of IOS links to lower-layer networks. This particular example illustrates the importance of understanding network layering; else we will not have a reliable method to plan and engineer the network to meet the availability objective. This point will be equally important to the IP backbone. A set of bundled links is also referred to as a *Shared Risk Link Group (SRLG)* in the telecommunications industry, since it refers to a group of links that are subject to a shared risk of disruption.

2.5.1.5 W-DCS Layer and Ethernet Layer

There are few restoration methods provided at the W-DCS layer itself. This is because most disruptions to a W-DCS link occurs from a disruption of (1) a W-DCS line card or (2) a component in a lower layer of which the link routes. Disruptions of type (1) are usually handled by providing 1:1 restorable intra-office links between the W-DCS and TDM node (IOS or ADM). Disruptions of type (2) are restored by the lower TDM layers. This only leaves failure or maintenance of the W-DCS itself as an unrestorable network disruption. However, a W-DCS is much less sophisticated than a router and less subject to failure.

Restoration of Layer 2 VPNs in an IP/MPLS backbone is discussed in Section 2.5.2. We note here that restoration in enterprise Ethernet networks is typically based on the Rapid Spanning Tree Protocol (RSTP). When enterprise Ethernet VPNs are connected over the IP backbone (such as VPLS), an enterprise customer who employs routing methods such as RSTP expects it to work in the extended network. By encapsulating the customer’s Ethernet frames inside pseudowires ensures that the client’s RTSP control packets are transported transparently across the wide area. For example, a client VPN may choose to restore local link disruptions by routing across other central offices or even distant metros. Since all this appears as one virtual network to the customer, such applications may be useful.

2.5.2 IP Backbone

There are two main restoration methods we describe for the IP layer: IGP reconfiguration and MPLS *Fast Reroute (FRR)*.

2.5.2.1 OSPF Failure Detection and Reconvergence

In a formal sense, the IGP reconvergence process responds to topology changes. Such topology changes are usually caused by four types of events:

1. Maintenance of an IP layer component
2. Maintenance of a lower-layer network component
3. Failure of an IP layer component (such as a router line card or common component)
4. Failure of a lower-layer network component (such as a link)

When network operations staff perform planned maintenance on an IP layer link, it is typical to raise the OSPF administrative weight of the link to ensure that all traffic is diverted from the link (this is often referred to as “costing out” the link). In the second case, most carriers have a maintenance procedure where organizations that manage the lower-layer networks schedule their daily maintenance events and inform the IP layer operations organization. The IP layer operations organization responds by costing out all the affected links before the lower-layer maintenance event is started.

In the first two cases (planned maintenance activity), the speed of the reconvergence process is usually not an issue. This is because the act of changing an IGP routing weight on a link causes LSAs to be issued. During the process of updating the link status and recomputation of the SPF tree, the affected links remain in service (i.e., “up”). Therefore, once the IGP reconfiguration process has settled, the routers can redirect packets to their new paths. While there may be a transient impact during the “costing out” period, in terms of transient loops and packet loss, the service impact is kept to a minimum by using this costing out technique to remove a link from the topology for performing maintenance.

In the last two cases (failures), once the affected links go down, packets may be lost or delayed until the reconvergence process completes. Such a disruption may be unacceptable to delay or loss-sensitive applications. This motivates us to examine how to reduce the time required for OSPF to converge from unexpected outages. This is the focus of the remainder of this section.

While most large IP backbones route over lower layers, such as DWDM, those do not provide restoration. Layer 1 failure detection is a key component of the IP layer restoration process. A key component of the overall failure recovery time in OSPF-based networks is the failure detection time. However, lower-layer failure detection mechanisms sometimes do not coordinate well with higher-layer mechanisms and do not detect disruptions that originate in the IP layer control plane. As a result, OSPF routers periodically exchange Hello messages to detect the loss of a link adjacency with a neighbor.

If a router does not receive a Hello message from its neighbor within a RouterDeadInterval, it assumes that the link to its neighbor has failed, or the neighbor router itself is down, and generates a new LSA to reflect the changed topology. All such LSAs generated by the routers affected by the failure are flooded throughout the network. This causes the routers in the network to redo the SPF

calculation and update the next-hop information in their respective forwarding tables. Thus, the time required to recover from a failure consists of: (1) the failure detection time, (2) LSA flooding time, (3) the time to complete the new SPF calculations and update the forwarding tables.

To avoid a false indication that an adjacency is down because of congestion related loss of Hello messages, the `RouterDeadInterval` is usually set to be four times the `HelloInterval` – the interval between successive Hello messages sent by a router to its neighbor. With the RFC suggested default values for these timers (`HelloInterval` value of 10 s and `RouterDeadInterval` value of 40 s), the failure detection time can take anywhere between 30 and 40 s. LSA flooding times consist of propagation delay and additional pacing delays inserted by the router. These pacing delays serve to rate-limit the frequency with which `LSUpdate` packets are sent on an interface. Once a router receives a new LSA, it schedules an SPF calculation. Since the SPF calculation using Dijkstra's algorithm (see e.g., [8]) constitutes a significant processing load, a router typically waits for additional LSAs to arrive for a time interval corresponding to `spfDelay` (typically 5 s) before doing the SPF calculation on a batch of LSAs. Moreover, routers place a limit on the frequency of SPF calculations (governed by a `spfHoldTime`, typically 10 s, between successive SPF calculations), which can introduce further delays.

From the description above, it is clear that reducing the `HelloInterval` can substantially reduce the Hello protocol's failure detection time. However, there is a limit to which the `HelloInterval` can be safely reduced. As the `HelloInterval` becomes smaller, there is an increased chance that network congestion will lead to loss of several consecutive Hello messages and thereby cause a false alarm that an adjacency between routers is lost, even though the routers and the link between them are functioning. The LSAs generated because of a false alarm will lead to new SPF calculations by all the routers in the network. This false alarm would soon be corrected by a successful Hello exchange between the affected routers, which then causes a new set of LSAs to be generated and possibly new path calculations by the routers in the network. Thus, false alarms cause an unnecessary processing load on routers and sometimes lead to temporary changes in the path taken by network traffic. If false alarms are frequent, routers have to spend considerable time doing unnecessary LSA processing and SPF calculations, which may significantly delay important tasks such as Hello processing, thereby leading to more false alarms.

False alarms can also be generated if a Hello message gets queued behind a burst of LSAs and thus cannot be processed in time. The possibility of such an event increases with the reduction of the `RouterDeadInterval`. Large LSA bursts can be caused by a number of factors such as simultaneous refresh of a large number of LSAs or several routers going down/coming up simultaneously. Choudhury [5] studies this issue and observes that reducing the `HelloInterval` lowers the threshold (in terms of number of LSAs) at which an LSA burst will lead to generation of false alarms. However, the probability of LSA bursts leading to false alarms is shown to be quite low.

Since the loss and/or delayed processing of Hello messages can result in false alarms, there have been proposals to give such packets prioritized treatment at the router interface as well as in the CPU processing queue [5]. An additional option is to consider the receipt of any OSPF packet (e.g., an LSA) from a neighbor as an indication of the good health of the router's adjacency with the neighbor. This provision can help avoid false loss of adjacency in the scenarios where Hello packets get dropped because of congestion, caused by a large LSA burst, on the link between two routers. Such mechanisms may help mitigate the false alarm problem significantly. However, it will take some time before these mechanisms are standardized and widely deployed.

It is useful to make a realistic assessment regarding how small the HelloInterval can be, to achieve faster detection and recovery from network failures while limiting the occurrence of false alarms. We summarize below the key results from [13]. This assessment was done via simulations on the network topologies of commercial ISPs using a detailed implementation of the OSPF protocol in the NS2 simulator. The work models all the important OSPF protocol features as well as various standard and vendor-introduced delays in the functioning of the protocol. These are shown in Table 2.4.

Goyal [13] observes that with the current default settings of the OSPF parameters, the network takes several tens of seconds before recovering from a failure. Since the main component in this delay is the time required to detect a failure using the Hello protocol, Goyal [13] examines the impact of lower HelloInterval values on failure detection and recovery times.

Table 2.5 shows typical results for failure detection and recovery times after a router failure. As expected, the failure detection time is within the range of three to four times the value of HelloInterval. Once a neighbor detects the router failure, it generates a new LSA about 0.5 s after the failure detection. The new LSA is flooded throughout the network and will lead to scheduling of an SPF calculation 5 s (spfDelay) after the LSA receipt. This is done to allow one SPF calculation to take care of several new LSAs. Once the SPF calculation is done, the router takes about 200 ms more to update the forwarding table. After including the LSA propagation and pacing delays, one can expect the failure recovery to take place about 6 s after the 'earliest' failure detection by a neighbor router.

Notice that many entries in Table 2.5 show the recovery to take place much sooner than 6 s after failure detection. This is partly an artifact of the simulation because the failure detection times reported by the simulator are the "latest" ones rather than the "earliest". In one interesting case (seed 2, HelloInterval 0.75 s), the failure recovery takes place about 2 s after the 'latest' failure detection. This happens because the SPF calculation scheduled by an earlier false alarm takes care of the LSAs generated because of router failure. There are also many cases in which failure recovery takes place more than 6 s after failure detection (notice entries for HelloInterval 0.25 s, seeds 1 and 3). Failure recovery can be delayed because of several factors. The SPF calculation frequency of the routers is limited by spfHoldTime (typically 10 s), which can delay the new SPF calculation in response to the router failure. The delay caused by spfDelay is also a contribution.

Table 2.4 Various delays affecting the operation of OSPF protocol

Standard configurable delays	
RxmtInterval	The time delay before an un-acked LSA is retransmitted. Usually 5 s.
HelloInterval	The time delay between successive Hello packets. Usually 10 s.
RouterDeadInterval	The time delay since the last Hello before a neighbor is declared to be down. Usually four times the HelloInterval.
Vendor-introduced configurable delays	
Pacing delay	The minimum delay enforced between two successive Link-State Update packets sent down an interface. Observed to be 33 ms. Not always configurable.
spfDelay	The delay between the shortest path calculation and the first topology change that triggered the calculation. Used to avoid frequent shortest path calculations. Usually 5 s.
spfHoldTime	The minimum delay between successive shortest path calculations. Usually 10 s.
Standard fixed delays	
LSRefreshTime	The maximum time interval before an LSA needs to be reflooded. Set to 30 min.
MinLSInterval	The minimum time interval before an LSA can be reflooded. Set to 5 s.
MinLSArrival	The minimum time interval that should elapse before a new instance of an LSA can be accepted. Set to 1 s.
Router-specific delays	
Route install delay	The delay between the shortest path calculation and update of forwarding table. Observed to be 0.2 s.
LSA generation delay	The delay before the generation of an LSA after all the conditions for the LSA generation have been met. Observed to be around 0.5 s.
LSA processing delay	The time required to process an LSA including the time required to process the Link-State Update packet before forwarding the LSA to the OSPF process. Observed to be less than 1 ms.
SPF calculation delay	The time required to do shortest path calculation. Observed to be $0.00000247x^2 + 0.000978$ s on Cisco 3600 series routers; x being the number of nodes in the topology.

Finally, the routers with a low degree of connectivity may not get the LSAs in the first try because of loss due to congestion. Such routers may have to wait for 5 s (RxmtInterval) for the LSAs to be retransmitted.

The results in Table 2.5 show that a smaller value of HelloInterval speeds up the failure detection but is not effective in reducing the failure recovery times beyond a limit because of other delays like spfDelay, spfHoldTime, and RxmtInterval. Failure recovery times improve as the HelloInterval reduces down to about 0.5 s. Beyond that, as a result of more false alarms, we find that the recovery times actually go up. While it may be possible to further speed up

Table 2.5 Failure detection time and failure recovery time for a router failure with different `HelloInterval` values

Hello interval (s)	Seed 1		Seed 2		Seed 3	
	FDT (s)	FRT (s)	FDT (s)	FRT (s)	FDT (s)	FRT (s)
10	32.08	36.60	39.84	46.37	33.02	38.07
2	7.82	11.68	7.63	12.18	7.79	12.02
1	3.81	9.02	3.80	8.31	3.84	10.11
0.75	2.63	7.84	2.97	5.08	2.81	7.82
0.5	1.88	6.98	1.82	6.89	1.79	6.85
0.25	0.95	10.24	0.84	6.08	0.99	13.41

the failure recovery by reducing the values of these delays, eliminating such delays altogether is not prudent. Eliminating `spfDelay` and `spfHoldTime` will result in potentially additional SPF calculations in a router in response to a single failure (or false alarm) as the different LSAs generated because of the failure arrive one after the other at the router. The resulting overload on the router CPUs may have serious consequences for routing stability, especially when there are several simultaneous changes in the network topology. Failure recovery below the range of 1–5 s is difficult with OSPF.

In summary, OSPF recovery time can be lowered by reducing the value of `HelloInterval`. However, too small a value of `HelloInterval` will lead to many false alarms in the network, which cause unnecessary routing changes and may lead to routing instability. The optimal value for the `HelloInterval` that will lead to fast failure recovery in the network, while keeping the false alarm occurrence within acceptable limits for a network, is strongly influenced by the expected congestion levels and the number of links in the topology. While the `HelloInterval` can be much lower than current default value of tens of seconds, it is not advisable to reduce it to the millisecond range because of potential false alarms. Further, it is difficult to prescribe a single `HelloInterval` value that will perform optimally in all cases. The network operator needs to set the `HelloInterval` conservatively taking into account both the expected congestion as well as the number of links in the network topology.

2.5.2.2 MPLS Fast Reroute

MPLS Fast Reroute (FRR) was designed to improve restoration performance using the additional protocol layer provided by MPLS LSPs [17]. Primary and alternate (backup) LSPs are established. Fast rerouting over the alternate paths after a network disruption is achieved using preestablished router forwarding table entries. Equipment suppliers have developed many flavors of FRR, some of which are not totally compliant with standardized MPLS FRR. This section provides an overview of the basic concept.

There are two basic varieties of backup path restoration in MPLS FRR, called *next-hop* and *next-next-hop*. The next-hop approach identifies a unidirectional link to be protected and a *backup* (or *bypass*) unidirectional LSP that routes around the

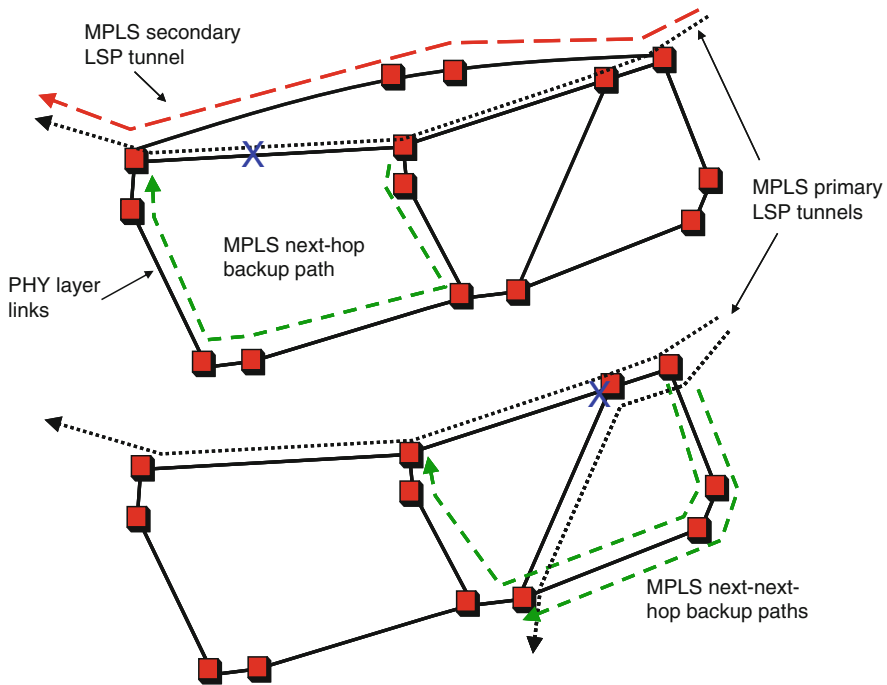


Fig. 2.26 Example of Fast Reroute backup paths

link if it fails. The protected link can be a router–router link adjacency or even another layer of LSP tunnel itself. The backup LSP routes over alternate links. The top graph in Fig. 2.26 illustrates a next-hop backup path for the potential failure of a given link (designated with an “X”). For now ignore the top path labeled “MPLS secondary LSP tunnel”, which will be discussed later. With the *next-next-hop* approach, the primary entities to protect are two-link working paths. The backup path is an alternate path over different links and routers than the protected entity. In general, a next-hop path is constructed to restore against individual link failures while next-next-hop paths are constructed to restore against both individual link failures and node failures. The trade-off is that next-hop paths are simpler to implement because all flows routing over the link can be rerouted similarly, whereas next-next-hop requires more LSPs and routing combinations. This is illustrated in the lower example of Fig. 2.26, wherein the first router along the path carries flows that terminate on different second hop routers, and therefore must create multiple backup LSPs that originate at that node.

We will briefly describe an implementation of the next-hop approach to FRR. A primary end-to-end path is chosen by RSVP. This path is characterized by the Forwarding Equivalence Class (FEC) discussed earlier and reflects packets that are to be corouted and have similar CoS queuing treatment and ability to be restored with FRR. Often, a mesh of fully connected end-to-end LSPs between the backbone routers (BRs) is created.

As discussed in earlier sections, an LSP is identified in forwarding tables by mappings of pairs of label and interface: (In-Label, In-Interface) \rightarrow (Out-Label, Out-Interface). An end-to-end LSP is provisioned (set up) by choosing and populating these entries at each intermediate router along the path by a protocol such as RSVP-TE. For the source router of the LSP, the “In-Label” variable is equivalent to the FEC. As a packet hops along routers, the labels are replaced according to the mapping until it reaches the destination router, in which case, the MPLS shim headers are *popped* and packets are placed on the final output port. With next-hop, *facility-based* FRR, a backup (or bypass) LSP is set up for each link. For example, consider a precalculated backup path to protect a link between routers A and B, say (A-1, B-1), where A-1 is the transmit interface at router A, B-1 is the receive interface at router B, and L-1 is the MPLS label for the path over this link. The forwarding table entries are of form (L-i, A-k) \rightarrow (L-1, A-1) at router A and (L-1, B-1) \rightarrow (L-j, B-s) at router B. When this link fails, a Layer 1 alarm is generated and forwarded to the router controller or line card at A and B. For packets arriving at router A, mapping entries in the forwarding table with the Out-Interface = A-1 have another (outer) layer of label *pushed* on the MPLS stack to coincide with the backup path. This action is preloaded into the forwarding table and triggered by the alarm. Forwarding continues along the routers of this backup LSP by processing the outer layer labels as with any MPLS packet. The backup path ends at router B and, therefore, when the packets arrive at router B, their highest (exterior) layer label is popped. Then, from the point of view of router B, after the outer label is popped, the MPLS header is left with (In-Label, In-Interface) = (L-1, B-1) and therefore the packets continue their journey beyond router B just as they would if link (A-1, B-1) were up. In this way, all LSPs that route over the particular link are rerouted (hence the term “facility based”). Various other specifications can be made to segregate the backup path to be pushed on given classes of LSPs, for example to provide restoration for some IP CoSs rather than others.

Another common implementation of next-hop FRR defines 1-hop pseudowires for each key link. Each pseudowire has defined a primary LSP and backup LSP (a capability found in most routers). If the link fails, a similar alarm mechanism causes the pseudowire to reroute over the backup LSP. When the primary LSP is again declared up, the pseudowire switches back to the primary path. An advantage of this method is that the pseudowire appears as a link to the IGP routing algorithm. Weights can be used to control how packets route over it or the underlying Layer 1 link. Section 2.6 illustrates this method for an IPTV backbone network.

MPLS FRR has been demonstrated to work very rapidly (less than 100 ms) in response to single-link (IP layer PHY link) failures by many vendors and carriers. Most FRR implementations behave similarly during the small interval immediately after the failure and before IGP reconvergence. However, implementations differ in what happens after IGP reconvergence. We describe two main approaches in the context of next-hop FRR here. In the first approach, the backup LSP stays in place until the link goes back into service and IGP reconverges back to its non-failure state. This is most common when a separate LSP or pseudowire is associated with each link in next-hop FRR. In this case, the link-LSP is rerouted onto its backup LSP and stays that way until the primary LSP is repaired.

In the second approach, FRR provides rapid restoration and then, after a short settling period, the network recomputes its paths [4]. Here, each primary end-to-end LSP is recomputed during the first IGP reconfiguration process after the failure. Since the IGP knows about the failed link(s), it reroutes the primary end-to-end LSPs around them and the backup LSPs become moot. This is illustrated in the three potential paths in the topmost diagram of Fig. 2.26. The IP flow routes along the primary LSP during the nonfailure state. Then, the given link fails and the path of the flow over the failed link deviates along the backup LSP, as shown by the lower dashed line. After the first IGP reconfiguration process, the end-to-end LSP path is recomputed, illustrated by the topmost dashed line.

When a failed component is repaired or a maintenance procedure is completed, the disrupted links are put back into service. The process to return the network to its nonfailure state is often called *normalization*. During the normalization process, LSAs are broadcast by the IGP and the forwarding tables are recalculated. The normalization process is often controlled by an MPLS route mechanism/timer. A similar procedure would occur for next-next hop.

The reason for the second approach is that while FRR enables rapid restoration, because these paths are segmental “patches” to the primary paths, the alternate route is often long and capacity-inefficient. With the first approach, IP flows continue routing over the backup paths until the repair is completed and alarms clear, which may span hours or days. Another reason is that if multiple link failures occur, then some of the backup FRR paths may fail; some response is needed to address this situation. These limitations of the first approach were early key inhibitors to implementation of FRR in large ISPs.

The key to implementing this second FRR strategy is that the switch from FRR backup paths to new end-to-end paths is *hitless* (i.e., negligible packet loss), else we may suffer three hits from each single failure (the failure itself, the process to reroute the end-to-end paths immediately after the failure, and then the process to revert to the original paths after repair). If the alternate end-to-end LSPs are presetup and the forwarding table changes implemented efficiently for most routers (often using pointers), this process is essentially hitless for most IP *unicast* (point-to-point) applications. However, we note that today’s *multicast* does not typically enjoy hitless switchover to the new forwarding table because most multicast trees are usually built via `join` and `prune` request messages issued backwards (upstream) from the destination nodes. However, it is expected that different implementations of multicast will fix this problem in the future. We discuss this again in Section 2.6 and refer the reader to [36] for more discussion of hitless multicast.

For the network design phase of implementing FRR, for next-hop FRR, each link (say L) along the primary path needs a predefined a backup path whose routing is diverse in lower layers. That is, the paths of all lower-layer connections that support the links of the backup path are disjoint from the path of the lower-layer connection for link L . The key is in predefining the backup tunnels. While next-next-hop paths can be also used to restore against single-link failures, the network becomes more complex to design if there is a high degree of lower-layer link overlap. More generally, the major difficulty for the FRR approach is defining the backup LSPs so

that the service paths can be rerouted, given a predefined set of lower-layer failures. Furthermore, when multiple lower-layer failures occur and MPLS backup paths fail, FRR does not work and the network must revert to the slower primary path recalculation approach (described in method 2 above).

2.5.3 Failures Across Multiple Layers

Now that the reader is armed with background on network layering and restoration methods, we are poised to delve deeper into the factors and carrier decision variables that shape the availability of the IP backbone.

Let us briefly revisit Fig. 2.9, which gives a simple example of the core ROADM Layer Diagram. Consider a backbone router (BR) in central office B with a link to one of the backbone routers in central office A. Furthermore, consider the remote access router (RAR) that is homed to the backbone router in office A. However, let us add a twist wherein the link between the RAR and BR routes over the IOS layer instead of directly onto the ROADM (DWDM layer) as pictured in Fig. 2.9. This can occur for RAR–BR links with lower bandwidth. This modification will illustrate more of the potential failure modes. In particular, we have constructed this simple example to illustrate several key points:

- Computing an estimate of the availability of the IP backbone involves analysis of many network layers.
- Network disruptions can originate from many different sources within each layer.
- Some lower layers may provide restoration and others do not; how does this affect the IP backbone?

Figure 2.27 gives examples of the types of individual component disruptions (“down events”) that might cause links to fail in this network example, but still only shows a few of the many disruptions that can originate at these layers. As one can see, this is a four-layer example; and, some of the layers are skipped. Note that for simplicity, we illustrate point-to-point DWDM systems at the DWDM layer; however, the concepts apply equally well for ROADMs. Some readers perhaps may think that the main source of network failures is fiber cuts and, therefore, the entire area of multi-layer restoration can be reduced to analyzing fiber cuts. However, this oversimplifies the problem. For example, an amplifier failure can often be as disruptive as a fiber cable cut and will likely result in the failure of multiple IP layer links. Furthermore, amplifier failures are more frequent. Let us examine the effect of some of the failures illustrated in Fig. 2.27.

IOS interface failure: The IOS network has restoration capability, as described in earlier sections. Consequently, the IOS layer reroutes its failed SONET STS-n connection that supports the RAR–BR link onto its restoration path. In this case, once the SONET alarms are detected by the two routers (the RAR and BR), they take the link out of service and generate appropriate LSAs to the correct IGP

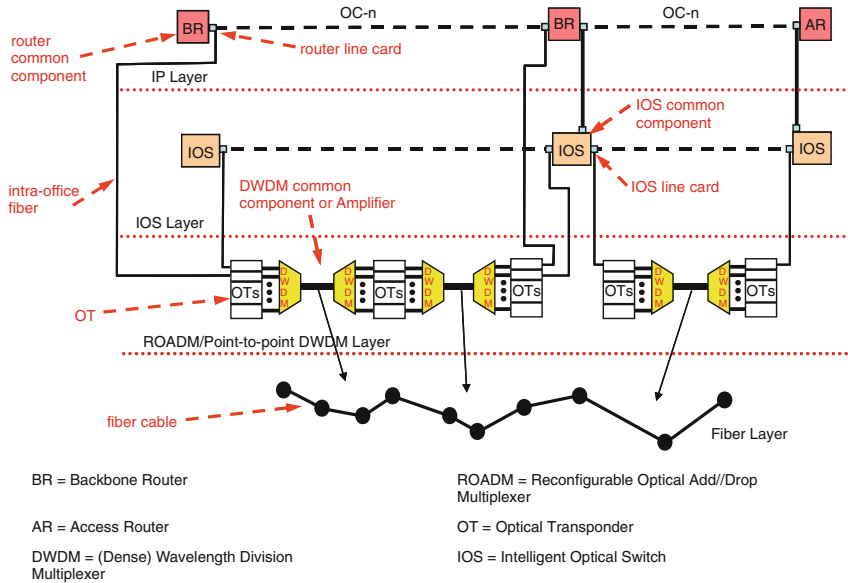


Fig. 2.27 Example of components disruptions (failure or maintenance activity) at multiple layers

administrative areas or control domains to announce the topology change. Assuming that the IOS-layer restoration is successful, the AR–BR link comes back after a short time (as specified in the IOS layer of Table 2.3) and the SONET alarm clears. After perhaps, an appropriate time-out on the routers to avoid link flapping, the link is brought back up by the router and the topology change is announced via LSAs. We note that in a typical AR/BR homing architecture, the LSAs from an AR–BR link are only announced in subareas and so do not affect unaffected ARs or BRs.

Fiber cut: In the core network, the probability of a fiber cut is roughly proportional to its length. They are less frequent than many of the other failures, but highly disruptive, where usually many simultaneous IP layer links fail because of the concentration of capacity enabled by DWDM.

Optical Transponder: OT failure is the most common of the failures shown in Fig. 2.27. However, a single OT failure only affects individual IP backbone links. Some of the more significant problems with OT failures are (1) performance degradation, where bit errors occasionally trip BER threshold crossing alerts and (2) there is a nonnegligible probability of multiple failures in the network, in which an OT fails while another major failure is in progress or vice versa.

DWDM terminal or amplifier: Amplifier failure is usually the most disruptive of failures because of its impact (multiple wavelengths) and sheer quantity, often placed every 50–100 miles, depending on the vintage and bit rate of the wavelengths of the DWDM equipment. Failure of the DWDM terminal equipment not associated with amplifiers and OTs is less probable because of the increased use of

passive (nonelectrical or powered) components. Note that in Fig. 2.27, for the OT, fiber cut, and amplifier failure, the affected connections at their respective layers are unrestored. Thus, the IP layer must reroute around its lost link capacity.

Intra-office fiber: These disruptions usually occur from maintenance, reconfiguration, and provisioning activity in the central office. This has been minimized over the years due to the use of fiber patch panels; however, when significant network capacity expansion or reconfiguration occurs, especially for the deployment of new technologies, architectures, or services, downtime from these class of failures typically spikes. However, it is typical to lump the intra-office fiber disruptions into the downtime for a linecard or port and model them as one unit.

Router: These network disruptions include failure of router line cards, failure of router common equipment, and maintenance or upgrade of all or parts of the router. Note that for these disruptions that originate at the IP layer, no lower-layer restoration method can help because rerouting the associated connections at the lower layers will not bring the affected link back up. However, in the dual-homing AR–BR architecture, all the ARs that home to the affected router can alternatively reroute through the mate BR.

The method of rerouting the AR traffic to the surviving AR–BR links differs per carrier. Usually, IGP reconfiguration is used. However, this can be unacceptably slow for some high-priority services, as evidenced by Table 2.3. Therefore, other faster techniques are sometimes used, such as Ethernet link load balancing or MPLS FRR.

We generalize some simple observations on multilayer restoration illustrated by Fig. 2.27 and its subsequent discussion:

1. Because of the use of express links, a single network failure or disruption at a lower layer usually results in multiple link failures at higher layers.
2. Failures that originate at an upper layer cannot be restored at a lower layer.
3. To meet most ISP network availability objectives, some form of restoration (even if rudimentary) must be provided in upper layers.

2.5.4 IP Backbone Network Design

Network design is covered in more detail in Chapter 5. However, to tie together the concepts of network layering, network failure modeling, and restoration, we provide a brief description of IP network design here to illustrate its importance in meeting network availability targets. In this section, we give a brief description about how these factors are accommodated in the network design. To illustrate this, we describe a very simplified network design (or network planning) process as follows. This process would occur every planning period or whenever major changes to the network occur:

1. Derive a traffic matrix.
2. Input the existing IP backbone topology and compute any needed changes. That is, determine the homing of AR locations to the BR locations and determine which BR pairs are allowed to have links placed between them.
3. Determine the routing of BR–BR links over the lower-layer networks (e.g., DWDM, IOS, fiber).
4. Route the traffic matrix over the topology and size the links. This results in an estimate of network cost across all the needed layers.
5. Resize the links by finding their maximum needed capacity over all possible events in the *Failure Set*, which models potential network disruptions (both component failures and maintenance activity). This step simulates each failure event, determining which IP layer link or nodes fail after lower-layer restoration, if it exists, is applied and determining the capacity needed after traffic is rerouted using IP layer restoration.
6. Re-optimize the topology by going back to step 2 and iterating with the objective of lowering network cost.

Note in steps 2 and 3 that most carriers are reluctant to make large changes to the existing IP backbone topology, since these can be very disruptive and costly events. Therefore, steps 2 and 3 usually incur small topology changes from one planning period to another planning period. We will not describe detailed algorithms for the above in detail here. Approaches to the above problem can be found in [22, 23].

The traffic matrix can come in a variety of forms, such as the peak 5-min average loads between AR-pairs or average loads, etc. Unfortunately, many organizations responsible for IP network design either have little or no data about their current or future traffic matrices. In fact, many engineers who manage IP networks expand their network by simply observing link loads. When a link load exceeds some threshold, they add more capacity. Given no knowledge or high uncertainty of the true, stochastic traffic matrix, this may be a reasonable approach. However, network failures and their subsequent restorations are the phenomena that cause the greatest challenges with such a simple approach. Because of the extensive rerouting that can occur after a network failure, there is no simple or intuitive parameter to determine the utilization threshold for each link. Traffic matrix estimation is discussed in detail in Chapter 5.

A missing ingredient in the above network design algorithm is we did not describe how to model the needed network availability for an ISP to achieve its SLAs. Theoretically, even if we assume the traffic matrix (present and/or future) is completely accurate, to achieve the network design availability objective, all the component failure modes and all the network layering must be modeled to design the IP backbone. The decision variables are the layers where we provide restoration (including what type of restoration should be used) and how much capacity should be deployed at each layer to meet the QoS objectives for the IP layer. This is further complicated by the fact that while network availability objectives for transport layers are often expressed in worst-case or average-case connection uptimes, IP backbone QoS objective often use packet-loss metrics.

However, we can approximate the packet loss constraints in large IP layer networks by establishing maximum link utilization targets. For example, through separate analysis it might be determined that every flow can achieve the objective maximum packet loss target by not exceeding 90% utilization on any 40 Gb/s link, with perhaps lower utilization maxima needed on lower-rate links. Then, one can model when this utilization condition is met over the set of possible failures, including subsequent restoration procedures. By modeling the probabilities of the failure set, one can compute a network availability metric appropriate for packet networks. The probabilities of events in the failure set can be computed using Markov models and the *Mean Time Between Failures (MTBF)* and the *Mean Time to Repair (MTTR)* of the component disruptions. These parameters are usually obtained from a combination of equipment-supplier specifications, network observation/data, and carrier policies and procedures.

A major stumbling block with this theoretical approach is that the failure event space is exponential in size. Even for very small networks and a few layers, it is intractable to compute all potential failures, let alone the subsequent restoration and network loss. An approach to probabilistic modeling to solve this problem is presented in more detail in Chapter 4 and in [28].

Armed with this background, we conclude this section by revisiting the issue of why we show the IP backbone routing over an unrestorable DWDM layer in the network layering of Fig. 2.3. This at first may seem counterintuitive because it is generally true that, per unit of capacity, the cost of links at lower layers is less than that of higher layers. Some of the reasons for this planning decision, which is consistent with most large ISPs, were hinted at in Section 2.5.3. We summarize them here.

1. Backbone router disruptions (failures or maintenance events) originate within the IP layer and cannot be restored at lower layers. Extra link capacity must be provided at the IP layer for such disruptions. Once placed, this extra capacity can then also be used for IP layer link failures that originate at lower layers. This obviates most of the cost advantages of lower-layer restoration.
2. Under nonfailure conditions, there is spare capacity available in the IP layer to handle uncertain demand. For example, restoration requirements aside, to handle normal service demand, IP layer links could be engineered to run below 80% utilization during peak intervals of the traffic matrix and well below that at off-peak intervals. If we allow higher utilization levels during network disruption events, then this provides an existing extra buffer during those events. Furthermore, there may be little appreciable loss during network disruptions during off-peak periods.

As QoS and CoS features are deployed in the IP backbone, there is yet another advantage to IP layer restoration. Namely, the IP layer can assign different QoS objectives to different service classes. For example, one such distinction might be to plan network restoration so that premium services receive better performance than best-effort services during network disruptions. In contrast, the DWDM layer cannot make such fine-grain distinctions; it either restores or does not restore the entire IP layer link, which carries a mixture of different classes of services.

2.6 IPTV Backbone Example

Some major carriers now offer nationwide digital television, high-speed Internet, and Voice-over-IP services over an IP network. These services typically include hundreds of digital television channels. Video content providers deliver their content to the service provider in digital format at select locations called *super hub offices (SHOs)*. This in turn requires that the carrier have the ability to deliver high-bandwidth IP streaming to its residential customers on a nationwide basis. If such content is delivered all the way to residential set-top boxes over IP, it is commonly called *IPTV*. There are two options to providing such an IPTV backbone. The first option is to create a virtual network on top of the IP backbone. Since video service consists mostly of streaming channels that are broadcast to all customers, IP multicast is usually the most cost-effective protocol to transport the content. However, users have high expectations for video service and even small packet losses negatively impact video quality. This requires the IP backbone to be able to transport multicast traffic at a very high level of network availability and efficiency. The first option results in a mixture of best-effort traffic and traffic with very high quality of service on the same IP backbone, which in turn requires comprehensive mechanisms for restoration and priority queuing.

Consequently, some carriers have followed the second option, wherein they create a separate overlay network on top of the lower-layer DWDM or TDM layers. In reality, this is another (smaller) IP layer network, with specialized traffic, network structure, and restoration mechanisms. We describe such an example in this section. Because of the high QoS objectives needed for broadcast TV services, the reader will find that this section builds on most of the previous material in this chapter.

2.6.1 Multicast-Based IPTV Distribution

Meeting the stringent QoS required to deliver a high-quality video service (such as low latency and loss) requires careful consideration of the underlying IP-transport network, network restoration, and video and packet recovery methods.

Figure 2.28 (borrowed from [9]) illustrates a simplified architecture for a network providing IPTV service. The SHO gathers content from the national video content providers, such as TV networks (mostly via satellite today) and distributes it to a large set of receiving locations, called *video hub offices (VHOs)*. Each VHO in turn feeds a metropolitan area. IP routers are used to transport the IPTV content in the SHO and VHOs. The combination of SHO and VHO routers plus the links that connect them comprise the IPTV backbone. The VHO combines the national feeds with local content and other services and then distributes the content to each metro area. The long-distance backbone network between the SHO and the VHO includes a pair of redundant routers that are associated with each VHO. This allows for protection against router component failures, router hardware maintenance, or software

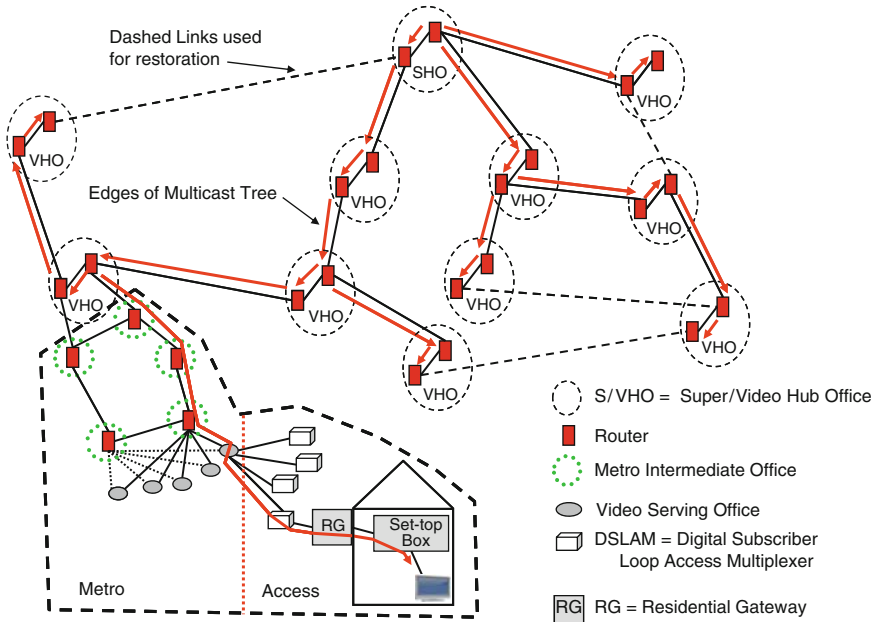


Fig. 2.28 Example nationwide IPTV network

upgrades. IP multicast is used for delivery as it provides economic advantages for the IPTV service to distribute video. With multicast, packets traverse each link at most once.

The video content is encoded using an encoding standard such as *H.264*. Video frames are packetized and are encapsulated in the *Real-Time Transport Protocol (RTP)* and UDP. In this example, PIM-SSM is used to support IP multicast over the video content. Each channel from the national live feed at the SHO is assigned a unique multicast group. There are typically hundreds of channels assigned to standard-definition (SD) (1.5 to 3 Mb/s) and high-definition (HD) (6 to 10 Mb/s) video signals plus other multimedia signals, such as “picture-in-picture” channels and music. So, the live feed can be multiple gigabits per second in aggregate bandwidth.

2.6.2 Restoration Mechanisms

The IPTV network can use various restoration methods to deliver the needed video QoS to end-users. For example, it can recover from relatively infrequent and short bursts of loss using a combination of video and packet recovery mechanisms and protocols, including the Society of Motion Picture and Television Engineers (SMPTE; www.smpte.org/standards) 2022–1 Forward Error Correction (FEC)

standard, retransmission approaches based on RTP/RTCP [33] and Reliable UDP (R-UDP) [31], and video player loss-concealment algorithms in conjunction with set-top box buffering. R-UDP supports retransmission-based packet-loss recovery. In addition to protecting against video impairments due to last-mile (loop) transmission problems in the access segment, a combination of these methods can recover from a network failure (e.g., fiber link or router line card) of 50 ms or less. Repairing network failures usually takes far more than 50 ms (potentially several hours), but when combined with link-based FRR, this restoration methodology could meet the stringent requirements needed for video against single-link failures.

Figure 2.29 (borrowed from [9]) illustrates how we might implement link-based FRR in an IPTV backbone by depicting a network segment with four node pairs that have defined virtual links (or pseudowires). This method is the pseudowire, next-hop FRR approach described in Section 2.5.2.2. For example, node pair E-C has a lower-layer link (such as SONET OC- n or Gigabit Ethernet) in each direction and a pseudowire in each direction (a total of four unidirectional logical links) used for FRR restoration. The medium dashed line shows the FRR backup path for the pseudowire E→C. Note that links such as E-A are for restoration and, hence, have no pseudowires defined. Pseudowire E→C routes over a primary path that consists of the single lower-layer link E→C (see the solid line in Fig. 2.29). If a failure occurs to a lower-layer link in the primary path such as C-E, then the router at node E attempts to switch to the backup path using FRR. The path from the root to node A will switch to the backup path at node E (E-A-B-C). Once it reaches node C, it will

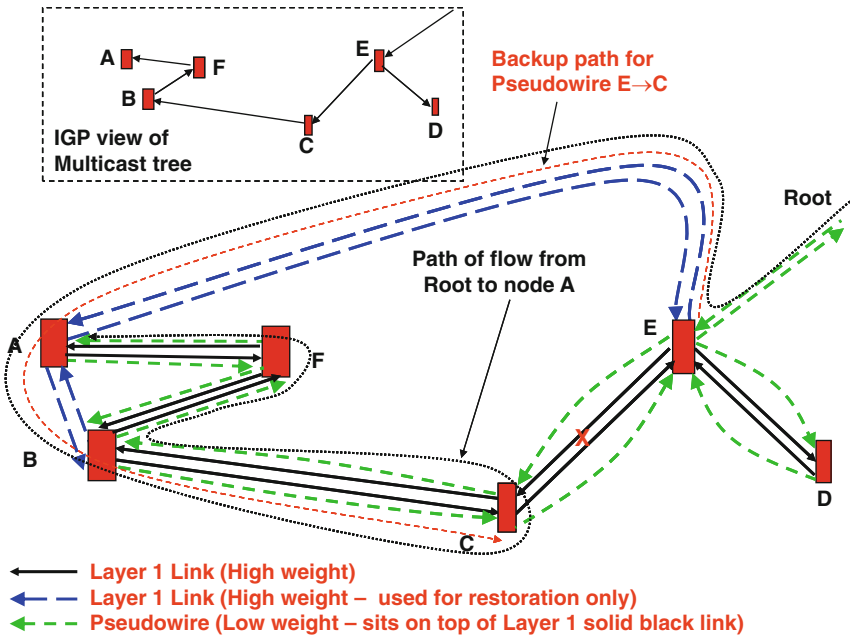


Fig. 2.29 Fast Reroute in IPTV backbone

continue on its previous (primary) path to node A (C-B-F-A). The entire path from E to A during the failure is shown by the outside dotted line. Although the path retraces itself between the routers B and C, the multicast traffic does not overlap because of the links' unidirectionality. Also, although the IGP view of the topology realizes that the lower-layer links between E and C have gone "down," because the pseudowire from E→C is still "up" and has the least weight, the shortest path tree remains unchanged. Consequently, the multicast tree remains unchanged. The IGP is unaware of the actual routing over the backup path. Note that these backup paths are precomputed, by analyzing all possible link failures in a comprehensive manner, a priori.

If we route the pseudowire FRR backup path on a lower-layer path that is diverse from its primary path, FRR operates rapidly (suppose around 50 ms), and we set the hold-down timers appropriately, IGP will not detect the effect of any single fiber or DWDM layer link failure. Therefore, the multicast tree will remain unaffected, reducing the outage time of any single-link failure from tens of seconds to approximately 50 ms. This order of restoration time is needed to achieve the stringent IPTV network availability objectives.

2.6.3 Avoiding Congestion from Traffic Overlap

A drawback of restoration using next-hop FRR is that since it reroutes traffic on a link-by-link basis, it can suffer traffic overlap during link failures, thus requiring more link capacity to meet the target availability. Links are deployed bidirectionally, and traffic overlap means that the packets of the same multicast flows travel over the same link (in the same direction) two or more times. If we avoid overlap, we can run the links at higher utilization and thus design more cost-effective networks. This requires that the multicast tree and backup paths be constructed so that traffic does not overlap.

To illustrate traffic overlap, Fig. 2.30a shows a simple network topology with node *S* as the source and nodes *d1* to *d8* as the destinations. Here, each router is connected by a pair of directed links (in opposite directions). The two links of the pair are assigned the same IGP weight and the multicast trees are derived from these weights. The Fig. 2.30a illustrates two sets of link weights. Figure 2.30b shows the multicast tree derived from the first set of weights. In this case, there exists a single-link failure that causes traffic overlap. For example, the dotted line shows the backup route for link *d1*–*d4*. If link *d1*–*d4* fails, then the rerouted traffic will overlap with other traffic on links *S*–*d2* and *d2*–*d6*, thereby resulting in congestion on those links. Client routers downstream of *d2* and *d6* will see impairments as a result of this congestion. It is desirable to avoid this congestion wherever possible by constructing a multicast tree such that the backup path for any single-link failure does not overlap with any downstream link on the multicast tree. This is achieved by choosing OSPF link weights suitably.

The tree derived from the second pair of weights is shown in Fig. 2.30c. In this case, the backup paths do not cause traffic overlap in response to any single-link

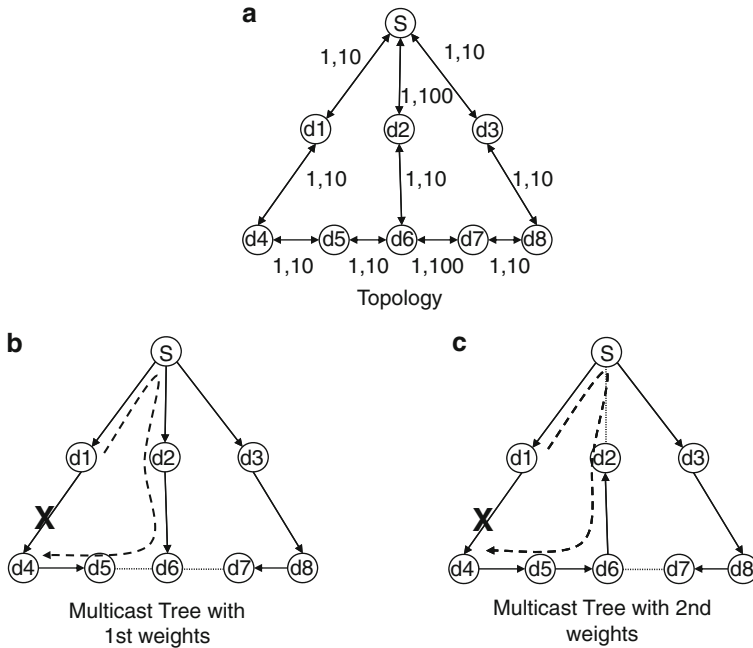


Fig. 2.30 Example of traffic overlap from single-link failure

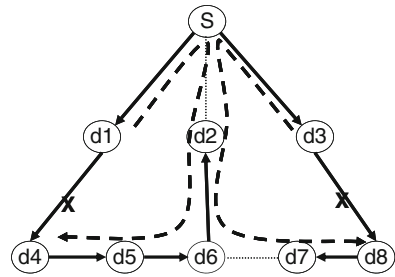
failure. The multicast tree link is now from $d6$ to $d2$. The backup path for link $d1-d4$ is the same as in Fig. 2.30b. Observe that traffic on this backup path does not travel in the same direction as any link of the multicast tree. An algorithm to define FRR backup paths and IGP weights so that the multicast tree does not overlap from any single failure can be found in [10].

2.6.4 Combating Multiple Concurrent Failures

The algorithm and protocol in [10] helps in avoiding traffic overlap of the multicast tree during single-link failures. However, multiple link failures can still cause overlap. An example is shown in Fig. 2.31. Assume that links $d1-d4$ and $d3-d8$ are both down. If the backup path for edge $d1-d4$ is $d1-S-d2-d6-d5-d4$ (as shown in Fig. 2.30b and in Fig. 2.31) and the backup path for edge $d3-d8$ is $d3-S-d2-d6-d7-d8$, traffic will overlap paths on edges $S-d2$ and $d2-d6$. There would be significant traffic loss due to congestion if the links of the network are sized to only handle a single stream of multicast traffic.

This situation essentially occurs because MPLS FRR occurs at Layer 2 and therefore the IGP is unaware of the FRR backup paths. Furthermore, the FRR backup paths are precalculated and there is no real-time (dynamic) accommodation for

Fig. 2.31 Example of traffic overlap from multiple link failures



different combinations of *multiple-link* failures. In reality, multiple (double and even triple) failures can happen. When they occur, they can have a large impact on the performance of the network.

Yuksel [36] describes an approach that builds on the FRR mechanism but limits its use to a short period. When a single link fails and a pseudowire's primary path fails, the traffic is rapidly switched over to the backup path as described above. However, soon afterwards, the router sets the virtual link weight to a high value and thus triggers the IGP reconvergence process – this is colloquially called “costing out” the link. Once IGP routing converges, a new PIM tree is rebuilt automatically. This avoids long periods where routing occurs over the FRR backup paths, which are unknown to the IGP. This ensures rapid restoration from single-link failures while allowing the multicast tree to dynamically adapt to any additional failures that might occur during a link outage. It is only during this short, transient period when FRR starts and IGP reconvergence finishes that another failure could expose the network to a path overlapping on the same link. The potential downside of this approach is that it incurs two more network reconvergence processes – that is, the period right after FRR has occurred and then again when the failure is repaired. If it is not carefully executed, this alternative approach can cause many new video interruptions due to small “hits” after single failures.

Yuksel [36] proposes a careful multicast recovery methodology to accomplish this approach, yet avoid such drawbacks. A key component of the method is the *make-before-break change* of the multicast tree – that is, the requirement to hitlessly switch traffic from the old multicast tree to the new multicast tree. When the failure is repaired, the method normalizes the multicast tree to its original shortest path tree again in a hitless manner. The key modification to the multicast tree-building process (pruning and joining nodes) is that the `prune` message to remove the branch to the previous parent is not sent until the router receives PIM-SSM data packets from its new parent for the corresponding (S, G) group. Another motivation for this modification is because current PIM-SSM multicast does not have an explicit acknowledgement to a `join` request. It is only through the receipt of a data packet on that interface that the node knows that the `join` request was successfully received and processed at the upstream node. The *soft-state* approach of IP Multicast (refresh the state by periodically sending join requests) is also used to ensure consistency. This principle is used to guide the tree reconfiguration process at a node in reaction to a

failure. In this way, routers do not lose data packets during the switchover period. Of course, this primarily works in the PIM-SSM case, where there is a single source.

As we can observe from the description above, building an IPTV backbone with high network availability builds on most of the protocols, multilayer failure models, and restoration machinery we have described in the previous sections of the chapter. In particular, given the underlying probabilities of network failures plus these complex failure and restoration mechanisms, such an approach must include the network design methodology to evaluate and estimate the theoretical network availability of the IPTV backbone. If such a methodology was not utilized, a carrier would run the risk of having its video customers dissatisfied with their video service because of inadequate network availability.

2.7 Summary

This chapter presents an overview of the layered network design that is typical in a large ISP backbone. We emphasized three aspects that influence the design of an IP backbone. The first aspect is that the IP network design is strongly influenced by its relationship with the underlying network layers (such as DWDM and TDM layers) and the network segments (core, metro, and access). ISP networks use a hierarchy of specialized routers, generally called access and backbone routers. At the edge of the network, the location of access routers, and the types of interfaces that they need to support are strongly influenced by the way the customers connect to the backbone through the metro network. In the core of a large carrier network, backbone routers are interconnected using DWDM transmission technology. As IP traffic is the dominant source of demand for the DWDM layer, the backbone demands drive requirements for the DWDM layer. The need for multiple DWDM links has driven the evolution of aggregate links in the core.

The second aspect is that ISP networks have evolved from traditional IP forwarding to support MPLS. The separation of routing and forwarding and the ability to support a routing hierarchy allow ISPs to support new functionality including Layer 2 and Layer 3 VPNs and flexible traffic engineering that could not be as easily supported in a traditional IP network.

Finally, this chapter provided an overview of the issues that affect IP network reliability, including the impact of network disruptions at multiple network layers and, conversely, how different network layers respond to disruptions through network restoration. We described how failures and maintenance events originate at various network layers and how they impact the IP backbone. We presented an overview of the performance of OSPF failure recovery to motivate the need for MPLS Fast Reroute. We summarized the interplay between network restoration and the network design process.

To tie these concepts together, we presented a “case study” of an IPTV backbone. An IPTV network can be thought of as an IP layer with a requirement for very high performance, essentially high network availability and low packet loss. This

requires the interlacing of multiple protocols, such as R-UDP, MPLS Fast Reroute, IP Multicast, and Forward Error Control. We described how lower-layer failures (including multiple failures) affect the IP layer and how these IP layer routing and control protocols respond. Understanding the performance of network restoration protocols and the overall availability of the given network design requires careful modeling of the types and likelihood of network failures, as well as the behavior of the restoration protocols. This chapter endeavored to lay a good foundation for reading the remaining chapters of this book.

We conclude by alerting the reader to an important observation about IP network design. Telecommunications and its technologies undergo constant change. Therefore, this chapter describes a point in time. The contents of this chapter are different from what they would have been 5 years ago. There will be further changes over the next 5 years and, consequently, the chapter written 5 years from now may look quite different.

References

1. AT&T (2003). Managed Internet Service Access Redundancy Options, from <http://www.pnetcom.com/AB-0027.pdf>. Accessed 15 April 2009.
2. Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., & Swallow, G. (2001). RSVP-TE: Extensions to RSVP for LSP Tunnels. IETF RFC 3209, Dec. <http://tools.ietf.org/html/rfc3209>. Accessed 29 January 2010.
3. Braden, R., Zhang, L., Berson, S., Herzog, S., & Jamin, S. (1997). Resource ReSer-Vation Protocol (RSVP) – Version 1 Functional Specification. IETF RFC 2205, Sept. <http://tools.ietf.org/html/rfc2205>. Accessed 29 January 2010.
4. Chiu, A., Choudhury, G., Doverspike, R., & Li, G. (2007). Restoration design in IP over re-configurable all-optical networks. NPC 2007, Dalian, P.R. China, September 2007.
5. Choudhury, G. (Ed.) (2005). Prioritized Treatment of Specific OSPF Version 2 Packets and Congestion Avoidance. IETF RFC 4222, Oct.
6. Ciena Core Director. http://www.ciena.com/products/products_coredirector_product_overview.htm. Accessed 13 April 2009.
7. Cisco (1999). Tag Switching in *Internetworking Technology Handbook*, Chapter 23, <http://www.cisco.com/en/US/docs/internetworking/technology/handbook/Tag-Switching.pdf>, accessed 12/26/09.
8. Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). Introduction to algorithms, second edition (pp. 595–601). Cambridge: MIT Press, New York: McGraw-Hill. ISBN 0–262–03293–7. Section 24.3: Dijkstra’s algorithm.
9. Doverspike R., Li, G., Oikonomou, K. N., Ramakrishnan, K. K., Sinha, R. K., Wang, D., et al. (2009). Designing a reliable IPTV network. *IEEE Internet Computing Magazine* May/June, pp. 15–22.
10. Doverspike, R., Li, G., Oikonomou, K., Ramakrishnan, K. K., & Wang, D. (2007). IP backbone design for multimedia distribution: architecture and performance. INFOCOM-2007, Anchorage Alaska April 2007.
11. Doverspike, R., & Magill, P. (2008). Commercial optical networks, overlay networks and services. In I. Kaminow, T. Li, & A. Willner, (Eds), Chapter 13 in *Optical fiber telecommunications VB*. San Diego, CA: Academic.
12. Feuer, M., Kilper, D., & Woodward, S. (2008). ROADMs and their system applications. In I. Kaminow, T. Li, & A. Willner, (Eds), Chapter 8 in *Optical fiber telecommunications VB*. San Diego, CA: Academic.

13. Goyal, M., Ramakrishnan K. K., & Feng W. (2003) "Achieving Faster Failure Detection in OSPF Networks," IEEE International Conference on Communications (ICC 2003), Alaska, May 2003.
14. IEEE 802.1Q-2005 (2005) Virtual Bridged Local Area Networks; ISBN 0-7381-3662-X.
15. IEEE: 802.1Qay – Provider Backbone Bridge Traffic Engineering. <http://www.ieee802.org/1/pages/802.1ay.html>. Accessed October 7, 2008.
16. IETF PWE3: Pseudo Wire Emulation Edge to Edge (PWE3) Working Group. <http://www.ietf.org/html.charters/pwe3-charter.html>. Accessed 7 Nov 2008.
17. IETF RFC 4090 (2005) Fast Reroute Extensions to RSVP-TE for LSP Tunnels. <http://www.ietf.org/rfc/rfc4090.txt>. May 2005. Accessed 7 Nov 2008.
18. ITU-T G.709, "Interfaces for the Optical Transport Network," March 2003.
19. ITU-T G.7713.2. Distributed Call and Connection Management: Signalling mechanism using GMPLS RSVP-TE.
20. Kalmanek, C. (2002). A Retrospective View of ATM. ACM Sigcomm CCR, Vol. 32, Issue 5, Nov, ISSN: 0146-4833.
21. Katz, D., Kompella, K., & Yeung, D. (2003). IETF RFC 3630: Traffic Engineering (TE) Extensions to OSPF Version 2. <http://tools.ietf.org/html/rfc3630>. Accessed 4 May 2009.
22. Klinkewicz, J. G. (2005). Issues in link topology design for IP networks. SPIE Conference on performance, quality of service and control of next-generation communication networks III, SPIE Vol. 6011, Boston, MA.
23. Klinkewicz, J. G. (2006). Why is IP network design so difficult? Eighth INFORMS telecommunications conference, Dallas, TX, March 30–April 1, 2006.
24. Kompella, K., & Rekhter, Y. (2007). IETF RFC 4761: Virtual private LAN service (VPLS) using BGP for auto-discovery and signaling. <http://tools.ietf.org/html/rfc4761>, accessed 12/26/09.
25. Lasserre, M., & Kompella, V. (2007). IETF RFC 4762: Virtual private LAN service (VPLS) using label distribution protocol (LDP) signaling. <http://tools.ietf.org/html/rfc4762>, accessed 12/26/09.
26. Moy, J. (1998). IETF RFC 2328: OSPF Version 2. <http://tools.ietf.org/html/rfc2328>, accessed 12/26/09.
27. Nortel. (2007). Adding scale, QoS and operational simplicity to Ethernet. <http://www.nortel.com/solutions/collateral/nn115500.pdf>, accessed 12/26/09.
28. Oikonomou, K., Sinha, R., & Doverspike, R. (2009). Multi-Layer Network Performance and Reliability Analysis. The International Journal of Interdisciplinary Telecommunications and Networking (IJITN), Vol. 1 (3), pp. 1–29, Sept.
29. Optical Internetworking Forum (OIF) (2008). OIF-UNI-02.0-Common-User Network Interface (UNI) 2.0 Signaling Specification: Common Part. <http://www.oiforum.com/public/documents/OIF-UNI-02.0-Common.pdf>.
30. Oran, D. (1990). IETF RFC 1142: OSI IS-IS intra-domain routing protocol. <http://tools.ietf.org/html/rfc1142>.
31. Partridge, C., & Hinden, R. (1990). Version 2 of the Reliable Data Protocol (RDP), IETF RFC 1151. April.
32. Perlman, R. (1999). Interconnections: Bridges, Routers, Switches, and Internetworking Protocols, 2e. Addison-Wesley Professional Computing Series.
33. Schulzrinne, H., Casner, S., Frederick, R., & Jacobson, V. (2003). RTP: A Transport Protocol for Real-Time Application, IETF RFC 3550. <http://www.ietf.org/rfc/rfc3550.txt>, accessed 12/26/09.
34. Sycamore Intelligent Optical Switch. (2009). <http://www.sycamorenet.com/products/sn16000.asp>. Accessed 13 April 2009.
35. Telcordia GR-253-CORE (2000) Synchronous Optical Network (SONET) Transport Systems: Common Generic Criteria.
36. Yuksel, M., Ramakrishnan, K. K., & Doverspike, R. (2008). Cross-layer failure restoration for a robust IPTV service. LANMAN-2008, Cluj-Napoca, Romania September.
37. Zimmermann, H. (1980). OSI reference model – the ISO model of architecture for open systems interconnection. *IEEE Transactions on Communications*, 28(Suppl. 4), 425–432.

Glossary of Acronyms and Key Terms

1:1	One-by-one (signal switched to restoration path on detection of failure)
1 + 1	One-plus-one (signal duplicated across both service path and restoration path; receiver chooses surviving signal upon detection of failure)
Access Network Segment	The feeder network and loop segments associated with a given metro segment
ADM	Add/Drop Multiplexer
Administrative Domain	Routing area in IGP
Aggregate Link	Bundles multiple physical links between a pair of routers into a single virtual link from the point of view of the routers. Also called bundled or composite link
AR	Access Router
AS	Autonomous System
ASBR	Autonomous System Border Router
ATM	Asynchronous Transfer Mode
AWG	Arrayed Waveguide Grating
B-DCS	Broadband Digital Cross-connect System (cross-connects at DS-3 or higher rate)
Backhaul	Using TDM connections that encapsulate packets to connect customers to packet networks
BER	Bit Error Rate
BGP	Border Gateway Protocol
BLSR	Bidirectional Line-Switched Ring
BR	Backbone Router
Bundled Link	See Aggregate Link
CE switch	Customer-Edge switch
Channelized	A TDM link/connection that multiplexes lower-rate signals into its time slots
CHOC Card	CHannelized OC- <i>n</i> card
CIR	Committed Information Rate
CO	Central Office
Composite Link	See Aggregate Link
Core Network Segment	Equipment in the POPs and network structures that connect them for intermetro transport and switching
CoS	Class of Service
CPE	Customer Premises Equipment

CSPF	Constraint-based Shortest Path First
DCS	Digital Cross-connect System
DDoS	Distributed Denial of Service (security attack on router)
DoS	Denial of Service (security attack on router)
DS-0	Digital Signal – level 0 a pre-SONET signal carrying one voice-frequency channel at 64 kb/s)
DS-1	Digital Signal – level 1 (a 1.544 Mb/s signal). A channelized DS-1 carries 24 DS0s
DS-3	Digital Signal – level 3 (a 44.736 Mb/s signal). A channelized DS-3 carries 28 DS1s
DWDM	Dense Wavelength-Division Multiplexing
E-1	European plesiosynchronous (pre-SDH) rate of 2.0 Mb/s
eBGP	External Border Gateway Protocol
EGP	Exterior Gateway Protocol
EIGRP	Enhanced Interior Gateway Routing Protocol
EIR	Excess Information Rate
EPL	Ethernet Private Line
FCC	Federal Communications Commission
FE	Fast Ethernet (100 Mb/s)
FEC	Forward Error Correction – bit-error recovery technique in TDM transmission and some IPs
FEC	Forwarding Equivalence Class – classification of flows defined in MPLS
Feeder Network	The portion of the access network between the loop and first metro central office
FRR	Fast Re-Route
FXC	Fiber Cross-Connect
Gb/s	Gigabits per second (1 billion bits per second)
GigE	Gigabit Ethernet (nominally 1 Gb/s)
GMPLS	Generalized MPLS
HD	High definition (short for HDTV)
HDTV	High-definition TV (television with resolution exceeding 720×1280)
Hitless	Method of changing network connections or routes that incur negligible loss
iBGP	Interior Border Gateway Protocol
IETF	Internet Engineering Task Force
IGP	Interior Gateway Protocol
Internet Route Free Core	Where MPLS removes external BGP information plus Layer 3 address lookup from the interior of the IP backbone
IGMP	Internet Group Management Protocol
Inter-office Links	Links whose endpoints are contained in different central offices

Intra-office Links	Links that are totally contained within the same central office
IOS	Intelligent Optical Switch
IP	Internet Protocol
IPTV	Internet Protocol television (i.e., entertainment-quality video delivered over IP)
IROU	Indefeasible Right of Use
IS-IS	Intermediate-System-to-Intermediate-System (IP routing and control plane protocol)
ISO	International Organization for Standardization (not an acronym)
ISP	Internet Service Provider
ITU	International Telecommunication Union
Kb/s	Kilobits per second (1,000 bits per second)
LAN	Local Area Network
LATA	Local Access and Transport Area
Layer <i>n</i>	A colloquial packet protocol layering model, with origins to the OSI reference model. Today, roughly Layer 3 corresponds to IP packets, Layer 2 to MPLS LSPs, pseudowires, or Ethernet-based VLANs, and Layer 1 to all lower-layer transport protocols
LDP	Label Distribution Protocol
LMP	Link Management Protocol
Local Loop	The portion of the access segment between the customer and feeder network. Also called “last mile”
LSA	Link-State Advertisement
LSDB	Link-State Database
LSP	Label Switched Path
LSR	Label Switch Router
MAC	Media Access Control
MAN	Metropolitan Area Network
Mb/s	Megabits per second (1 Million bits per second)
MEMS	Micro-Electro-Mechanical Systems
Metro Network Segment	The network layers of the equipment located in the central offices of a given metropolitan area
MPEG	Moving Picture Experts Group
MPLS	Multiprotocol Label Switching
MSO	Multiple System Operator (typically coaxial cable companies)
MSP	Multi-Service Platform – A type of ADM enhanced with many forms of interfaces
MTBF	Mean Time Between Failure

MTSO	Mobile Telephone Switching Office
MTTR	Mean Time to Repair
Multicast	Point-to-multipoint flows in packet networks
N-DCS	Narrowband Digital Cross-connect System (cross-connects at DS0 rate)
n -degree	A ROADM that can fiber to more than three different
ROADM	ROADMS (also called multidegree ROADM)
Next-hop	Method in MPLS FRR that routes around a down link
Next-next-hop	Method in MPLS FRR that routes around a down node
Normalization	Step in network restoration after all failures are repaired to bring the network back to its normal state
NTE	Network Terminating Equipment
OC- n	Optical Carrier – level n (designation of optical transport of a SONET STS- n)
ODU	Optical channel Data Unit – protocol data unit in ITU OTN
O-E-O	Optical-to-Electrical-to-Optical
OIF	Optical Internetworking Forum
OL	Optical Layer
OSPF	Open Shortest Path First
OSPF-TE	Open Shortest Path First – Traffic Engineering
OSS	Operations Support System
OT	Optical Transponder
OTN	Optical Transport Network – ITU optical protocol
P Router	Provider Router
PBB-TE	Provider Backbone Bridge – Traffic Engineering
PBT	Provider Backbone Transport
PE Router	Provider-Edge Router
PIM	Protocol-Independent Multicast
PL	Private Line
P-NNI	Private Network-to-Network Interface (ATM routing protocol)
POP	Point Of Presence
PPP	Point-to-Point Protocol
PPPoE	Point-to-Point Protocol over Ethernet
Pseudowire	A virtual connection defined in the IETF PWE3 that encapsulates higher-layer protocols
PVC	Permanent Virtual Circuit
PWE3	Pseudo-Wire Emulation Edge-to-Edge
QoS	Quality of Service
RAR	Remote Access Router
RD	Route Distinguisher
Reconvergence	IGP process to update network topology and adjust routing tables

RIB	Router Information Base
ROADM	Reconfigurable Optical Add/Drop Multiplexer
RR	Route Reflector
RSTP	Rapid Spanning Tree Protocol
RSVP	Resource Reservation Protocol
RT	Route Target (also Remote Terminal in metro TDM networks)
RD	Route Distinguisher
RTP	Real-Time Protocol
SD	Standard Definition (television with resolution of about 640×480)
SDH	Synchronous Digital Hierarchy (a synchronous optical networking standard used outside North America, documented by the ITU in G.707 and G.708)
Serving CO	The first metro central office to which a given customer homes
SHO	Super Hub Office
SLA	Service Level Agreement
SRLG	Shared Risk Link Group
SONET	Synchronous Optical Network (a synchronous optical networking standard used in North America, documented in GR-253-CORE from Telcordia)
SONET/SDH self-healing rings	Typically UPSR or BLSR rings
SPF	Shortest Path First
STS- n	Synchronous Transport Signal – level n (a signal level of the SONET hierarchy with a data rate of $n \times 51.84$ Mb/s)
SVC	Switched Virtual Circuit
TCP	Transmission Control Protocol
TDM	Time Division Multiplexing
UDP	User Data Protocol
UNI	User-Network Interface
Unicast	Point-to-point flows in packet networks
UPSR	Unidirectional Path-Switched Ring
VHO	Video Hub Office
VLAN	Virtual Local Area Network
VoD	Video on Demand
VoIP	Voice-over-Internet Protocol
VPLS	Virtual Private LAN Service (i.e., Transparent LAN Service)
VPN	Virtual Private Network

WAN	Wide Area Network
Wavelength continuity	A restriction in DWDM equipment that a through connection must be optically cross-connected to the same wavelength on both fibers
W-DCS	Wideband Digital Cross-connect System (cross-connects at DS-1, SONET VT- n or higher rate)
DWDM	Wavelength-Division Multiplexing

Guide to Reliable Internet Services and Applications

Kalmanek, C.R.; Misra, S.; Yang, Y.R. (Eds.)

2010, XIV, 644 p. 168 illus., Hardcover

ISBN: 978-1-84882-827-8