Contents

Introduction 2
Traffic Engineering 2
Applications for Traffic Engineering 3
Looking Ahead 3
The Past: Traditional Router Cores 4
Traffic Engineering Capabilities 4
Limitations of a Traditional Routed Core for Traffic Engineering 5
The Present: IP Overlay Networks 6
Operation of an IP Overlay Network 6
Benefits of the IP-over-ATM Model in Service Provider Networks 8
Limitations of the IP-over-ATM Model in the OC-48 Optical Internet 9
The Future: The Juniper Networks Architecture for Traffic Engineering 11
Components of the Traffic Engineering Solution 11
Packet-Forwarding Component 12
LSR Packet Forwarding Based on Label Swapping 12
Example of a How Packet Traverses an MPLS Backbone 13
Benefits of MPLS 14
Information Distribution Component 14
Path Selection Component 15
Signaling Component 17
Flexible LSP Calculation and Configuration 18
Operational Requirements for a Successful Traffic Engineering Solution 19
Benefits of the Juniper Networks Traffic Engineering Architecture 20
Conclusion 21
References 22
Introduction

The central challenge facing Internet Service Providers (ISPs) is keeping customers happy and sustaining high rates of growth. At a fundamental level, meeting these challenges requires that an ISP provision a number of circuits of various bandwidths over a geographic area. In other words, the ISP must deploy a physical topology that meets the needs of the customers connected to its network.

After the network is deployed, the ISP must map customer traffic flows onto the physical topology. In the early 1990s, mapping traffic flows onto a physical topology was not approached in a particularly scientific way. Instead, the mapping occurred as a by-product of routing configuration—traffic flows simply followed the shortest path calculated by the ISP’s Interior Gateway Protocol (IGP). The limitations of this haphazard mapping were often resolved by overprovisioning bandwidth as individual links began to experience congestion. Today as ISP networks grow larger, as the circuits supporting IP grow faster, and as the demands of customers become greater, the mapping of traffic flows onto physical topologies needs to be approached in a fundamentally different way so that the offered load can be supported in a controlled and efficient manner.

Traffic Engineering

The task of mapping traffic flows onto an existing physical topology is called traffic engineering and is currently a topic of intense discussion within the Internet Engineering Task Force (IETF) and large ISPs. If a traffic engineering “application” implements the right set of features, it should provide ISPs precise control over the placement of traffic flows within its routed domain. Specifically, traffic engineering provides the ability to move traffic flows away from the shortest path selected by the IGP and onto a potentially less congested physical path across the service provider’s network (Figure 1).

![Figure 1: Traffic Engineering Path vs. IGP Shortest Path across a Service Provider’s Network](image-url)
Traffic engineering is a powerful tool that can be used by ISPs to balance the traffic load on the various links, routers, and switches in the network so that none of these components is overutilized or underutilized. In this way, ISPs can exploit the economies of the bandwidth that has been provisioned across the entire network. Traffic engineering should be viewed as assistance to the routing infrastructure that provides additional information used in forwarding traffic along alternate paths across the network.

**Applications for Traffic Engineering**

Traffic engineering has become an important issue within the ISP community because of the unprecedented growth in customer demand for network resources, the mission-critical nature of IP applications, and the increasingly competitive nature of the Internet marketplace. Existing IGPs can actually contribute to network congestion because they do not take bandwidth availability and traffic characteristics into account when building their forwarding tables. ISPs understand that traffic engineering can be leveraged to significantly enhance the operation and performance of their networks. They intend to use traffic engineering capabilities to:

- Route primary paths around known bottlenecks or points of congestion in the network.
- Provide precise control over how traffic is rerouted when the primary path is faced with single or multiple failures.
- Provide more efficient use of available aggregate bandwidth and long-haul fiber by ensuring that subsets of the network do not become overutilized while other subsets of the network along potential alternate paths do not become underutilized.
- Make an ISP more competitive within its market by maximizing operational efficiency, resulting in lower operational costs.
- Enhance the traffic-oriented performance characteristics of the network by minimizing packet loss, minimizing prolonged periods of congestion, and maximizing throughput.
- Enhance statistically bounded performance characteristics of the network (such as loss ratio, delay variation, and transfer delay) that will be required to support the forthcoming multiservices Internet.
- Provide more options, lower costs, and better service to their customers.

**Looking Ahead**

Traffic engineering has become an extremely important tool for ISPs as they struggle to keep pace with the ever-increasing volume of Internet traffic. To enhance the reader’s understanding of traffic engineering and its critical role in supporting future Internet growth, this white paper begins with a description of how traffic engineering has traditionally been performed in router-based cores. The paper then moves to a discussion of the techniques, benefits, and limitations of traffic engineering as it is performed in today’s ATM and Frame Relay “overlay” networks.

After describing today’s popularly deployed solutions, the white paper introduces a new traffic engineering approach that has been specifically designed to operate in the core of an optical Internet—an environment where Dense Wave-Division Multiplexing (DWDM), OC-48 and OC-192 interfaces, IP-over-SONET, IP-over-glass, and Internet backbone routers make up the basic infrastructure of the core. The final section describes the approach of Juniper
Networks and the IETF to designing a traffic engineering solution based on the emerging Multiprotocol Label Switching (MPLS) and Resource Reservation Protocol (RSVP) technologies.

The Past: Traditional Router Cores

In the early 1990s, ISP networks were composed of routers interconnected by leased lines—T1 (1.5-Mbps) and T3 (45-Mbps) links. As the Internet started its growth spurt, the demand for bandwidth increased faster than the speed of individual network links. ISPs responded to this challenge by simply provisioning more links to provide additional bandwidth. At this point, traffic engineering became increasingly important to ISPs so that they could efficiently use aggregate network bandwidth when multiple parallel or alternate paths were available.

Traffic Engineering Capabilities

In router-based cores, traffic engineering was achieved by simply manipulating routing metrics. Metric-based control was adequate because Internet backbones were much smaller in terms of the number of routers, number of links, and amount of traffic. Also, in the days before the tremendous popularity of the any-to-any WWW, the Internet’s topological hierarchy forced traffic to flow across more deterministic paths and events on the network (for example, John Glenn and the Starr Report) did not create temporary hot spots.

Figure 2 illustrates how metric-based traffic engineering operates. Assume that Network A sends a large amount of traffic to Network C and Network D. With the metrics in Figure 2, Links 1 and 2 might become congested because both the Network A-to-Network C and the Network A-to-Network D flows go over those links. If the metric on Link 4 were change to 2, the Network A-to-Network D flow would be moved to Link 4, but the Network A-to-Network C flow would stay on Links 1 and 2. The result is that the “hot spot” is fixed without breaking anything else on the network.

Figure 2: Metric-Based Traffic Control
Metric-based traffic controls provided an adequate traffic engineering solution until around 1994 or 1995. At this point, some ISPs began to reach a size at which they did not feel comfortable moving forward with either metric-based traffic controls or router-based cores. It became increasingly difficult to ensure that a metric adjustment in one part of a huge network did not create a new problem in another part of the network. As for router-based cores, they did not offer the high-speed interfaces or deterministic performance that ISPs required as they planned to grow their core networks.

**Limitations of a Traditional Routed Core for Traffic Engineering**

Traditional routed cores had a number of limitations in terms of providing scalable support for traffic engineering:

- Traditional software-based routers had the potential to become traffic bottlenecks under heavy load because their aggregate bandwidth and packet-processing capabilities were limited.

- Traffic engineering based on metric manipulation was not scalable. As ISP networks became more richly connected (that is, bigger, more thickly meshed, and more redundant), it was more difficult to ensure that a metric adjustment in one part of the network did not cause problems in another part of the network. Traffic engineering based on metric manipulation offered a trial-and-error approach rather than a scientific solution to an increasingly complex problem.

- IGP route calculation was topology driven and based on a simple additive metric such as the hop count or an administrative value. IGPs did not distribute information such as bandwidth availability or traffic characteristics. This meant that the traffic load on the network was not taken into account when the IGP calculated its forwarding table. As a result, traffic was not evenly distributed across the network’s links, causing inefficient use of expensive resources. Some of the links could become congested, while other links remained underutilized. This might have been satisfactory in a sparsely connected network, but in a richly connected network it was necessary to control the paths that traffic took in order to load the links relatively equally.
The Present: IP Overlay Networks

Around 1994 or 1995, the volume of Internet traffic reached a point that ISPs were required to migrate their networks to support trunks that were larger than T3 (45 Mbps). Fortunately, at this time OC-3 ATM interfaces (155 Mbps) became available for switches and routers. To obtain the required speed, ISPs were forced to redesign their networks so that they could make use of the higher speeds supported by a switched (ATM or Frame Relay) core. Some ISPs transitioned from a network of DS-3 point-to-point links to routers with OC-3 ATM SAR interfaces at the edge and OC-3 ATM switches in the core. Then, after a period of 6 to 9 months, the links between core ATM switches were upgraded to OC-12 (622 Mbps). Other ISPs began by increasing the mesh of their DS-3 Frame Relay networks. When they began the transition from Frame Relay to ATM, they relied on OC-3 at the edge but immediately deployed OC-12 interswitch links in the core (see Figure 3).

Figure 3: Typical Physical Topology for the Core of a Large ISP Network in 1997 and 1998

![Typical Physical Topology for the Core of a Large ISP Network](image)

Operation of an IP Overlay Network

When IP runs over an ATM network, routers surround the edge of the ATM cloud. Each router communicates with every other router by a set of Permanent Virtual Circuits (PVCs) that are configured across the ATM physical topology. The PVCs function as logical circuits, providing connectivity between edge routers. The routers do not have direct access to information describing the physical topology of the underlying ATM infrastructure supporting the PVCs. The routers have knowledge only of the individual PVCs that appear to them as simple point-to-point circuits between two routers. Figure 4 illustrates how the physical topology of an ATM core differs from the logical IP overlay topology.
For large ISPs, the ATM core is completely owned and operated by the ISP and is dedicated to supporting Internet backbone service. This core infrastructure is entirely separate from the carrier’s other private data services. Because the network is fully owned by the ISP and dedicated to IP service, all traffic flows across the ATM core utilizing the unspecified bit rate (UBR) ATM class of service—there is no policing, no traffic shaping, no peak cell rate, and no sustained cell rate. ISPs simply use the ATM switched infrastructure as a high-speed transport without relying on ATM’s traffic and congestion control mechanisms. There is little reason for them to use these advanced “features” because each ISP owns its own backbone and they do not need to police themselves.

The physical paths for the PVC overlay are typically calculated by an offline configuration utility on an as-needed basis—when congestion occurs, a new trunk is added, or a new POP is deployed. Some ATM switch vendors offer proprietary techniques for routing PVCs online while taking some traffic engineering concerns into account. However, these solutions are immature, and an ISP frequently has to resort to full offline path calculation to achieve the desired results. The PVC paths and attributes are globally optimized by the configuration utility based on link capacity and historical traffic patterns. The offline configuration utility can also calculate a set of secondary PVCs that is ready to respond to failure conditions. Finally, after the globally optimized PVC mesh has been calculated, the supporting configurations are downloaded to the routers and ATM switches to implement the single or double full-mesh logical topology (see Figure 5).
The distinct ATM and IP networks “meet” when the ATM PVCs are mapped to router subinterfaces. Subinterfaces on a router are associated with ATM PVCs, and then the routing protocol works to associate IP prefixes (routes) with the subinterfaces. In practice, the offline configuration utility generates both router and switch configurations, making sure that the PVC numbers are consistent and the proper mappings occur.

Finally, ATM PVCs are integrated into the IP network by running the IGP across each of the PVCs to establish peer relationships and exchange routing information. Between any two routers, the IGP metric for the primary PVC is set such that it is more preferred than the backup PVC. This guarantees that the backup PVC is used only when the primary PVC is not available. Also, if the primary PVC becomes available after an outage, traffic is automatically returned to the primary PVC from the backup.

Benefits of the IP-over-ATM Model in Service Provider Networks

In the mid-1990s, ATM switches offered a solution when ISPs required more bandwidth to handle ever-increasing traffic loads. The ISPs who decided to migrate to ATM-based cores continued to experience growth and in the process discovered that ATM PVCs provided a tool that offered precise control over traffic as it flowed across their networks. ISPs have come to rely on the high-speed interfaces, deterministic performance, and PVC functionality that ATM switches provide to successfully managing the operation of their networks.

When compared to traditional software-based routers, ATM switches provided higher-speed interfaces and significantly greater aggregate bandwidth, thereby eliminating the potential for router bottlenecks in the core of the network. Together, the speed and bandwidth provided deterministic performance for ISPs at a time when router performance was unpredictable.

An ATM-based core fully supported traffic engineering because it could explicitly route PVCs. Routing PVCs is done by provisioning an arbitrary virtual topology on top of the network’s physical topology in which PVCs are routed to precisely distribute traffic across all links so that they are evenly utilized. This approach eliminates the traffic magnet effect of least-cost routing, which results in overutilized and underutilized links. The traffic engineering capabilities supported by ATM PVCs made the ISPs more competitive within their market, permitting them to provide lower costs and better service to their customers.
Per-PVC statistics provided by the ATM switches facilitate the monitoring of traffic patterns for optimal PVC placement and management. Network designers initially provision each PVC to support specific traffic engineering objectives, and then they constantly monitor the traffic load on each PVC. If a given PVC begins to experience congestion, the ISP has the information it needs to remedy the situation by modifying either the virtual or physical topology to accommodate shifting traffic loads.

**Limitations of the IP-over-ATM Model in the OC-48 Optical Internet**

Over the past few years, ATM switches have empowered ISPs, allowing them to expand market share and increase their profitability. In the mid-1990s, ATM switches were selected for their unparalleled ability to provide high-speed interfaces, deterministic performance, and traffic engineering through the use of explicitly routed PVCs. Today, however, the once unique features of ATM switches are also supported by Internet backbone routers. The latest advances in routing technology are causing ISPs to re-evaluate their willingness to continue tolerating the limitations of the overlay model—the administrative expense, equipment expense, operational stability, and scale.

One of the fundamental limitations of an ATM-based core is that it requires the management of two different networks, an ATM infrastructure and a logical IP overlay. By running an IP network over an ATM network, an ISP not only increases the complexity of its network, but also doubles its overhead because it must manage and coordinate the operation of two separate networks. Also, routing and traffic engineering occur on different sets of systems—routing executes on the routers and traffic engineering runs on the ATM switches—so it is very difficult to fully integrate traffic engineering with routing. Recent technological advances allow Internet backbone routers to provide the high-speed links and deterministic performance formerly found only in ATM switches. When considering a future migration to OC-48 speeds, ISPs must determine whether it is in their best interest to continue with a costly and complex design when the same functionality can now be achieved with a single set of equipment in a fully integrated router-based core.

ATM router interfaces have not kept pace with the latest increases in optical bandwidth. The fastest commercially available ATM SAR router interface is OC-12. OC-48 packet-over-SONET (POS) router interfaces are available today, but OC-48 ATM router interfaces are not likely to be available in the near future. Soon, OC-192 POS interfaces (~10 Gbps) will be available for routers, but OC-192 ATM router interfaces might never be commercially available because of the expense and complexity of implementing the SAR function at these high speeds. The limits in SAR scaling mean that ISPs attempting to increase the speed of their networks using the IP-over-ATM model will have to purchase large ATM switches and routers with a large number of slower ATM interfaces. This will dramatically increase the expense and complexity of growing the network, which will only compound as ISPs consider a future migration to OC-192.

A cell tax is introduced when packet-oriented protocols, such as IP, are carried over an ATM infrastructure. Assuming a 20% overhead for ATM when accounting for framing and realistic distribution of packets sizes, on a 2.488-Gbps OC-48 link, 1.99 Gbps is available for customer data and 498 Mbps—almost a full OC-12—is required for the ATM overhead. When 10-Gbps OC-192 interfaces become available, 1.99-Gbps—almost a full OC-48—of the link’s capacity will be consumed by ATM overhead. When faced with the challenge of migrating their networks to OC-48 and OC-192 speeds, ISPs must determine whether continuing to pay the ATM cell tax puts them at a competitive disadvantage when router-based cores offer the alternative of using the wasted overhead capacity for customer traffic.
A network that deploys a full mesh of ATM PVCs exhibits the traditional “n-squared” problem. For relatively small or moderately sized networks, the “n-squared” problem is not a major issue. But for core ISPs with hundreds of attached routers, the challenge can be quite significant. For example, when growing a small network from five to six routers, an ISP is required to increase the number of simplex PVCs from 20 to 30. However, increasing the number of attached routers from 200 to 201 requires the addition of 400 new simplex PVCs—an increase from 39,800 to 40,200 PVCs. It should be emphasized that these numbers do not include backup PVCs or additional PVCs for networks running multiple services that require more than one PVC between any two routers. A number of operational challenges are caused by the “n-squared” problem:

- New PVCs need to be mapped over the physical topology.
- The new PVCs must be tuned so that they have minimal impact on existing PVCs.
- The large number of PVCs might exceed the configuration and implementation capabilities of the ATM switches.
- The configuration of each switch and router in the core must be modified.

In the mid 1990s, a full mesh of PVCs was required to eliminate interior router hops because of their slow speed interfaces and lack of deterministic performance. Today, Internet backbone routers have overcome these historic limitations, leaving the “n-squared” problem a legacy of the IP-over-ATM architecture.

Deploying a full mesh of PVCs also stresses the IGP. This stress results from the number of peer relationships that must be maintained, the challenge of processing “n-cubed” link-state updates in the event of a failure, and the complexity of performing the Dijkstra calculation over a topology containing a significant number of logical links. Anytime the topology results in a full mesh, the impact on the IGP is a suboptimal topology that is extremely difficult to maintain. And, as an ATM core expands, the “n-squared” stress on the IGP compounds. In a router-based core, the “n-squared” stress on the IGP is eliminated. As with the cell tax, IGP stress is a heritage of the IP-over-ATM model.

Traffic engineering based on the overlay model requires the presence of a Layer 2 technology that supports switching and PVCs. On a mixed-media network, the dependency on a specific Layer 2 technology (that is, ATM) to support traffic engineering can severely constrain possible solutions. If an ISP wants to perform traffic engineering across a POS or photonic network, the Layer 2 transport cannot perform traffic engineering because it has been eliminated. The growth of mixed-media networks and the goal of reducing the number of layers between IP and the glass require that traffic engineering capabilities be supported at Layer 3 to provide an integrated approach. As ISPs continue to build out their networks based on the optical internetworking model, the limitations of the IP-over-ATM architecture become significantly more restrictive.

In summary, the fundamental assumptions supporting the original deployment of ATM-based cores are no longer valid. There are numerous disadvantages for continuing to follow the IP-over-ATM model when other alternatives are available. High-speed interfaces, deterministic performance, and traffic engineering using PVCs no longer distinguish ATM switches from Internet backbone routers. Furthermore, the deployment of a router-based core solves a number of inherent problems with the ATM model—the complexity and expense of coordinating two sets of equipment, the bandwidth limitations of ATM SAR interfaces, the cell tax, the “n-squared” PVC problem, the IGP stress, the limitation of not being able to operate over a mixed-media infrastructure, and the disadvantages of not being able to seamlessly integrate Layer 2 and Layer 3. The topics of high-speed interfaces and deterministic router
performance have been discussed in the Juniper Networks white paper *Internet Backbone Routers and Evolving Internet Design*. The remainder of this paper focuses on how traffic engineering can best be performed over a router-based core.

**The Future: The Juniper Networks Architecture for Traffic Engineering**

ISPs planning to migrate to higher speeds should carefully examine the possible alternatives so that their past bandwidth and traffic engineering decisions do not constrain the future growth and operation of their networks. For high-performance backbones operating at OC-48 speeds, the issues are changing so rapidly that it is not always possible to continue with the same strategy and just make minor (or major) enhancements in hopes of tuning the solution. Eventually all technologies reach a point in their life cycles where they can no longer scale and discerning network architects realize when it is time to step back, start over, and examine new solutions.

It is clear that any router-based traffic engineering approach must provide a level of functionality equivalent to the current IP-over-ATM model. ISPs have learned to rely on ATM’s high-speed interfaces, deterministic performance, and PVC capabilities for traffic engineering and they will settle for nothing less. The router-based approach must successfully overcome the limitations of the existing overlay model while providing an integrated solution that eliminates the additional complexity and expense of coordinating and managing two different networks. Finally, the proposed solution must offer the option of automating the traffic engineering process so that ISPs can provide enhanced customer service and reliability while reducing operational costs.

To meet ISP requirements in a reasonable amount of time, any future approach should leverage existing work performed by IETF working groups. If components of the solution already exist, they should be enthusiastically embraced rather than ignored while engineers stubbornly attempt to “reinvent the wheel.” New technologies that need to be developed should be relatively simple and unambiguous so that they can be implemented and deployed with minimal risk. Ultimately, the inventiveness of the new traffic engineering solution should come from combining relatively simple and easily deployable technologies to create a robust solution capable of scaling with the growth of the optical Internet.

**Components of the Traffic Engineering Solution**

In pursuing router-based traffic engineering solutions, Juniper Networks has been actively involved in the Multiprotocol Label Switching (MPLS) and related working groups of the IETF. Our strategy for traffic engineering using MPLS involves four functional components:

- Packet-forwarding
- Information distribution
- Path selection
- Signaling

Each functional component is an individual module. The Juniper Networks traffic engineering architecture provides clean interfaces between each of the four functional components. This combination of modularity and clean interfaces provides the flexibility of allowing individual components to be changed as needed if a better solution becomes available.
Packet-Forwarding Component

The packet-forwarding component of the Juniper Networks traffic engineering architecture is Multiprotocol Label Switching (MPLS). MPLS is responsible for directing a flow of IP packets along a predetermined path across a network. This path is called a label-switched path (LSP). LSPs are similar to ATM PVCs in that they are simplex in nature; that is, the traffic flows in one direction from the head-end router to a tail-end router. Duplex traffic requires two LSPs; that is, one LSP to carry traffic in each direction. An LSP is created by the concatenation of one or more label-switched hops, allowing a packet to be forwarded from one label-switching router (LSR) to another LSR across the MPLS domain (see Figure 6). An LSR is a router that supports MPLS-based forwarding.

Figure 6: LSP across an MPLS Domain

When a head-end LSR receives an IP packet, it adds an MPLS header to the packet and forwards it to the next LSR in the LSP. The labeled packet is forwarded along the LSP by each LSR until it reaches the tail end of the LSP, at which point the MPLS header is removed and the packet is forwarded based on Layer 3 information such as the IP destination address. The key point in this scheme is that the physical path of the LSP is not limited to what the IGP would choose as the shortest path to reach the destination IP address.

LSR Packet Forwarding Based on Label Swapping

The packet-forwarding process at each LSR is based on the concept of label swapping. This concept is similar to what occurs at each ATM switch in a PVC. Each MPLS packet carries a 4-byte encapsulation header that contains a 20-bit fixed-length label field. When a packet containing a label arrives at an LSR, the LSR examines the label and uses it as an index into its MPLS forwarding table. Each entry in the forwarding table contains an interface–inbound label pair that is mapped to a set of forwarding information that is applied to all packets arriving on the specific interface with the same inbound label (Figure 7).
Figure 8 illustrates the operation of the label-swapping algorithm used by an LSR. A packet is received on interface 3 of the LSR, containing a label of 21. The LSR, using the information from the forwarding table shown in Figure 7, replaces the label with a value of 18 and forwards the packet out interface 4 to the next hop LSR.

**Figure 8: LSR Forwards a Packet**

![Diagram of LSR Forwards a Packet](image)

**Example of a How Packet Traverses an MPLS Backbone**

This section describes how an IP packet is processed as it traverses an MPLS backbone network. We will examine the operations performed on the packet at three distinct points in the network: as the packet arrives at the ingress edge of the MPLS backbone, as the packet is forwarded by each LSR along the LSP, and as the packet reaches the egress edge of the MPLS backbone (see Figure 9).

**Figure 9: MPLS Backbone Network**

![Diagram of MPLS Backbone Network](image)

At the ingress edge of the MPLS backbone, the IP header is examined by the head-end LSR. Based on this analysis, the packet is classified, assigned a label, encapsulated in an MPLS header, and forwarded toward the next hop in the LSP. MPLS provides a tremendous amount of flexibility in the way that an IP packet can be assigned to an LSP. For example, in the Juniper Networks traffic engineering implementation, all packets arriving at the head-end LSR that are destined to exit the MPLS domain at the same egress LSR are forwarded along the same LSP.

Once the packet begins to traverse the LSP, each LSR uses the label to make the forwarding decision. Remember that the MPLS forwarding decision is made without consulting the original IP header. Rather, the incoming interface and label are used as lookup keys into the MPLS forwarding table. The old label is replaced with a new label, and the packet is forwarded to the next hop along the LSP. This process is repeated at each LSR in the LSP until the packet reaches the tail-end LSR.

When the packet arrives at the egress LSR, the label is removed and the packet exits the MPLS domain. The packet is then forwarded based on the destination IP address contained in the packet’s original IP header according to the traditional shortest path calculated by the IP routing protocol.

In this section we have not described how labels are assigned and distributed to the LSRs along the LSP. We will discuss this important task when describing the signaling component in the Juniper Networks traffic engineering architecture.
Benefits of MPLS

It is commonly believed that MPLS in and of itself significantly enhances the forwarding performance of LSRs. It is more accurate to say that exact-match lookups, such as those performed by MPLS and ATM switches, have historically been faster than the longest match lookups performed by IP routers. However, recent advances in silicon technology allow ASIC-based route lookup engines to run just as fast as MPLS or ATM VPI/VCI lookup engines.

The real benefit of MPLS technology is that it provides a clean separation between routing (that is, control) and forwarding (that is, moving data). This separation allows the deployment of a single forwarding algorithm—MPLS—that can be used for multiple services and traffic types. In the future, as ISPs see the need to develop new revenue-generating services, the MPLS forwarding infrastructure can remain the same while new services are built by simply changing the way that packets are assigned to an LSP. For example, packets could be assigned to an LSP based on a combination of the destination subnetwork and application type, a combination of the source and destination subnetworks, a specific QoS requirement, an IP multicast group, or a Virtual Private Network (VPN) identifier. In this manner, new services can easily be migrated to operate over the common MPLS forwarding infrastructure.

Information Distribution Component

Because traffic engineering requires detailed knowledge about the network topology as well as dynamic information about network loading, a primary requirement for the new traffic engineering model is a framework for information distribution. This component can easily be implemented by defining relatively simple extensions to the IGP so that link attributes are included as part of each router’s link-state advertisement. IS-IS extensions can be supported by the definition of new Type Length Values (TLVs), while OSPF extensions can be implemented with Opaque LSAs. The standard flooding algorithm used by the link-state IGP ensures that link attributes are distributed to all routers in the ISP’s routing domain.

Each LSR maintains network link attributes and topology information in a specialized Traffic Engineering Database (TED). The TED is used exclusively for calculating explicit paths for the placement of LSPs across the physical topology. A separate database is maintained so that the subsequent traffic engineering computation is independent of the IGP and the IGP’s link-state database. Meanwhile, the IGP continues its operation without modification, performing the traditional shortest-path calculation based on information contained in the router’s link-state database (See Figure 10).

Figure 10: Information Distribution Component
Some of the traffic engineering extensions to be added to the IGP link-state advertisement include:

- Maximum link bandwidth
- Maximum reservable link bandwidth
- Current bandwidth reservation
- Current bandwidth usage
- Link coloring

Path Selection Component

After network link attributes and topology information are flooded by the IGP and placed in the TED, each head-end LSR uses the TED to calculate the paths for its own set of LSPs across the routing domain (See Figure 11). The path for each LSP can be represented by either a strict or loose explicit route. An explicit route is a preconfigured sequence of LSRs that should be part of the physical path of the LSP. If the ingress LSR specifies all the LSRs in the LSP, the LSP is said to be identified by a strict explicit route. However, if the head-end LSR specifies only some of the LSRs in the LSP, the LSP is described by a loose explicit route. Support for strict and loose explicit routes allows the path selection process to be given broad latitude whenever possible, but to be constrained when necessary.

Figure 11: Head-End LSR Calculates Explicit Routes

The head-end LSR determines the physical path for each LSP by applying a Constrained Shortest Path First (CSPF) algorithm to the information in the TED (Figure 12). CSPF is a shortest-path-first algorithm that has been modified to take into account specific restrictions when calculating the shortest path across the network. Input into the CSPF algorithm includes:

- Topology link-state information learned from the IGP and maintained in the TED
- Attributes associated with the state of network resources (such as total link bandwidth, reserved link bandwidth, available link bandwidth, and link color) that are carried by IGP extensions and stored in the TED
- Administrative attributes required to support traffic traversing the proposed LSP (such as bandwidth requirements, maximum hop count, and administrative policy requirements) that are obtained from user configuration
As CSPF considers each candidate node and link for a new LSP, it either accepts or rejects a specific path component based on resource availability or whether selecting the component violates user policy constraints. The output of the CSPF calculation is an explicit route consisting of a sequence of LSR addresses that provides the shortest path through the network that meets the constraints. This explicit route is then passed to the signaling component, which establishes forwarding state in the LSRs along the LSP. The CSPF algorithm is repeated for each LSP that the head-end LSR is required to generate.

Despite the reduced management effort resulting from online path calculation, an offline planning and analysis tool is still required to optimize traffic engineering globally. Online calculation takes resource constraints into account and calculates one LSP at a time. The challenge with this approach is that it is not deterministic. The order in which an LSP is calculated plays a critical role in determining its physical path across the network. LSPs that are calculated early in the process have more resources available to them than LSPs calculated later in the process because previously calculated LSPs consume network resources. If the order in which the LSPs are calculated is changed, the resulting set of physical paths for the LSPs also can change.

An offline planning and analysis tool simultaneously examines each link’s resource constraints and the requirements of each ingress-to-egress LSP. While the offline approach can take several hours to complete, it performs global calculations, compares the results of each calculation, and then selects the best solution for the network as a whole. The output of the offline calculation is a set of LSPs that optimizes utilization of network resources. After the offline calculation is completed, the LSPs can be established in any order because each is installed following the rules for the globally optimized solution.
Signaling Component

Because the information residing in the head-end LSR’s TED about the state of the network at any given time is out-of-date, CSPF computes a path that is thought to be acceptable. However, the path is not known to be workable until the LSP is actually established by the signaling component (see Figure 13). The signaling component, which is responsible for establishing LSP state and label distribution, relies on a number of extensions to the Resource Reservation Protocol (RSVP):

- The Explicit Route Object allows an RSVP PATH message to traverse an explicit sequence of LSRs that is independent of conventional shortest-path IP routing. Recall that the explicit route can be either strict or loose.
- The Label Request Object permits the RSVP PATH message to request that intermediate LSRs provide a label binding for the LSP that it is establishing.
- The Label Object allows RSVP to support the distribution of labels without having to change its existing mechanisms. Because the RSVP RESV message follows the reverse path of the RSVP PATH message, the Label Object supports the distribution of labels from downstream nodes to upstream nodes.

RSVP is ideal for use as a signaling protocol to establish LSPs:

- RSVP is the standard resource reservation protocol for the Internet, and it was specifically designed to support enhancements by the addition of new object types.
- RSVP’s soft state can reliably establish and maintain LSPs in an MPLS environment.
- RSVP allows network resources to be explicitly reserved and allocated to a given LSP.
- RSVP allows the establishment of explicitly routed LSPs that provide traffic engineering and load-balancing capabilities equivalent to those currently provided by ATM and Frame Relay.
- Edge-to-edge RSVP signaling across an MPLS domain scales because the number of LSPs is related to the number of edge LSRs in the domain rather than to the number of entries in the routing table or the number of end system traffic flows.
Flexible LSP Calculation and Configuration

Recall that the essence of traffic engineering is mapping traffic flows onto a physical topology. This means that the heart of the issue for performing traffic engineering with MPLS is determining the physical path for each LSP. This path can be determined offline by a configuration utility or online by constraint-based routing. Independent of the way that the physical path is calculated, the forwarding state can be installed across the network using the signaling capabilities of RSVP (see Figure 14).

Figure 14: Offline and Online LSP Calculation and Configuration

The Juniper Networks strategy for traffic engineering over MPLS supports a number of different ways to route and configure an LSP:

- An ISP can calculate the full path for the LSP offline and individually configure each LSR in the LSP with the necessary static forwarding state. This is analogous to how some ISPs currently configure their IP-over-ATM cores.

- ISPs can calculate the full path for the LSP offline and statically configure the head-end LSR with the full path. The head-end LSR then uses RSVP as a dynamic signaling protocol to install forwarding state in each LSR along the LSP.

- ISPs can rely on constraint-based routing to perform dynamic online LSP calculation. In constraint-based routing, the network administrator configures the constraints for each LSP and then the network itself determines the path that best meets those constraints. As discussed earlier, the Juniper Networks strategy for traffic engineering allows the head-end LSR to calculate the entire LSP based on certain constraints and then to initiate signaling across the network.

- ISPs can calculate a partial path for an LSP offline and statically configure the head-end LSR with a subset of the LSRs in the path and then permit online calculation to determine the complete path. For example, imagine an ISP has a topology that includes two east-west paths across the United States: one in the north through Chicago and one in the south through Dallas. Now imagine that the ISP wants to establish an LSP between a router in New York and one in San Francisco. The ISP can configure the partial path for the LSP to include a single loose-routed hop of an LSR in Dallas, and the result is that the LSP is routed along the southern path. The head-end LSR uses CSPF to compute the complete path and uses RSVP to install the forwarding state along the LSP just as above.
ISPs can configure the head-end LSR with no constraints whatsoever. In this case, normal IGP shortest-path routing is used to determine the path of the LSP. This configuration does not provide any value in terms of traffic engineering. However, the configuration is easy and it might be useful in situations when services such as Virtual Private Networks (VPNs) are needed.

In all these cases, any number of LSPs can be specified as backups for the primary LSP. To establish backup LSPs in the event of a failure, two or more approaches can be combined. For example, the primary path might be explicitly computed offline, the secondary path might be constraint based, and the tertiary path might be unconstrained. If a circuit on which the primary LSP is routed fails, the head-end LSR notices the outage from error notifications received from downstream LSRs or by RSVP soft-state expiration. The head-end LSR then can dynamically forward traffic to a hot-standby LSP or call on RSVP to create forwarding state for a new backup LSP.

Operational Requirements for a Successful Traffic Engineering Solution

To provide a powerful and user-friendly tool, the Juniper Networks traffic engineering architecture is required to support a wide range of customer requirements. The implementation needs to allow network operators to:

- Perform a number of specific operations on LSPs that are critical to the traffic engineering process:
  - Establish an LSP.
  - Activate an LSP, causing it to start forwarding traffic.
  - Deactivate an LSP, causing it to stop forwarding traffic.
  - Modify the attributes of an LSP (such as bandwidth, hop limit, and CoS) to manage its behavioral characteristics.
  - Reroute an LSP, causing it to change its physical path through the network.
  - Tear down an LSP, causing the network to reclaim all resources allocated to the LSP.
- Configure a loose or strict explicit route for an LSP. Recall that support for this option allows the path selection process to be given broad latitude whenever possible and be constrained when necessary.
- Instantiate an order for alternate physical paths supporting a given LSP. For example, a list of paths can be created in which the first path in the list is considered the primary path. If the primary path cannot be established, then secondary paths are tried in the order listed.
- Permit or disable the reoptimization of an LSP “on the fly.”
- Specify a class of resources that must be explicitly included or excluded from the physical path of an LSP. A resource class can be seen as a “color” assigned to a link such that the set of links with the same color belong to the same class. For example, network policy could state that a given LSP should not traverse gold-colored links.
- Establish LSPs in a priority order so that an LSR attempts to establish higher-priority LSPs before establishing lower-priority LSPs.
- Determine whether one LSP can preempt another LSP from a given physical path based on the LSP’s attributes and priorities. Preemption allows the network to tear down existing LSPs to support an LSP that is being established.
- Automatically obtain a solution to the LSP placement problem by manipulating constraint-based routing parameters.
- Gain access to accounting and traffic statistics at the per-LSP level. These statistics can be used to characterize traffic, optimize performance, and plan capacity.

**Benefits of the Juniper Networks Traffic Engineering Architecture**

The Juniper Networks traffic engineering architecture offers many of the benefits provided by the current IP-over-ATM model:

- High-speed optical interfaces are supported.
- Explicit paths allow network administrators to specify the exact physical path that an LSP takes across the service provider’s network.
- Support for dynamic failover to a precalculated, hot-standby, backup LSP.
- Because of the similarity between LSPs and connection-oriented virtual circuits, fitting LSPs into existing offline network planning and analysis tools is relatively straightforward. The output of these tools can be converted into the configuration necessary to establish the physical paths for LSPs.
- Per-LSP statistics can be used as input to network planning and analysis tools to identify bottlenecks and trunk utilization and plan for future expansion.
- Constraint-based routing provides enhanced capabilities that allow an LSP to meet specific performance requirements before it is established.

In addition to supporting and extending many of the benefits of the overlay model, the Juniper Networks traffic engineering architecture eliminates the scalability problems associated with the current overlay model allowing ISPs to continue to grow their networks to OC-48 speeds and beyond:

- The architecture provides an integrated solution that merges the separate Layer 2 and Layer 3 networks from the overlay model into a single network. This integration eliminates the management burden of coordinating the operation of two distinct networks, permits routing and traffic engineering to occur on the same platform, and reduces the operational cost of the network. In addition, LSPs are derived from the IP state rather than the Layer 2 state so the network can better reflect the needs of IP services. As ISPs continue to grow, the Juniper Networks traffic engineering strategy will eliminate the need to purchase, deploy, manage, and debug two different sets of equipment because the same functionality will be provided on a single integrated network.
- The architecture is not limited to OC-12 links due to the technical challenges of developing OC-48 ATM SAR router interfaces. This means that the absence of high-speed ATM SAR router interfaces will not prevent an ISP from increasing the speed of its network to OC-48 and beyond.
- Because ATM is no longer required as the Layer 2 technology, the cell tax is completely eliminated. This means that the 15-25% of the provisioned bandwidth that was previously consumed carrying ATM overhead is now available to carry additional customer traffic.
- An MPLS routed core does not exhibit ATM’s “n-squared” PVC problem, which stresses the IGP and results in complex configuration issues. Modern Internet backbone routers no longer exhibit the performance issues that caused ISPs to deploy a full-mesh design in the first place.
Conclusion

For a number of years, the core of Internet has experienced exponential growth. Today, the rapidly increasing volume of traffic is forcing some ISPs to double the capacity of their networks every three months. ISPs who have succeeded in increasing their market share in this ever-changing environment are those who had the insight and flexibility to transition their backbones to new technologies capable of supporting expanding customer demand.

In the early 1990s, ISPs depended on metric manipulation to manage the placement of traffic flows across their router-based cores. Metric-based traffic controls provided a satisfactory traffic engineering solution until the mid-1990s when the size and redundancy of core topologies began to limit the scalability of the solution. Also, around 1994 or 1995, ISPs desperately needed to grow their networks, deploy fatter pipes, and obtain deterministic performance from intermediate systems. At this time, OC-3 and OC-12 interfaces became available for core ATM switches and routers to provide the needed bandwidth. The mid-1990s represented a critical juncture in the evolution of the ISP marketplace. ISPs who recognized the limitations of their existing infrastructures and quickly responded by redesigning their networks to an overlay model were able to smoothly expand their market share and increase profitability.

In the late 1990s, ISPs are once again facing a similar crossroad as they plan the migration of their networks to OC-48 speeds and beyond. Continuing to follow the IP-over-ATM model could result in considerable expense and increased complexity. The Juniper Networks traffic engineering architecture provides the traffic management features of an ATM core while eliminating ATM’s performance and scalability limitations. ISPs who examine the constraints of their existing IP-over-ATM solutions and consider the benefits of the emerging MPLS/RSVP alternative will understand that their past traffic engineering decisions could impact the future growth and profitability of their networks.
References

Internet Drafts


Next-Generation Networks 1998: Conference Tutorials and Sessions


