Network Working Group
Request for Comments: 2816
Category: Informational

A. Ghanwani
Nortel Networks
W. Pace
IBM
V. Srinivasan
CoSine Communications
A. Smith
Extreme Networks
M. Seaman
Telseon

May 2000

A Framework for Integrated Services Over Shared and Switched IEEE 802 LAN Technologies

Status of this Memo

Copyright Notice

Abstract

This memo describes a framework for supporting IETF Integrated Services on shared and switched LAN infrastructure. It includes background material on the capabilities of IEEE 802 like networks with regard to parameters that affect Integrated Services such as access latency, delay variation and queuing support in LAN switches. It discusses aspects of IETF's Integrated Services model that cannot easily be accommodated in different LAN environments. It outlines a functional model for supporting the Resource Reservation Protocol (RSVP) in such LAN environments. Details of extensions to RSVP for use over LANs are described in an accompanying memo [14]. Mappings of the various Integrated Services onto IEEE 802 LANs are described in another memo [13].

Contents

1. Introduction

The Internet has traditionally provided support for best effort traffic only. However, with the recent advances in link layer technology, and with numerous emerging real time applications such as video conferencing and Internet telephony, there has been much interest for developing mechanisms which enable real time services over the Internet. A framework for meeting these new requirements was set out in RFC 1633 [8] and this has driven the specification of various classes of network service by the Integrated Services working group of the IETF, such as Controlled Load and Guaranteed Service [6,7]. Each of these service classes is designed to provide certain Quality of Service (QoS) to traffic conforming to a specified set of parameters. Applications are expected to choose one of these classes according to their QoS requirements. One mechanism for end stations to utilize such services in an IP network is provided by a QoS signaling protocol, the Resource Reservation Protocol (RSVP) [5] developed by the RSVP working group of the IETF. The IEEE under its Project 802 has defined standards for many different local area network technologies. These all typically offer the same MAC layer datagram service [1] to higher layer protocols such as IP although they often provide different dynamic behavior characteristics -- it is these that are important when considering their ability to support real time services. Later in this memo we describe some of the relevant characteristics of the different MAC layer LAN technologies. In addition, IEEE 802 has defined standards for bridging multiple LAN segments together using devices known as "MAC Bridges" or "Switches" [2]. Recent work has also defined traffic classes, multicast filtering, and virtual LAN capabilities for these devices [3,4]. Such LAN technologies often constitute the last hop(s) between users and the Internet as well as being a primary building block for entire campus networks. It is therefore necessary to provide standardized mechanisms for using these technologies to support end-to-end real time services. In order to do this, there must be some mechanism for resource management at the data link layer. Resource management in this context encompasses the functions of admission control, scheduling, traffic policing, etc. The ISSLL (Integrated Services

over Specific Link Layers) working group in the IETF was chartered with the purpose of exploring and standardizing such mechanisms for various link layer technologies.

2. Document Outline

This document is concerned with specifying a framework for providing Integrated Services over shared and switched LAN technologies such as Ethernet/IEEE 802.3, Token Ring/IEEE 802.5, FDDI, etc. We begin in Section 4 with a discussion of the capabilities of various IEEE 802 MAC layer technologies. Section 5 lists the requirements and goals for a mechanism capable of providing Integrated Services in a LAN. The resource management functions outlined in Section 5 are provided by an entity referred to as a Bandwidth Manager (BM). The architectural model of the BM is described in Section 6 and its various components are discussed in Section 7. Some implementation issues with respect to link layer support for Integrated Services are examined in Section 8. Section 9 discusses a taxonomy of topologies for the LAN technologies under consideration with an emphasis on the capabilities of each which can be leveraged for enabling Integrated Services. This framework makes no assumptions about the topology at the link layer. The framework is intended to be as exhaustive as possible; this means that it is possible that all the functions discussed may not be supportable by a particular topology or technology, but this should not preclude the usage of this model for it.

3. Definitions

The following is a list of terms used in this and other ISSLL documents.

- Link Layer or Layer 2 or L2: Data link layer technologies such as Ethernet/IEEE 802.3 and Token Ring/IEEE 802.5 are referred to as Layer 2 or L2.

- Link Layer Domain or Layer 2 Domain or L2 Domain: Refers to a set of nodes and links interconnected without passing through a L3 forwarding function. One or more IP subnets can be overlaid on a L2 domain.

- Layer 2 or L2 Devices: Devices that only implement Layer 2 functionality as Layer 2 or L2 devices. These include IEEE 802.1D [2] bridges or switches.

- Internetwork Layer or Layer 3 or L3: Refers to Layer 3 of the ISO OSI model. This memo is primarily concerned with networks that use the Internet Protocol (IP) at this layer.

- Layer 3 Device or L3 Device or End Station: These include hosts and routers that use L3 and higher layer protocols or application programs that need to make resource reservations.

- Segment: A physical L2 segment that is shared by one or more senders. Examples of segments include: (a) a shared Ethernet or Token Ring wire resolving contention for media access using CSMA or token passing; (b) a half duplex link between two stations or switches; (c) one direction of a switched full duplex link.

- Managed Segment: A managed segment is a segment with a DSBM (designated subnet bandwidth manager, see [14]) present and responsible for exercising admission control over requests for resource reservation. A managed segment includes those interconnected parts of a shared LAN that are not separated by DSBMs.

- Traffic Class: Refers to an aggregation of data flows which are given similar service within a switched network.

- Subnet: Used in this memo to indicate a group of L3 devices sharing a common L3 network address prefix along with the set of segments making up the L2 domain in which they are located.

- Bridge/Switch: A Layer 2 forwarding device as defined by IEEE 802.1D [2]. The terms bridge and switch are used synonymously in this memo.

4. Frame Forwarding in IEEE 802 Networks

4.1. General IEEE 802 Service Model

The user_priority is a value associated with the transmission and reception of all frames in the IEEE 802 service model. It is supplied by the sender that is using the MAC service and is provided along with the data to a receiver using the MAC service. It may or may not be actually carried over the network. Token Ring/IEEE 802.5 carries this value encoded in its FC octet while basic Ethernet/IEEE 802.3 does not carry it. IEEE 802.12 may or may not carry it depending on the frame format in use. When the frame format in use is IEEE 802.5, the user_priority is carried explicitly. When IEEE 802.3 frame format is used, only the two levels of priority (high/low) that are used to determine access priority can be recovered. This is based on the value of priority encoded in the start delimiter of the IEEE 802.3 frame.

NOTE: The original IEEE 802.1D standard [2] contains the specifications for the operation of MAC bridges. This has recently been extended to include support for traffic classes and dynamic multicast filtering [3]. In this document, the reader should be aware that references to the IEEE 802.1D standard refer to [3], unless explicitly noted otherwise.

IEEE 802.1D [3] defines a consistent way for carrying the value of the user_priority over a bridged network consisting of Ethernet, Token Ring, Demand Priority, FDDI or other MAC layer media using an extended frame format. The usage of user_priority is summarized below. We refer the interested reader to the IEEE 802.1D specification for further information.

If the user_priority is carried explicitly in packets, its utility is as a simple label enabling packets within a data stream in different classes to be discriminated easily by downstream nodes without having to parse the packet in more detail.

Apart from making the job of desktop or wiring closet switches easier, an explicit field means they do not have to change hardware or software as the rules for classifying packets evolve; e.g. based on new protocols or new policies. More sophisticated Layer 3 switches, perhaps deployed in the core of a network, may be able to provide added value by performing packet classification more accurately and, hence, utilizing network resources more efficiently and providing better isolation between flows. This appears to be a good economic choice since there are likely to be very many more desktop/wiring closet switches in a network than switches requiring Layer 3 functionality.

The IEEE 802 specifications make no assumptions about how user_priority is to be used by end stations or by the network. Although IEEE 802.1D defines static priority queuing as the default mode of operation of switches that implement multiple queues, the user_priority is really a priority only in a loose sense since it depends on the number of traffic classes actually implemented by a switch. The user_priority is defined as a 3 bit quantity with a value of 7 representing the highest priority and a value of 0 as the lowest. The general switch algorithm is as follows. Packets are queued within a particular traffic class based on the received user_priority, the value of which is either obtained directly from the packet if an IEEE 802.1Q header or IEEE 802.5 network is used, or is assigned according to some local policy. The queue is selected based on a mapping from user_priority (0 through 7) onto the number of available traffic classes. A switch may implement one or more traffic classes. The advertised IntServ parameters and the switch's admission control behavior may be used to determine the mapping from

user_priority to traffic classes within the switch. A switch is not precluded from implementing other scheduling algorithms such as weighted fair queuing and round robin.

IEEE 802.1D makes no recommendations about how a sender should select the value for user_priority. One of the primary purposes of this document is to propose such usage rules, and to discuss the communication of the semantics of these values between switches and end stations. In the remainder of this document we use the term traffic class synonymously with user_priority.

4.2. Ethernet/IEEE 802.3

There is no explicit traffic class or user_priority field carried in Ethernet packets. This means that user_priority must be regenerated at a downstream receiver or switch according to some defaults or by parsing further into higher layer protocol fields in the packet. Alternatively, IEEE 802.1Q encapsulation [4] may be used which provides an explicit user_priority field on top of the basic MAC frame format.

For the different IP packet encapsulations used over Ethernet/IEEE 802.3, it will be necessary to adjust any admission control calculations according to the framing and padding requirements as shown in Table 1. Here, "ip_len" refers to the length of the IP packet including its headers.

Table 1: Ethernet encapsulations

Encapsulation	Framing Overhead bytes/pkt	IP MTU bytes
IP EtherType (ip_len<=46 bytes)	64-ip_len	1500
(1500>=ip_len>=46 bytes)	18	1500
IP EtherType over 802.1D/Q (ip_len<=42)	64-ip_len	1500*
(1500>=ip_len>=42 bytes)	22	1500*
IP EtherType over LLC/SNAP (ip_len<=40)	64-ip_len	1492
(1500>=ip_len>=40 bytes)	24	1492

*Note that the packet length of an Ethernet frame using the IEEE 802.1Q specification exceeds the current IEEE 802.3 maximum packet length values by 4 bytes. The change of maximum MTU size for IEEE 802.1Q frames is being accommodated by IEEE 802.3ac [21].

4.3. Token Ring/IEEE 802.5

The Token Ring standard [6] provides a priority mechanism that can be used to control both the queuing of packets for transmission and the access of packets to the shared media. The priority mechanisms are implemented using bits within the Access Control (AC) and the Frame Control (FC) fields of a LLC frame. The first three bits of the AC field, the Token Priority bits, together with the last three bits of the AC field, the Reservation bits, regulate which stations get access to the ring. The last three bits of the FC field of a LLC frame, the User Priority bits, are obtained from the higher layer in the user_priority parameter when it requests transmission of a packet. This parameter also establishes the Access Priority used by the MAC. The user_priority value is conveyed end-to-end by the User Priority bits in the FC field and is typically preserved through Token Ring bridges of all types. In all cases, 0 is the lowest priority.

Token Ring also uses a concept of Reserved Priority which relates to the value of priority which a station uses to reserve the token for its next transmission on the ring. When a free token is circulating, only a station having an Access Priority greater than or equal to the Reserved Priority in the token will be allowed to seize the token for transmission. Readers are referred to [14] for further discussion of this topic.

A Token Ring station is theoretically capable of separately queuing each of the eight levels of requested user_priority and then transmitting frames in order of priority. A station sets Reservation bits according to the user_priority of frames that are queued for transmission in the highest priority queue. This allows the access mechanism to ensure that the frame with the highest priority throughout the entire ring will be transmitted before any lower priority frame. Annex I to the IEEE 802.5 Token Ring standard recommends that stations send/relay frames as follows.

Table 2: Recommended use of Token Ring User Priority

Application	User Priority
Non-time-critical data	0
-	1
-	2
-	3
LAN management	4
Time-sensitive data	5
Real-time-critical data	6
MAC frames	7

To reduce frame jitter associated with high priority traffic, the annex also recommends that only one frame be transmitted per token and that the maximum information field size be 4399 octets whenever delay sensitive traffic is traversing the ring. Most existing implementations of Token Ring bridges forward all LLC frames with a default access priority of 4. Annex I recommends that bridges forward LLC frames that have a user_priority greater than 4 with a reservation equal to the user_priority (although IEEE 802.1D [3] permits network management override this behavior). The capabilities provided by the Token Ring architecture, such User Priority and Reserved Priority, can provide effective support for Integrated Services flows that require QoS guarantees.

For the different IP packet encapsulations used over Token Ring/IEEE 802.5, it will be necessary to adjust any admission control calculations according to the framing requirements as shown in Table 3.

Table 3: Token Ring encapsulations

Encapsulation	Framing Overhead bytes/pkt	IP MTU bytes
IP EtherType over 802.1D/Q	29	4370*
IP EtherType over LLC/SNAP	25	4370*

*The suggested MTU from RFC 1042 [13] is 4464 bytes but there are issues related to discovering the maximum supported MTU between any two points both within and between Token Ring subnets. The MTU reported here is consistent with the IEEE 802.5 Annex I recommendation.

4.4. Fiber Distributed Data Interface

The Fiber Distributed Data Interface (FDDI) standard [16] provides a priority mechanism that can be used to control both the queuing of packets for transmission and the access of packets to the shared media. The priority mechanisms are implemented using similar mechanisms to Token Ring described above. The standard also makes provision for "Synchronous" data traffic with strict media access and delay guarantees. This mode of operation is not discussed further here and represents area within the scope of the ISSLL working group that requires further work. In the remainder of this document, for the discussion of QoS mechanisms, FDDI is treated as a 100 Mbps Token Ring technology using a service interface compatible with IEEE 802 networks.

4.5. Demand Priority/IEEE 802.12

IEEE 802.12 [19] is a standard for a shared 100 Mbps LAN. Data packets are transmitted using either the IEEE 802.3 or IEEE 802.5 frame format. The MAC protocol is called Demand Priority. Its main characteristics with respect to QoS are the support of two service priority levels, normal priority and high priority, and the order of service for each of these. Data packets from all network nodes (end hosts and bridges/switches) are served using a simple round robin algorithm.

If the IEEE 802.3 frame format is used for data transmission then the user_priority is encoded in the starting delimiter of the IEEE 802.12 data packet. If the IEEE 802.5 frame format is used then the user_priority is additionally encoded in the YYY bits of the FC field in the IEEE 802.5 packet header (see also Section 4.3). Furthermore, the IEEE 802.1Q encapsulation with its own user_priority field may also be applied in IEEE 802.12 networks. In all cases, switches are able to recover any user_priority supplied by a sender.

The same rules apply for IEEE 802.12 user_priority mapping in a bridge as with other media types. The only additional information is that normal priority is used by default for user_priority values 0 through 4 inclusive, and high priority is used for user_priority levels 5 through 7. This ensures that the default Token Ring user_priority level of 4 for IEEE 802.5 bridges is mapped to normal priority on IEEE 802.12 segments.

The medium access in IEEE 802.12 LANs is deterministic. The Demand Priority mechanism ensures that, once the normal priority service has been preempted, all high priority packets have strict priority over packets with normal priority. In the event that a normal priority packet has been waiting at the head of line of a MAC transmit queue

for a time period longer than PACKET_PROMOTION (200 - 300 ms) [19], its priority is automatically promoted to high priority. Thus, even normal priority packets have a maximum guaranteed access time to the medium.

Integrated Services can be built on top of the IEEE 802.12 medium access mechanism. When combined with admission control and bandwidth enforcement mechanisms, delay guarantees as required for a Guaranteed Service can be provided without any changes to the existing IEEE 802.12 MAC protocol.

Since the IEEE 802.12 standard supports the IEEE 802.3 and IEEE 802.5 frame formats, the same framing overhead as reported in Sections 4.2 and 4.3 must be considered in the admission control computations for IEEE 802.12 links.

5. Requirements and Goals

This section discusses the requirements and goals which should drive the design of an architecture for supporting Integrated Services over LAN technologies. The requirements refer to functions and features which must be supported, while goals refer to functions and features which are desirable, but are not an absolute necessity. Many of the requirements and goals are driven by the functionality supported by Integrated Services and RSVP.

5.1. Requirements

- Resource Reservation: The mechanism must be capable of reserving resources on a single segment or multiple segments and at bridges/switches connecting them. It must be able to provide reservations for both unicast and multicast sessions. It should be possible to change the level of reservation while the session is in progress.

- Admission Control: The mechanism must be able to estimate the level of resources necessary to meet the QoS requested by the session in order to decide whether or not the session can be admitted. For the purpose of management, it is useful to provide the ability to respond to queries about availability of resources. It must be able to make admission control decisions for different types of services such as Guaranteed Service, Controlled Load, etc.

- Flow Separation and Scheduling: It is necessary to provide a mechanism for traffic flow separation so that real time flows can be given preferential treatment over best effort flows. Packets of real time flows can then be isolated and scheduled according to their service requirements.

- Policing/Shaping: Traffic must be shaped and/or policed by end stations (workstations, routers) to ensure conformance to negotiated traffic parameters. Shaping is the recommended behavior for traffic sources. A router initiating an ISSLL session must have implemented traffic control mechanisms according to the IntServ requirements which would ensure that all flows sent by the router are in conformance. The ISSLL mechanisms at the link layer rely heavily on the correct implementation of policing/shaping mechanisms at higher layers by devices capable of doing so. This is necessary because bridges and switches are not typically capable of maintaining per flow state which would be required to check flows for conformance. Policing is left as an option for bridges and switches, which if implemented, may be used to enforce tighter control over traffic flows. This issue is further discussed in Section 8.

- Soft State: The mechanism must maintain soft state information about the reservations. This means that state information must periodically be refreshed if the reservation is to be maintained; otherwise the state information and corresponding reservations will expire after some pre-specified interval.

- Centralized or Distributed Implementation: In the case of a centralized implementation, a single entity manages the resources of the entire subnet. This approach has the advantage of being easier to deploy since bridges and switches may not need to be upgraded with additional functionality. However, this approach scales poorly with geographical size of the subnet and the number of end stations attached. In a fully distributed implementation, each segment will have a local entity managing its resources. This approach has better scalability than the former. However, it requires that all bridges and switches in the network support new mechanisms. It is also possible to have a semi-distributed implementation where there is more than one entity, each managing the resources of a subset of segments and bridges/switches within the subnet. Ideally, implementation should be flexible; i.e. a centralized approach may be used for small subnets and a distributed approach can be used for larger subnets. Examples of centralized and distributed implementations are discussed in Section 6.

- Scalability: The mechanism and protocols should have a low overhead and should scale to the largest receiver groups likely to occur within a single link layer domain.

- Fault Tolerance and Recovery: The mechanism must be able to function in the presence of failures; i.e. there should not be a single point of failure. For instance, in a centralized implementation, some mechanism must be specified for back-up and recovery in the event of failure.

- Interaction with Existing Resource Management Controls: The interaction with existing infrastructure for resource management needs to be specified. For example, FDDI has a resource management mechanism called the "Synchronous Bandwidth Manager". The mechanism must be designed so that it takes advantage of, and specifies the interaction with, existing controls where available.

5.2. Goals

- Independence from higher layer protocols: The mechanism should, as far as possible, be independent of higher layer protocols such as RSVP and IP. Independence from RSVP is desirable so that it can interwork with other reservation protocols such as ST2 [10]. Independence from IP is desirable so that it can interwork with other network layer protocols such as IPX, NetBIOS, etc.

- Receiver heterogeneity: this refers to multicast communication where different receivers request different levels of service. For example, in a multicast group with many receivers, it is possible that one of the receivers desires a lower delay bound than the others. A better delay bound may be provided by increasing the amount of resources reserved along the path to that receiver while leaving the reservations for the other receivers unchanged. In its most complex form, receiver heterogeneity implies the ability to simultaneously provide various levels of service as requested by different receivers. In its simplest form, receiver heterogeneity will allow a scenario where some of the receivers use best effort service and those requiring service guarantees make a reservation. Receiver heterogeneity, especially for the reserved/best effort scenario, is a very desirable function. More details on supporting receiver heterogeneity are provided in Section 8.

- Support for different filter styles: It is desirable to provide support for the different filter styles defined by RSVP such as fixed filter, shared explicit and wildcard. Some of the issues with respect to supporting such filter styles in the link layer domain are examined in Section 8.

- Path Selection: In source routed LAN technologies such as Token Ring/IEEE 802.5, it may be useful for the mechanism to incorporate the function of path selection. Using an appropriate path selection mechanism may optimize utilization of network resources.

5.3. Non-goals

This document describes service mappings onto existing IEEE and ANSI defined standard MAC layers and uses standard MAC layer services as in IEEE 802.1 bridging. It does not attempt to make use of or describe the capabilities of other proprietary or standard MAC layer protocols although it should be noted that published work regarding MAC layers suitable for QoS mappings exists. These are outside the scope of the ISSLL working group charter.

5.4. Assumptions

This framework assumes that typical subnetworks that are concerned about QoS will be "switch rich"; i.e. most communication between end stations using integrated services support is expected to pass through at least one switch. The mechanisms and protocols described will be trivially extensible to communicating systems on the same shared medium, but it is important not to allow problem generalization which may complicate the targeted practical application to switch rich LAN topologies. There have also been developments in the area of MAC enhancements to ensure delay deterministic access on network links e.g. IEEE 802.12 [19] and also proprietary schemes.

Although we illustrate most examples for this model using RSVP as the upper layer QoS signaling protocol, there are actually no real dependencies on this protocol. RSVP could be replaced by some other dynamic protocol, or the requests could be made by network management or other policy entities. The SBM signaling protocol [14], which is based upon RSVP, is designed to work seamlessly in the architecture described in this memo.

There may be a heterogeneous mix of switches with different capabilities, all compliant with IEEE 802.1D [2,3], but implementing varied queuing and forwarding mechanisms ranging from simple systems with two queues per port and static priority scheduling, to more complex systems with multiple queues using WFQ or other algorithms.

The problem is decomposed into smaller independent parts which may lead to sub-optimal use of the network resources but we contend that such benefits are often equivalent to very small improvement in network efficiency in a LAN environment. Therefore, it is a goal that the switches in a network operate using a much simpler set of

information than the RSVP engine in a router. In particular, it is assumed that such switches do not need to implement per flow queuing and policing (although they are not precluded from doing so).

A fundamental assumption of the IntServ model is that flows are isolated from each other throughout their transit across a network. Intermediate queuing nodes are expected to shape or police the traffic to ensure conformance to the negotiated traffic flow specification. In the architecture proposed here for mapping to Layer 2, we diverge from that assumption in the interest of simplicity. The policing/shaping functions are assumed to be implemented in end stations. In some LAN environments, it is reasonable to assume that end stations are trusted to adhere to their negotiated contracts at the inputs to the network, and that we can afford to over-allocate resources during admission control to compensate for the inevitable packet jitter/bunching introduced by the switched network itself. This divergence has some implications on the types of receiver heterogeneity that can be supported and the statistical multiplexing gains that may be exploited, especially for Controlled Load flows. This is discussed in Section 8.7 of this document.

6. Basic Architecture

The functional requirements described in Section 5 will be performed by an entity which we refer to as the Bandwidth Manager (BM). The BM is responsible for providing mechanisms for an application or higher layer protocol to request QoS from the network. For architectural purposes, the BM consists of the following components.

6.1. Components

6.1.1. Requester Module

The Requester Module (RM) resides in every end station in the subnet. One of its functions is to provide an interface between applications or higher layer protocols such as RSVP, ST2, SNMP, etc. and the BM. An application can invoke the various functions of the BM by using the primitives for communication with the RM and providing it with the appropriate parameters. To initiate a reservation, in the link layer domain, the following parameters must be passed to the RM: the service desired (Guaranteed Service or Controlled Load), the traffic descriptors contained in the TSpec, and an RSpec specifying the amount of resources to be reserved [9]. More information on these parameters may be found in the relevant Integrated Services documents [6,7,8,9]. When RSVP is used for signaling at the network layer, this information is available and needs to be extracted from the RSVP PATH and RSVP RESV messages (See [5] for details). In addition to

these parameters, the network layer addresses of the end points must be specified. The RM must then translate the network layer addresses to link layer addresses and convert the request into an appropriate format which is understood by other components of the BM responsible admission control. The RM is also responsible for returning the status of requests processed by the BM to the invoking application or higher layer protocol.

6.1.2. Bandwidth Allocator

The Bandwidth Allocator (BA) is responsible for performing admission control and maintaining state about the allocation of resources in the subnet. An end station can request various services, e.g. bandwidth reservation, modification of an existing reservation, queries about resource availability, etc. These requests are processed by the BA. The communication between the end station and the BA takes place through the RM. The location of the BA will depend largely on the implementation method. In a centralized implementation, the BA may reside on a single station in the subnet. In a distributed implementation, the functions of the BA may be distributed in all the end stations and bridges/switches as necessary. The BA is also responsible for deciding how to label flows, e.g. based on the admission control decision, the BA may indicate to the RM that packets belonging to a particular flow be tagged with some priority value which maps to the appropriate traffic class.

6.1.3. Communication Protocols

The protocols for communication between the various components of the BM system must be specified. These include the following:

- Communication between the higher layer protocols and the RM: The BM must define primitives for the application to initiate reservations, query the BA about available resources, change change or delete reservations, etc. These primitives could be implemented as an API for an application to invoke functions of the BM via the RM.

- Communication between the RM and the BA: A signaling mechanism must be defined for the communication between the RM and the BA. This protocol will specify the messages which must be exchanged between the RM and the BA in order to service various requests by the higher layer entity.

- Communication between peer BAs: If there is more than one BA in the subnet, a means must be specified for inter-BA communication. Specifically, the BAs must be able to decide among themselves

about which BA would be responsible for which segments and bridges or switches. Further, if a request is made for resource reservation along the domain of multiple BAs, the BAs must be able to handle such a scenario correctly. Inter-BA communication will also be responsible for back-up and recovery in the event of failure.

6.2. Centralized vs. Distributed Implementations

Example scenarios are provided showing the location of the components of the bandwidth manager in centralized and fully distributed implementations. Note that in either case, the RM must be present in all end stations that need to make reservations. Essentially, centralized or distributed refers to the implementation of the BA, the component responsible for resource reservation and admission control. In the figures below, "App" refers to the application making use of the BM. It could either be a user application, or a higher layer protocol process such as RSVP.

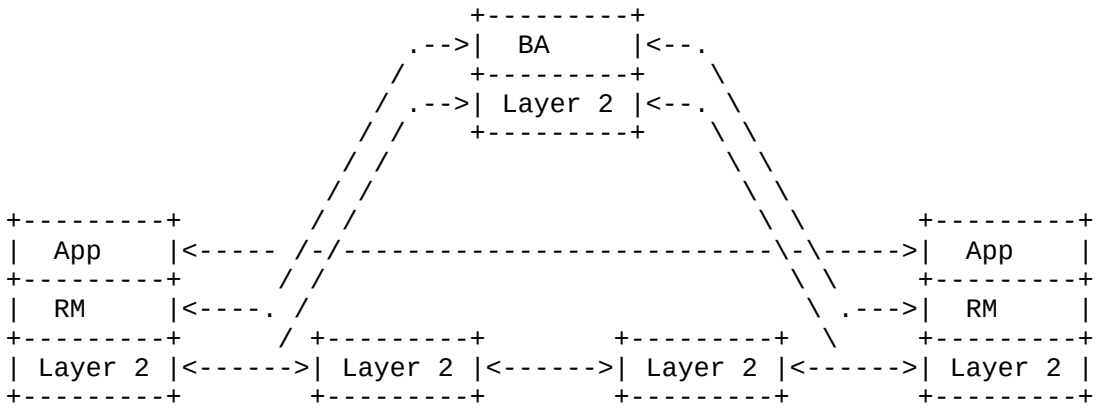

Figure 1: Bandwidth Manager with centralized Bandwidth Allocator

Figure 1 shows a centralized implementation where a single BA is responsible for admission control decisions for the entire subnet. Every end station contains a RM. Intermediate bridges and switches in the network need not have any functions of the BM since they will not be actively participating in admission control. The RM at the end station requesting a reservation initiates communication with its BA. For larger subnets, a single BA may not be able to handle the reservations for the entire subnet. In that case it would be necessary to deploy multiple BAs, each managing the resources of a

non-overlapping subset of segments. In a centralized implementation, the BA must have some knowledge of the Layer 2 topology of the subnet e.g., link layer spanning tree information, in order to be able to reserve resources on appropriate segments. Without this topology information, the BM would have to reserve resources on all segments for all flows which, in a switched network, would lead to very inefficient utilization of resources.

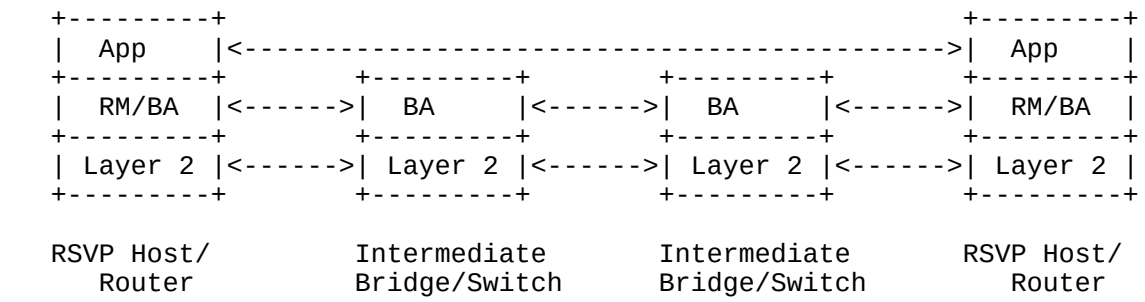

Figure 2: Bandwidth Manager with fully distributed Bandwidth Allocator

Figure 2 depicts the scenario of a fully distributed bandwidth manager. In this case, all devices in the subnet have BM functionality. All the end hosts are still required to have a RM. In addition, all stations actively participate in admission control. With this approach, each BA would need only local topology information since it is responsible for the resources on segments that are directly connected to it. This local topology information, such as a list of ports active on the spanning tree and which unicast addresses are reachable from which ports, is readily available in today's switches. Note that in the figures above, the arrows between peer layers are used to indicate logical connectivity.

7. Model of the Bandwidth Manager in a Network

In this section we describe how the model above fits with the existing IETF Integrated Services model of IP hosts and routers. First, we describe Layer 3 host and router implementations. Next, we describe how the model is applied in Layer 2 switches. Throughout we indicate any differences between centralized and distributed implementations. Occasional references are made to terminology from the Subnet Bandwidth Manager specification [14].

7.1. End Station Model

7.1.1. Layer 3 Client Model

We assume the same client model as IntServ and RSVP where we use the term "client" to mean the entity handling QoS in the Layer 3 device at each end of a Layer 2 Domain. In this model, the sending client is responsible for local admission control and packet scheduling onto its link in accordance with the negotiated service. As with the IntServ model, this involves per flow scheduling with possible traffic shaping/policing in every such originating node.

For now, we assume that the client runs an RSVP process which presents a session establishment interface to applications, provides signaling over the network, programs a scheduler and classifier in the driver, and interfaces to a policy control module. In particular, RSVP also interfaces to a local admission control module which is the focus of this section.

The following figure, reproduced from the RSVP specification, depicts the RSVP process in sending hosts.

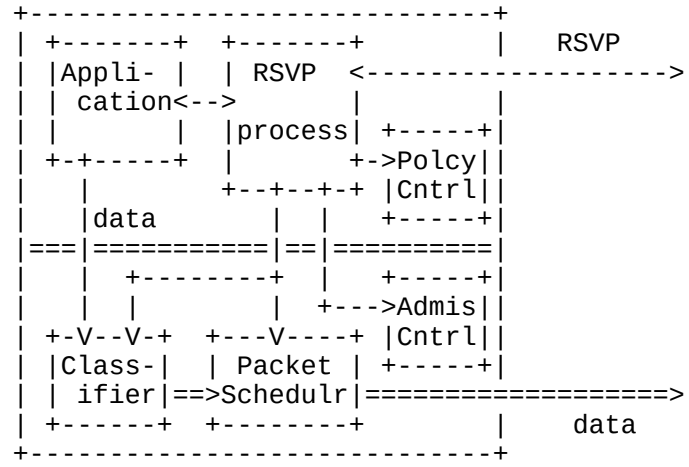

Figure 3: RSVP in Sending Hosts

7.1.2. Requests to Layer 2 ISSLL

The local admission control entity within a client is responsible for mapping Layer 3 session establishment requests into Layer 2 semantics.

The upper layer entity makes a request, in generalized terms to ISSLL of the form:

"May I reserve for traffic with <traffic characteristic> with <performance requirements> from <here> to <there> and how should I label it?"

where

<traffic characteristic> = Sender Tspec (e.g. bandwidth, burstiness, MTU)
<performance requirements> = FlowSpec (e.g. latency, jitter bounds)
<here> = IP address(es)
<there> = IP address(es) - may be multicast

7.1.3. At the Layer 3 Sender

The ISSLL functionality in the sender is illustrated in Figure 4.

The functions of the Requester Module may be summarized as follows:

- Maps the endpoints of the conversation to Layer 2 addresses in the LAN, so that the client can determine what traffic is going where. This function probably makes reference to the ARP protocol cache for unicast or performs an algorithmic mapping for multicast destinations.

- Communicates with any local Bandwidth Allocator module for local admission control decisions.

- Formats a SBM request to the network with the mapped addresses and flow/filter specs.

- Receives a response from the network and reports the admission control decision to the higher layer entity, along with any negotiated modifications to the session parameters.

- Saves any returned user_priority to be associated with this session in a "802 header" table. This will be used when constructing the Layer 2 headers for future data packets belonging to this session. This table might, for example, be indexed by the RSVP flow identifier.

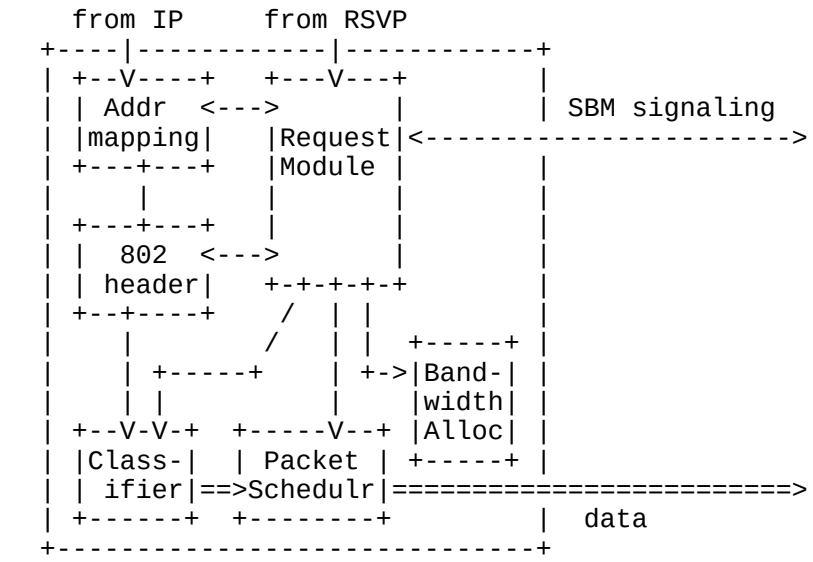

Figure 4: ISSLL in a Sending End Station

The Bandwidth Allocator (BA) component is only present when a distributed BA model is implemented. When present, its function is basically to apply local admission control for the outgoing link bandwidth and driver's queuing resources.

7.1.4. At the Layer 3 Receiver

The ISSLL functionality in the receiver is simpler and is illustrated in Figure 5.

The functions of the Requester Module may be summarized as follows:

- Handles any received SBM protocol indications.
- Communicates with any local BA for local admission control decisions.
- Passes indications up to RSVP if OK.
- Accepts confirmations from RSVP and relays them back via SBM signaling towards the requester.

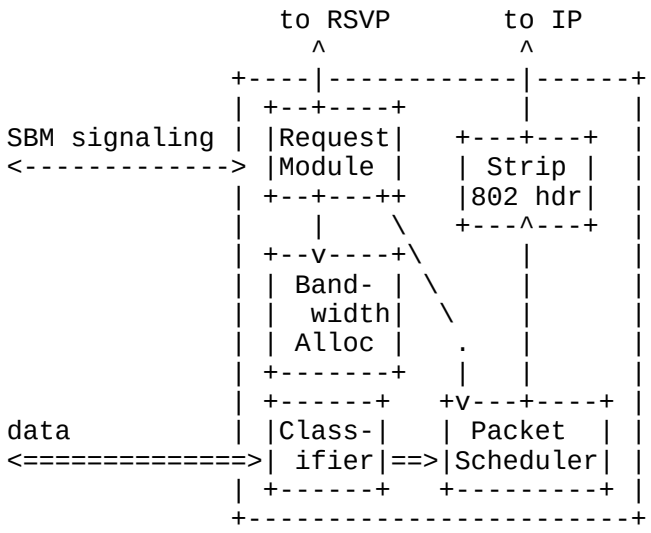

Figure 5: ISSLL in a Receiving End Station

- May program a receive classifier and scheduler, if used, to identify traffic classes of received packets and accord them appropriate treatment e.g., reservation of buffers for particular traffic classes.

- Programs the receiver to strip away link layer header information from received packets.

The Bandwidth Allocator, present only in a distributed implementation applies local admission control to see if a request can be supported with appropriate local receive resources.

7.2. Switch Model

7.2.1. Centralized Bandwidth Allocator

Where a centralized Bandwidth Allocator model is implemented, switches do not take part in the admission control process. Admission control is implemented by a centralized BA, e.g., a "Subnet Bandwidth Manager" (SBM) as described in [14]. This centralized BA may actually be co-located with a switch but its functions would not necessarily then be closely tied with the switch’s forwarding functions as is the case with the distributed BA described below.

7.2.2. Distributed Bandwidth Allocator

The model of Layer 2 switch behavior described here uses the terminology of the SBM protocol as an example of an admission control protocol. The model is equally applicable when other mechanisms, e.g. static configuration or network management, are in use for admission control. We define the following entities within the switch:

- Local Admission Control Module: One of these on each port accounts for the available bandwidth on the link attached to that port. For half duplex links, this involves taking account of the resources allocated to both transmit and receive flows. For full duplex links, the input port accountant's task is trivial.

- Input SBM Module: One instance on each port performs the "network" side of the signaling protocol for peering with clients or other switches. It also holds knowledge about the mappings of IntServ classes to user_priority.

- SBM Propagation Module: Relays requests that have passed admission control at the input port to the relevant output ports' SBM modules. This will require access to the switch's forwarding table (Layer-2 "routing table" cf. RSVP model) and port spanning tree state.

- Output SBM Module: Forwards requests to the next Layer 2 or Layer 3 hop.

- Classifier, Queue and Scheduler Module: The functions of this module are basically as described by the Forwarding Process of IEEE 802.1D (see Section 3.7 of [3]). The Classifier module identifies the relevant QoS information from incoming packets and uses this, together with the normal bridge forwarding database, to decide at which output port and traffic class to enqueue the packet. Different types of switches will use different techniques for flow identification (see Section 8.1). In IEEE 802.1D switches this information is the regenerated user_priority parameter which has already been decoded by the receiving MAC service and potentially remapped by the forwarding process (see Section 3.7.3 of [3]). This does not preclude more sophisticated classification rules such as the classification of individual IntServ flows. The Queue and Scheduler implement the

output queues for ports and provide the algorithm for servicing the queues for transmission onto the output link in order to provide the promised IntServ service. Switches will implement one or more output queues per port and all will implement at least a basic static priority dequeuing algorithm as their default, in accordance with IEEE 802.1D.

- Ingress Traffic Class Mapping and Policing Module: Its functions are as described in IEEE 802.1D Section 3.7. This optional module may police the data within traffic classes for conformance to the negotiated parameters, and may discard packets or re-map the user_priority. The default behavior is to pass things through unchanged.

- Egress Traffic Class Mapping Module: Its functions are as described in IEEE 802.1D Section 3.7. This optional module may perform re-mapping of traffic classes on a per output port basis. The default behavior is to pass things through unchanged.

Figure 6 shows all of the modules in an ISSLL enabled switch. The ISSLL model is a superset of the IEEE 802.1D bridge model.

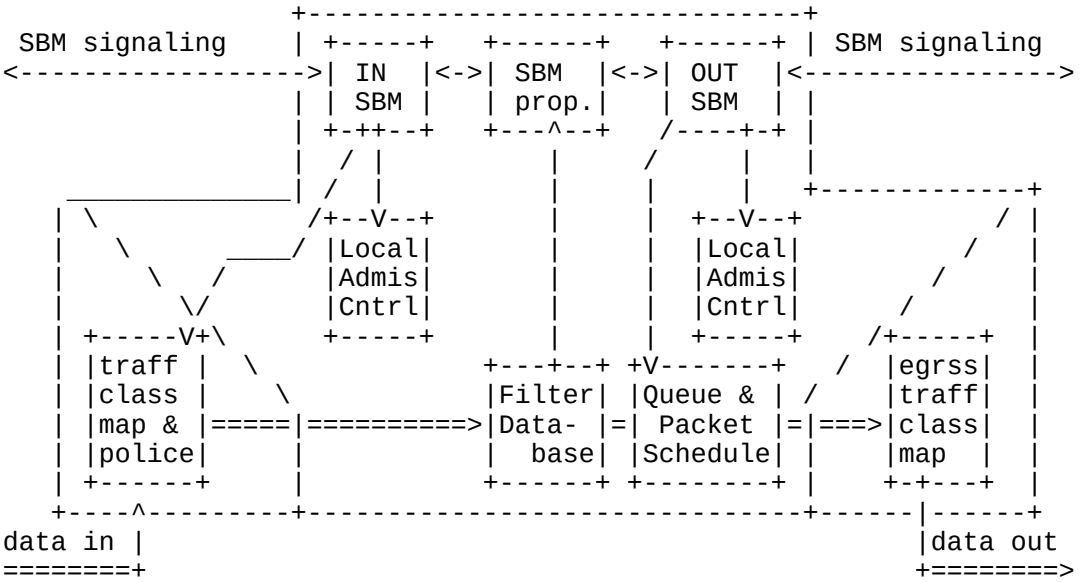

Figure 6: ISSLL in a Switch

7.3. Admission Control

On receipt of an admission control request, a switch performs the following actions, again using SBM as an example. The behavior is different depending on whether the "Designated SBM" for this segment is within this switch or not. See [14] for a more detailed specification of the DSBM/SBM actions.

- If the ingress SBM is the "Designated SBM" for this link, it either translates any received user_priority or selects a Layer 2 traffic class which appears compatible with the request and whose use does not violate any administrative policies in force. In effect, it matches the requested service with the available traffic classes and chooses the "best" one. It ensures that, if this reservation is successful, the value of user_priority corresponding to that traffic class is passed back to the client.

- The ingress DSBM observes the current state of allocation of resources on the input port/link and then determines whether the new resource allocation from the mapped traffic class can be accommodated. The request is passed to the reservation propagator if accepted.

- If the ingress SBM is not the "Designated SBM" for this link then it directly passes the request on to the reservation propagator.

- The reservation propagator relays the request to the bandwidth accountants on each of the switch’s outbound links to which this reservation would apply. This implies an interface to routing/forwarding database.

- The egress bandwidth accountant observes the current state of allocation of queuing resources on its outbound port and bandwidth on the link itself and determines whether the new allocation can be accommodated. Note that this is only a local decision at this switch hop; further Layer 2 hops through the network may veto the request as it passes along.

- The request, if accepted by this switch, is propagated on each output link selected. Any user_priority described in the forwarded request must be translated according to any egress mapping table.

- If accepted, the switch must notify the client of the user_priority to be used for packets belonging to that flow. Again, this is an optimistic approach assuming that admission control succeeds; downstream switches may refuse the request.

- If this switch wishes to reject the request, it can do so by notifying the client that originated the request by means of its Layer 2 address.

7.4. QoS Signaling

The mechanisms described in this document make use of a signaling protocol for devices to communicate their admission control requests across the network. The service definitions to be provided by such a protocol e.g. [14] are described below. We illustrate the primitives and information that need to be exchanged with such a signaling protocol entity. In all of the examples, appropriate delete/cleanup mechanisms will also have to be provided for tearing down established sessions.

7.4.1. Client Service Definitions

The following interfaces can be identified from Figures 4 and 5.

- SBM <-> Address Mapping

  This is a simple lookup function which may require ARP protocol interactions or an algorithmic mapping. The Layer 2 addresses are needed by SBM for inclusion in its signaling messages to avoid requiring that switches participating in the signaling have Layer 3 information to perform the mapping.

  l2_addr = map_address(ip_addr)

- SBM <-> Session/Link Layer Header

  This is for notifying the transmit path of how to add Layer 2 header information, e.g. user_priority values to the traffic of each outgoing flow. The transmit path will provide the user_priority value when it requests a MAC layer transmit operation for each packet. The user_priority is one of the parameters passed in the packet transmit primitive defined by the IEEE 802 service model.

  bind_l2_header(flow_id, user_priority)

- SBM <-> Classifier/Scheduler

  This is for notifying transmit classifier/scheduler of any additional Layer 2 information associated with scheduling the transmission of a packet flow. This primitive may be unused in some implementations or it may be used, for example, to provide information to a transmit scheduler that is performing per traffic

class scheduling in addition to the per flow scheduling required by IntServ; the Layer 2 header may be a pattern (in addition to the FilterSpec) to be used to identify the flow's traffic.

bind_l2schedulerinfo(flow_id, , l2_header, traffic_class)

- SBM <-> Local Admission Control

  This is used for applying local admission control for a session e.g. is there enough transmit bandwidth still uncommitted for this new session? Are there sufficient receive buffers? This should commit the necessary resources if it succeeds. It will be necessary to release these resources at a later stage if the admission control fails at a subsequent node. This call would be made, for example, by a segment's Designated SBM.

  status = admit_l2session(flow_id, Tspec, FlowSpec)

- SBM <-> RSVP

  This is outlined above in Section 7.1.2 and fully described in [14].

- Management Interfaces

  Some or all of the modules described by this model will also require configuration management. It is expected that details of the manageable objects will be specified by future work in the ISSLL WG.

7.4.2. Switch Service Definitions

The following interfaces are identified from Figure 6.

- SBM <-> Classifier

  This is for notifying the receive classifier of how to match incoming Layer 2 information with the associated traffic class. It may in some cases consist of a set of read only default mappings.

  bind_l2classifierinfo(flow_id, l2_header, traffic_class)

- SBM <-> Queue and Packet Scheduler

  This is for notifying transmit scheduler of additional Layer 2 information associated with a given traffic class. It may be unused in some cases (see discussion in previous section).

bind_l2schedulerinfo(flow_id, l2_header, traffic_class)

- SBM <-> Local Admission Control

  Same as for the host discussed above.

- SBM <-> Traffic Class Map and Police

  Optional configuration of any user_priority remapping that might be implemented on ingress to and egress from the ports of a switch. For IEEE 802.1D switches, it is likely that these mappings will have to be consistent across all ports.

  bind_l2ingressprimap(inport, in_user_pri, internal_priority)
  bind_l2egressprimap(outport, internal_priority, out_user_pri)

  Optional configuration of any Layer 2 policing function to be applied on a per class basis to traffic matching the Layer 2 header. If the switch is capable of per flow policing then existing IntServ/RSVP models will provide a service definition for that configuration.

  bind_l2policing(flow_id, l2_header, Tspec, FlowSpec)

- SBM <-> Filtering Database

  SBM propagation rules need access to the Layer 2 forwarding database to determine where to forward SBM messages. This is analogous to RSRR interface in Layer 3 RSVP.

  output_portlist = lookup_l2dest(l2_addr)

- Management Interfaces

  Some or all of the modules described by this model will also require configuration management. It is expected that details of the manageable objects will be specified by future work in the ISSLL working group.

8. Implementation Issues

As stated earlier, the Integrated Services working group has defined various service classes offering varying degrees of QoS guarantees. Initial effort will concentrate on enabling the Controlled Load [6] and Guaranteed Service classes [7]. The Controlled Load service provides a loose guarantee, informally stated as "the same as best effort would be on an unloaded network". The Guaranteed Service provides an upper bound on the transit delay of any packet. The

extent to which these services can be supported at the link layer will depend on many factors including the topology and technology used. Some of the mapping issues are discussed below in light of the emerging link layer standards and the functions supported by higher layer protocols. Considering the limitations of some of the topologies, it may not be possible to satisfy all the requirements for Integrated Services on a given topology. In such cases, it is useful to consider providing support for an approximation of the service which may suffice in most practical instances. For example, it may not be feasible to provide policing/shaping at each network element (bridge/switch) as required by the Controlled Load specification. But if this task is left to the end stations, a reasonably good approximation to the service can be obtained.

8.1. Switch Characteristics

There are many LAN bridges/switches with varied capabilities for supporting QoS. We discuss below the various kinds of devices that that one may expect to find in a LAN environment.

The most basic bridge is one which conforms to the IEEE 802.1D specification of 1993 [2]. This device has a single queue per output port, and uses the spanning tree algorithm to eliminate topology loops. Networks constructed from this kind of device cannot be expected to provide service guarantees of any kind because of the complete lack of traffic isolation.

The next level of bridges/switches are those which conform to the more recently revised IEEE 802.1D specification [3]. They include support for queuing up to eight traffic classes separately. The level of traffic isolation provided is coarse because all flows corresponding to a particular traffic class are aggregated. Further, it is likely that more than one priority will map to a traffic class depending on the number of queues implemented in the switch. It would be difficult for such a device to offer protection against misbehaving flows. The scope of multicast traffic may be limited by using GMRP to only those segments which are on the path to interested receivers.

A next step above these devices are bridges/switches which implement optional parts of the IEEE 802.1D specification such as mapping the received user_priority to some internal set of canonical values on a per-input-port basis. It may also support the mapping of these internal canonical values onto transmitted user_priority on a per-output-port basis. With these extra capabilities, network administrators can perform mapping of traffic classes between specific pairs of ports, and in doing so gain more control over admission of traffic into the protected classes.

Other entirely optional features that some bridges/switches may support include classification of IntServ flows using fields in the network layer header, per-flow policing and/or reshaping which is essential for supporting Guaranteed Service, and more sophisticated scheduling algorithms such as variants of weighted fair queuing to limit the bandwidth consumed by a traffic class. Note that it is advantageous to perform flow isolation and for all network elements to police each flow in order to support the Controlled Load and Guaranteed Service.

8.2. Queuing

Connectionless packet networks in general, and LANs in particular, work today because of scaling choices in network provisioning. Typically, excess bandwidth and buffering is provisioned in the network to absorb the traffic sourced by higher layer protocols, often sufficient to cause their transmission windows to run out on a statistical basis, so that network overloads are rare and transient and the expected loading is very low.

With the advent of time-critical traffic such over-provisioning has become far less easy to achieve. Time-critical frames may be queued for annoyingly long periods of time behind temporary bursts of file transfer traffic, particularly at network bottleneck points, e.g. at the 100 Mbps to 10 Mbps transition that might occur between the riser to the wiring closet and the final link to the user from a desktop switch. In this case, however, if it is known a priori (either by application design, on the basis of statistics, or by administrative control) that time-critical traffic is a small fraction of the total bandwidth, it suffices to give it strict priority over the non-time-critical traffic. The worst case delay experienced by the time-critical traffic is roughly the maximum transmission time of a maximum length non-time-critical frame -- less than a millisecond for 10 Mbps Ethernet, and well below the end to end delay budget based on human perception times.

When more than one priority service is to be offered by a network element e.g. one which supports both Controlled Load as well as Guaranteed Service, the requirements for the scheduling discipline become more complex. In order to provide the required isolation between the service classes, it will probably be necessary to queue them separately. There is then an issue of how to service the queues which requires a combination of admission control and more intelligent queuing disciplines. As with the service specifications themselves, the specification of queuing algorithms is beyond the scope of this document.

8.3. Mapping of Services to Link Level Priority

The number of traffic classes supported and access methods of the technology under consideration will determine how many and what services may be supported. Native Token Ring/IEEE 802.5, for instance, supports eight priority levels which may be mapped to one or more traffic classes. Ethernet/IEEE 802.3 has no support for signaling priorities within frames. However, the IEEE 802 standards committee has recently developed a new standard for bridges/switches related to multimedia traffic expediting and dynamic multicast filtering [3]. A packet format for carrying a user_priority field on all IEEE 802 LAN media types is now defined in [4]. These standards allow for up to eight traffic classes on all media. The user_priority bits carried in the frame are mapped to a particular traffic class within a bridge/switch. The user_priority is signaled on an end-to-end basis, unless overridden by bridge/switch management. The traffic class that is used by a flow should depend on the quality of service desired and whether the reservation is successful or not. Therefore, a sender should use the user_priority value which maps to the best effort traffic class until told otherwise by the BM. The BM will, upon successful completion of resource reservation, specify the value of user_priority to be used by the sender for that session’s data. An accompanying memo [13] addresses the issue of mapping the various Integrated Services to appropriate traffic classes.

8.4. Re-mapping of Non-conforming Aggregated Flows

One other topic under discussion in the IntServ context is how to handle the traffic for data flows from sources that exceed their negotiated traffic contract with the network. An approach that shows some promise is to treat such traffic with "somewhat less than best effort" service in order to protect traffic that is normally given "best effort" service from having to back off. Best effort traffic is often adaptive, using TCP or other congestion control algorithms, and it would be unfair to penalize those flows due to badly behaved traffic from reserved flows which are often set up by non-adaptive applications.

A possible solution might be to assign normal best effort traffic to one user_priority and to label excess non-conforming traffic as a lower user_priority although the re-ordering problems that might arise from doing this may make this solution undesirable, particularly if the flows are using TCP. For this reason the controlled load service recommends dropping excess traffic, rather than re-mapping to a lower priority. This is further discussed below.

8.5. Override of Incoming User Priority

In some cases, a network administrator may not trust the user_priority values contained in packets from a source and may wish to map these into some more suitable set of values. Alternatively, due perhaps to equipment limitations or transition periods, the user_priority values may need to be re-mapped as the data flows to/from different regions of a network.

Some switches may implement such a function on input that maps received user_priority to some internal set of values. This function is provided by a table known in IEEE 802.1D as the User Priority Regeneration Table (Table 3-1 in [3]). These values can then be mapped using an output table described above onto outgoing user_priority values. These same mappings must also be used when applying admission control to requests that use the user_priority values (see e.g. [14]). More sophisticated approaches are also possible where a device polices traffic flows and adjusts their onward user_priority based on their conformance to the admitted traffic flow specifications.

8.6. Different Reservation Styles

In the figure above, SW is a bridge/switch in the link layer domain. S1, S2, S3, R1 and R2 are end stations which are members of a group associated with the same RSVP flow. S1, S2 and S3 are upstream end stations. R1 and R2 are the downstream end stations which receive traffic from all the senders. RSVP allows receivers R1 and R2 to specify reservations which can apply to: (a) one specific sender only (fixed filter); (b) any of two or more explicitly specified senders (shared explicit filter); and (c) any sender in the group (shared wildcard filter). Support for the fixed filter style is straightforward; a separate reservation is made for the traffic from each of the senders. However, support for the other two filter styles has implications regarding policing; i.e. the merged flow from the different senders must be policed so that they conform to traffic parameters specified in the filter’s RSpec. This scenario is further complicated if the services requested by R1 and R2 are different. Therefore, in the absence of policing within bridges/switches, it may be possible to support only fixed filter reservations at the link layer.

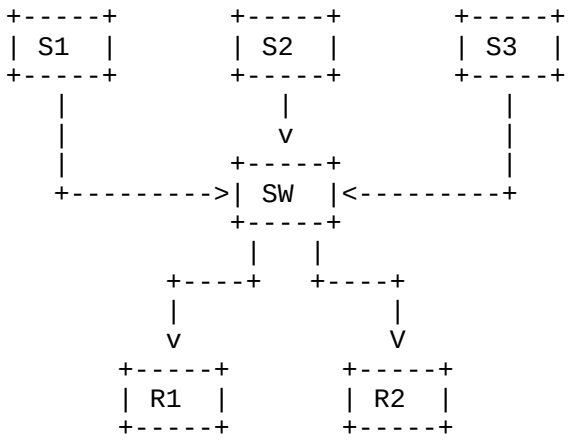

Figure 7: Illustration of filter styles

8.7. Receiver Heterogeneity

At Layer 3, the IntServ model allows heterogeneous receivers for multicast flows where different branches of a tree can have different types of reservations for a given multicast destination. It also supports the notion that trees may have some branches with reserved flows and some using best effort service. If we were to treat a Layer 2 subnet as a single network element as defined in [8], then all of the branches of the distribution tree that lie within the subnet could be assumed to require the same QoS treatment and be treated as an atomic unit as regards admission control, etc. With this assumption, the model and protocols already defined by IntServ and RSVP already provide sufficient support for multicast heterogeneity. Note, however, that an admission control request may well be rejected because just one link in the subnet is oversubscribed leading to rejection of the reservation request for the entire subnet.

As an example, consider Figure 8, SW is a Layer 2 device (bridge/switch) participating in resource reservation, S is the upstream source end station and R1 and R2 are downstream end station receivers. R1 would like to make a reservation for the flow while R2 would like to receive the flow using best effort service. S sends RSVP PATH messages which are multicast to both R1 and R2. R1 sends an RSVP RESV message to S requesting the reservation of resources.

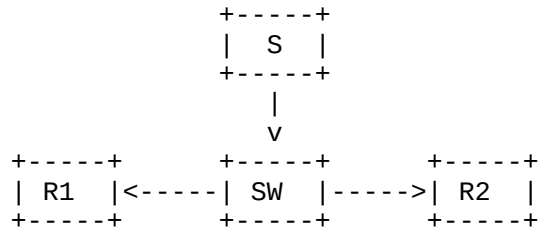

Figure 8: Example of receiver heterogeneity

If the reservation is successful at Layer 2, the frames addressed to the group will be categorized in the traffic class corresponding to the service requested by R1. At SW, there must be some mechanism which forwards the packet providing service corresponding to the reserved traffic class at the interface to R1 while using the best effort traffic class at the interface to R2. This may involve changing the contents of the frame itself, or ignoring the frame priority at the interface to R2.

Another possibility for supporting heterogeneous receivers would be to have separate groups with distinct MAC addresses, one for each class of service. By default, a receiver would join the "best effort" group where the flow is classified as best effort. If the receiver makes a reservation successfully, it can be transferred to the group for the class of service desired. The dynamic multicast filtering capabilities of bridges and switches implementing the IEEE 802.1D standard would be a very useful feature in such a scenario. A given flow would be transmitted only on those segments which are on the path between the sender and the receivers of that flow. The obvious disadvantage of such an approach is that the sender needs to send out multiple copies of the same packet corresponding to each class of service desired thus potentially duplicating the traffic on a portion of the distribution tree.

The above approaches would provide very sub-optimal utilization of resources given the expected size and complexity of the Layer 2 subnets. Therefore, it is desirable to enable switches to apply QoS differently on different egress branches of a tree that divide at that switch.

IEEE 802.1D specifies a basic model for multicast whereby a switch makes multicast forwarding decisions based on the destination address. This would produce a list of output ports to which the packet should be forwarded. In its default mode, such a switch would use the user_priority value in received packets, or a value regenerated on a per input port basis in the absence of an explicit value, to enqueue the packets at each output port. Any IEEE 802.1D switch which supports multiple traffic classes can support this operation.

If a switch selects per port output queues based only on the incoming user_priority, as described by IEEE 802.1D, it must treat all branches of all multicast sessions within that user_priority class with the same queuing mechanism. Receiver heterogeneity is then not possible and this could well lead to the failure of an admission control request for the whole multicast session due to a single link being oversubscribed. Note that in the Layer 2 case as distinct from the Layer 3 case with RSVP/IntServ, the option of having some receivers getting the session with the requested QoS and some getting it best effort does not exist as basic IEEE 802.1 switches are unable to re-map the user_priority on a per link basis. This could become an issue with heavy use of dynamic multicast sessions. If a switch were to implement a separate user_priority mapping at each output port, then, in some cases, reservations can use a different traffic class on different paths that branch at such a switch in order to provide multiple receivers with different QoS. This is possible if all flows within a traffic class at the ingress to a switch egress in the same traffic class on a port. For example, traffic may be forwarded using user_priority 4 on one branch where receivers have performed admission control and as user_priority 0 on ones where they have not. We assume that per user_priority queuing without taking account of input or output ports is the minimum standard functionality for switches in a LAN environment (IEEE 802.1D) but that more functional Layer 2 or even Layer 3 switches (i.e. routers) can be used if even more flexible forms of heterogeneity are considered necessary to achieve more efficient resource utilization. The behavior of Layer 3 switches in this context is already well standardized by the IETF.

9. Network Topology Scenarios

The extent to which service guarantees can be provided by a network depend to a large degree on the ability to provide the key functions of flow identification and scheduling in addition to admission control and policing. This section discusses some of the capabilities of the LAN technologies under consideration and provides a taxonomy of possible topologies emphasizing the capabilities of each with regard to supporting the above functions. For the

technologies considered here, the basic topology of a LAN may be shared, switched half duplex or switched full duplex. In the shared topology, multiple senders share a single segment. Contention for media access is resolved using protocols such as CSMA/CD in Ethernet and token passing in Token Ring and FDDI. Switched half duplex, is essentially a shared topology with the restriction that there are only two transmitters contending for resources on any segment. Finally, in a switched full duplex topology, a full bandwidth path is available to the transmitter at each end of the link at all times. Therefore, in this topology, there is no need for any access control mechanism such as CSMA/CD or token passing as there is no contention between the transmitters. Obviously, this topology provides the best QoS capabilities. Another important element in the discussion of topologies is the presence or absence of support for multiple traffic classes. These were discussed earlier in Section 4.1. Depending on the basic topology used and the ability to support traffic classes, we identify six scenarios as follows:

1. Shared topology without traffic classes.
2. Shared topology with traffic classes.
3. Switched half duplex topology without traffic classes.
4. Switched half duplex topology with traffic classes.
5. Switched full duplex topology without traffic classes.
6. Switched full duplex topology with traffic classes.

There is also the possibility of hybrid topologies where two or more of the above coexist. For instance, it is possible that within a single subnet, there are some switches which support traffic classes and some which do not. If the flow in question traverses both kinds of switches in the network, the least common denominator will prevail. In other words, as far as that flow is concerned, the network is of the type corresponding to the least capable topology that is traversed. In the following sections, we present these scenarios in further detail for some of the different IEEE 802 network types with discussion of their abilities to support the IntServ services.

9.1. Full Duplex Switched Networks

On a full duplex switched LAN, the MAC protocol is unimportant as as access is concerned, but must be factored into the characterization parameters advertised by the device since the access latency is equal to the time required to transmit the largest packet. Approximate values for the characteristics on various media are provided in the following tables. These delays should be also be considered in the context of the speed of light delay which is approximately 400 ns for typical 100 m UTP links and 7 us for typical 2 km multimode fiber links.

Table 4: Full duplex switched media access latency

Type	Speed	Max Pkt Length	Max Access Latency
Ethernet	10 Mbps	1.2 ms	1.2 ms
	100 Mbps	120 us	120 us
	1 Gbps	12 us	12 us
Token Ring	4 Mbps	9 ms	9 ms
	16 Mbps	9 ms	9 ms
FDDI	100 Mbps	360 us	8.4 ms
Demand Priority	100 Mbps	120 us	120 us

Full duplex switched network topologies offer good QoS capabilities for both Controlled Load and Guaranteed Service when supported by suitable queuing strategies in the switches.

9.2. Shared Media Ethernet Networks

Thus far, we have not discussed the difficulty of dealing with allocation on a single shared CSMA/CD segment. As soon as any CSMA/CD algorithm is introduced the ability to provide any form of Guaranteed Service is seriously compromised in the absence of any tight coupling between the multiple senders on the link. There are a number of reasons for not offering a better solution to this problem.

Firstly, we do not believe this is a truly solvable problem as it would require changes to the MAC protocol. IEEE 802.1 has examined research showing disappointing simulation results for performance guarantees on shared CSMA/CD Ethernet without MAC enhancements. There have been proposals for enhancements to the MAC layer protocols, e.g. BLAM and enhanced flow control in IEEE 802.3. However, any solution involving an enhanced software MAC running above the traditional IEEE 802.3 MAC, or other proprietary MAC protocols, is outside the scope of the ISSLL working group and this document. Secondly, we are not convinced that it is really an interesting problem. While there will be end stations on shared segments for some time to come, the number of deployed switches is steadily increasing relative to the number of stations on shared segments. This trend is proceeding to the point where it may be satisfactory to have a solution which assumes that any network communication requiring resource reservations will take place through at least one switch or router. Put another way, the easiest upgrade to existing Layer 2 infrastructure for QoS support is the installation of segment switching. Only when this has been done is it worthwhile to investigate more complex solutions involving

admission control. Thirdly, the core of campus networks typically consists of solutions based on switches rather than on repeated segments. There may be special circumstances in the future, e.g. Gigabit buffered repeaters, but the characteristics of these devices are different from existing CSMA/CD repeaters anyway.

Table 5: Shared Ethernet media access latency

Type	Speed	Max Pkt Length	Max Access Latency
Ethernet	10 Mbps	1.2 ms	unbounded
	100 Mbps	120 us	unbounded
	1 Gbps	12 us	unbounded

9.3. Half Duplex Switched Ethernet Networks

Many of the same arguments for sub optimal support of Guaranteed Service on shared media Ethernet also apply to half duplex switched Ethernet. In essence, this topology is a medium that is shared between at least two senders contending for packet transmission. Unless these are tightly coupled and cooperative, there is always the chance that the best effort traffic of one will interfere with the reserved traffic of the other. Dealing with such a coupling would require some form of modification to the MAC protocol.

Not withstanding the above argument, half duplex switched topologies do seem to offer the chance to provide Controlled Load service. With the knowledge that there are exactly two potential senders that are both using prioritization for their Controlled Load traffic over best effort flows, and with admission control having been done for those flows based on that knowledge, the media access characteristics while not deterministic are somewhat predictable. This is probably a close enough useful approximation to the Controlled Load service.

Table 6: Half duplex switched Ethernet media access latency

Type	Speed	Max Pkt Length	Max Access Latency
Ethernet	10 Mbps	1.2 ms	unbounded
	100 Mbps	120 us	unbounded
	1 Gbps	12 us	unbounded

9.4. Half Duplex Switched and Shared Token Ring Networks

In a shared Token Ring network, the network access time for high priority traffic at any station is bounded and is given by (N+1)*THTmax, where N is the number of stations sending high priority traffic and THTmax is the maximum token holding time [14]. This assumes that network adapters have priority queues so that reservation of the token is done for traffic with the highest priority currently queued in the adapter. It is easy to see that access times can be improved by reducing N or THTmax. The recommended default for THTmax is 10 ms [6]. N is an integer from 2 to 256 for a shared ring and 2 for a switched half duplex topology. A similar analysis applies for FDDI.

Table 7: Half duplex switched and shared Token Ring media access latency

Type	Speed	Max Pkt Length	Max Access Latency
Token Ring	4/16 Mbps shared	9 ms	2570 ms
	4/16 Mbps switched	9 ms	30 ms
FDDI	100 Mbps	360 us	8 ms

Given that access time is bounded, it is possible to provide an upper bound for end-to-end delays as required by Guaranteed Service assuming that traffic of this class uses the highest priority allowable for user traffic. The actual number of stations that send traffic mapped into the same traffic class as Guaranteed Service may vary over time but, from an admission control standpoint, this value is needed a priori. The admission control entity must therefore use a fixed value for N, which may be the total number of stations on the ring or some lower value if it is desired to keep the offered delay guarantees smaller. If the value of N used is lower than the total number of stations on the ring, admission control must ensure that the number of stations sending high priority traffic never exceeds

this number. This approach allows admission control to estimate worst case access delays assuming that all of the N stations are sending high priority data even though, in most cases, this will mean that delays are significantly overestimated.

Assuming that Controlled Load flows use a traffic class lower than that used by Guaranteed Service, no upper bound on access latency can be provided for Controlled Load flows. However, Controlled Load flows will receive better service than best effort flows.

Note that on many existing shared Token Rings, bridges transmit frames using an Access Priority (see Section 4.3) value of 4 irrespective of the user_priority carried in the frame control field of the frame. Therefore, existing bridges would need to be reconfigured or modified before the above access time bounds can actually be used.

9.5. Half Duplex and Shared Demand Priority Networks

In IEEE 802.12 networks, communication between end nodes and hubs and between the hubs themselves is based on the exchange of link control signals. These signals are used to control access to the shared medium. If a hub, for example, receives a high priority request while another hub is in the process of serving normal priority requests, then the service of the latter hub can effectively be preempted in order to serve the high priority request first. After the network has processed all high priority requests, it resumes the normal priority service at the point in the network at which it was interrupted.

The network access time for high priority packets is basically the time needed to preempt normal priority network service. This access time is bounded and it depends on the physical layer and on the topology of the shared network. The physical layer has a significant impact when operating in half duplex mode as, e.g. when used across unshielded twisted pair cabling (UTP) links, because link control signals cannot be exchanged while a packet is transmitted over the link. Therefore the network topology has to be considered since, in larger shared networks, the link control signals must potentially traverse several links and hubs before they can reach the hub which has the network control function. This may delay the preemption of the normal priority service and hence increase the upper bound that may be guaranteed.

Upper bounds on the high priority access time are given below for a UTP physical layer and a cable length of 100 m between all end nodes and hubs using a maximum propagation delay of 570 ns as defined in

[19]. These values consider the worst case signaling overhead and assume the transmission of maximum sized normal priority data packets while the normal priority service is being preempted.

Table 8: Half duplex switched Demand Priority UTP access latency

Type	Speed	Max Pkt Length	Max Access Latency
Demand Priority	100 Mbps, 802.3 pkt, UTP	120 us	254 us
	802.5 pkt, UTP	360 us	733 us

Shared IEEE 802.12 topologies can be classified using the hub cascading level "N". The simplest topology is the single hub network (N = 1). For a UTP physical layer, a maximum cascading level of N = 5 is supported by the standard. Large shared networks with many hundreds of nodes may be built with a level 2 topology. The bandwidth manager could be informed about the actual cascading level by network management mechanisms and can use this information in its admission control algorithms.

In contrast to UTP, the fiber optic physical layer operates in dual simplex mode. Upper bounds for the high priority access time are given below for 2 km multimode fiber links with a propagation delay of 10 us.

For shared media with distances of up to 2 km between all end nodes and hubs, the IEEE 802.12 standard allows a maximum cascading level of 2. Higher levels of cascaded topologies are supported but require a reduction of the distances [15].

The bounded access delay and deterministic network access allow the support of service commitments required for Guaranteed Service and Controlled Load, even on shared media topologies. The support of just two priority levels in 802.12, however, limits the number of services that can simultaneously be implemented across the network.

Table 9: Shared Demand Priority UTP access latency

Type	Speed	Max Pkt Length	Max Access Latency	Topology
Demand Priority	100 Mbps, 802.3 pkt	120 us	262 us	N = 1
		120 us	554 us	N = 2
		120 us	878 us	N = 3
		120 us	1.24 ms	N = 4
		120 us	1.63 ms	N = 5
Demand Priority	100 Mbps, 802.5 pkt	360 us	722 us	N = 1
		360 us	1.41 ms	N = 2
		360 us	2.32 ms	N = 3
		360 us	3.16 ms	N = 4
		360 us	4.03 ms	N = 5

Table 10: Half duplex switched Demand Priority fiber access latency

Type	Speed	Max Pkt Length	Max Access Latency
Demand Priority	100 Mbps, 802.3 pkt, fiber	120 us	139 us
	802.5 pkt, fiber	360 us	379 us

Table 11: Shared Demand Priority fiber access latency

Type	Speed	Max Pkt Length	Max Access Latency	Topology
Demand Priority	100 Mbps, 802.3 pkt	120 us	160 us	N = 1
		120 us	202 us	N = 2
Demand Priority	100 Mbps, 802.5 pkt	360 us	400 us	N = 1
		360 us	682 us	N = 2

10. Justification

An obvious concern is the complexity of this model. It essentially does what RSVP already does at Layer 3, so why do we think we can do better by reinventing the solution to this problem at Layer 2?

The key is that there are a number of simple Layer 2 scenarios that cover a considerable portion of the real QoS problems that will occur. A solution that covers the majority of problems at significantly lower cost is beneficial. Full RSVP/IntServ with per flow queuing in strategically positioned high function switches or routers may be needed to completely resolve all issues, but devices implementing the architecture described in herein will allow for a significantly simpler network.

11. Summary

This document has specified a framework for providing Integrated Services over shared and switched LAN technologies. The ability to provide QoS guarantees necessitates some form of admission control and resource management. The requirements and goals of a resource management scheme for subnets have been identified and discussed. We refer to the entire resource management scheme as a Bandwidth Manager. Architectural considerations were discussed and examples were provided to illustrate possible implementations of a Bandwidth Manager. Some of the issues involved in mapping the services from higher layers to the link layer have also been discussed. Accompanying memos from the ISSLL working group address service mapping issues [13] and provide a protocol specification for the Bandwidth Manager protocol [14] based on the requirements and goals discussed in this document.

References

[1] IEEE Standards for Local and Metropolitan Area Networks: Overview and Architecture, ANSI/IEEE Std 802, 1990.

[2] ISO/IEC 10038 Information technology - Telecommunications and information exchange between systems - Local area networks - Media Access Control (MAC) Bridges, (also ANSI/IEEE Std 802.1D-1993), 1993.

[3] ISO/IEC 15802-3 Information technology - Telecommunications and information exchange between systems - Local and metropolitan area networks - Common specifications - Part 3: Media Access Control (MAC) bridges (also ANSI/IEEE Std 802.1D-1998), 1998.

[4] IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks, IEEE Std 802.1Q-1998, 1998.

[5] Braden, B., Zhang, L., Berson, S., Herzog, S. and S. Jamin, "Resource Reservation Protocol (RSVP) - Version 1 Functional Specification", RFC 2205, September 1997.

[6] Wroclawski, J., "Specification of the Controlled Load Network Element Service", RFC 2211, September 1997.

[7] Shenker, S., Partridge, C. and R. Guerin, "Specification of Guaranteed Quality of Service", RFC 2212, September 1997.

[8] Braden, R., Clark, D. and S. Shenker, "Integrated Services in the Internet Architecture: An Overview", RFC 1633, June 1994.

[9] Wroclawski, J., "The Use of RSVP with IETF Integrated Services", RFC 2210, September 1997.

[10] Shenker, S. and J. Wroclawski, "Network Element Service Specification Template", RFC 2216, September 1997.

[11] Shenker, S. and J. Wroclawski, "General Characterization Parameters for Integrated Service Network Elements", RFC 2215, September 1997.

[12] Delgrossi, L. and L. Berger (Editors), "Internet Stream Protocol Version 2 (ST2) Protocol Specification - Version ST2+", RFC 1819, August 1995.

[13] Seaman, M., Smith, A. and E. Crawley, "Integrated Service Mappings on IEEE 802 Networks", RFC 2815, May 2000.

[14] Yavatkar, R., Hoffman, D., Bernet, Y. and F. Baker, "SBM Subnet Bandwidth Manager): Protocol for RSVP-based Admission Control Over IEEE 802-style Networks", RFC 2814, May 2000.

[15] ISO/IEC 8802-3 Information technology - Telecommunications and information exchange between systems - Local and metropolitan area networks - Common specifications - Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications, (also ANSI/IEEE Std 802.3-1996), 1996.

[15] ISO/IEC 8802-5 Information technology - Telecommunications and information exchange between systems - Local and metropolitan area networks - Common specifications - Part 5: Token Ring Access Method and Physical Layer Specifications, (also ANSI/IEEE Std 802.5-1995), 1995.

[17] Postel, J. and J. Reynolds, "A Standard for the Transmission of IP Datagrams over IEEE 802 Networks", STD 43, RFC 1042, February 1988.

[18] C. Bisdikian, B. V. Patel, F. Schaffa, and M Willebeek-LeMair, The Use of Priorities on Token Ring Networks for Multimedia Traffic, IEEE Network, Nov/Dec 1995.

[19] IEEE Standards for Local and Metropolitan Area Networks: Demand Priority Access Method, Physical Layer and Repeater Specification for 100 Mb/s Operation, IEEE Std 802.12-1995.

[20] Fiber Distributed Data Interface MAC, ANSI Std. X3.139-1987.

[21] ISO/IEC 15802-3 Information technology - Telecommunications and information exchange between systems - Local and metropolitan area networks - Specific requirements - Supplement to Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications - Frame Extensions for Virtual Bridged Local Area Network (VLAN) Tagging on 802.3 Networks, IEEE Std 802.3ac-1998 (Supplement to IEEE 802.3 1998 Edition), 1998.

Security Considerations

Implementation of the model described in this memo creates no known new avenues for malicious attack on the network infrastructure. However, readers are referred to Section 2.8 of the RSVP specification [5] for a discussion of the impact of the use of admission control signaling protocols on network security.

Acknowledgements

Much of the work presented in this document has benefited greatly from discussion held at the meetings of the Integrated Services over Specific Link Layers (ISSLL) working group. We would like to acknowledge contributions from the many participants via discussion at these meetings and on the mailing list. We would especially like to thank Eric Crawley, Don Hoffman and Raj Yavatkar for contributions via previous Internet drafts, and Peter Kim for contributing the text about Demand Priority networks.

Authors' Addresses

Anoop Ghanwani
Nortel Networks
600 Technology Park Dr
Billerica, MA 01821, USA

Phone: +1-978-288-4514
EMail: aghanwan@nortelnetworks.com

Wayne Pace
IBM Corporation
P. O. Box 12195
Research Triangle Park, NC 27709, USA

Phone: +1-919-254-4930
EMail: pacew@us.ibm.com

Vijay Srinivasan
CoSine Communications
1200 Bridge Parkway
Redwood City, CA 94065, USA

Phone: +1-650-628-4892
EMail: vijay@cosinecom.com

Andrew Smith
Extreme Networks
3585 Monroe St
Santa Clara, CA 95051, USA

Phone: +1-408-579-2821
EMail: andrew@extremenetworks.com

Mick Seaman
Telseon
480 S. California Ave
Palo Alto, CA 94306
USA

Email: mick@telseon.com

Full Copyright Statement

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.