

Quality of Service Differentiated Services and Multiprotocol Label Switching

Brian Williams, Ericsson Australia

March 2000

The requirements for datacom networks are changing. Corporate use of the Internet is placing new demands on the network for service guarantees in both reliability and service quality. When your business depends on communication, you cannot afford a service that fails to deliver. The data networks of today simply do not offer any guarantees that your service-level requirements can be met without some degradation at any time, day or night, irrespective of other users of the network.

To meet these requirements, the network must be enhanced with new technologies that offer the network operator capabilities for controlling its behavior. Together, the capabilities offered by the combination of Differentiated Services and Multi-Protocol Label Switching enhance the ability of the network operator to control the network to deliver service according to customized service contracts.

The Need for Quality of Service (QoS)

Over time, the needs and uses of the Internet have changed. No longer is it the province of governments and research institutes. More and more, the Internet is becoming a medium for business communication. Along with corporate use of the Internet come requirements for a new paradigm. Rather than the existing “best effort” paradigm, it is necessary for the Internet to support service-level agreements that guarantee a specified level of goodput and network reliability, irrespective of the usage level and individual network failures.

For corporate customers, there is a range of applications that operate across the network. Many of these applications do not have any strict service-level requirements; however, there are applications that are mission critical, and the service delivery to these applications is crucial. Along with the growing importance of these data services, there is also a change in the types of applications that are available. The traditional range of non-real-time applications (e.g., e-mail and ftp) is being extended to include real-time interactive applications such as voice and video services and the World Wide Web. The real-time nature of these new applications places additional demands on the network.

Quality of Service for Different Applications

To operate the network efficiently, every application used on that network must be considered in terms of its requirements to operate effectively. While this may appear to be a daunting task, the reality is that applications typically can be grouped into a relatively small number of classes, with the applications in each class having similar requirements on the network. One class of applications has no requirements beyond that of the traditional “best effort” network. However, other classes of applications introduce new requirements (see Figure 1). For these other classes, it is necessary to determine what requirements must be met to ensure the applications perform satisfactorily.

New Application Requirements

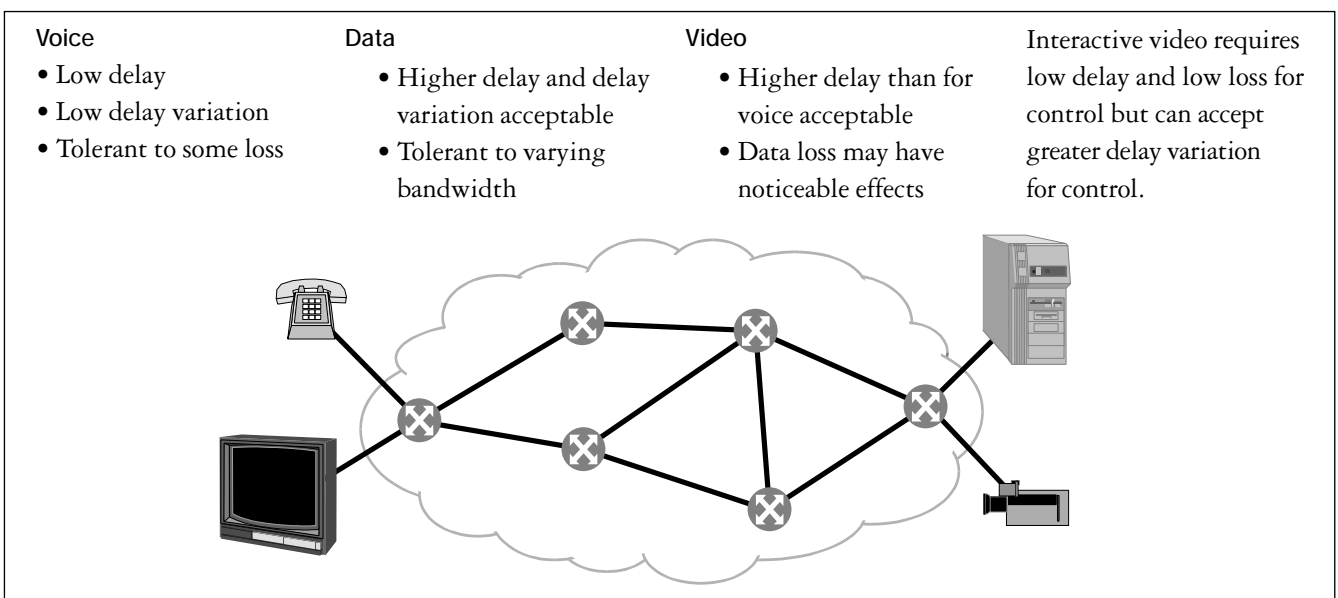


Figure 1. Different applications have different QoS requirements.

Technologies for Quality of Service

Some technologies [e.g., Type Of Service (TOS) and Integrated Service (Int-Serv)] were created in an attempt to provide some quality-of-service (QoS) control. Limitations in these technologies have restricted their use, and they have been unable to provide a framework for provision of service to meet service-level agreements.

One of these early technologies allowed the network to distinguish between network control traffic and user traffic. Based on the TOS byte defined in the IP header, this technique provides a coarse-grained service classification and a small number of service classes. Here, the service classes were defined very early when the exact needs of applications and users were unclear. It has been revealed since that the definition of the classes was not well-suited for providing the required range of services. Because of this, the byte is often not fully supported in routers and hosts.

Another technique that was developed is Int-Serv. Int-Serv provides a well-defined, end-to-end service between hosts for both point-to-point and point-to-multipoint applications. In Int-Serv, the application initiates a session on demand with the network using the Resource Reservation Signaling Protocol (RSVP). This session identifies the service requirements of the application, including information such as bandwidth and delay, and the source of the data.

The objective of Int-Serv has dictated the properties of the traditional RSVP protocol, such as the soft-state nature and the merging of resource requests. While these aspects of Int-Serv make it powerful, enabling it to guarantee the minimum requirements of an application will be met, it also imposes a high price in terms of the processing power and signaling required. Traditional RSVP requires a state machine that includes timers for each session and a classifier in each router, which makes both memory and processing capacity expensive. In an Internet backbone router, there can be many of these sessions with individual users and hosts, and these routers do not have the necessary resources to deal with all of the sessions. This limitation generally restricts Int-Serv

deployment to the edge of the network and tunnels it through the backbone. This reduces the effectiveness, since there is now no guarantee that the tunneled part with the remainder of the session meets the end-to-end requirements.

Differentiated Services

It is obvious that something different is required to enable services on the Internet given today's state-of-the-art router capabilities. This has led to the development of Differentiated Services (Diff-Serv).

Unlike Int-Serv, the objective of Diff-Serv is not to provide an end-to-end service for the host/application. Rather, the goal is to create a set of "building blocks" that provide a foundation for building end-to-end services throughout the network. Diff-Serv takes an approach similar to TOS but is better targeted to meet the needs of today's applications. Diff-Serv is sufficiently scalable that it can be supported in routers at the core of the Internet since it avoids per-session states. Instead, each packet carries information about the service class.

How Diff-Serv Works

Diff-Serv is a strategy for providing QoS across a network through a set of "building blocks" that can be used together in order to achieve an end-customer service. One of these key building blocks is the Per Hop Behavior (PHB) (Figure 2).

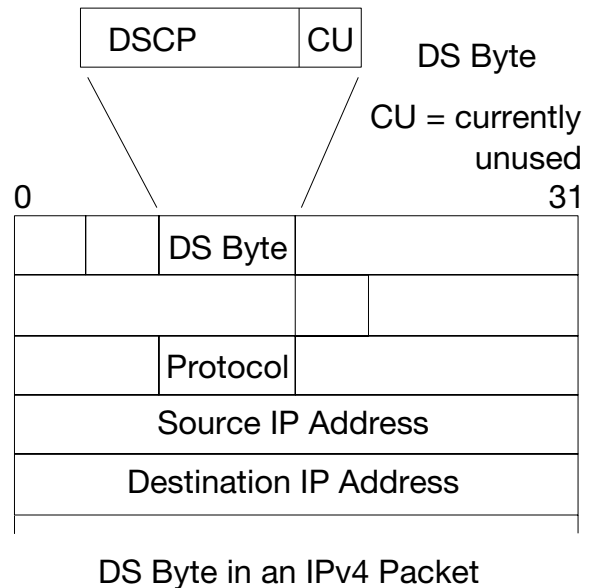


Figure 2. The PHB is indicated by the (DSCP) in the IP header.

Diff-Serv defines a number of “data treatments,” known as a PHB, that can be applied to the packets in each node. The PHB is used to identify the “treatment” that will be given to the packet within the node. This “treatment” includes selection of the queue and scheduling discipline to apply at the egress interface and congestion thresholds. For example, a “treatment” can select a queue with a high-scheduling priority but a low threshold for congestion, and a congestion management scheme. If a packet is given a similar “treatment” at each node throughout the network, then the effect on packets across the network end-to-end can be identified.

The packets are marked to identify the treatment that the packets must receive using the DS byte. The DS byte replaces the TOS byte in the IP header. This byte was selected because it originally was intended to be used to indicate service information, but, as mentioned earlier, its use has been limited.

Within the DS byte, a Diff-Serv CodePoint (DSCP) field has been defined. The value in this field identifies the “behavior” or “treatment” to be applied to the packet within that node in the network. As the packet progresses through the network, each node applies the same “forwarding treatment” to the packet.

The DS byte contains 6 bits for the DSCP, plus 2 bits that are currently unused and reserved for the future. The 6 bits are used as an indexed table to identify the PHB, rather than used as bit fields. This allows for 64 independent codepoints. These are mapped in the node to determine the “treatment” to apply. Depending on this mapping, it is possible to have multiple codepoints selecting the same behavior, allowing local use of existing behaviors for different purposes.

The Defined Per Hop Behaviors

The defined PHBs are:

Expedited Forwarding (EF)

The EF PHB provides a low-loss, low-jitter, and low-delay handling within the node. To provide the low delay, the EF handling further defines that the maximum aggregated reception rate of data for this PHB at any time must be no greater than the minimum transmission rate available for this PHB per egress (see Figure 3). This requirement ensures that there is no queue buildup occurring for this service within the node (apart from synchronous data arrival compared to the packet transmission period), which minimizes the delay and the delay variation.

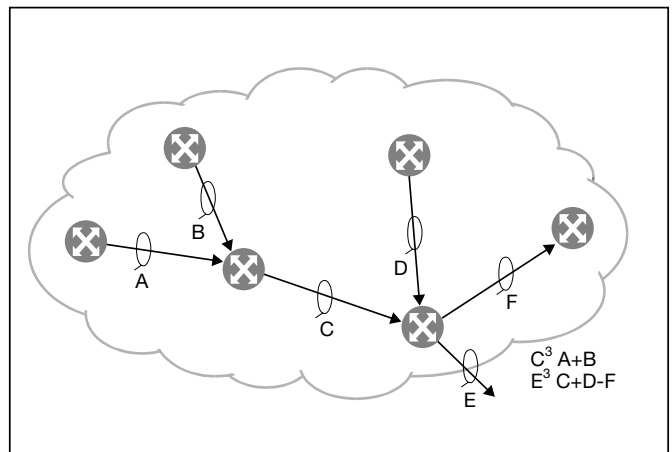


Figure 3. For EF PHB, the sum of the ingress rates must not exceed the egress rate in each node.

Assured Forwarding (AF)

The PHB group defines N independent forwarding classes (4 defined currently) denoted as AF1 to AFn. Within each of these forwarding classes, there are also M subclasses (3 defined currently) for probability of delivery. The higher level of delivery probability should have a greater probability than the second level of getting data through in times of congestion. Likewise, the second level should have a better probability for delivery than the third level. Each forwarding class within this group is configured independently for resources such as buffer space and minimum egress capacity that should be ensured by the scheduling mechanism (see text box — Scheduling Mechanisms).

Default Behavior (DE)

The DE PHB identifies the existing “best effort” traffic. The behavior defines that the node will deliver as many of these packets as possible, as soon as possible. Of course, other defined behaviors have greater requirements on timeliness of delivery, so the DE definition of “as many as possible” and “as soon as possible” allows deference to these other behaviors.

Other PHBs

The Class Selector (CS) codepoints create a set of codepoints for backward compatibility with the precedence field of the IPv4 TOS byte that is now used as the DS byte. These PHBs ensure that routers implementing TOS will provide compatible behavior to routers employing Diff-Serv for these values. Apart from the specific codepoints that have already been defined, there are also spare codepoints. Some spare codepoints have been left for definition of more assured forwarding PHBs, if required in the future. In addition, certain ranges of codepoints have been set aside for local and experimental use. These may be used by a network operator as they see fit or by network equipment vendors to implement additional classes.

Traffic Conditioning: The Other Side of Diff Serv

The DSCP identifies the “behavior” to be applied in each node. Within a Diff-Serv domain, the DSCP must be set in each packet for the nodes to determine the required “behavior.” Therefore, the node at the ingress edge of the Diff-Serv domain must ensure that this field is set appropriately in each packet. This is just part of the role of traffic conditioning.

It has been noted that there can be other conditions applicable to the use of a PHB. For example, the EF PHB has a requirement that the egress capacity of a link for this class is greater than the traffic rate into it. Since the PHB definition can specify the admissible traffic profiles, the ingress node of the Diff-Serv domain actually provides more functions than simply marking the DS byte. All of the functions that are combined in the role of traffic conditioning are then examined in turn.

Since the bearer service class is dependent on the individual Service Level Agreement (SLA), the traffic conditioning is performed independently for each logical access. For each logical interface into the node, there is one instantiation of the traffic-conditioning function.

Scheduling Mechanisms

The node must decide how packets from different PHBs are to be scheduled out on the link. The scheduling mechanism may consider priority order between classes, but it must control access to link bandwidth for each class.

A strict priority mechanism between two or more classes aims to provide the lowest possible delay for the highest priority class. This mechanism sends the data from the highest priority class before sending data for the next class. This could lead to starvation of lower priority classes, so the traffic level must be shaped to limit the used bandwidth.

Weighted Round Robyn (WRR) aims to give a weighted access to the available bandwidth to each class, ensuring a minimum allocation and distribution. The scheduling services each class in a round-robin manner according to the weights. If one or more classes is not using its full allocation, then the unused capacity is distributed to the other classes according to their weighting. A class can be given a lower effective delay by giving it a higher weighting than the traffic level it is carrying.

Weighted Fair Queueing (WFQ) similarly aims to distribute available bandwidth over a number of weighted classes. The scheduling mechanism uses a combination of weighting and timing information to select which queue to service. The weighting effectively again controls the ration of bandwidth distribution between classes under congestion and can also indirectly control delay for underutilized classes.

Class Based Queueing (CBQ) is a more general term for any mechanism that is based on the class. CBQ can allow the unused capacity to be distributed according to a different algorithm than a minimum bandwidth weighting. For example, there could be a different weighting which is configured for this excess capacity, or it could be dependent on the traffic load in each class, or some other mechanism such as priority.

Classifier

The data has to be classified for the PHB at the boundary of the Diff-Serv domain in a classifier function. The classification typically considers information from the IP header such as protocol, source and destination IP address, and source and destination port, although it can go even further into the protocol if required to identify an application.

The classifier applies a filter to each packet. This filter defines the conditions that the IP header must match to be accepted. If the filter accepts the traffic, then the profile attached to that filter is applied to that traffic.

Meter

After the data has passed the classifier, it passes through a metering engine. The metering engine calculates the traffic level, which is compared against the customer's contract/SLA profile. The traffic profile can include such aspects as the agreed average data rates, maximum data rate, and the maximum data burst at the maximum rate. The metering engine polices the traffic level for each service-class agreement and can take one of a number of actions if the level exceeds the agreed parameters.

Note that service agreements may contain multiple clauses for different applications, so the meter may examine the same data in the IP header as the classifier.

Marker

When the data has been classified and the rate has been determined, the selected PHB is marked into the DS byte.

The reason that the PHB is not marked directly from the classifier is that the PHB is dependent on the data rate. There are groups of related PHBs that can be applied to data that do not result in packet reordering. Although the timeliness of the delivery cannot be varied (because it would cause reordering), these groups of related PHBs provide different levels of delivery probability. The probability for delivery is dependent on the level of network congestion, and the various classes are impacted differently by that congestion. For example, traffic rates up to a first threshold would all be marked as class 1. Traffic making up the data rate between the first and second threshold would be

marked as class 2, while traffic in excess of the second threshold would be marked as class 3. Under network congestion, traffic in class 3 is more likely to be discarded than traffic in class 2, which is likewise more likely to be discarded than traffic in class 1. The AF service classes have multiple levels of drop preference that can be used in this manner.

However, it is not until the data has passed the metering engine that the final PHB can be determined.

Customers can choose to perform most of the traffic conditioning function in their own networks, rather than in the operator's network. (When the customer performs this action, the network provider for the service may offer some discount.) However, the network must still verify the marking and shaping by the customer, and hence much of the same functionality is still required in the operator's network.

Some Examples of Classification Filters

Basic traffic between two corporate nodes:

- Destination address is within corporate network range using a netmask

Traffic from a voicemail server:

- Source address is the specific host address of the server
- Source port number is the specific voice service port number
- Protocol is UDP

Traffic for a voice over IP call between two parties:

- Source address is A-party's IP address
- Destination address is B-party's IP address
- Source port number is the A-party's port number for this application
- Destination port number is the B-party's port number for this application
- Protocol is UDP

All filters must be considered to find the "best" match, not the first match.

Shaper

Once the data has been metered and marked, the router enforcing the policy knows whether the data rate is within the allowed traffic profile or whether the traffic profile has been exceeded. If the data rate is within the profile, then the data can be routed and scheduled toward the destination along with other traffic. If, on the other hand, the data rate is in excess of the profile, then there are two possible actions. The first of these is to “shape” the traffic. The data is forwarded on to the normal routing/scheduling processes at a rate defined by the customer’s traffic profile. Since this rate is lower than the rate at which data has been received, it is necessary to buffer this excess data and then send it on later as the profile permits. This can be used to smooth out small rate excesses if any previous traffic shaping is not sufficiently accurate or if traffic delay variation has changed the timing.

Dropper

The “dropper” performs the other action that can be applied for traffic in excess of the profile. The excess data can simply be discarded. This occurs if the buffering for the shaper is full. In some cases, the service itself may not tolerate any excess data rate; therefore, the allocation

of buffers to the service may be very low. This would result in almost all data in excess of the profile being discarded immediately.

While this action appears drastic, it may be necessary for some PHBs such as EF, for which there are strict requirements on the traffic levels throughout the internal network.

IP Bearer Services

The PHB in combination with traffic conditioning is used to define IP Bearer Services (IBS) within the network (see Figure 4). Each of the two elements in the bearer service plays a critical role in the creation of the IBS. It is possible to use the same PHB, yet apply different parameters for the traffic conditioners to create a separate bearer service with very different characteristics.

Besides the different traffic conditioners, the performance of the IBS is also dependent on the engineering of that service. For example, increasing the capacity offered to one IBS beyond its traffic level can give that service a lower latency. Depending on the allocation of buffers, it can also offer a lower probability of congestion (see text box — Congestion Control).

Congestion Control

Early congestion mechanisms were very simple, typically based on discarding packets when buffers filled up to a threshold level. This protected the routers but gave rise to an unexpected phenomenon known as global synchronization. When TCP packets are discarded, the TCP scheduling algorithm responds by lowering its transmission rate, then building it up again. When core routers overload, they drop packets from many hosts, leading to many TCP sessions backing off and ramping up their transmission rates synchronously again. This can lead to a “sawtooth” pattern of underutilization and congestion.

To achieve higher average utilization, Random Early Detection (RED) was introduced to stop the synchronized effect. Instead of discarding all traffic when a threshold is reached, a random level of drops is started as buffer utilization increases towards the maximum. The aim is to make some proportion of TCP sessions back off before hitting congestion.

Many variants of RED have now been proposed; for example, there is Weighted RED (WRED), Fair RED (FRED), RED with In-Out (RIO), and adaptive RED. It is not clear which RED variant will give the best results for a backbone router. However, it is clear that the network must have a congestion control mechanism that provides drop-class differentiation.

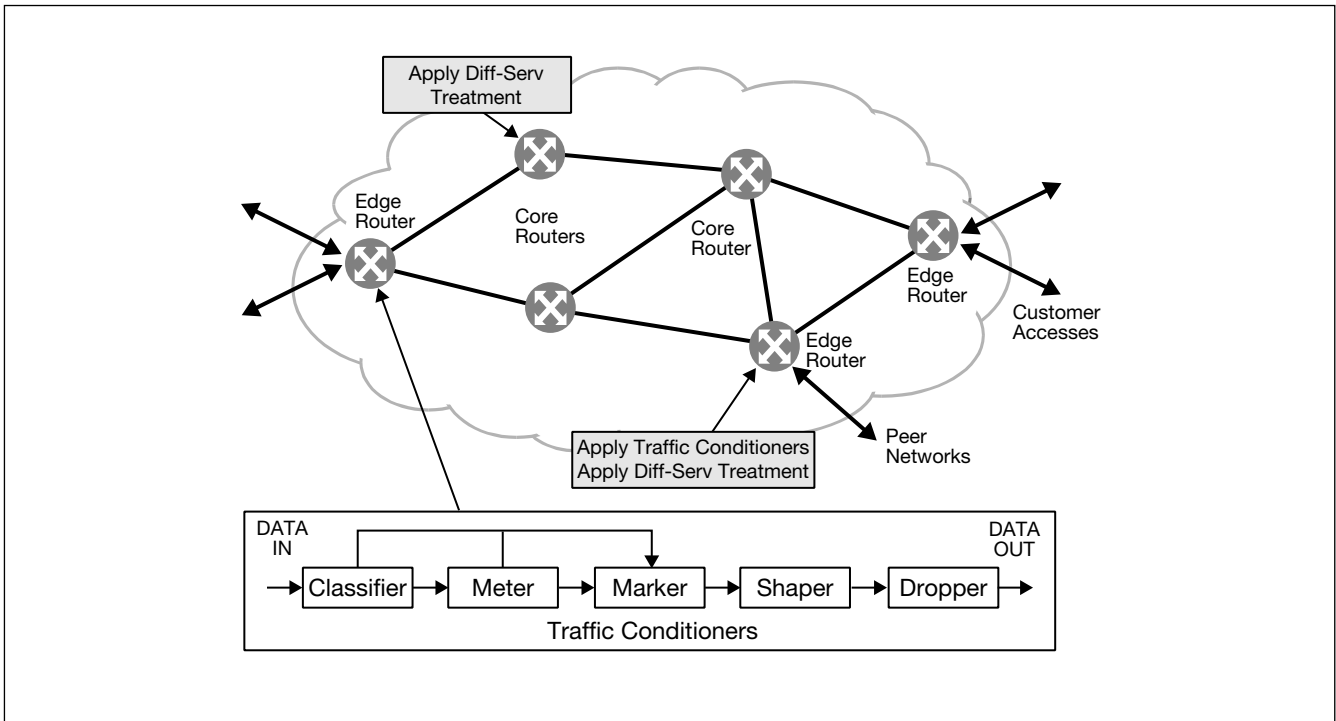


Figure 4. The edge and core routers perform specific roles.

All of these elements combined create a number of controls the network operator can use to create different levels of service, thereby enabling the network operator to provide customized service-level agreements to the customer. A number of examples of different IBS that can be created follow:

Emulated Leased Line (ELL)

The primary goal of this service is to provide an equivalent to a leased line between stub links in an IP-VPN solution. It can also be useful for other applications with strict QoS requirements such as throughput, loss, delay, and jitter. One such application is Voice over IP (VoIP) trunks between Voice Gateways (VoGW). This service provides quantitative guarantees on delay and congestion and uses the EF PHB. The traffic conditioning controls the traffic ingress to the network and only permits traffic within an allowed peak rate to pass to a specific destination.

Emulated Virtual Leased Line (EVLL)

The goal of the EVLL service is similar to the ELL service but with looser delay, jitter, and loss commitments. Although the service has looser characteristics, it can still be a quantitative service. This service can be used, for example, in real-time streaming and interactive applications. This service uses one AF PHB, with the traffic conditioners being defined point-to-point like ELL. The service permits burstier data under something like a token bucket model, and excess data is permitted; however, it is dropped if there is congestion.

Better than Best Effort (BBE)

This service creates one or more service classes that are given preferential treatment compared to the basic “best effort” service. Each class is allocated some share of the network resources (via weights in schedulers, drop classes, etc.), and better service (e.g., lower delay, lower drop rate) is provided in classes with lower loading. This service typically uses one or more of the AF PHB and is not restricted to any specific destinations, since it is only aiming to provide a qualitative gain over the lower classes. The traffic conditioning permits a bursty profile for the user, and excess data is permitted, but with a higher drop probability.

Best Effort (BE)

This is the traditional “best effort” service that is currently offered. Data to any destination is allowed and is handled in the best way that the network allows, given its other service commitments. There are typically no guarantees on throughput and loss rates, although the operator is expected to dimension the network resources to provide a certain minimum level of network throughput. This service typically uses the DE PHB, and the traffic conditioners are similar to those used for the BBE services.

Other Services

In addition to the aforementioned general service classes, network operators can create service classes for a specific application that they wish to support that may have rather stringent requirements. For example, a voice application has very specific requirements on delay and jitter through the network; therefore, the network operator may create a “voice” service. This service may have similar restrictions to the ELL service, but the network operator can differentiate the voice service from the ELL service based on other characteristics such as performance guarantee under network failure conditions.

Again, it is important to note that the end-customer services are built based on a combination of factors including:

- PHB to be applied to the traffic.
- Traffic conditioners to be applied at the network edge:
 - Classifying
 - Metering
 - Marking
 - Shaping
 - Dropping
- Network engineering to control the quality of the service. This includes aspects of network dimensioning to control the delay, jitter, and loss rate under normal conditions as well as controlling the network adaptation to failure conditions.
- Service Provision Strategy Network policy controlling access and provisioning of services.

These services have the service classification marked at the ingress. In many cases, applications require packets in both directions to receive the same type of service. For example, a premium service would suffer significantly increased round-trip times if TCP acknowledgements were not marked for that service. To enable two-way service, it is necessary to use the Diff-Serv mechanisms on both ends of the session and to mark the reverse direction packets with the appropriate service class, too. That is, the profiles must be set up to establish the relevant DSCP not only for the traffic from this node, but also for the traffic back to this node from the destination for the application.

Receiver Control

So far, only traffic conditioning at the ingress of the network has been mentioned. Where there are low-speed links (which is common in scenarios such as mobile access) or links with high utilization, control may be provided by the receiver of the data. Although the originator of the data may have identified the relevant priority of the data, the receiver may want to override the original classification when there is congestion as the data is aggregated on the final link towards the customer. This is critical for enabling the customer to defeat “denial of service” attacks, when their ingress link is being flooded maliciously. Customers can control filters in the network to apply to the received data, enabling them to select which packets are the most critical and should be given priority.

Remarking the DSCP

Between network operators, there are service-level agreements that define what level of traffic they are willing to exchange. The agreement must also define how the different services offered by the two operators interrelate. Since the operators might have different data treatment definitions and different IP bearer services, the DSCP may need to be re-marked at the boundary of a domain to select the most applicable PHB for the following domain. This re-marking could be performed either at the egress of the domain or the ingress of the following domain.

Multi-Protocol Label Switching (MPLS)

Another new technology being developed for use in the Internet is MPLS.

The basics of MPLS are discussed in the white paper “The Future of IP Backbone Technology” (EN/LZT 108 2098 R2). Network operators can use MPLS to provide a number of different services.

Topology-Driven MPLS

One use of MPLS is to create topology-driven paths through the network. These paths allow the IP traffic between different nodes in the network to be routed only at the ingress edge of the MPLS domain. After the first node, the data is then forwarded based purely on the attached label, rather than a routing analysis performed on the IP header. If this label is encoded into the header of various Layer 2 forwarding technologies such as ATM and frame relay, these switches then can participate in the network packet forwarding using the MPLS label.

With topology-driven MPLS, Label Switch Paths (LSPs) are created between pairs of nodes in the network. Hence, the LSPs are established according to the network node topology. The LSPs are initiated based on the routing information within the nodes; therefore, the path to the destination address will be the same for the LSP as for routed packets. If the network consists of router nodes at the edge and a core of MPLS ATM switches, then MPLS LSPs are established between each pair of routers at the edge through the ATM switches. These paths then create direct connections between the MPLS domain edge routers (which are not directly connected).

Explicit Paths and Traffic Engineering

Besides creating paths for traffic according to the network topology, MPLS also can be used to create LSPs throughout the network for specific purposes. In this case, the LSPs to be created are strictly controlled, since they are supposed to support a specific need in the network such as traffic engineering.

Each LSP created through the network is established by signaling. This signaling carries information about the required characteristics for the LSP. Since each node in the network must ensure that its part of the connection meets those requirements, it is possible to ensure that the entire LSP also meets the requirements.

The requirement characteristics of an LSP can include:

- Bandwidth, including sustained and peak data rates and maximum burst sizes
- Delay and delay variation
- Path selection

The path through the network created by MPLS can be controlled using the path selection capability of explicit routing (see Figure 5). With explicit routing, the path does not need to follow the normal routing path for that destination. Instead, the path to be taken by the LSP is specified in the signaling. In strict explicit routing, each node the LSP passes through is identified. In loose explicit routing, only selected nodes are explicitly identified. The route is “pinned” at a node through which it must pass, but the LSP may pass through other nodes between the pinned points.

Service providers can leverage the combination of the explicit routing and LSP characteristics to provide traffic engineering and dedicated service usage. Traffic engineering enhances the service delivery across the network by optimizing the usage of network resources. For example, traffic can be diverted around network congestion “hot spots” and through other nodes that are not congested.

Another use for explicit LSPs is for specific service usage, rather than for general service-class usage. For example, a customer may need a Virtual Leased Line (VLL) with even greater availability than that provided by the general Virtual Leased Line service offered by the network operator, such as for VoIP trunks between VoGWs. The Virtual Leased Line can be provided through a dedicated explicit LSP.

Both of these aspects can be considered part of the general task of traffic engineering within the network. Traffic engineering aims to optimize service delivery throughout the network by improving the network utilization. This optimization must consider aspects such as individual service-level requirements for customers. With MPLS, the operator has a choice as to whether a specific part of a service-level agreement will be provided over shared routing infrastructure, shared explicit paths with well defined characteristics, or dedicated paths.

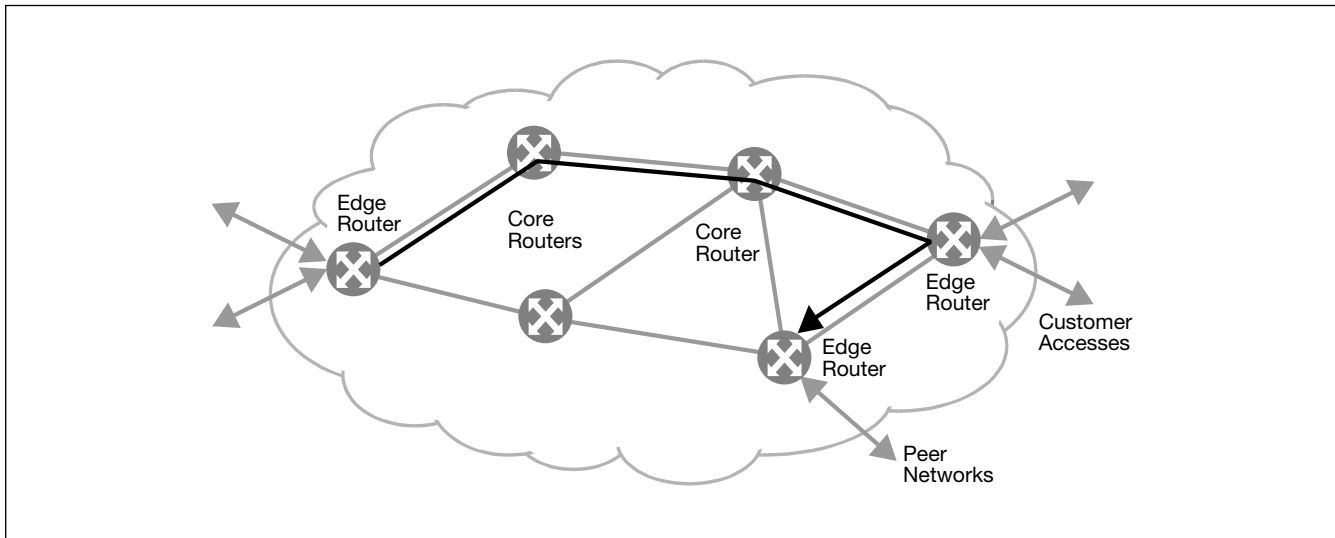


Figure 5. Explicit Routing allows a path to be directed through specific nodes, irrespective of the shortest path derived from IGP.

Within a network, multiple LSPs can be created between a pair of nodes. One reason to do this is to provide some of the capabilities of traffic engineering with alternative paths for redundancy and load distribution. Another reason is to provide trunks with different characteristics to support a range of services. For example, one LSP could be created to provide leased line services and would generally support traffic using an EF PHB. Another LSP could be established for one subclass of BBE service that would carry traffic from a set of AF PHB codepoints.

The node at the ingress edge of the LSP performs an additional role. This node controls which traffic is permitted to use this LSP (see Figure 6). In the case of an access (AS domain ingress) node, it may be the premium PHB traffic from a customer destined for a specific network. The IBS management system that controls the network resources is also responsible for the filter profiles of the edge routers that determine traffic eligibility to use an LSP.

The classifier element required to select traffic applicable for the LSP is a very similar function to that required for Diff-Serv. It is interesting to note that both functions are performed at the ingress of the domain. If the DS and MPLS domains are identical, then the same function within the ingress node may be used to perform both the DS traffic conditioning as well as the MPLS eligibility determination.

Since an LSP is extended to the MPLS domain egress, use of this LSP ensures that the data sent into it will receive the same service treatment throughout the entire LSP. Since Diff-Serv requires the same behavior at each node to get the end-to-end treatment, there is again excellent synergy between MPLS and Diff-Serv.

In many cases, an LSP carries an aggregation of many customers' flows within the network. However, it is also possible to use MPLS LSPs for other purposes.

Since MPLS creates paths through a network, and data on these paths is *not* routed at each node, MPLS effectively creates "tunnels" through the network.

These tunnels have a well-defined entrance, a well-defined exit, and a gate to control what is allowed into the tunnel. Once in the tunnel, there are no branch exits since the data is not routed at intermediate nodes. Since only the network operator can create paths, malicious users cannot create additional tunnel entrances to merge traffic into network tunnels or disrupt the network. The tunnel entrances and exits exist within the operator's network, but the entrance and exit ramps may also be extended out further via stub links to the customer's domain. In other words, data leaving the tunnel is directed down the exit ramp. Data entering the tunnel must pass the gating criteria; this can include the requirement that the data was received from the entrance ramp. Even if the data is received from the entrance ramp, however, there are still further gating

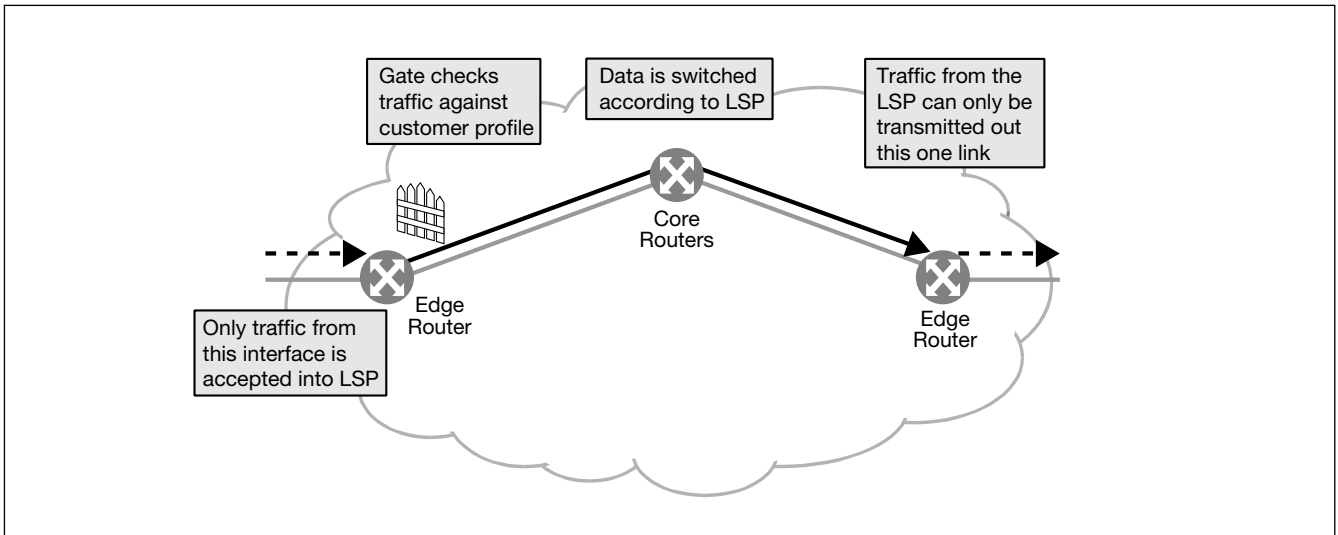


Figure 6. LSP is used within the network, with restrictions on ingress and egress traffic.

criteria in the form of traffic-conditioning actions (which are applied at the actual tunnel entrance). These properties allow the LSPs to support additional functions.

Consider the case of a Virtual Private Network (VPN). If the customer has a private network with its own private addressing plan, the private addresses cannot be used within the operator's network. The network operator can provide a VPN capability to link together the customer's sites using Virtual Network Routers (VNRs) at the edge of the network. These routers appear as part of the customer's network and hide the operator's network. If the data between two of the customers' sites is placed into a tunnel by the VNR at the ingress access node, it can be passed through the network to the VNR at the egress node. The tunnel endpoints are not simply the access nodes, but rather the VNRs within these nodes. The virtual routers may perform routing within the customer's virtual private network. Alternatively, the VNR may not need to be involved in the routing of the customer data if there are dedicated tunnels to the other sites and extended entrance and exit ramps to the customer.

Specific applications such as VPN can also benefit from the QoS capabilities of the MPLS application. As is done for traffic engineering, LSPs with specific characteristics can be established to provide support for service-level agreements within the VPN.

Mapping MPLS Classes to ATM Switching Hardware

As mentioned earlier, MPLS can be supported on different link layer technologies. ATM is a technology that provides high-speed forwarding using a label-switching paradigm. Even before MPLS was developed, this capability of ATM hardware made it a common choice for high-speed backbone networks. Furthermore, ATM is designed to support multiple service classes with different transit delay and delay variation requirements through the network. The capabilities of ATM hardware contribute to its ability to provide an excellent platform for an MPLS network.

LSPs in an MPLS environment are similar in many ways to ATM connections, but there are also some differences. Let us examine more closely how MPLS and ATM differ in the network, and what effects this has on the ability of an ATM switch to support MPLS services.

Earlier in this paper, there was discussion of two modes of MPLS: topology-driven and explicit paths. These two different uses of MPLS must be examined individually.

First, let us examine the topology-driven MPLS connections. In a traditional routed network, the neighboring nodes exchange information through a routing protocol that is propagated through the network. Data is routed through the intermediate nodes toward the destination node according to the routing table in each node. The MPLS application uses the network topology information learned via the routing protocol to establish cut-through LSPs toward the destination networks. Routing information still is exchanged only between neighbors, but now data paths exist toward non-neighboring nodes. The route for the cut-through paths is controlled by normal routing decisions within the network. Hence, as next-hop routing conditions change, the paths are re-established.

Since these cut-through paths share the same links as the direct neighbor paths, there must be some resource sharing of the link between the traffic to the directly connected neighbors and these cut-through paths.

In ATM, each connection is characterized throughout the network. That is, each connection has a set of parameters that control the connection characteristics such as peak cell rate, maximum burst size, and delay. On a link that contains multiple connections of different classes, the scheduling of the data is typically performed according to a priority for the service class and the traffic shaping for that individual connection. Traffic on higher priority classes is typically policed at the ingress to ensure the network links have sufficient capacity for the connections, and the lower priority connections are not starved out.

In the case of MPLS, it is not feasible to determine parameters for each individual path, since this varies over both short and long periods according to the fractal nature of the traffic. Hence, it is not feasible to apply traffic shaping for each connection to control the scheduling. Likewise, the traffic at the ingress cannot be policed because of the variation in traffic. Instead, the traffic can be characterized for each link between pairs of neighboring nodes, and this characterization is

then applicable for the aggregate of MPLS connections over that link between the nodes. Scheduling then is applied to the aggregate of connections on the link that share the same service class according to some form of weighted fair queuing. This is to ensure that the appropriate share of the link capacity is allocated to each service class and to prevent starvation of any service class. Of course, excess capacity can be shared further between service classes that are temporarily oversubscribed for their portion.

The weighted access to link capacity shared by multiple connections of the same service class is one of the secrets to providing good support for MPLS on ATM hardware. These capabilities are even more important when the node is supporting both ATM and MPLS applications at the same time over a single link. Both the MPLS and ATM services must be serviced to meet their commitments without interfering with each other. This is a capability known as Ships In the Night (SIN).

MPLS Protocol for Topology-Driven Mode

Topology-driven mode enables switching hardware (e.g., ATM switches) to perform forwarding through the network. Thus, it is applicable in networks in which the switching capacity greatly exceeds that of the routing capacity. Networks based on traditional routing hardware may not use topology-driven mode since the difference in forwarding capacity could be minor.

There is only one protocol proposed for establishing topology-driven LSPs within a network. This protocol is known as the Label Distribution Protocol (LDP). This protocol provides the basic capabilities for controlling establishment of paths to destination networks.

LDP uses a general message format that allows additional parameters to be introduced into the protocol at a later stage. One parameter that has been recently proposed is the PFC (Per Hop Behavior Forwarding Class) parameter. This parameter is used to specify a PFC, which will control the scheduling behavior within the node for a LSP. This is a similar concept to the

PHB supported in a Diff-Serv router. Establishing multiple LSPs between nodes within the domain supports multiple Diff-Serv PHBs, and therefore provides a fully Diff-Serv-compliant domain.

The signaling scheme allows some flexibility in the options for establishing LSPs for support of multiple service classes. One option is to have a separate LSP established for each service class required toward that destination edge. Another mechanism is to use one LSP for a group of service classes such as the group of assured forwarding classes AF1x. In this case, the subclasses for the different drop probabilities are mapped to the LSP using the CLP bit in the ATM header. Although this bit only provides two drop classes, there is a draft proposal for standardizing the mapping from the drop subclasses to the CLP bit.

For explicit paths, another factor is introduced. The characteristics required for the LSP are indicated via the signaling used in the establishment of the LSPs. This information can be used by the node to ensure that sufficient resources are available to satisfy the request and to allocate the resources required to the connection. This ensures that services cannot be overallocated within the network causing the network to fail to meet its service commitments.

MPLS Protocol for Explicit Mode

There are two protocols currently proposed for establishing explicit LSPs: namely CR-LDP and RSVP.

The Constraint Based Routing LDP (CR-LDP) protocol is an extension of the LDP protocol, which provides the necessary information for control of the LSP routing and specifies the other characteristics required for the connection, such as data rates. It provides additional parameters over the basic LDP protocol for:

- CDR = committed data rate
- PDR = peak data rate
- CBT = committed burst tolerance

The RSVP protocol for MPLS has adaptations from traditional RSVP to address the scalability issues that make RSVP for Int-Serv unsuitable for core network nodes. The other adaptations to RSVP provide the label distribution and traffic engineering capabilities. The main aspects of RSVP for specifying the flow characteristics are retained. The controlled load service

of RSVP is used, with the characteristics required by the connection defined by the RSVP flowspec parameters.

Note that both the RSVP and CR-LDP protocols provide the capability to control the route of the LSP. That is, the LSP can be forced to take a specific path through the network.

MPLS and ATM Together

If the ATM node is running SIN with a combination of ATM and MPLS traffic, it is often important for the network operator to understand the relationship between ATM service categories and MPLS service classes within the node.

Although there is not necessarily a one-to-one mapping between MPLS service classes and ATM service classes, there will typically be a number of MPLS classes with an equivalent ATM service class, at least in terms of priority handling within the switch. A typical relationship between the different services in an ATM switch with SIN is shown in the text box — Service Mapping Example.

Example of Mapping Between Services	
ATM Service Category	Diff-Serv PHB
CRB	EF
rt-VBR	AF priority 1
nrt-VBR	AF priority 2, CS
	AF priority 3
ABR, GFR, UBR W-UBR	AF priority 4

Of course, the actual mapping required by a network operator providing both ATM and MPLS services may be different from that shown. Some comparisons of the different classes highlighted in this table are appropriate.

ATM connections are generally excellent for providing well-defined connections with quantified characteristics across an ATM network. They are typically point-to-point connections (although point-to-multipoint connections are also possible). Typical characteristics of ATM service classes are shown in the table on the following page.

Diff-Serv support is provided by establishing LSPs with a specified PFC, which in turn controls the characteristics of the data path. The IP Bearer Services EF and one or two AF classes typically have characteristics similar to the ATM CBR and rt-VBR connections. More than likely they would have very similar restraints, such as the point-to-point nature and strict traffic profiles that are applied. There typically would be relatively small buffers allocated to these services.

The other AF classes typically would have greater buffer allocations to allow for the less predictable nature of the traffic and the greater delay that the lower priority classes experience, with correspondingly looser quantified or only qualitative levels of service quality. There also would be different levels within the classes for congestion control. Finally, the BE service would exist to provide the same service class provided today.

The ATM service classes provide scheduling, shaping, and policing mechanisms to meet the ATM service-class requirements. These same mechanisms are used to provide the traffic conditioning and scheduling control for the IP traffic within the node.

The Diff-Serv classes by themselves control the treatment of data within the nodes, creating multiple service classes. However, to meet all aspects of customer SLAs, Diff-Serv capabilities of the nodes are used in conjunction with MPLS and ATM connections to provide the necessary network engineering dimension.

The CR-LDP and RSVP service classes provide the same capabilities; that is, to support explicit paths through the network to provide service assurances. The service classes provide the same sort of assurances that ATM can provide (where similar traffic conditioners are applied).

Typical Characteristics for ATM Service Categories	
CBR	<ul style="list-style-type: none"> - Quantified low latency and delay variation (high priority) - Small buffer allocation - Peak rate defined (excess traffic discarded) - Shaped/policed at network edge
rt-VBR	<ul style="list-style-type: none"> - Quantified latency and delay variation (not as low as for CBR) - Larger (though still small) buffer allocation - Sustainable and peak rates defined (excess traffic discarded) - Shaped/policed at network edge
nrt-VBR	<ul style="list-style-type: none"> - Even greater buffer allocation (medium priority) - Greater business permitted - Sustainable and peak rates defined - Excess traffic may not be discarded
UBR	<ul style="list-style-type: none"> - Not qualified for delay and delay variation (low priority) - Typically weighted access to unallocated bandwidth
ABR	<ul style="list-style-type: none"> - Tight feedback control loop to control the data rate based on available network buffering resources - Typically weighted access to unallocated bandwidth
GFR (New Service Category)	<ul style="list-style-type: none"> - Guaranteed frame rates for data traffic - Typically weighted access to unallocated bandwidth

Applications for QoS

We have mentioned the need for service-level agreements between the customer and the network provider. These agreements specify the amount of traffic of each service class that is permitted for that customer between different addresses.

For service classes that include traffic to any destination, the LSPs created in the links between each node can be dimensioned according to the expected traffic levels in that link in the network. In this manner, the network can be engineered to cater to the traffic within the network according to the service-level agreements.

MPLS can offer advantages over ATM even in small networks, since MPLS can provide closer alignment of service classes to Diff-Serv than the ATM service categories. Also, the dynamic nature of the IP traffic makes appropriate allocation of the capacity between nodes required for ATM mesh connections difficult to configure efficiently. The shared access to bandwidth permitted with MPLS connection styles (as shown in Figure 7) provides a more effective network configuration.

MPLS using RSVP or CR-LDP can create LSPs throughout the network that are dimensioned to provide the agreed traffic levels for point-to-point service classes. This allows MPLS to be used for traffic engineering — a task that previously would have required an ATM switching layer. The traffic engineering and VPN applications that will make use of QoS will be covered in upcoming white papers.

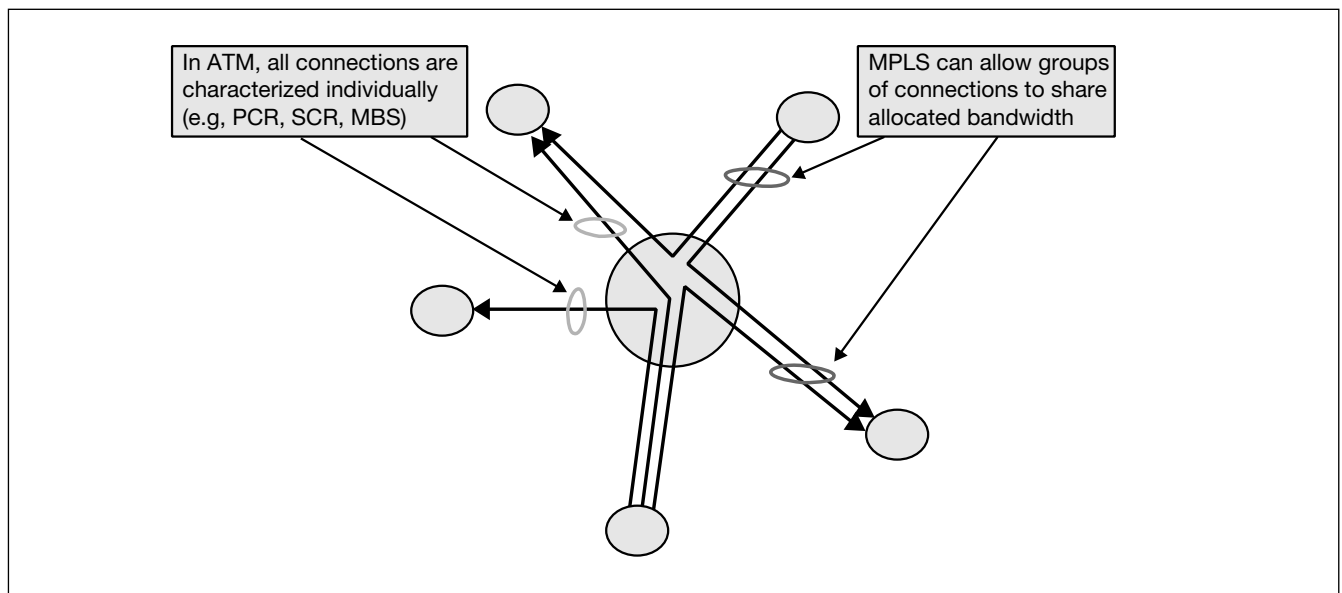


Figure 7. MPLS-style connections allow sharing of configured link bandwidth, unlike ATM-style connections. This can provide more efficient utilization.

Conclusion

Diff-Serv is a basic building block for providing QoS within the Internet. MPLS has good synergy with Diff-Serv because of some similarities in their elements, such as the role of the domain edge and the application of a treatment throughout the domain. The combination of MPLS and Diff-Serv enables the operator to provide a network capable of supporting services with defined characteristic requirements throughout the network and an ability to deliver them according to service-level agreements.

This is enhanced further by the ability of MPLS LSPs to be used for specific support of other services such as VPNs. These functions, along with other features such as policy servers for network control, will enable the delivery of new network services into the future, meeting new customer needs.

TERM	MEANING	
ABR	Available Bit Rate	An ATM service class with tight feedback to control the rate at which the stations transmit to maximize utilization while minimizing congestion
AF	Assured Forwarding	PHBs defined in Diff-Serv with multiple levels of relative delay and probability of delivery
ATM	Asynchronous Transfer Mode	
BB	Bandwidth Broker	A policy control point that determines policy for allocation of network resources
BBE	Better than Best Effort	An end-user service that provides better probability of delivery and/or lower latency than the traditional "best effort" service. There may be multiple levels of this service
BS	Better Service	A defined treatment available within a Diff-Serv domain
CAC	Connection Admission Control	The node performs checks to ensure the requested resources in that node are available before allowing the service to be accepted
CBQ	Class Based Queuing	A scheduling algorithm controlled according to the service class of the data
CBR	Constant Bit Rate	
CBT	Committed Burst Tolerance	
CDR	Committed Data Rate	
CR-LDP	Constraint Based Routed LDP	
CS	Class Selector	
DE	Default Behavior	The traditional "best effort" handling within a node
Diff-Serv	Differentiated Services	
EF	Expedited Forwarding	
ELL	Emulated Leased Line	
ER	Explicit Rate	
EVL	Emulated Virtual Leased Line	
GFR	Guaranteed Frame Rate	
IBS	IP Bearer Service	A bearer service for IP traffic conditioners applied at the ingress to regulate usage of the service within the network
Int-Serv	Integrated Services	
LDP	Label Distribution Protocol	
LSP	Label Switch Path	
MPLS	Multiprotocol Label Switching	
NC	Network Control	
nrt-VBR	Non-Real-Time Variable Bit Rate	
PCP	Policy Control Point	The element in the network that makes policy decisions
PDR	Peak Data Rate	
PEP	Policy Enforcement Point	The network element that enforces policy as determined by the PCP
PHB	Per Hop Behavior	A treatment to be applied to data at a node
PS	Policy Server	
RSVP	Resource Reservation Protocol	
rt-VBR	Real-Time Variable Bit Rate	
SIN	Ships in the Night	When ATM and MPLS applications are used together in a node, they should not interfere with each other
SLA	Service Level Agreement	
SVC	Switched Virtual Connection	An ATM connection that is controlled by the host
UBR	Unspecified Bit Rate	
VC	Virtual Connection	
VNR	Virtual Network Router	A public network router appears to be part of the private network and participates in the private network routing, while retaining separation of the two networks
VP	Virtual Path	
WFQ	Weighted Fair Queuing	A scheduling mechanism that allocates available capacity between service classes according to a defined weighting
WFS	Weighted Fair Share	The share of the available capacity allocated to this particular service class
VoIP	Voice over IP	
WWW	World Wide Web	
W-UBR	Weighted UBR	

© Ericsson Datacom Inc., 2000. All rights reserved.
The Ericsson logotype is a registered trademark and IPulse,
MINI-LINK, and PHONE DOUBLER are trademarks
of Telefonaktiebolaget LM Ericsson. All other brand and
product names are trademarks or registered trademarks
of their respective holders. Information in this document is
subject to change without notice. Ericsson Datacom Inc.
assumes no responsibility for any errors that may appear in
this document.

Printed in U.S.A.