

PROCEEDINGS OF SPIE

[SPIDigitalLibrary.org/conference-proceedings-of-spie](https://spiedigitallibrary.org/conference-proceedings-of-spie)

Current status on quality of service in IP

Angela Chiu

Angela L. Chiu, "Current status on quality of service in IP," Proc. SPIE 3842, Internet II: Quality of Service and Future Directions, (5 November 1999); doi: 10.1117/12.368318

SPIE.

Event: Photonics East '99, 1999, Boston, MA, United States

Current Status of Quality of Service in IP

Angela L. Chiu¹

AT&T Labs, 200 Laurel Ave. Room C4-3A22, Middletown, NJ 07733

ABSTRACT

The Internet has grown rapidly in the last several years. This is largely due to the simple, flexible, and robust connectionless nature of the Internet Protocol (IP). The Internet architecture has been successful up to this point, and the best effort service paradigm has been adequate. However, with increasing demands of supporting voice, video, mission critical data on IP, the best effort paradigm without differentiating traffic according to application requirements can not meet the market demands. In this paper, we summarize and compare two major models in IP Quality of Service (QoS): Integrated Services Model and Differentiated Services Model proposed in Internet Engineering Task Force (IETF), the main Internet standard committee. We also discuss other related areas in IETF, e.g., Multiprotocol Label Switching and Resource Reservation Protocol in terms of their impacts on supporting QoS in IP.

Keywords: IP Quality of Service, Integrated Services, Differentiated Services, Resource Reservation Protocol, Multiprotocol Label Switching

1. INTRODUCTION

The Internet has grown rapidly in the last several years. This is largely due to the simple, flexible, and robust connectionless nature of the Internet Protocol. The Internet architecture has been successful up to this point, and the best effort service paradigm has been adequate. The current Internet architecture evolved from a network that was providing reliable connectivity, despite link failures, to thousands of users. Now, the Internet is an international network with millions of users. However, some of the control mechanisms that were originally implemented to prevent congestion collapse and depend on cooperative users are still in use. The network infrastructure did not incorporate any means for differentiation other than by severely under-utilizing specific portions of the network. Over-provisioning network bandwidth and keeping the network lightly loaded in order to support differentiated services for select customers is not a cost-effective solution, and cannot be achieved at all times. Furthermore, for a commercial network, relying on cooperation by many disparate users for congestion control is not desirable. It cannot be assumed that users will cooperatively slow down their transmission rates when significant congestion is detected. Assuming cooperative behavior leads to several problems, especially when the network has no means of isolating traffic from different users or enforcing cooperation [12]. Moreover, the best effort service paradigm does not permit users to obtain better service, no matter how critical their requirements are and no matter how much they may be willing to pay for better service. Clearly, with the increased use of the Internet for commercial purposes, this model of no differentiation is not satisfactory since the ISPs do not have the means to meet an existing market demand.

During the last few years, introducing Quality of Service into the current Internet architecture has been intensively discussed and explored within the industry and standard committees. A few working groups in IETF have been formed to address QoS related issues. In general, QoS approaches can be divided into two main categories: flow-based and class-based. Flow-based QoS is designed to provide QoS assurance for each individual IP flow. An IP flow is defined as a set of packets matching a particular profile. A profile can be defined for example in terms of source/destination IP addresses, IP protocol, source/destination port numbers, and Type of Service (ToS) requirements. A typical example of flow-based QoS is the Integrated Services (Intserv) model [22][23] proposed by the Intserv working group of IETF. It provides fine granularity and flexibility in specifying flows. However, it requires storing and processing states on a per-flow basis at each network node and an end-to-end signaling protocol such as Resource Reservation Protocol (RSVP) [4]. This raises concerns regarding the scalability of the approach due to the fact that the number of active flows in a large backbone network can be quite large. In order to avoid the scalability

¹ Correspondence: Email: alchiu@att.com, Tel. (732) 420-2290.

problem with flow-based QoS, class-based QoS, which is also referred as Class of Service (CoS), is proposed to provide differentiated service for each class. An example is the Differentiated Services (Diffserv) model [2] proposed by the Diffserv working group of IETF. Generally speaking, in the class-based QoS model packets are marked as different classes and get served accordingly. It is much more scalable, but it does not have the fine granularity of the flow-based approach. There is no QoS assurance for each individual flow within a class. Note that it is possible and sometimes necessary for a network to have both flow-based and class-based QoS, as we will see in the later sections.

This document attempts to provide a survey on current activities and efforts on providing QoS in IP services in the Internet community. The rest of the document is organized as follows. In sections 2 and 3, we first describe the two approaches the IETF proposed for providing QoS in the Internet: Intserv and Diffserv, respectively. In section 4, we summarize QoS approaches in IP over ATM. In section 5, we describe a relatively new direction in IP called Multiprotocol Label Switching (MPLS) and its role in providing QoS. Finally, we summarize in section 6.

2. INTEGRATED SERVICES

There are two main service models of QoS proposed by the Intserv working group: guaranteed service and controlled-load service. The QoS signaling protocol proposed is RSVP for supporting either of the two service models. We describe each of these related efforts in the following subsections.

2.1 Guaranteed Service

Guaranteed service [22] guarantees that packets will arrive within the guaranteed delivery time and will not be discarded due to queue overflows, provided the flow's traffic stays within its specified traffic parameters. This service is intended for applications which need a firm guarantee that a packet will arrive no later than a certain time after it was transmitted by its source. For example, some audio and video "play-back" applications are intolerant of any packet arriving after their play-back time. Applications that have hard real-time requirements will also require guaranteed service. This service does not attempt to minimize the jitter (the difference between the minimal and maximal packet delays); it merely controls the maximal queueing delay.

Guaranteed service is invoked by specifying the desired traffic parameters (TSpec) and service (RSpec) to the network element using RSVP.

2.2 Controlled-Load Service

Controlled-load service [23] provides the client data flow with a quality of service closely approximating the QoS that same flow would receive from an unloaded network element, but uses capacity (admission) control to assure that this service is received even when the network element is overloaded.

The controlled load service is intended to support broad classes of applications which have been developed for use in today's Internet, but are highly sensitive to overloaded conditions. Important members of this class are the "adaptive real-time applications". These applications have been shown to work well on lightly loaded networks, but to degrade quickly under overloaded conditions. The controlled-load service does not accept or make use of specific target values for control parameters such as delay or loss. Instead, acceptance of a request for controlled-load service is defined to imply a commitment by the network element to provide the requestor with service closely equivalent to that provided to uncontrolled (best-effort) traffic under lightly loaded conditions.

A flow receiving controlled-load service at a network element may expect to experience:

- Small average packet queueing delay over all timescales significantly larger than the "burst time". The burst time is defined as the time required for the flow's maximum size data burst to be transmitted at the flow's requested transmission rate, where the burst size and rate are given by the flow's TSpec.
- Small congestion loss over all timescales significantly larger than the "burst time" defined above. In this context, congestion loss includes packet losses due to shortage of any required processing resource,

such as buffer space or link bandwidth. Although occasional congestion losses may occur, any substantial sustained loss represents a failure of the admission control algorithm.

The controlled-load service is invoked by specifying the data flow's desired traffic parameters (TSpec) to the network element using RSVP.

2.3 RSVP

RSVP [4] was proposed as a resource reservation setup protocol designed for Intserv, but not exclusively from being used by other applications. A host uses RSVP to request a specific Quality of Service from the network, on behalf of an application data stream. RSVP carries the request through the network, visiting each node the network uses to carry the stream. At each node, RSVP attempts to make a resource reservation for the stream.

To make a resource reservation at a node, the RSVP daemon communicates with two local decision modules, admission control and policy control. Admission control determines whether the node has sufficient available resources to supply the requested QoS. Policy control determines whether the user has administrative permission to make the reservation. If either check fails, the RSVP program returns an error notification to the application process that originated the request. If both checks succeed, the RSVP daemon sets parameters in a packet classifier and packet scheduler to obtain the desired QoS. The packet classifier determines the QoS class for each packet and the scheduler orders packet transmission to achieve the promised QoS for each stream.

An elementary RSVP reservation request consists of a "flowspec" together with a "filter spec"; this pair is called a "flow descriptor". The flowspec specifies a desired QoS. The filter spec, together with a session specification, defines the set of data packets, i.e., the "flow" to receive the QoS defined by the flowspec. The flowspec is used to set parameters in the node's packet scheduler or other link layer mechanism, while the filter spec is used to set parameters in the packet classifier. Data packets that are addressed to a particular session but do not match any of the filter specs for that session are handled as best-effort traffic.

The flowspec in a reservation request generally includes a service class and two sets of numeric parameters: (1) an "Rspec" (R for 'reserve') that defines the desired QoS, and (2) a "Tspec" (T for 'traffic') that describes the data flow. See RFC2205 for further details on Tspec, Rspec and the reservation models/styles.

RSVP is designed for multicast applications as well as unicast ones. It can scale to very large multicast groups because it uses receiver-oriented reservation requests that merge as they progress up the multicast tree. The reservation for a single receiver does not need to travel to the source of a multicast tree; rather it travels only until it reaches a reserved branch of the tree.

RSVP maintains "soft" states² in routers and hosts with sources sending PATH messages periodically. This provides graceful support for dynamic membership changes and automatic adaptation to routing changes.

RSVP is also designed to utilize the robustness of current Internet routing algorithms. RSVP does not perform its own routing; instead it uses underlying routing protocols to determine where it should carry reservation requests. As routing changes paths to adapt to topology changes, RSVP adapts its reservation to the new paths wherever reservations are in place. RSVP runs over IP, both IPv4 and IPv6.

Recently, IETF started a new working group called RSVP admission policy (RAP) to address an important aspect of admission control. That is, network managers and service providers must be able to monitor, control, and enforce use of network resources and services based on policies derived from criteria such as the identity of users and applications, traffic/bandwidth requirements, security considerations, and time-of-day/week. The framework document [25] focuses on policy-based control over admission control using RSVP as an example of the QoS signaling mechanism. Even though the focus of the work is on RSVP-based admission control, the document outlines a framework that can provide policy-based admission control in other QoS contexts.

Moreover, the RAP working group proposed the Common Open Policy Service (COPS) Protocol [3], a simple client/server model for supporting policy control over QoS Signaling Protocols. It is a simple query and response protocol that can be used to exchange policy information between a policy server (Policy Decision Point or PDP) and its clients (Policy Enforcement Points or PEPs). Policy clients are expected to

² States are only valid for some fixed time interval.

be RSVP routers that must exercise policy-based admission control over RSVP usage. It assumes that at least one policy server exists in each controlled administrative domain.

2.4 Remarks

Intserv is a flow-based QoS architecture that provides each individual flow with certain level of guaranteed service or controlled load service. Several barriers to deploy Intserv in today's network have been recognized. The reliance of Intserv on per-flow state and per-flow processing is an impediment to its deployment in the Internet at large, and in particular in large carrier networks. Additionally, the fairly complex RSVP signaling is supposed to originate from hosts, which as of yet are not RSVP enabled in large numbers. Recently, attention has shifted to Differentiated Services. Diffserv promises to expedite the realization of QoS enabled networks by offering a significantly simpler alternative to Intserv, which eliminates scalability concerns and which can be implemented and managed in large networks, without requiring end-to-end deployment.

3. DIFFERENTIATED SERVICES

The main idea of Diffserv is to minimize state and per-flow information in core routers by placing all packets in fairly broad classes at the edge of the network. A single byte in each packet is used to do this, called the DS byte (the ToS [Type of Service] byte in IPv4 and the Traffic Class byte in IPv6), which can be set by the end station or the ingress edge router to indicate the class of service desired. Thus, for example, the edge router could detect individual flows based on IP address or application type, and set the DS byte appropriately. All other routers (i.e., the core routers) now only need to read the DS byte in each packet and take the appropriate action, such as placing the packet in a certain outgoing queue, thereby achieving aggregated flows by traffic class. This minimizes the work of the core routers and should facilitate scaling.

In the architecture document [2], it is proposed to implement an end-to-end service by concatenating per-hop behaviors along the routing path, together with service provisioning, traffic classification and conditioning. Per-hop behavior (PHB) is defined as the forwarding behavior a packet receives at a given network node, as implemented by queue service disciplines and queue management disciplines.

Service provisioning can be either static or dynamic. Static provisioning can be achieved by negotiating service level agreement (SLA) and traffic conditioning agreement (TCA) between a user and a provider during service subscription time. Dynamic provisioning can be achieved by using some signaling protocol to signal service requirements in real time, e.g., RSVP.

The packet classification policy identifies the subset of traffic which may receive a differentiated service by being conditioned and/or mapped to one or more behavior aggregates. Traffic conditioning performs metering, shaping, policing and/or re-marking to ensure that the traffic entering the DS domain conforms to the rules specified in the TCA, in accordance with the domain's service provisioning policy. The extent of traffic conditioning required is dependent on the specifics of the service offering, and may range from simple ToS byte re-marking to complex policing and shaping operations.

In the following subsections, we describe the proposals on ToS Byte definition and PHBs.

3.1 A Differentiated Services Definition of the ToS Byte

The authors in [2] propose to redefine the structure for the 8-bit IPv4 TOS octet and the 8-bit IPv6 Class octet [IPv6], and re-label the field as the "DS field". The first six bits of the DS field are used as a codepoint (DSCP) to select the PHB that the packet experiences at each node. A two-bit currently unused (CU) field is reserved for other purposes. The value of the CU bits is ignored by PHB selections.

The DS field structure is presented below:

0	1	2	3	4	5	6	7	where:
+---+---+---+---+---+---+---+---+DSCP: Diffserv codepoint								
DSCP						CU		CU: currently unused
+---+---+---+---+---+---+---+---+								

In a DSCP value notation 'xxxxxx' (where 'x' may equal '0' or '1') used in this document, the left-most bit signifies bit 0 of the DS field (as shown above), and the right-most bit signifies bit 5.

With some exceptions for codepoints 'xxx000'³, the mapping of codepoints to PHB must be configurable. Nodes MAY rewrite the DS field as needed to provide a desired local or end-to-end service. See [2] for further discussion of re-marking. Specifications of DS field translations at DS boundaries are the subject of service level agreements between providers and users. Standardized PHBs allow providers to build their services from a well-known set of packet forwarding treatments that can be expected to be present in the equipment of many vendors.

Packets received with an unrecognized codepoint should be forwarded as if they were marked for the Default behavior, and their codepoints should not be changed. Such packets MUST NOT cause the network node to malfunction.

3.2 Operational Model: Per-Hop Behavior

The differentiated services model is that a router has a (possibly large) set of parameters that can be used to control how packets are scheduled onto an output interface (e.g., N separate queues with settable priorities, queue lengths, round-robin weights, drop algorithm, drop preference weights and thresholds, etc). Router vendors are free to offer whatever parameters and capabilities they feel are useful or marketable. An ISP is free to configure those parameters in whatever way it feels is appropriate for their service offerings and traffic engineering objectives. The router's capabilities and its particular configuration determines the different ways packets can be treated. The DSCP marking then selects which treatment a particular packet receives. There is no requirement that different ISPs use the same router mechanisms or configuration or that any value of the DSCP map to a particular packet forwarding behavior. However, to the extent that the per-hop behaviors are used to implement end-to-end services, it may be undesirable for the "same" per-hop behavior to be selected by different DSCP values in different ISPs. To avoid this, it is expected that over time certain common per-hop behaviors will evolve (ones that are particularly useful for implementing end-to-end services) and that these will be associated with particular code points in the DS byte. It is not suggested that PHB patterns be selected and assigned without adherence to a larger context.

3.3 Examples of Per-Hop Behavior

Only those per-hop behaviors that both have some degree of orthogonality and have been implemented, deployed, and shown to be useful should be standardized. Detailed requirements on PHB are described in [2]. Two per-hop behaviors are already in widespread use and have been proposed for standardization: Default (DE) is the common, best-effort forwarding available in today's Internet and already standardized in RFC-1812. Expedited Forwarding (EF) is a "high priority" behavior typically used for network control traffic such as routing updates. The code points chosen for these are compatible with existing practice: PHB field value = 000000 for Default (DE), and 101100 for Expedited Forwarding (EF). Another PHB group called Assured Forwarding (AF) [9] have also been standardized. The three behaviors are defined as follows.

- Default: an incoming packet goes to the tail of a queue that is serviced in FIFO order when the output link is free. Thus if a continuous stream of DE-marked packets is sent into a device, packets will emerge in the same order they enter. It is not required that all arriving packets are seen at the output, but it is required that the aggregate of packets with this marking are not completely starved. Delay requirements are as soon as possible, and the corresponding bandwidth requirements are as much as possible (subject to the other constraints of the router's scheduling system). The Default behavior is designed to closely approximate the best-effort behavior of existing routers.
- Expedited Forwarding [11]: The EF PHB is defined as a forwarding treatment for a particular Diffserv aggregate where the departure rate of the aggregate's packets from any Diffserv node must equal or exceed a configurable rate. The EF traffic should receive this rate independent of the intensity of any other traffic attempting to transit the node. It should average at least the configured rate when

³ The DS field values of 'xxx000|xx', or DSCP = 'xxx000' and CU subfield unspecified, are reserved as a set of Class Selector Codepoints.

measured over any time interval equal to or longer than a packet time at the configured rate. (Behavior at time scales shorter than a packet time at the configured rate is deliberately not specified.) The configured minimum rate must be settable by a network administrator (using whatever mechanism the node supports for non-volatile configuration). Several types of queue scheduling mechanisms, e.g., priority queueing or weighted fair queueing [19], may be employed to deliver the forwarding behavior in the EF PHB. The Expedited Forwarding behavior may be used to implement services requiring low delay and low jitter, and may also be useful for network control traffic.

- **Assured Forwarding:** AF PHB group is a means for a provider DS domain to offer different levels of delivery assurances for IP packets received from a customer DS domain. Four AF classes are defined, where each AF class is in each DS node allocated a certain amount of forwarding resources (buffer space, bandwidth). IP packets that wish to use the services provided by the AF PHB group are assigned by the customer or the provider DS domain into one or more of these AF classes according to the subscribed services. Within each AF class IP packets are marked (again by the customer or the provider DS domain) with one of three possible drop precedence values. In case of congestion, the drop precedence of a packet determines the relative importance of the packet within the AF class. A congested DS node tries to protect packets with a lower drop precedence value from being lost by preferably discarding packets with a higher drop precedence value. It is recommended that the discard mechanism is based on a RED-like [8] algorithm with three configurable levels of drop precedence and a configurable averaging function (interval). A DS node must not reorder AF packets of the same microflow when they belong to the same AF class regardless of their drop precedence. There are no timing requirements (delay or delay variation) associated with the forwarding of AF packets.

In terms of constructing an end-to-end service with each of the above three PHBs, here are some examples. A Basic service can be provided using the Default PHB where no service guarantees are provided besides the normal connectivity guarantee, plus some performance objectives. Meanwhile, a Virtual Leased Line (VLL) service can be provided by cascading EF PHBs at each hop of the network where bandwidth as well as low delay and jitter can be guaranteed. Finally, an Assured service can be provided using AF PHBs where qualitative service levels are provided to an AF class with each drop priority. Performance issues have been studied by authors in [10] where simulation results show that such a service is not a quantifiable one and does not provide clearly defined and consistent rate guarantees as well as a fair share of the excess bandwidth. It can only provide service levels in relatively terms as it intends to. More studies are needed in this area.

3.4 Diffserv and COPS

COPS (Common Open Policy Service Protocol) can also be used for Diffserv by adding a new ClientType to COPS, the Diffserv ClientType, plus Diffserv-specific data objects and lists of filters. Many issues regarding fault tolerance, authentication, dynamic provisioning and scalability remain to be addressed.

3.5 Remarks

The Differentiated Services model is simpler and more scalable than the Integrated Service model, and it may be suitable for wide range of applications. However in order to achieve some meaningful level of end-to-end QoS for certain applications that require QoS assurance for each individual flow, careful provisioning and traffic engineering are critical which has not been addressed in detail by the Diffserv documents. There are also needs to study the tradeoffs between service levels, network capacity, impacts of QoS traffic on non-QoS traffic, necessary tool supports. One of our studies [5] addresses most aspects of these areas.

Recently, there have been some efforts to combine Intserv and Diffserv for providing end-to-end QoS to meet the needs of large ISPs who manage the transit networks of the Internet, and the users of QoS applications and hosts, who are the ISPs' ultimate customers. It is believed that in the coming years, there will be a proliferation of applications that depend on QoS and of hosts which are capable of QoS signaling. The authors of a framework document [intdiff] envision the deployment of Diffserv capable core networks and Intserv capable stub networks at the periphery. The framework allows each to proceed at its own pace, providing immediate incremental benefits in areas of the network in which one or the other is deployed and

additional benefits where both are deployed. The framework builds upon current work in IETF including Diffserv and RSVP. RSVP can still be used to signal the end station's QoS needs to the edge router, but the edge router would translate this request into the appropriate class of service by setting the DS byte. More commonly, the per-flow information would be entered administratively into the edge router.

4. QOS IN IP OVER ATM

ATM has been considered to be one of the major protocols over which IP can operate. With ATM underneath, IP can take advantage of the high speed hardware switching, the efficient statistical multiplexing, and the relatively mature QoS on-demand features that ATM offers. Much interest has been generated with regards to how these desirable features of ATM can best be exploited in an IP network. Various approaches have been proposed in standards and implemented by the industry. They are: classical IP over ATM [14] from the IPATM working group of IETF, LAN emulation [13] from the ATM Forum, [15] from the Routing Over Large Clouds (ROLC) working group of the IETF, and the Multiprotocol Over ATM (MPOA) [17] from the ATM Forum.

In some early schemes for IP over ATM deployment, leveraging potential benefits of ATM was not the primary concern. In fact the "classical" models of the "Classical IP over ATM" and "LAN emulation" largely negate the potential benefits of ATM in exchange for maintaining the classical IP paradigm in order to facilitate easy migration to ATM. For example, the "classical IP over ATM" model separates an ATM network into logical IP subnets (LISs) interconnected by IP routers. Direct ATM connections between IP hosts in separate LISs are then prohibited even if the underlying ATM topology is capable of supporting them. Instead, an ATM virtual connection (VC) originating within a given LIS can only extend as far as a router at the LIS boundary where the contents of the received ATM cells must then be reassembled into IP packets. Each of them is then subjected to an IP forwarding decision and resegmentation into ATM cells to be sent along the next intra-LIS VC along the journey. The reassembling and re-segmentation at each router along the path severely restricts the potential benefits of ATM, namely those of the high-bit-rate low-latency path with QoS support. For example, if the application sending the packet had any particular QoS requirements for the flow of IP packets, the QoS provided by the ATM network would be lost at each router hop.

To eliminate this inefficiency, the IETF developed the Next Hop Resolution Protocol (NHRP) as a means of facilitating inter-LIS VCs in order to utilize the potential benefits of ATM which are lost with the classical models. NHRP is an inter-LIS address resolution mechanism that maps a destination's IP address to the destination's ATM address in cases where the destination resides within the same ATM cloud as the source. In cases where the destination resides outside the ATM cloud containing the source, NHRP returns the ATM address of the source ATM cloud's egress router closest to the destination. Once the source received the NHRP response, it can then open a direct cut-through VC to the destination⁴ using standard ATM signaling/routing protocols.

Like classical IP over ATM and NHRP, MPOA is based on the overlay model and uses client/server interactions to:

- Integrate the IETF's NHRP with support for LANE to provide LANE hosts with the benefits of NHRP shortcuts;
- Use both routing and bridging information to locate the optimal exit interface from the ATM cloud for both routed and bridged packets;
- Add support for virtual routing, which separates routers into two components: 1) a more expensive routing engine that maintains IP routing tables and runs routing protocols and 2) a less expensive forwarding engine that caches routes when necessary from the routing engine. Many forwarding engines can be supported by a single routing engine, and the two components use ATM to intercommunicate.

MPOA is heavily based on LANE but allows direct interconnection of hosts (regardless of whether or not they are in the same IP address prefix or on the same logical subnet) through use of NHRP shortcuts.

⁴ Or to the closest egress router to the destination in cases where the destination resides outside the ATM cloud.

MPOA therefore essentially is a combination of LANE and NHRP, combining the virtual networking ability of LANE with the address resolution ability of NHRP. MPOA will be most useful in environments that use both bridging and routing on LAN's in combination with both LANE and NHRP over ATM.

All of these approaches obscure the real topology of the underlying network from the internetwork layer routing protocol. To IP the physical network becomes a large opaque cloud which results in some significant problems. These problems include duplication of functionality in routing, scalability limitation for a large IP network, challenges in supporting higher than OC48 ATM SAR, inefficiency for multicasting, increased management and manual configuration requirements, and difficulty in extending the approach to QoS routing and bandwidth management.

5. MPLS AND QOS

5.1 MPLS Overview

The current IP uses destination-based hop-by-hop routing. At each hop, a router makes decision on the next best hop to forward a packet based on the destination IP address from the IP header independently of other packets. This routing paradigm has been proved to be very simple and robust, however, it has the following limitations:

- It has limiting forwarding capacity due to the "longest prefix matching" on the destination IP address.
- It is insensitive to traffic conditions in the network and is therefore inefficient in the way it utilizes network capacity.
- Destination-based routing is vulnerable to security attacks by anyone who learns destinations of other users. The common approach today of filtering incoming packets for disallowed destinations is prone to provisioning errors and costly to implement.
- With destination-based routing, routing based on bandwidth availability and traffic conditions is costly to implement and likely to lead to instabilities.
- Destination-based routing does not allow packets belonging to different classes of services to be routed and protected separately to meet their quality of service needs.

To address these limitations, many vendors have proposed various solutions to integrate routing with switching under the scope of "IP Switching" (or Label Switching). They are IP Switch from Nokia [18], CSR (Cell Switch Router) from Toshiba [16], Tag Switch from Cisco [20], ARIS (Aggregated Routed Based IP Switch) from IBM [24], and IP Navigator from Lucent [1]. Recently, IETF formed the MPLS working group to combine various approaches into one common standard [7] [21] that can run over any media (e.g., PPP, frame relay, ATM). With MPLS, each packet is associated with a label that identifies its Forwarding Equivalence Class (FEC). Forwarding Equivalence Class can be defined by a common egress router and optionally by other characteristics (e.g., a common VPN or class of service). The label is encoded in connection identifier when available (e.g., ATM VPI/VCI and FR DLCI), otherwise a "shim" is inserted between IP header and link-layer header. Based on the label of an incoming packet, a router determines the packet's next hop as well as the label to use on outgoing interface and performs the label swapping. In order to provide scalability, multiple incoming labels may be mapped to same outgoing labels, which is called stream merging. Label Switched Paths (LSPs) can be established based on the IP address of a router that is downstream on the hop-by-hop route. On the other hand, LSPs can also be established using explicit routing determined manually or by a separate routing protocol. Explicit routing can take the bandwidth availability and traffic conditions into account as well as providing failure protection and restoration using path protection and/or link/node protection. But no procedure has been standardized yet. MPLS has many potential advantages compared with the traditional hop-by-hop destination based routing. They are as follows:

- MPLS increases routers' forwarding capacity by switching based on short labels. Although this was one of the main motivations of introducing MPLS, however, this advantage is diminishing due to many new ASIC designed fast and scalable prefix matching algorithms.
- MPLS provides a way to perform traffic engineering (TE) by overlaying logical connectivity on top of the physical topology to distribute loads more efficiently and in a more scalable way comparing with

the traditional IP over ATM overlay model, which has been used for traffic engineering for large IP backbones during the last few years.

- MPLS is connection-oriented, thus is capable of providing faster and more reliable protection and restoration than what current Interior Gateway Protocol (IGP) (e.g., OSPF and IS-IS) rerouting can provide.
- MPLS can map traffic with different class of service requirements onto different LSPs, which are routed and protected independently.
- MPLS provides added security if packets sent on outgoing interfaces are limited on those from particular connections.
- MPLS labels may be stacked to implement tunnels, which is useful in constructing virtual private networks (VPNs).

In the next subsection, we focus on the enhancements MPLS brings to QoS in terms of scalability and availability.

5.2 MPLS enhancements to QoS

Although traffic engineering for general traffic is regarded as one immediate application of MPLS, it is unclear whether it will still be useful once fast provision of lightpaths at the WDM layer becomes available. On the other hand, MPLS has great potentials in enhancing other services, one of them is Quality of Service. Class of service (CoS) based on Diffserv architecture can be enhanced in scalability by performing per-class traffic engineering as well as availability by providing faster protection and restoration to high priority classes than normal IGP rerouting.

First, we outline main circumstances where per-class traffic engineering and protection are needed as below.

- Relative ratio of traffic volume among different classes varies significantly across different pairs of backbone routers such that FEC belonging to different classes of service may need to travel on different paths due to bandwidth shortage for certain class.
- FEC belonging to different classes may have different constraints. For example, traffic in a Virtual Leased Line (VLL) class that is designed for supporting real-time application such as voice and video needs to travel on a low delay/jitter path, while traffic in a Enhanced service class that is designed for high-end non-real time traffic needs to travel on a low-loss path.
- FEC with different classes may require different protection levels. For example, VLL service may need near SONET protection level using link/node protection or fast reroute), Assured service may need pre-established path protection, while Basic service may rely on dynamic path protection or IGP reroute. Since the faster protection, the more it costs in bandwidth over-provisioning, per-class protection can provide cost effective protection to each class at the level it requires.

For supporting per-class TE and protection, Label-Inferred-PHS LSP (L-LSP) has been proposed⁵ where one LSP is established per <FEC, SA⁶> between two LSR neighbors. This method relies on explicit signaling of the PHS⁷ at label establishment time so that, after label establishment, the LSP can infer from the label value the PHS to be applied to a labeled packet. Drop precedence to be applied by the LSR to the labeled packet is conveyed inside the packet MPLS Shim header using the EXP field for PPP links, or by the CPI bit for ATM links or the DE bit for FR links.

⁵ "MPLS Support of Differentiated Services over PPP Links", F. Faucheur, S. Davari, R. Krishnan, P. Vanannen, and B. Davie, IETF draft, <http://www.ietf.org/draft-lefaucheur-mpls-diff-ppp-00.txt>, June 1999.

⁶ Scheduling Aggregate (SA) is defined as a set of Behavior Aggregates (BA) which need to preserve the correct packet ordering relationships. The term SA may be changed later once consensus is reached in the MPLS working group.

⁷ Per Hop Scheduling (PHS) is referred as a set of one or more PHBs applied to the set of BAs forming a given SA.

One the other hand, when per-class TE or protection is not needed, traffic engineering and/or protection can be performed for all classes of traffic as an aggregate. This can be achieved using the recently proposed EXP-Inferred-PHS LSPs (E-LSP) where up to 8 Behavior Aggregate (BAs) can be supported via a single LSP for PPP links. In this method, the DSCP value gets mapped entirely into the EXP field of the MPLS Shim Header at the Edge of the MPLS PPP Cloud (thus encoding both drop precedence and PHS/scheduling information). In other words, both PHS and drop precedence are conveyed in each labeled packet using the EXP field of the MPLS Shim Header.

Another important functionality of E-LSPs that is currently missing from the draft that proposes it is that E-LSPs can also be used for performing traffic engineering and protection at per-SA level by combining packets to FEC mapping with DSCP filtering at headend LSRs, i.e., only allow the corresponding SA in a FEC to be mapped on the E-LSP. There are two limitations of this approach: one is that it only supports up to 8 BAs, the other is that it does not support any per-SA group resource allocation or admission control based on signaled per-SA request. Instead, admission control and resource allocation need to rely on either some centralized system (online and/or offline) or being performed at the aggregate level. In practice, this approach allows network operators to perform a large set of traffic engineering and protection functionality without relying on L-LSPs which need certain extensions to current protocols. Many vendors have already provided early implementation of E-LSPs, and planned to provide DSCP filtering at edges of a MPLS/Diffserv domain. Since the mapping between EXP and PHB are pre-configured locally in the MPLS/Diffserv domain without relying on signaling, E-LSPs are also capable of performing per-SA-group traffic engineering and protection if the corresponding group of SAs share the same criterion. For example, both the pipe-based virtual leased line class and hose-based premium real time class may require the same level of fast protection using pre-configured, pre-reserved alternate path protection in conjunction with fast link/node protection. Then they can use the same E-LSP by mapping only the corresponding group of SAs onto the E-LSP at the headend LSR. This is more flexible and requires less signaling and configuration than the per-SA only traffic engineering and protection using L-LSPs.⁸

Per-class traffic engineering and protection can be achieved using two options: one is to use manually configured ER-LSPs, denoted as MC-LSPs, by specifying each hop on the path by a network operator; another is to use constraint-based routed ER-LSPs, denoted as CR-LSPs, where the path selection is done at the headend of each LSP using constraint-based routing based on criterion specified in terms of bandwidth and delay requirement plus some path attributes. The major requirement of the first option is having an offline or online planning tool that is aware of resource allocation of all Diffserv classes and provide path selections to the MC-LSPs. With the second option, path selection relies on some online algorithm provided by vendors. There are some main issues and concerns regarding the effectiveness and predictability of the algorithms as well as the ability to satisfy multiple constants beside bandwidth requirement that have not been fully addressed.

Most of activities in MPLS working group and vendors have been concentrating on providing flow-based QoS with MPLS. To the best of our knowledge, the only product that supports both class-based and flow-based QoS is Lucent pre-standard ATM-based product IP Navigator [1]. IP Navigator uses MPTs (Multipoint-to-Point Trees) to carry class-based QoS traffic as well as best effort traffic, while using NHRP shortcuts to carry relatively long-lived QoS flows. It triggers onset and abatement of NHRP shortcuts using two static pre-determined parameters. In [6], we improved the triggering scheme by introducing a dynamic triggering scheme called SDDTS that adjusts the two parameters according to some states of the system. The algorithm provides robust and consistent near-optimal performance under various traffic scenarios.

6. SUMMARY

In this paper, we summarized main approaches and activities in standards committees and the industry for providing quality of service in IP. The two main models standardized in IETF, Intserv and Diffserv, each has its own strength and weakness. Due to scalability advantages, we foresee that Diffserv is likely to be

⁸ Authors of the draft have agreed to add this additional functionality of E-LSPs in their next version as suggested by us.

provided by main ISPs in the near future, while Intserv may be realized in enterprise networks and can be extended to edges of ISP backbones.

ACKNOWLEDGEMENTS

The author would like to thank Seyhan Civanlar and Z. Bo Tang for their valuable comments and suggestions.

REFERENCES

- [1] H. Ahmed, R. Callon, A. Makis, and J. Moy, "IP Switching for Scalable IP Services," *Proc. of the IEEE*, Vol. 85, No. 12, pp. 1984-1997, Dec. 1997.
- [2] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services", *RFC2475*, Dec. 1998.
- [3] J. Boyle, R. Cohen, D. Durham, S. Herzog, R. Rajan, and A. Sastry, "The COPS (Common Open Policy Service) Protocol", *Internet-draft*, draft-ietf-rap-cops-07.txt, Aug. 1998.
- [4] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation Protocol (RSVP) - Version 1 Functional Specification", *RFC 2205*, Sept. 1997.
- [5] A. Chiu, S. Civanlar, A. Greenberg, and Z. Tang, "Method and Apparatus for an Architecture and Design of IP Quality of Service", *patent application in pending*, Sept. 1999.
- [6] A. Chiu and X. Mang, "A State Dependent Dynamic Triggering Scheme for Flow-Based QoS Shortcuts", *ITC16*, Edinburgh, U.K., June 1999.
- [7] R. Callon, P. Doolan, N. Feldman, A. Fredette, G. Swallow, and A. Viswanathan, "A Framework for Multiprotocol Label Switching", *Internet-draft*, draft-ietf-mpls-framework-04.txt, July 1999.
- [8] S. Floyd and V. Jacobson, "Random Early Detection Gateways for Congestion Avoidance", *IEEE/ACM Transactions on Networking*, Vol. 1 Nov. 4., p. 387-413, Aug. 1993.
- [9] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, "Assured Forwarding PHB Group", *RFC2597*, June 1999.
- [10] J. Ibanez and K. Nichols, "Preliminary Simulation Evaluation of An Assured Service", *Internet-Draft*, draft-ibanez-diffserv-assured-eval-00.txt, Aug. 1998.
- [11] V. Jacobson, K. Nichols, and K. Poduri, "An Expedited Forwarding PHB", *RFC2598*, June 1999.
- [12] V. P. Kumar, T. V. Lakshman, and D. Stiliadis, "Beyond Best Effort: Router Architectures for the Differentiated Services of Tomorrow's Internet", *IEEE Communication Magazine*, May 1998.
- [13] "LAN Emulation over ATM", *The ATM Forum version 1.0*, Jan. 1995.
- [14] M. Laubach, "Classical IP and ARP over ATM", *RFC1577*, Jan. 1994.
- [15] J. V. Luciani, D. Katz, D. Piscitello, and B. Cole, "NBMA Next Hop Resolution Protocol (NHRP)", *IETF Internet Draft*, draft-ietf-rolc-nhrp-10.txt, Oct. 1996.
- [16] S. Masuzawa, K. Magami, A. Mogi, T. Jinmei, Y. Katsube, and H. Esaki, "Architecture of Cell Switch Router and Prototype System Implementation" (to be published), *IEICE Transactions on Communications*.
- [17] "Multi-Protocol over ATM (MPOA) version 1.0", ATM Forum, AF-AF-MPOA-0087.000, 1997.
- [18] P. Newman, G. Minshall, and T. Lyon, "Flow labeled IP: A Connectionless Approach to ATM," *Proc. IEEE INFOCOM*, pp. 1251-1260, Mar. 1996.
- [19] A. K. Parekh and R. G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single-Node Case", *IEEE/ACM Trans. On Networking*, June, 1993.
- [20] Y. Rekhter, B. Davie, D. Katz, E. Rosen, and G. Swallow, "Cisco Systems' Tag Switching Architecture Overview", *RFC2105*, Feb. 1997.
- [21] E. Rosen, A. Viswanathan, and R. Callon, "Multiprotocol Label Switching Architecture", *Internet-draft*, draft-ietf-mpls-arch-06.txt, Aug. 1999.

- [22] S. Shenker, C. Partridge, and R. Guerin, "Specification of Guaranteed Quality of Service", *RFC2212*, Sept. 1997.
- [23] J. Wroclawski, "Specification of the Controlled-Load Network Element Service", *RFC2211*, Sept. 1997.
- [24] R. Woundy, A. Viswanathan, N. Feldman, and R. Boivie, "ARIS: Aggregate Route-Based IP Switching", *Internet-Draft*, draft-woundy-arip-switching-00.txt, Nov. 1996.
- [25] R. Yavatkar, D. Pendarakis, and R. Guerin, "A Framework for Policy-based Admission Control", *Internet-draft*, draft-ietf-rap-framework-03.txt, April 1999.