QoS Support in MPLS Networks

MPLS/Frame Relay Alliance White Paper

May 2003

By: Victoria Fineberg, Consultant
fineberg@illinois alumni.org

Abstract

MPLS is sometimes used synonymously with QoS, but more accurately, it is a QoS-*enabling* technology that forces application flows into connection-oriented paths and provides mechanisms for traffic engineering and bandwidth guarantees along these paths. Furthermore, when an MPLS network supports DiffServ, traffic flows can receive class-based admission, differentiated queue servicing in the network nodes, preemption priority, and other network treatment that provide bases for QoS guarantees. The IETF work in this area has been augmented by the MPLS/Frame Relay Alliance Implementation Agreement which extends MPLS to the user-network interface, and thus serves as a foundation for implementing QoS end-to-end. This paper describes various QoS and MPLS mechanisms and analyzes their applicability.

Table of Contents

1. Introduction

Multiprotocol Label Switching (MPLS) is frequently mentioned among major Quality of Service (QoS) technologies for packet networks. While it is certainly true that MPLS plays a vital role in enabling QoS, QoS is not a fundamental feature of MPLS. More accurately, MPLS provides a connection-oriented environment that enables Traffic Engineering (TE) of packet networks. Traffic-engineered networks can guarantee bandwidth for various flows, which is a necessary condition for QoS. In order to control latency and jitter of time-sensitive applications (which is another major QoS requirement), MPLS-TE must be combined with technologies that provide traffic flows with their class-specific treatment, such as Resource reservation Protocol with Tunneling Extensions (RSVP-TE) signaling, and Differentiated Services (DiffServ)-based forwarding. This paper discusses various architecture and implementation aspects of MPLS-enabled packet backbones, as well as QoS features of the MPLS User-to-Network Interface (UNI) defined by the MPLS/Frame Relay Alliance.

Section 1 begins with a review of industry drivers for QoS; it then provides some background on QoS and presents major concepts and definitions; it concludes with a discussion of the necessary conditions for QoS in packet networks. Section 2 first reviews the initial IP-QoS models, Integrated Services (IntServ) and DiffServ, and then describes MPLS and its use for Traffic Engineering, including MPLS-TE and RSVP-TE. With this background, the paper proceeds to the section 3 discussion of MPLS support of DiffServ, i.e., combining two complementary technologies to achieve QoS-enabled networks. Section 4 provides an analysis of practical queue management models based on the MPLS/Frame Relay Alliance work in this area. Section 5 describes another important technology for QoS in MPLS networks, the MPLS UNI (which is also being defined by the MPLS/Frame Relay Alliance), and its potential to create a QoS-enabled interface. Section 6 summarizes the main ideas discussed in the paper.

1.1 QoS Drivers

In the beginning, IP networks existed without any explicit QoS mechanisms. The TCP/IP-based Internet was not planned for providing voice or other services that have stringent requirements for bandwidth, delay and jitter. TCP was defined with FTP, SMTP, TELNET and other types of data communications in mind, where a best effort service was deemed adequate. It featured variable-size sliding windows for flow control, slow-start, multiplicative decrease congestion avoidance which reduces the congestion window by half for every loss, and timer backoff for adjustment of the timeout intervals for receiving acknowledgements. The basic mechanism for dealing with congestion, i.e., offered traffic load being higher than available bandwidth, was packet discarding. In this "QoS environment",, Service Providers (SP) added bandwidth to reduce congestion levels as traffic volume in the Internet grew. One consequence has been that SP Capital Expenditures (CAPEX) have been driven by traffic volumes, which have not necessarily correlated to service revenues, resulting in a difficult business model.

However, lately SP focus has switched towards deploying traffic management and QoS mechanisms. There are several reasons for that. First, with ever-increasing traffic volume in their networks, SPs discovered that it is difficult to alleviate congestion with bandwidth alone. Changes in traffic distribution and node / link failures could result in unpredictable congestion patterns, and significant overprovisioning across an entire network became prohibitively expensive. Second, bottlenecks were frequently happening in the access, where over-provisioning is not economical. Third, the recent economic downturn in the networking industry has caused companies to reduce the new CAPEX that would be required for adding bandwidth, and to focus instead on getting better performance from their existing infrastructure. Fourth, and very important, industry is supplementing the business model of providing just transport with the value-added enhanced-services approach. In the long run, converged networks that offer voice, data and video services are expected to be easier to operate and manage than existing parallel networks, thus significantly impacting SP's Operating Expenditures (OPEX). But in order to offer packet-based audio/video conferencing and multimedia services to business users (the most lucrative market segment) converged networks must offer flawless quality and strict support for Service Level Agreements (SLA).

These considerations have lead to strong SP interest in providing assured QoS in their networks.

1.2 Main Definitions

Before discussing QoS mechanisms and their relation to MPLS, it is useful to review several key definitions of QoS and related concepts. While no single "official" definition of QoS exists, the following definitions are considered in this paper as authoritative.

Yoram Bernet [Ber] has distinguished between active and passive QoS definitions. A *passive* definition describes the service quality experienced by traffic transiting a network, whereas an *active* definition refers to the mechanisms that control the service quality experienced by traffic transiting a network. Bernet's active definition of *Network QoS* is "The capability to control traffic-handling mechanisms in the network such that the network meets the service needs of certain applications and users subject to network policies." We will refer to the relevant traffic-handling mechanisms as "QoS mechanisms."

Jerry Ash provides in [TE-QoS] an extensive set of Traffic Engineering and QoS-related definitions from a SP's point of view. It addresses various TE methods such as traffic management, capacity management and network planning. While capacity management and network planning ensure future performance of the network, *traffic management* refers to optimization of the existing network resources under various conditions, including load shifts and failures. Traffic management encompasses control of routing functions, which include call routing (number or name translation to a routing address), connection routing, routing table management, QoS resource management, and dynamic transport routing.

In [TE-QoS], *QoS* is defined as "a set of service requirements to be met by the network while transporting a connection or flow; the collective effect of service performance which determine the degree of satisfaction of a user of the service." This definition is "passive" based on the distinction outlined in [Ber], but the following definition of the *QoS Resource Management* is active: "network functions which include class-of-service identification, routing table derivation, connection admission, bandwidth allocation, bandwidth protection, bandwidth reservation, priority routing, and priority queuing." In this paper, we will be using "QoS mechanisms" and "QoS resource management functions" interchangeably.

In summary, we will be talking about *QoS* as the service needs of various applications, and about *QoS mechanisms* / *QoS resource management functions* as network control mechanisms that allow a network to satisfy QoS.

The service needs of different applications can be represented as a set of parameters, including bandwidth, delay, jitter, packet loss, preemption and some others. For example, voice and multimedia applications are very sensitive to delay and jitter, whereas some data applications may require very low packet loss. We will refer to these parameters as *QoS variables*.

The final terminology comment is regarding the distinction between QoS and Class of Service (CoS), which is sometimes referred to as *Qualitative* QoS, because it does not provide absolute (quantifiable) performance guarantees. This paper uses the term "QoS" as defined in this section and refers to CoS as a method of assigning traffic flows to separate classes and providing class-based differentiated services. Sections 2.2 and 3.2 discuss CoS features of the DiffServ architecture and the DS-TE model, respectively.

1.3 Necessary Conditions for QoS

With QoS as defined above, let us now consider the fundamental requirements that must be met in order to achieve it. In order to provide QoS for more demanding types of applications (e.g., voice, multimedia), a network must satisfy two necessary conditions.

The first condition is that *bandwidth must be guaranteed* for an application under various circumstances, including congestion and failures.

The second condition is that as an application flow traverses the network, it must receive the appropriate *class-based treatment*, including scheduling and packet discarding. We can think about these two conditions as orthogonal. A flow may get sufficient bandwidth but get delayed on the way (the first condition is met but not the second). Alternatively, a flow may be appropriately serviced in most network nodes but get terminated or severely distorted by occasional lack of bandwidth (the second condition is met but not the first). Therefore, it is necessary to satisfy both of these conditions in order to achieve the hard QoS guarantees that are required by service providers and their customers.

This paper analyzes various IP-QoS approaches for their compliance with these conditions. The two orthogonal concepts come together in the DiffServ-aware MPLS Traffic Engineering (DS-TE) framework that is at present being defined by the IETF Traffic Engineering Working Group (TE-WG) in the Internet drafts [DSTE-REQ], [DSTE-PRO], [DSTE-RUS], and [DSTE-MAM] as described in section 3.2 below.

2. Initial QoS and TE Models

As the internetworking community started realizing the need for QoS mechanisms in packet networks, several approaches emerged. IntServ, together with the signaling protocol RSVP, provided the first genuine QoS architecture. However, upon observing the scalability and operational problems of IntServ with RSVP, the IETF defined the DiffServ architecture, which in its basic form did not require a signaling protocol. Later, MPLS was introduced as a connection-oriented approach to connectionless IP-based networks, and it has enabled Traffic Engineering. This section reviews these earlier architectures and provides the background for the latest scheme for a scalable guaranteed QoS described in the next section.

2.1 IntServ with RSVP

[IntServ] has defined the requirements for QoS mechanisms in order to satisfy two goals: (1) to serve real-time applications and (2) to control bandwidth-sharing among different traffic classes. Two types of service were defined to comply with the IntServ architecture: Guaranteed Service and Controlled Load Service, both focusing on an individual application's requirements.

Guaranteed Service was defined to provide an assured level of bandwidth, a firm end-to-end delay bound, and no queuing loss; and it was intended for real-time applications such as voice and video. The *Controlled Load Service* definition did not include any firm quantitative guarantees but rather "the appearance of a lightly loaded network." It was intended for applications that could tolerate a limited amount of loss and delay, including adaptive real-time applications. By design, Controlled Load Service provided better performance than the Best-Effort treatment, because it would not noticeably deteriorate as the network load increased.

In order to achieve their stated goals and provide the proposed services, the IntServ models included various traffic parameters such as *rate* and *slack term* for Guaranteed Service; and *average rate*, *peak rate* and *burst size* for Controlled Load Service. To install these parameter values in a network and to provide service guarantees for the real-time traffic, the Resource Reservation Protocol [RSVP] was developed as a signaling protocol for reservations and explicit admission control.

The IntServ architecture has satisfied both necessary conditions for the network QoS, i.e., it provided the appropriate bandwidth and queuing resources for each application flow (a "microflow"). However, the IntServ implementations with RSVP required the per-microflow state and signaling at every hop. This added significant complexity to network

operation and was widely considered unscalable. Therefore, the IntServ model was implemented only in a limited number of networks, and the IETF moved to develop DiffServ as an alternative QoS approach with minimal complexity.

2.2 DiffServ

The DiffServ architecture has assumed an opposite approach to that of IntServ. It defined Classes of Service (CoS), called *Aggregates*, and QoS resource management functions with node-based, or *Per-Hop,* operation. The CoS definitions include a Behavior Aggregate (BA) which has specific requirements for scheduling and packet discarding, and an Ordered Aggregate (OA) which performs classification based on scheduling requirements only, and may include several drop precedence values. Thus, an OA is a coarser classification than a BA and may include several BAs. The node behavior definitions correspond to the CoS definitions. A Per Hop Behavior (PHB) is offered to a BA, whereas a PHB Scheduling Class (PSC) serves an OA; PHB mechanisms include scheduling and packet discarding, whereas PSC only concerns scheduling.

The DiffServ model is based on redefining the meaning of the 8-bit ToS field in the IP header. The original ToS definition was not widely implemented, and now the field is split into the 6-bit DiffServ Code Point (DSCP) value and the 2-bit Explicit Congestion Notification (ECN) part, as shown in Figure 1 below.

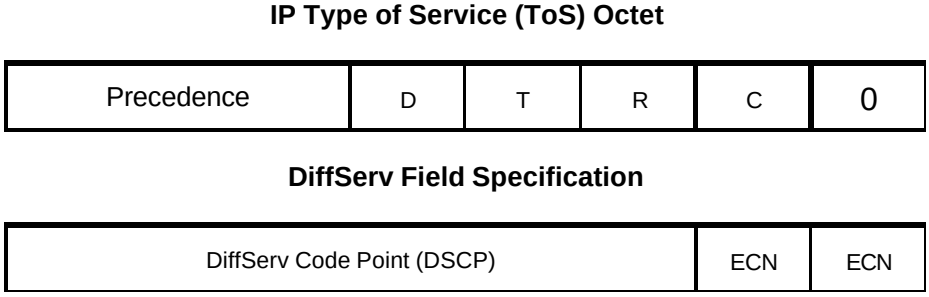

Figure 1. Relationship between ToS and DiffServ / ECN

In Figure 1, the letters indicate the following: D = Delay, T = Throughput, R = Reliability, C = Cost, ECN = Explicit Congestion Notification.

The value of the DSCP field is used to specify a BA (i.e., a class), which is used by DiffServ-compliant nodes for choosing the appropriate PHB (i.e., a queue servicing treatment). Fourteen PHBs have been defined, including one for Expedited Forwarding (EF), twelve for Assured Forwarding (AF), and one for Default, or Best Effort, PHB. The twelve AF PHBs are divided into four PSCs, and each of the AF PSCs consists of three sub-behaviors related to different packet discarding treatment.

In summary, the DiffServ model allows the network to classify (combine) microflows into flow aggregates (BAs) and then to offer to these aggregates differentiated treatment in each DiffServ-capable node. This treatment is reflected in the queue servicing

mechanisms which include scheduling and packet discarding. PHB is reflected in both scheduling and discarding, whereas PSC applies only to scheduling.

In the introductory section, we mentioned the two necessary conditions for QoS: guaranteed bandwidth, and class-related scheduling and packet discarding treatment. The DiffServ architecture satisfies the second condition, but not the first.

2.3 MPLS

2.3.1 MPLS Terminology

We are assuming that a reader is already familiar with the basic operation of Multiprotocol Label Switching (MPLS) or can refer to [MPLS-ARCH] and [MPLS-WP]. In this section, we briefly mention some MPLS terminology that we use elsewhere in the paper, and then we describe MPLS-TE and RSVP-TE.

We use the term Label Edge Router (LER) to designate an edge Label Switching Router (LSR), because this allows us to make a further distinction between the Ingress LER (I-LER) and the Egress LER (E-LER). Note that some documents refer to these nodes as Head-End and Tail-End, respectively.

To tag traffic flows and direct them into connection-oriented Label Switched Paths (LSPs), MPLS uses labels which are fields in MPLS "shim" headers as illustrated in Figure 2 below.

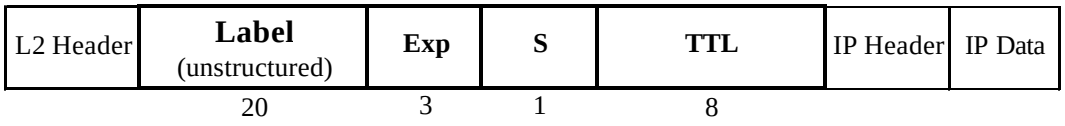

Figure 2. MPLS "shim" header

MPLS labels are assigned based on the traffic flow's Forwarding Equivalency Class (FEC). FECs are destination-based flexible packet groupings. For example, they may be formed based on the MPLS domain E-LERs, or customer access routers, or even on the individual flow destinations. This flexibility in forming FECs is one of the important benefits MPLS brings to routing.

Specification of LSPs is made by using extended IP routing protocols, and distribution of labels along these paths is accomplished by label distribution protocols. Label distribution can be accompanied by bandwidth reservations for specific LSPs. Note that the term "LDP" can be used in two ways, as a general term to indicate a label distribution protocol and as a specific protocol called LDP, defined in RFC3036 [LDP].

2.3.2 MPLS-TE

The label switching approach was initially conceived in order to improve router performance, but this motivation has diminished with advances in router design and achievement of line-speed forwarding of native IP packets. But later the most important advantage of the MPLS architecture over the native IP forwarding has become apparent: the connection-oriented nature of MPLS allows SPs to implement TE in their networks and achieve a variety of goals, including bandwidth assurance, diverse routing, load balancing, path redundancy, and other services that lead to QoS.

[TE-REQ] describes issues and requirements for Traffic Engineering implementation in MPLS networks. It provides a general definition of TE as a set of mechanisms for performance optimization of operational networks in order to achieve specific performance objectives and describes how MPLS supports TE by enabling control and measurement mechanisms.

[TE-REQ] uses the concept of an MPLS Traffic Trunk (TT) which is an aggregation of traffic flows of the same class that are placed inside an LSP. The principal distinction between a TT and an LSP is that a TT is an aggregated traffic flow, whereas an LSP is a *path* a TT takes through a network. For example, during a recovery process, a TT may be using a different LSP. [TE-REQ] describes a framework for mapping TTs onto LSPs by addressing three sets of capabilities:

1. TT attributes
2. resource attributes that constrain placement of TTs, and
3. a constraint-based routing (CR) approach that allows the selection of LSPs for TTs.

TT attributes of particular interest are traffic parameters, priority, and preemption. *Traffic parameter attributes* may include values of peak rates, average rates, burst sizes and other resource requirements of a traffic trunk that can be used for resource allocation and congestion avoidance. The *priority attribute* allows the CR process to establish an order in which path selection is done so that higher priority TTs will have an earlier opportunity to claim network resources than lower priority TTs. The *preemption attribute* determines whether a TT can or cannot preempt *and* can or cannot be preempted by another TT.

Resource attributes are topology state parameters such as *Maximum Allocation Multiplier* (MAM) which allows a network operator to allocate more or less resources than the link capacity in order to achieve the goals of overbooking or overprovisioning, respectively; and *Resource Class Attributes* which allow a network operator to classify network resources (e.g., “satellite,” “intercontinental,” etc.) and then apply to them resource-class based policies.

Constraint-based Routing (CR), sometimes referred to as "QoS routing," enables a demand-driven, resource reservation-aware routing environment in which an I-LER automatically determines explicit routes for each TT it handles.

CR requires several network capabilities which include:

- traffic-engineering extensions to Interior Gateway Protocols (IGPs) OSPF and IS-IS, i.e., OSPF-TE and ISIS-TE defined in [OSPF-TE] and [ISIS-TE] respectively, to carry additional information about the maximum link bandwidth, maximum reservable bandwidth, current bandwidth reservation at each *priority* level, and other values - to allow the network management system to discover paths that meet TT constraints, resource availability and load balancing and recovery objectives
- algorithms that select feasible paths based on the information obtained from IGP-TEs (e.g., by pruning ineligible links and running a SPF algorithm on the remaining links resulting in a Constrained Shortest Path First (CSPF)) and generate explicit routes
- label distribution by a traffic-engineering-enabled protocol such as RSVP-TE [RSVP-TE]; RSVP-TE carries information about the explicit path identified by CR algorithms and several objects which contain signaling setup and holding priority attributes, preemption attribute, and some others
- a bandwidth management or admission control function in each node that performs accounting of used and still available resources in the node, and provides this information to IGP-TE and RSVP-TE.

With these mechanisms in place, MPLS-TE allows an SP to create stable paths with bandwidth reservation and traffic-engineer them for various network objectives. In order to guarantee bandwidth along these paths, MPLS-TE reservations must be supplemented with mechanisms that protect flows from interfering with each other during bursts beyond their reserved values. These mechanisms may include flow policing, overprovisioning, or queuing discipline that enforces fair sharing of links in the presence of contending traffic flows. Of the two necessary conditions for QoS: guaranteed bandwidth and differentiated servicing – MPLS-TE addresses the first condition, and RSVP-TE provides the means for controlling delay and delay variation for time-sensitive flows.

2.3.3 RSVP-TE

RSVP-TE is widely used for label distribution in networks that require Traffic Engineering and QoS. RSVP-TE is defined in [RSVP-TE] as a set of tunneling extensions to the original RSVP protocol described in section 2.1 above. RSVP-TE was developed for a variety of network applications, only one of which is Traffic Engineering. Thus, the "TE" part of RSVP-TE is properly interpreted as "Tunneling Extensions," rather than Traffic Engineering. Also, several different notations exist to refer to the protocol defined in [RSVP]; this paper follows the terminology of [RSVP-TE] which calls the original RSVP "Standard RSVP".

RSVP-TE operates on RSVP-capable routers where tunneling extensions allow the creation of explicitly routed LSPs, provide smooth rerouting, preemption, and loop detection. Some of the major differences between the Standard RSVP and RSVP-TE protocols include the following:

- Standard RSVP provides signaling between pairs of hosts; RSVP-TE provides signaling between pairs of LERs.
- Standard RSVP applies to single host-to-host flows; RSVP-TE creates a state for a traffic trunk. An LSP tunnel usually aggregates multiple host-to-host flows and thus reduces the amount of RSVP state in the network.
- Standard RSVP uses standard routing protocols operating on the destination address; RSVP-TE uses extended IGPs and constraint-based routing (CR).

But just like Standard RSVP, RSVP-TE can support various IntServ service models and distribute various traffic conditioning parameters such as, for example, *average rate*, *peak rate* and *burst size* for Controlled Load Service. These features allow networks with MPLS-TE and RSVP-TE to provide various services with strict QoS requirements. One shortcoming of this solution is lack of a packet discard mechanism. A technology addressing this issue and providing another approach to QoS guarantees is described in section 3 below.

3. MPLS with DiffServ

3.1 MPLS Support of DiffServ

Now, that both DiffServ and MPLS have been reviewed, we can discuss a technology that *combines* these two approaches in order to guarantee QoS. Let us recall that DiffServ provides a QoS treatment to traffic aggregates. It is a scalable and operationally simple solution as it does not require per-flow signaling and state. However, it cannot guarantee QoS, because it does not influence a packet path, and therefore, during a congestion or failure, even high-priority packets do not get guaranteed bandwidth.

MPLS, on the other hand, can force packets into specific paths and - in combination with constraint-based routing - can guarantee bandwidth for FECs. But in its basic form MPLS does not specify class-based differentiated treatment of flows.

Combining the DiffServ-based classification and PHBs with MPLS-based TE leads to true QoS in packet backbones. The mechanisms for MPLS support of DiffServ are described in RFC3270 [MPLS-DiffServ].

[MPLS-DiffServ] defines two types of LSPs: E-LSPs and L-LSPs. In an E-LSP, a label is used as the indication of the FEC destination, and the 3-bit Exp field is used as the indication of the class of a flow in order to select its PHB, including both scheduling and drop priority. Note that DiffServ uses 6 bits to define BAs and the corresponding PHBs, whereas E-LSP has only 3 bits for this function.

In an L-LSP, a label is used as the indication of both the FEC destination and its scheduling priority. The Exp field in an L-LSP is used only for the indication of the drop priority.

Mappings between IP headers with DiffServ and MPLS shim headers for E-LSP and L-LSP are shown in Figures 3 and 4, respectively. In these figures, the term "5-tuple" refers to the five fields in an IP packet header, including source and destination IP addresses, source and destination TCP or UDP ports, and a protocol that can be used for defining a FEC. All other terminology is based on the DiffServ architecture described in section 2.2 above.

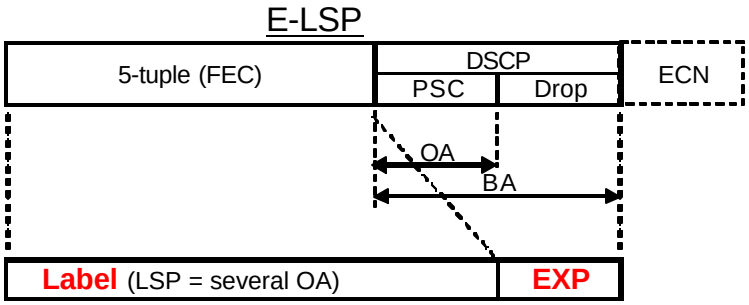

Figure 3. Mapping between an IP header and an MPLS shim header for an E-LSP

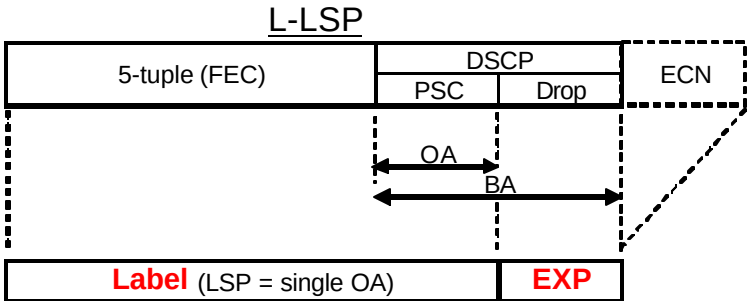

Figure 4. Mapping between an IP header and an MPLS shim header for an L-LSP

Note that Figures 3 and 4 represent mappings between portions of the native IP header, and the Label and EXP parts of the MPLS shim header. They are *not to-scale* and do not represent the complete structure of either header.

Each type of LSP has its advantages and disadvantages. E-LSPs are easier to operate, and are more scalable because they preserve labels and use the EXP field for DiffServ features. But considering that MPLS signaling reserves bandwidth on a per-LSP basis, the bandwidth is reserved for the entire LSP without the PSC-based granularity, and there may be insufficient bandwidth in queues serving some particular PSCs.

L-LSPs, on the other hand, are more cumbersome to provision, because more labels are needed to tag all PSCs of all FECs. But (because a label carries the scheduling

information) when bandwidth is reserved for a given L-LSP, it is associated with the priority queue to which this LSP belongs.

The next two figures illustrate how routing and QoS improve network routing by using basic MPLS and then DiffServ Support of MPLS.

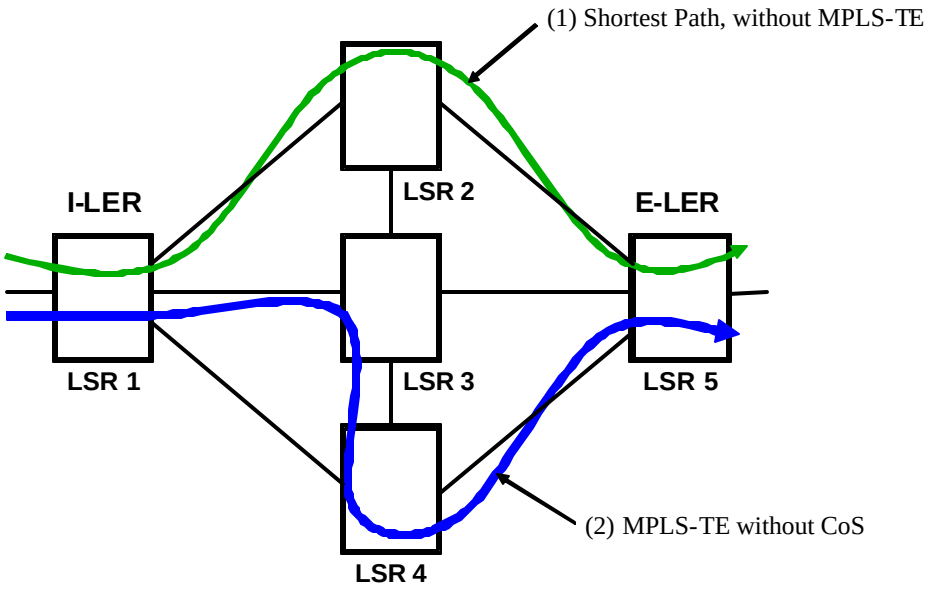

Figure 5. Packet flow in MPLS without DiffServ

Figure 5 illustrates the difference between a path taken by packets that follow shortest path routing (1) and a traffic-engineered path (2). Path (2) may have been chosen because it has sufficient bandwidth to serve a given FEC, but this bandwidth is not associated with any specific class of service, and thus priority traffic (for example, VoIP) may not have sufficient bandwidth for its particular queue.

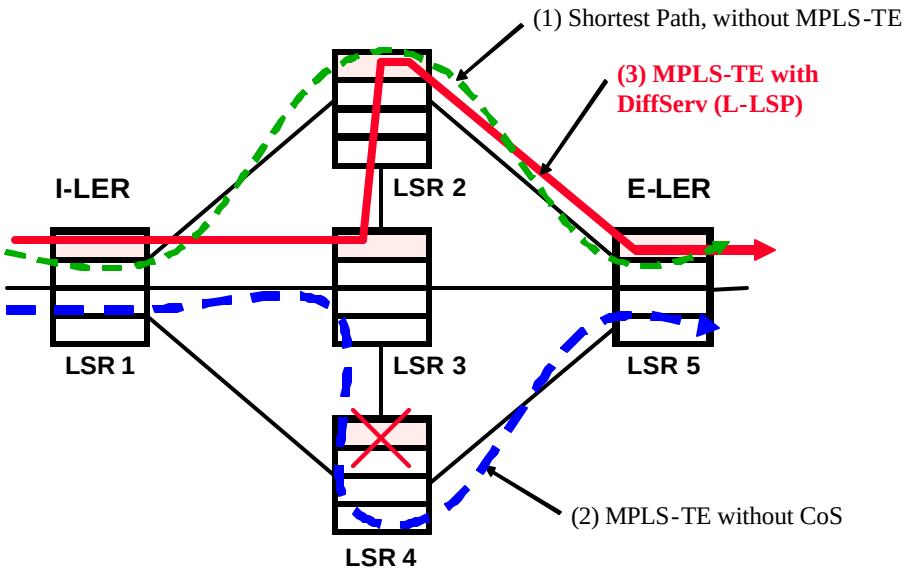

Figure 6. Packet flow in MPLS with DiffServ

Figure 6 illustrates an improvement on the architecture illustrated in Figure 5. Paths (1) and (2) of the previous figure are shown here in dashed lines for reference. In this architecture, MPLS support of DiffServ technology is deployed, and bandwidth reservations can be made with respect to specific priority queues. Let us assume that VoIP traffic uses queue-0, which is the top queue in every LSR.

LSR-4 may have sufficient bandwidth across all of its queues, but it does not have enough bandwidth in queue-0, and therefore, path (2) will not provide QoS that is appropriate for the VoIP traffic. That is why we crossed the VoIP queue on LSR-4. But if an L-LSP is used with queue-0-specific bandwidth reservations, then traffic can be routed along path (3) via LSR-3 and LSR-2, and VoIP can be delivered with guaranteed QoS.

In summary, MPLS support of DiffServ satisfies *both* necessary conditions for QoS: guaranteed bandwidth and differentiated queue servicing treatment. MPLS satisfies the first condition, i.e., it forces applications flows into the paths with guaranteed bandwidth; and along these paths, DiffServ satisfies the second condition by providing differentiated queue servicing.

Note that MPLS support of DiffServ is still simpler and more scalable than IntServ with Standard RSVP. IntServ requires per-microflow signaling and per-microflow states in each router. In contrast, LSPs may themselves be aggregations of many microflows and thus require less signaling. Additionally, routers do not keep per-flow states. Instead, LSRs keep aggregated information on the bandwidth availability for all LSPs or for each priority queue.

3.2 DiffServ-Aware MPLS Traffic Engineering

In section 3.1 above we described MPLS operation in networks where LSRs are DiffServ-enabled. But in order to achieve this functionality – and the resultant QoS – these networks have to be carefully engineered with TE applied on a per-class basis as opposed to the aggregated TE described in section 2.3.2 above.

This section describes the work at present in-progress in the IETF TE-WG to define DiffServ-aware MPLS Traffic Engineering (DS-TE) based on the Internet drafts [DSTE-REQ], [DSTE-PRO], [DSTE-RUS], and [DSTE-MAM]. The DS-TE model is described in this paper despite its pre-standard status, because it is an essential component of the QoS-enabling capabilities of MPLS networks. Considering that the referenced drafts are evolving as we speak, the reader is encouraged to refer to the latest versions of the DS-TE work.

The essential goal of DS-TE is to guarantee bandwidth separately for each type of traffic in order to improve and optimize its compliance with QoS requirements. The DS-TE model modifies the existing, aggregate-based TE model by enabling a more-granular, CoS-based TE, where a Class of Service (CoS) is defined by the model as a set of Ordered Aggregates (OA) generalized from the link level to the network level. In the DS-TE model, the CoS-based bandwidth guarantee is achieved by two new network functions:

1. separate bandwidth reservations for different sets of traffic classes and
2. admission-control procedures applied on a per-class basis.

To describe these two functions, the DS-TE model introduces two new concepts:

1. Class-Type (CT) is a grouping of Traffic Trunks (TT) based on their CoS values so that they share the same bandwidth reservation, and where a single CT can represent one of more classes; and
2. Bandwidth Constraint (BC) is a limit on the percentage of a link's bandwidth that a particular CT or a group of CTs may take up.

The relationships between CTs and BCs are defined in the Bandwidth Constraint Models (BC Models). At present, the TE-WG has defined two BC Models:

1. Maximum Allocation Model (MAM) [DSTE-MAM] assigns a BC to each CT (as illustrated in Figure 7 below); and
2. Russian Dolls Model (RDM) [DSTE-RUS] assigns BC to groups of CTs in such a way that a CT with the strictest QoS requirements (e.g., CT_7 for VoIP) receives its own bandwidth reservation, BC_7; a CT with the next strictest QoS requirements, CT_6, shares bandwidth reservation BC_6 with CT_7 ($BC_6 > BC_7$); and so on, up to CT_0 (e.g., Best Effort traffic) which shares BC_0 (i.e., the entire link bandwidth) with all other types of traffic (as illustrated in Figure 8 below).

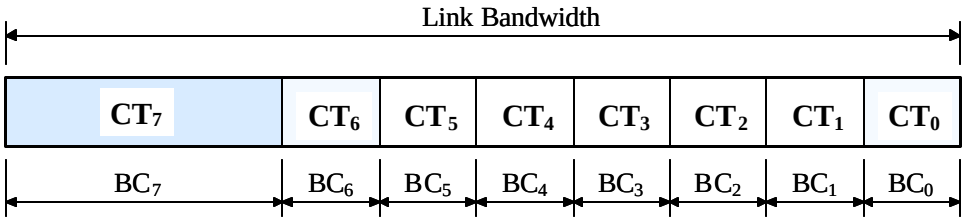

Figure 7. Maximum Allocation Model (MAM)

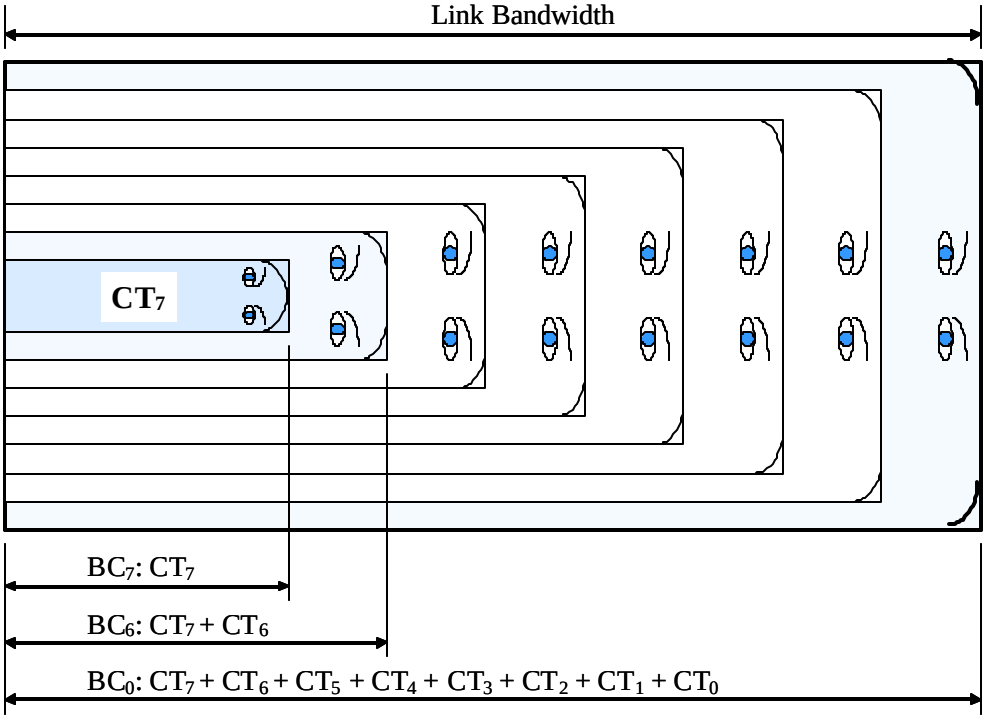

Figure 8. Russian Dolls Model (RDM)

The DS-TE model also defines a mechanism that allows the release of shared bandwidth occupied by lower priority traffic when higher priority traffic arrives. It introduces the concept of Traffic Engineering Class (TE-Class), where a TE-Class is defined by two parameters: Class-Type (CT) and preemption priority (p). Two or more TE-Classes may contain the same CT with different p values, or different CTs with the same p values, thus enabling preemption and preservations of LSPs within and between CTs.

In order to implement DS-TE, the IGPs (OSPF-TE and ISIS-TE) and the LDP (RSVP-TE) must be extended beyond the currently defined MPLS-TE-based extensions to carry additional information as described in [DSTE-PRO]. Note that [DSTE-PRO] does not repeat the definitions, TLVs and objects defined in [OSPF-TE], [ISIS-TE], and [RSVP-TE], but it only introduces new components and modifications.

For extended IGPs it defines additional sub-TLVs that carry values of BC and Unreserved Bandwidth for each TE-Class. Likewise, RSVP-TE is extended by specifying in Path messages a new "CLASSTYPE object" which includes a CT field. Thus, the extended protocols allow an LSR to manage accounting and decision-making on a per-class basis. For example, an LSR can calculate the bandwidth utilized by all existing traffic trunks on the per-CT and per-preemption priority basis, make a decision on whether to admit a new TT that is being set up by a Path message, and compute unreserved bandwidth values to be used by IGPs.

The DS-TE model provides a lot of flexibility for the implementation of traffic engineering. For example, it allows assignment of CTs to an LSR scheduler queue and, when a scheduler enforces bandwidth, the scheduler adjusts the bandwidth parameters of each queue to the reservation state of the traffic grouping it services. The adjustment of the schedulers can be made dynamically, as reservations by new class-based LSPs increase and decrease, or statically, by aligning scheduler configuration with properly anticipated loads.

Various reasons for setting maximum bandwidth allocations for different types of traffic could apply. For example, in order to provide greater control over delay variation to real-time traffic that shares a link with more bursty data traffic, the percentage of the link used by the real-time traffic is kept low (often below 50%). This maximum percentage is then the BC that would be associated with the real-time CT. While the use of BCs over and above class-sensitive bandwidth reservations is optional in DS-TE, a network operator that pre-configures guaranteed bandwidth parameters on its scheduler queues might decide to assign BCs to all the CTs it carries. The purpose of this would be to engineer the traffic reservations of the CTs to conform to the pre-configured guaranteed bandwidth parameters.

A definition of the relationship between CoSs and CTs is also very open to the Network Operator's implementations. Certain traffic classes may not be different enough from one another to warrant separate PSCs and bandwidth guarantees. An example would be a scenario where two kinds of packet voice, say VoIP and uncompressed VoATM, are carried over an MPLS network. The two classes of real-time traffic could well be assigned different scheduler treatments (PSCs), yet a single bandwidth percentage limit could be applied to both together. They would then appropriately be treated as one Class-Type. Nevertheless, it still may be desirable to put them into two different TE-Classes that have the same CT value but different preemption priority values p.

Thus the DS-TE model provides the capability to flexibly engineer and guarantee network resources on a per-class basis and enables the high-granularity QoS functionality described in section 3.1.

4. Practical Implementation of Queue Management in MPLS-DiffServ

In [MPLS-DiffServ] and the DS-TE drafts, the IETF has outlined a comprehensive architecture for MPLS support of DiffServ, and SPs have to analyze this architecture in order to chose the solutions that are optimal for their environments.

The MPLS Forum has conducted some analysis of queue management disciplines in the network nodes that support E-LSPs and L-LSPs. [QMgmt] makes a practical assumption that although queues may be managed before and after the LSR switching fabric, it is possible to engineer LSPs so that contention for the switching fabric will be minimal and queuing will take place after the switching fabric and before the egress interface, as shown in Figure 9 below.

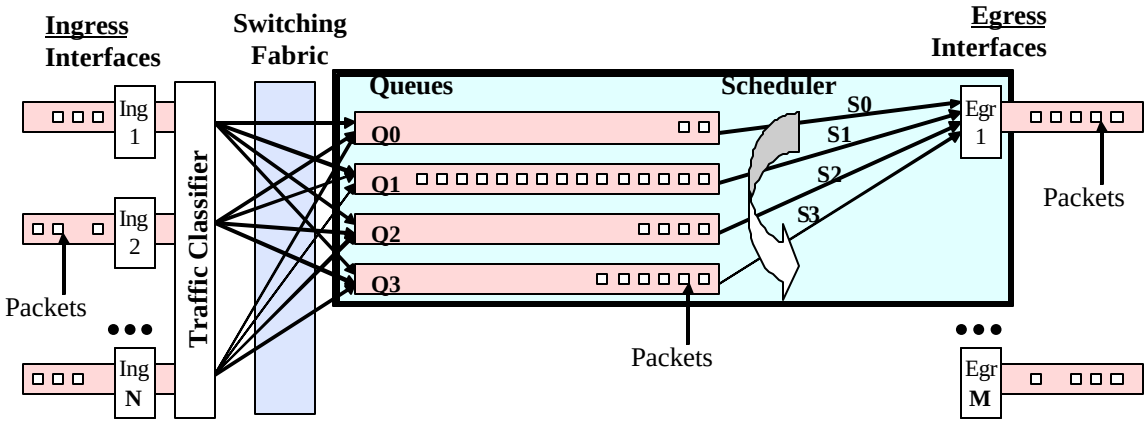

Figure 9. LSR architecture with queuing and scheduling

In Figure 9, packets arrive on one of the *N* Ingress Interfaces. They get classified and forwarded to the appropriate Egress Interface where they are put into a corresponding priority queue. In general, there could be *M* Egress Interfaces. Figure 9 shows an expanded view of the Egress Interface-1. The interface supports four priority queues, 0 through 3, and the Scheduler pulls out packets from each queue based on its priority. Small white squares in various parts of the LSR represent packets.

Figure 9 represents a simplified view of egress queuing, but the reality of the E-LSPs and L-LSPs is more complex as described in [QMgmt]. For example, queues could be allocated for each CoS of each LSP. In this scenario, if an E-LSP supported four different scheduling behaviors and there were two such E-LSPs, the LSR would have eight different priority queues, one for each CoS of each LSP. This arrangement would be very difficult to manage, and it is not scalable.

An alternative arrangement is to have a small set of queues for each egress interface. Packets that arrive at these queues generally represent different LSPs, but the queues are managed with respect to packet discarding and scheduling as if they represented a single flow. [QMgmt] recommends this approach for the initial phase of MPLS support of DiffServ. With this approach, an L-LSP represents a single priority and the allocation of

resources to an L-LSP leads to guaranteed QoS as was discussed above. An E-LSP, on the other hand, may represent several scheduling priorities; allocation of resources to an E-LSP does not imply that every CoS will get queue-specific resources. In more mature networks, more granular queue assignments and priorities may be possible, and the E-LSPs may also be used for guaranteed servicing.

5. QoS Features of the MPLS UNI

The discussions above described and analyzed various approaches for providing QoS in MPLS networks. But in practical network architectures, MPLS networks are provided in SP domains and do not extend to customer premises. One reason for this is that traditional LSP signaling is handled by the LERs; and if LERs were located on customer premises, SPs would have to give customers some control over their backbones. Another reason is that customers have different levels of sophistication and resources, and they may not wish to be involved in managing MPLS.

But, as described in [E2E-QoS], in order to assure End-to-End QoS, QoS mechanisms have to be provided in all parts of the constituent networks, and they have to be properly mapped at the interfaces between different networks. In fact, as we discussed in the Introduction, the access part of the network may be the weakest link in terms of QoS, and must be carefully designed for QoS interoperability.

The MPLS Forum has assumed the leading role in defining an MPLS-based interface between customer premises and MPLS-based service provider networks. This interface is called MPLS User to Network Interface (UNI), and it is described in [MPLS-UNI].

[MPLS-UNI] describes signaling over the UNI as shown in Figure 10 below.

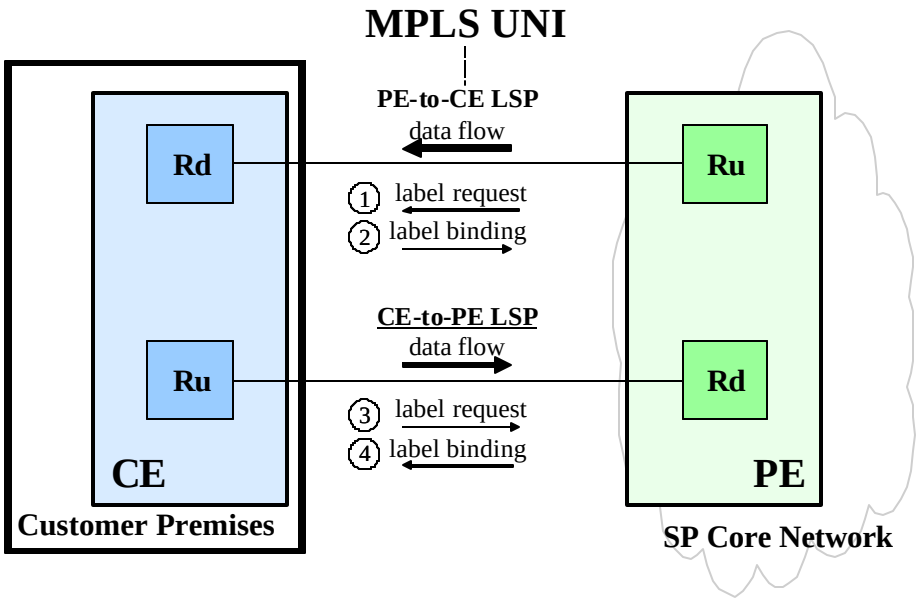

Figure 10. MPLS UNI

In Figure 10, Ru and Rd are functional representations of the upstream and downstream routers as defined in [MPLS-Arch]. The Provider Edge (PE) and Customer Edge (CE) notation is adopted from [VPN] to provide some clarity when referring to the provider and customer devices. Considering that an LSP is always unidirectional, in the PE-to-CE data flow direction the PE is Ru and the CE is Rd; and in the CE-to-PE data flow direction, the PE is Rd and the CE is Ru.

Important parts of the MPLS UNI definition are the aim to minimize the CE functionality and the fact that LSPs defined across the UNI do not extend into the SP networks, but are cross-connected at the PE with the provider's LSPs. [MPLS-UNI] uses an extended LDP protocol for signaling which is based on the Downstream on Demand approach. The signaling process starts with the establishment of a TCP connection between the PE and the CE. Then the PE sends to the CE a label request for the PE-to-CE LSP. Here, the CE is a "downstream" node, and when it receives the request (the "demand"), it provides label bindings for all attributes that were specified in the label request. At this point, in order to signal the CE-to-PE LSP, the CE must send a request to the PE, which it does, using exactly the same attributes that were provided in the PE's request. Thus the CE behavior is quite simple and decision-free. The signaling process is completed when the PE responds with the label binding for the CE-to-PE LSP.

The present version of [MPLS-UNI] addresses only a PVC (Permanent Virtual Connection) approach which is characterized by SP-initiated provisioning and CE passivity. Future work will extend this architecture to MPLS UNI SVCs (Switched Virtual Connections) where CEs will be able to signal requests for establishment or termination of LSPs.

While [MPLS-UNI] describes the signaling over the MPLS UNI that leads to the establishment of bi-directional LSPs, ongoing work at the MPLS/Frame Relay Alliance also looks into service connections that can ride on top of these LSPs. The MPLS UNI service connections will provide customers with an MPLS access to their VPN services across the provider network as well as to the other services offered by the SP, such as Internet, VoIP gateways, and others. These services can be automatically provisioned by the SP to a customer, with a minimal customer involvement. The MPLS UNI functions similarly to the ATM and Frame Relay connections where customer virtual connections are associated with the LSPs and not with the interfaces, and thus, for example, several VPNs could be supported over a single interface. Figure 11 below illustrates this concept.

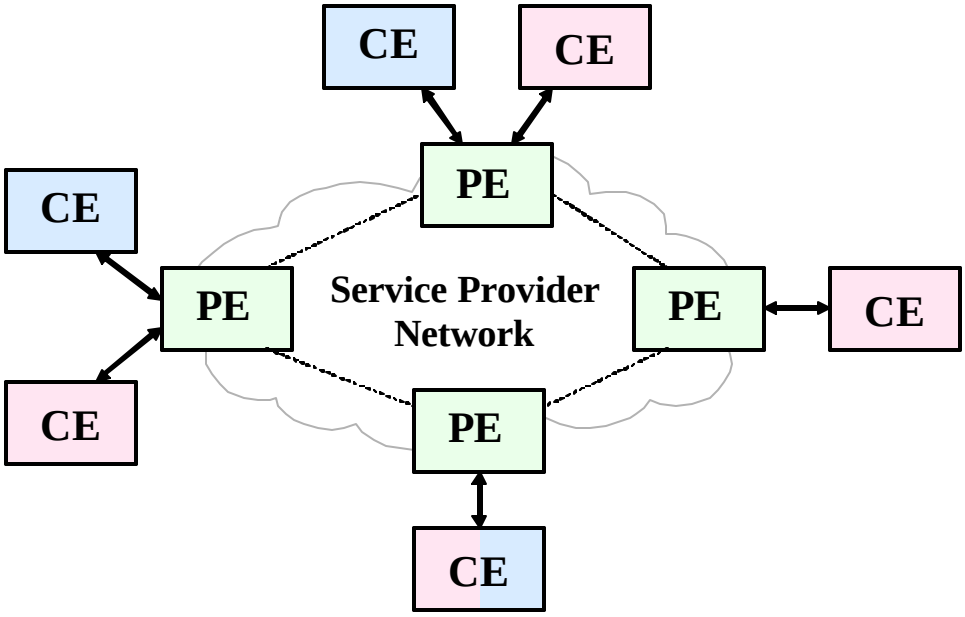

Figure 11. VPNs served over the MPLS UNI

In Figure 11, the PE on the left supports two different CEs that belong to two different VPNs (diagonally-striped blue and vertically-striped red). The PE on the bottom has a single interface to a CE, but this CE is participating in both VPNs which are distinguished by the LSPs defined over this UNI.

The MPLS/Frame Relay Alliance work on the UNI services will also describe various service attributes that could be configured by the PE in the CE. One of these attributes will be the QoS profile which will include the QoS class (e.g., real-time), QoS marking, bandwidth, delay and jitter reservations, availability requirements and other parameters. The profile will allow an SP to provide a necessary QoS for the flows across the UNI, as well as to map it into the network LSPs. Thus, QoS will be preserved and properly mapped at the interface between the customer and provider networks.

While the MPLS PVC UNI service environment is managed entirely by the PE, the MPLS SVC UNI will also allow the CE to request modifications of the service

6. Summary

The initial approaches to packet network QoS which primarily focused on throwing in bandwidth are now being replaced with sophisticated mechanisms that allow SPs to provision and operate their networks more precisely. This phenomenon is forced by two recent drivers: (1) reduction in the CAPEX for acquiring ever more bandwidth and (2) generation of additional revenues by providing value-added services with stricter SLAs.

Advanced network services, such as VoIP, require hard QoS guarantees. While the IntServ architecture offered such guarantees, it was not scalable or practical to operate and manage. The DiffServ architecture has provided a scalable alternative but it had the

drawback of providing no guarantees. Recent IETF work on combining the DiffServ and MPLS technologies in a packet network leads to enabling hard QoS assurances; and these guarantees come with better scalability and reduced complexity in comparison with IntServ. These improvements are a result of the stacking hierarchies and FEC aggregations characteristic of MPLS networks as well as the aggregated states maintained by the DiffServ-supporting nodes.

While MPLS support of DiffServ is defined by the IETF [MPLS-DiffServ], service providers must still work out practical uses of this architecture. The MPLS/Frame Relay Alliance supports the industry in defining scalable scenarios, for example, by analyzing various aspects of E-LSP and L-LSP provisioning, as outlined in [QMgmt]. Likewise the DiffServ-aware TE principles that are at present being defined by the TE-WG will require some analysis for practical network implementations.

Another major area addressed by the MPLS/Frame Relay Alliance is the extension of the MPLS network boundaries to the customer premises without service providers losing control over their network provisioning and operations. The MPLS UNI provides an important solution for extending MPLS to the PE-CE interface [MPLS-UNI] and activating QoS and various services over this interface. The MPLS UNI enables QoS interoperability between the customer and the SP domains, thus supporting End-to-End QoS objectives [E2E-QoS].

7. Acknowledgements

The author thanks MPLS/Frame Relay Alliance members for the support of this work. Special appreciation to Sandy Goldfless for good ideas and extensive comments related to MPLS-TE and DS-TE and to David Drury, David Sinicrope, Bernard da Costa, and Dan Proch for constructive reviews and suggestions.

8. References

[Ber] Y. Bernet, “Networking Quality of Service and Windows® Operating Systems,” New Riders, 2001.

[DiffServ-Arch] S. Blake, et al, “An Architecture for Differentiated Services,” RFC2475, Dec. 1998.

[DiffServ-Def] K. Nichols, et al, “Definition of the Differentiated Services Field (DS field) in the IPv4 and IPv6 headers,” RFC2474, Dec. 1998.

[DiffServ-Term] D. Grossman, “New Terminology and Clarifications for DiffServ,” RFC3260, Apr. 2002.

[DSTE-PRO] F. Le Faucheur, “Protocol Extensions for Support of Diff-Serv-aware MPLS Traffic Engineering,” draft-ietf-tewg-diff-te-proto-03.txt, Feb 2002.

[DSTE-REQ] F. Le Faucheur, "Requirements for Support of Diff-Serv-aware MPLS Traffic Engineering," draft-ietf-tewg-diff-te-reqts-07.txt, Feb. 2003.

[DSTE-RUS] F. Le Faucheur, "Russian Dolls Bandwidth Constraints Model for Diff-Serv-aware MPLS Traffic Engineering," draft-ietf-tewg-diff-te-russian-02.txt, Mar. 2003.

[DSTE-MAM] F. Le Faucheur, "Maximum Allocation Bandwidth Constraints Model for Diff-Serv-aware MPLS Traffic Engineering," draft-lefaucheur-diff-te-mam-00.txt, Feb. 2003.

[E2E-QoS] V. Fineberg, "A Practical Architecture for Implementing End-to-End QoS in an IP Network," IEEE Communications Magazine, Jan. 2002.

[IntServ] R. Braden, et al, "Integrated Services in the Internet Architecture: an Overview," RFC1633, Jun. 1994.

[ISIS-TE] H. Smit, T. Li, "IS-IS Extensions for Traffic Engineering," draft-ietf-isis-traffic-04.txt, Dec. 2002.

[LDP] L. Andersson, et al, "LDP Specification," RFC3036, Jan. 2001.

[MPLS-Arch] E. Rosen, A. Viswanathan, R. Callon, "Multiprotocol Label Switching Architecture," RFC3031, Jan. 2001.

[MPLS-DiffServ] F. Le Faucheur, et al, "MPLS Support of Differentiated Services," RFC3270, May 2002.

[MPLS-UNI] A. Malis, D. Sinicrope, "MPLS PVC UNI Implementation Agreement: Baseline Text," MPLS Forum draft, mplsforum2002.75, May 2002.

[MPLS-WP] White Papers related to MPLS, MPLS Resource Center, http://www.mplsrc.com/

[OSPF-TE] D. Katz, D. Yeung, K. Kompella, "Traffic Engineering Extensions to OSPF Version 2," draft-katz-yeung-ospf-traffic-09.txt, Oct. 2002.

[QMgmt] P. Schicker, "Class of Services Queue Management in Switches/Routers," MPLS Forum draft, mplsforum2002.035, Mar. 5, 2002.

[RSVP] R. Braden, "Resource Reservation Protocol (RSVP)," RFC2205, Sep. 1997.

[RSVP-TE] D. Awduche, et al, "RSVP-TE: Extensions to RSVP for LSP Tunnels," RFC3209, Dec. 2001.

[TE-QoS] G. Ash, "Traffic Engineering & QoS Methods for IP-, ATM-, & TDM-Based Multiservice Networks," draft-ietf-tewg-qos-routing-04.txt, Oct. 2001.

[TE-REQ] D. Awduche, et al, “Requirements for Traffic Engineering over MPLS,” RFC2702, Sep. 1999

[VPN] Eric C. Rosen, et al, “BGP/MPLS VPNs,” draft-ietf-ppvpn-rfc2547bis-03.txt, Oct. 2002.

9. Acronyms

AF	Assured Forwarding
ATM	Asynchronous Transfer Mode
BA	Behavior Aggregate
BC	Bandwidth Constraint
BC Model	Bandwidth Constraint Model (e.g., RDM, MAM)
BE	Best Effort
BW	Bandwidth
CAPEX	Capital Expenditure
CE	Customer Edge
CoS	Class of Service
CR	Constraint-based Routing
CSPF	Constrained Shortest Path First
CT	Class Type
DiffServ	Differentiated Services
DS	Differentiated Services
DSCP	DiffServ Code Point
DS-TE	DiffServ-aware MPLS Traffic Engineering
E2E	End-to-End
ECN	Explicit Congestion Notification
EF	Expedited Forwarding
E-LER	Egress LER
E-LSP	EXP-Inferred-PSC LSP
EXP	Experimental field
FE	Fast Ethernet
FEC	Forwarding Equivalency Class
FTP	File Transfer Protocol
IETF	Internet Engineering Task Force
IGP	Interior Gateway Protocol
I-LER	Ingress LER
IntServ	Integrated Services
IP	Internet Protocol
IS-IS	Intermediate System to Intermediate System protocol
LDP	Label Distribution Protocol
LER	Label Edge Router
L-LSP	Label-Only-Inferred-PSC LSP
LSP	Label Switched Path
LSR	Label Switching Router
MAM	Maximum Allocation Multiplier
MAM	Maximum Allocation Model
MPLS	Multiprotocol Label Switching
MPLS-TE	Traffic Engineering used in non-DiffServ-aware MPLS networks, applied on aggregated basis

OA	Ordered Aggregate
OPEX	Operating Expenditures
OSPF	Open Shortest Path First
P	Provider router (SP's router)
PE	Provider Edge
PHB	Per-Hop Behavior
PSC	PHB Scheduling Class
PVC	Permanent Virtual Connection
QoS	Quality of Service
Rd	down-stream Router
RDM	Russian Dolls Model
RFC	Request for Comments (an IETF document)
RSVP	Resource reservation Protocol
Ru	up-stream Router
S	Stacking bit
SLA	Service Level Agreements
SLS	Service Level Specification
SMTP	Simple Mail Transfer Protocol
SP	Service Provider
SPF	Shortest Path First
SVC	Switched Virtual Connection
TA	Traffic Aggregate
TCP	Transmission Control Protocol
TDM	Time Division Multiplexing
TE	Traffic Engineering
TE-Class	set of {CT, p} where p is a preemption priority value associated with this TE-Class
TELNET	a TCP/IP standard protocol for remote terminal service
TE-WG	IETF Traffic Engineering Working Group
TLV	Type Length Value
TT	Traffic Trunk
TTL	Time To Live
UDP	User Datagram Protocol
UNI	User-to-Network Interface
VoATM	Voice over ATM
VoIP	Voice over IP
VPN	Virtual Private Network