Internet Draft
Expires: September 1994
File: draft-braden-rsvp-01.ps

Lixia Zhang
Xerox PARC
Bob Braden
USC-ISI
Deborah Estrin
USC/USC-ISI
Shai Herzog
USC-ISI
Sugih Jamin
USC
March 1994

Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification

DRAFT March 18, 1994

Status of Memo

Abstract

This memo describes version 1 of RSVP, a resource reservation setup protocol designed for an integrated services Internet. RSVP provides receiver-initiated setup of resource reservations for multicast or unicast data flows, with good scaling and robustness properties.

1 Introduction

This memo describes RSVP, a resource reservation setup protocol designed for an integrated services Internet [RSVP93,ISInt93]. An application invokes RSVP to request a specific quality of service (*QoS*) for a data stream. Hosts and routers use RSVP to deliver these requests to the routers along the path(s) of the data stream and to maintain router and host state to provide the requested service. This generally requires reserving resources in those nodes.

At each *node* (i.e., router or host) along the path, RSVP passes a new resource reservation request to an *admission control* routine, to determine whether there are sufficient resources available. If there are, the node reserves the resources and updates its *packet scheduler* and *classifier* control parameters to provide the requested QoS [ISInt93].

The objectives and general justification for RSVP design are presented in [RSVP93,ISInt93]. In summary, RSVP has the following attributes:

- RSVP supports multicast or unicast data delivery and adapts to changing group membership as well as changing routes.
- RSVP is not itself a routing protocol; it is designed to use the existing unicast and multicast routing protocols to determine the data path(s).
- RSVP reserves resources for simplex data streams.
- RSVP is receiver-oriented, i.e., the receiver of a data flow is responsible for the initiation and maintenance of the resource reservation used for that flow.
- RSVP maintains "soft state" in the routers, enabling it to gracefully support dynamic membership changes and automatically adapt to routing changes.
- RSVP provides several "reservation styles" (see below) with different reservation models to fit a variety of applications.
- RSVP provides transparent operation through routers that do not support it.

There are two aspects to RSVP, its reservation model and its protocol mechanisms. The RSVP protocol mechanisms provide a general facility for creating and maintaining distributed reservation state across a mesh of multicast delivery paths. These mechanisms treat the reservation parameters as opaque data, except for certain well-defined operations, and simply pass them to the traffic control modules (admission control, packet scheduler, and classifier) for interpretation. Although the RSVP protocol mechanisms are independent of the encoding of these parameters, the encodings must be defined in the reservation model that is presented to an application (see section 3.6.1).

In order to efficiently accommodate heterogeneous receivers and dynamic group membership, RSVP makes the receivers responsible for requesting resource reservations [RSVP93]. Each receiver can request a reservation that is tailored to its particular requirement, and RSVP will deliver this request to the routers along the reverse path(s) to the sender(s).

Sections 2.1 and 2.2 of this memo summarize the RSVP reservation model, while Sections 2.3 describes the protocol mechanisms. Sections 2.4 gives examples of both model and mechanism, and Section 2.5 summarizes the model of RSVP seen by a host. Section 3 presents the functional specification for RSVP.

2 RSVP Overview

2.1 RSVP Reservation Model

Figure 1 illustrates a single multicast distribution *session*. The arrows indicate data flowing from senders S1 and S2 to receivers R1, R2, and R3, and the cloud represents the distribution mesh created by the multicast routing protocol. Multicast distribution replicates each data packet from a sender Si and forwards a copy to every receiver Rj. Each sender Si and receiver Rj may correspond to a unique Internet host, or there may be multiple senders (e.g., multiple TV cameras) and/or receivers in a single host.

RSVP reserves resources for *simplex* data streams, i.e., it reserves resources in only one direction on a link, so that senders are logically distinct from receivers. However, the same host may act as both sender and receiver.

```
       Senders                                Receivers

                   ---------------------
                  (                     ) ===> R1
          S1 ===> (    Multicast        )
                  (                     ) ===> R2
                  (    distribution     )
          S2 ===> (                     )
                  (    by Internet      ) ===> R3
                  (_____________________)
```

Figure 1: Multicast Distribution Session

All data packets in a session are addressed to the same IP destination address DestAddress. For multicast delivery, DestAddress is the multicast group address to which the data is addressed. For unicast delivery, DestAddress is simply the unicast address of the single receiver. RSVP identifies a session by DestAddress plus a 32-bit stream identifier called the *destination stream id* (DestSID). We use the term *session socket* for the (DestAddress, DestSID) pair that defines a session. RSVP treats each session independently. In the rest of this document, a particular session (hence, session socket) is always implied even if not stated.

Depending upon the reservation style and the session state already in place, a new or modified reservation request may or may not result in a call to admission control at each node [IsInt93]. If an admission control call fails, the reservation is rejected and an RSVP error message is sent to the receiver(s) responsible for it.

A single RSVP resource reservation request is defined by a *flowspec* together with a *filterspec*; this pair is called a *Flow Descriptor*. The flowspec specifies the desired QoS in a quantitative manner, e.g., the tolerable delay, the average throughput, the maximum burstiness, etc [ISInt93,IServ93]. It is used to set the paramters used in the packet scheduling mechanism in the node (router or host). The filterspec defines the set of data packets to receive this service.

RSVP makes no assumption about the format of the filterspec carried in reservation requests. RSVP treats filters as a number of uninterpreted bytes of data that need to be exchanged among only the applications and the traffic control module at each node. A filter may select data packets from particular senders Si, or it may further reduce the data packet subset based on flow demultiplexing fields such as a UDP port, or perhaps on some hierarchical encoding bits within the application layer. For all packets that are addressed to a particular session, only those that can pass the filter(s) of that session will be forwarded according to the flowspec; the rest will be either dropped or sent as best-effort traffic.

2.2 Reservation Styles

In addition to the Flow Descriptors, each RSVP reservation request also specifies a *reservation style*. The following reservation styles have been defined so far.

1. Wildcard-Filter (WF) Style

   A Wildcard-Filter (WF) style reservation creates a single resource "pipe" along each link, and this pipe can be used by data packets from all senders for the given session. The "size" of this pipe is the maximum of the resource requests for that link from all receivers, independent of the number of senders using it.

   The term 'wildcard' refers to a filterspec that selects all senders. A WF reservation automatically extends to new senders to the group, as they appear.

2. Fixed-Filter (FF) Style

   A Fixed-Filter (FF) style reservation request from a receiver selects particular sender(s) and specifies corresponding flowspec(s).

   FF reservation requests from different receivers that select the same sender will share a single reservation for that sender (indeed, due to multicasting there is only one stream of data packets from any Si in a particular router, regardless of the number of receivers Rj downstream). The shared reservation for Si will be the maximum of the individual flowspecs of reservation requests from different receivers Rj. However, FF reservations for different senders are distinct; they do NOT share a common pipe. The total reservation on a link for a given session is the cumulative total of the reservations for each requested sender.

   A Fixed-Filter reservation request from a particular receiver Rj generally contains a list of Flow Descriptors, each consisting of a filterspec (specifying some sender Si) and a corresponding flowspec.

If a receiver making FF style reservations changes its filterspec during the session, this is treated as a new reservation that is subject to admission control and may fail. The receiver may also modify the flowspecs, again subject to admission control.

3. Dynamic-Filter (DF) Style

   Using the Dynamic-Filter (DF) style, each receiver creates N distinct reservations to carry flows from up to N different senders. A DF style reservation also specifies a list of K filterspecs, for some K in the range: $0 <= K <= N$, defining particular senders to use these reservations.

   Regardless of K, N reservations are made, using a common flowspec. Like the FF style, the DF style causes distinct reservations for different senders.

   During the life time of a reservation the receiver may specify the same value of N and the same common flowspec but a different selection of senders (by changing the filterspecs), without a new admission control check. This is known as *channel switching*, in analogy with a television set. Each DF style reservation may be said to be "owned" by the receiver that established it and is permitted to switch channels (change filterspecs) used by that reservation.

   If a receiver using the DF reservation style changes the number N of distinct reservations or the common flowspec, this request is treated as a new reservation that is subject to admission control and may fail.

The essential difference between the FF and DF styles is that DF allows dynamic channel switching without admission control. Once a DF-style reservation has been made, the receiver may switch channels without danger of an admission denial due to limited resources (unless the route changes to a lower-capacity path or new senders appear). In order to provide assured channel switching, each node along the path must reserve enough bandwidth for *all* the possible senders Si, even though some of this bandwidth may be unused at any one time.

WF reservations are appropriate for a number of multicast applications that have application-level constraints which prohibit all data sources from transmitting simultaneously; one example is audio conferences, where a limited number of people talk at once. Thus, each receiver might issue a WF reservation request for twice one audio channel (to allow some over-speaking). On the other hand, the FF and DF styles create independent reservations for the flows from different senders; this is required for video signals, whose 'silence' periods, if any, are uncoordinated among different senders.

2.3 RSVP Protocol Mechanisms

2.3.1 RSVP Messages

RSVP messages are sent as IP datagrams; thus, RSVP occupies the place of a transport protocol in the protocol stack. However, like ICMP and IGMP, RSVP is really an Internet control protocol;

it does not carry any application data, and its messages are processed by the routers in the path.

Each receiver host sends RSVP reservation (*Resv*) messages into the Internet, carrying Flow Descriptors requesting the desired reservation; see Figure 2. These reservation messages must follow the reverse of the routes the data packets will use, all the way upstream to all the senders. If a reservation request fails at any node, an RSVP error message is returned to the receiver; however, RSVP sends no positive acknowledgment messages to indicate success. *Resv* messages are finally delivered to the sender hosts, so that the senders can set up appropriate traffic control parameters for the first hop.

In order to route the *Resv* messages in the reverse direction from each receiver to all selected senders for a given session socket, each sender must transmit RSVP *Path* messages forward along the uni-/multicast routes provided by the routing protocol(s). These *Path* messages store *path state* in all the intermediate routers, effectively combining all the routing trees given by the routing protocol for the same DestAddress.

```
Sender                                        Receiver

             _____________________
   Path -->  (                     )
 Si =======> (    Multicast        ) Path -->
   <-- Resv  (                     ) =========> Rj
             (    distribution     ) <-- Resv
             (_____________________)
```

Figure 2: RSVP Messages

The minimum information content of a *Path* message is a list of sender IP addresses, since this is required for routing *Resv* messages. However, *Path* messages may carry the following additional information:

- A template describing the format of data packets that the sender will originate.

  This template takes the form of two bitstrings forming a (value, mask) pair. Zero mask bits represent "don't care" (variable) bits in data packets. If present, this template is used by RSVP to validate the filters in a *Resv* message. Without such a template in the path state, there will be no feedback (except poor service) to the receiver that sets an impossible filter by mistake.

- A flowspec defining an upper bound on the traffic that will be generated.

  This flowspec can be used by RSVP to prevent over-reservation on the non-shared links starting at the sender.

A (template, flowspec) pair in a *Path* message is called a *Sender Descriptor*.

2.3.2 Soft State

To maintain reservation state, RSVP keeps “soft state” in router and host nodes. RSVP soft state is created and periodically refreshed by *Path* and *Resv* messages, and it can be removed at each node by explicit *Teardown* messages. RSVP also has a timer-driven cleanup procedure if no message is received within a *cleanup timeout* interval.

When the route changes, the next *Path* message will initialize the path state on the new route and future *Resv* messages will establish reservation state accordingly, while the state on the now-unused segment of the route will time out. Thus, whether a message is “new” or a “refresh” is determined separately at each node, depending upon the existence of state at that node. (This document will use the term “refresh message” in this effective sense, to indicate an RSVP message that does not modify the existing state at the node in question.)

RSVP sends all its messages as IP datagrams without any reliable enhancement. Periodic transmission of refresh messages by hosts and routers should replace any lost RSVP messages.

If the set of senders Si or receivers Rj changes, or if any of the receivers’ reservation requests change, the RSVP state is adjusted accordingly. To modify a reservation, a receiver simply starts sending the new values; it is not necessary, although we do recommend, to “close” the old reservation first. RSVP believes the latest *Path* and *Resv* messages (ignoring the possibility of reordering).

When a *Resv* message is received at a router or sender host, the RSVP module checks whether the message is a new or modified reservation request, or whether it simply refreshes an existing reservation. A new or modified request is passed to the admission control module for a decision. If the reservation is accepted, RSVP sets up (or modifies) the reservation and filter state. It also forwards the *Resv* message to the next reverse-hop router(s) or sender host(s), as determined by the path state. If RSVP on the node rejects the reservation request due to admission control failure or to some processing error, it discards the Resv message and returns a RSVP error message to the originating receiver host. If the request modifies a previous reservation, RSVP may immediately remove the old state, or it may simply let the old state time out since it is no longer being refreshed.

BEGIN CHANGES
Figure 3 illustrates RSVP’s model of a router node. Each data stream arrives from a *previous hop* through a corresponding *incoming interface* and departs through one or more *outgoing interface(s)*. Since the same host may be hosting both sender and receiver applications for a given session, the same physical interface may act in both the incoming and outgoing roles (for different data streams).

The interfaces shown in Figure 3 may be physical interfaces (e.g., to point-to-point links), or they may be logical interfaces that reach multiple nodes through the same physical interface. Multiple previous hops and/or next hops through a given physical interface can result from either the connected network being a shared medium (e.g., an Ethernet), or because there are non-RSVP routers in the path to the next RSVP hop (see Section 3.5). It is generally necessary for RSVP to

```
 Previous     Incoming               Outgoing         Next
 Hops         Interfaces             Interfaces       Hops

 -----             ---------------------             -----
|     | data -->  |                     |  data -->  |     |
|     |-----------|                     |------------|     |
|_____|  <-- Resv |                     |   <-- Resv |_____|
        Path -->  |                     |  Path -->
                  |       Router        |
 -----            |                     |            -----
|     | data -->  |                     |   data --> |     |
|     |-----------|                     |------------|     |
|_____|  <-- Resv |                     |   <-- Resv |_____|
        Path -->  |_____________________|   Path -->
```

Figure 3: Router Using RSVP

track both logical and physical interfaces on both the incoming and outgoing sides.

END CHANGES

2.3.3 Merging RSVP Messages

To control protocol overhead, RSVP supports the *merging* of multiple *Path* and *Resv* messages for the same session socket. Those messages that cause a state change are forwarded without delay, while the rest may be merged into fewer messages, perhaps only one. This is accomplished by each node saving the state of received messages, deleting the messages, and then periodically generating new *Path* and *Resv* messages from the saved state in their place, which carry cumulative information from all the "merged" messages. Thus, inside the network refresh messages are created hop-by-hop, at a rate determined by a *refresh period*.

Messages that modify the state in a node ("new" messages) are forwarded without delay. Thus, the refresh period does not affect the rate at which new state propagates from end to end.

For *Path* messages, merging implies collecting together the Sender Descriptors from multiple incoming messages into a single outgoing *Path* message. For *Resv* messages, merging implies that only the essential (e.g., the largest) reservation requests need be forwarded, once per refresh period; redundant messages are *purged*. A successful reservation request will propagate as far as the closest point(s) along the sink tree to the sender(s) where a reservation level equal or greater than that being requested has been made. At that point, the merging process will drop it in favor of another equal, or a larger, reservation.

As a result of merging, the number of RSVP messages increases slower than linearly with the number of senders and receivers in a session. There is no merging across sessions, however, therefore the number of RSVP messages will increase linearly with the number of sessions.

Although flowspecs are opaque to RSVP, merging requires that RSVP be able to determine whether two flowspecs, fls1 and fls2, are identical or whether one is "larger" than the other, i.e. whether one represents at least as strict a request (and hence represents at least as large a resource commitment) as the other.

However a flowspec may be a complex multi-dimensional vector, thus the relationship Fls1 ">" Fls2 may not be well defined in general. New research effort is needed to resolve this issue. For the time being, to allow successful merging of multiple, different reservations for the same sender, RSVP requires that a comparison operation be clearly and uniquely defined throughout the network.

BEGIN CHANGES
Given two incompatible flowspecs Fls1 and Fls2 (e.g. Fls1 asks for a higher throughput and a longer tolerable delay, Fls2 asks for a lower thoughput but shorter delay), here we list a few possible solutions to this unsettled issue: (1) do not merge, instead make independent reservations for each; or (2) find a third flowspec Fls3 that dominates both Fls1 and Fls2 (e.g. combine two flowspecs by picking the higher throughput and the lower delay); or (3) compare the throughput requirement only.

This mechanism, reserving for the highest demand, allows an application to increase an existing reservation immediately. Decreasing a reservation has to be handled more cautiosly, however; a *Resv* message that is received with an apparently decreased reservation might be temporary, caused by the loss of a contributing *Resv* message. Therefore, an RSVP should not "believe" a reservation decrease until the cleanup timeout has passed.

END CHANGES

The refresh period and the cleanup timeout must obey the following general principles:

A. The cleanup timeout interval should be long enough to avoid reservation-flapping due to route-flapping.

If route flapping does occur, persistent state will allow duplicate reservations to be created along the alternate paths. This duplication is desirable to prevent interruptions of service quality due to route flapping. (a *route-pinning* scheme that prevents route-flapping for RSVP data streams is under study.)

B. The refresh period must be long enough to keep RSVP overhead at an acceptable level.

C. The refresh period should be short enough to allow quick adaptation to route and multicast membership changes.

Applications may differ in their sensitivity to service outages, and therefore they should be able to adjust the refresh period for their session state.

2.3.4 Teardown

BEGIN CHANGES

As an optimization to release resources quickly, RSVP *teardown* messages removes path and reservation state without waiting for the cleanup timeout period. RSVP messages are not delivered reliably, but if a teardown message is lost, state will eventially time out.

Teardown may be initiated either by an end system (sender or receiver), or by a router as the result of state timeout. A router may also initiate a teardown message as the result of router or link failures detected by the routing protocol. A teardown, once initiated, will be forwarded hop-by-hop without delay.

There are two types of RSVP Teardown message, *PathTear* and *ResvTear*. A *PathTear* message travels towards all receivers downstream from its point of initiation and tears down path state along the way, while an *ResvTear* message tears down reservation state and travels towards all senders upstream from its point of initiation.

A particular reservation on a node may be shared among multiple senders and/or receivers, but it must apply to a unique next hop (and outgoing interface). The receipt of an *ResvTear* message implies that the corresponding reservation state has been removed downstream, so that the reservation can safely be deleted locally. Again, the local node will only forward the teardown message upstream when the state named in the message has been entirely removed locally. As a result, an *ResvTear* message will prune the reservation state back (only) as far as possible. Note that the *ResvTear* message will cease to be forwarded at the same node where merging suppresses forwarding of the corresponding *Resv* messages.

Consider the router configuration shown in Figure 4 below. Assume that there are reservations for source S1 on both outgoing interfaces (c) and (d), and that the receiver R1 wants to tear down its reservation state for S1. R1's *ResvTear* message arriving through interface (c) indicates that all reservation state for (this session and) sender S1 has been removed downstream. The current node therefore removes the S1 reservation state from interface (c). However, since there will still be an S1 reservation on interface (d), the *ResvTear* message will not be forwarded any further.

However, if the outgoing interface connects to a shared medium or if there is a non-RSVP router immediately downstream, then there may be multiple next-hop RSVP nodes downstream that are reached through the same outgoing interface, say (c). Then a single reservation may be shared among multiple next hops. RSVP must tag each reservation with the next hop(s) from which the

Resv messages came, for use by teardown to avoid deleting shared state.

Deletion of path state, whether as the result of a teardown message or because of timeout, may force adjustments in related reservation state to maintain consistency in the local node. Consider the path state for a sender S; the related reservation state would be as follows.

- Wildcard-Filter style: If S is the only sender to the session, delete the reservation.
- Fixed-Filter style: Delete reservations made for S.
- Dynamic-Filter style: Reduce total reservation if it now exceeds the total number of remaining senders.

END CHANGES

2.4 Examples

We use the following notation for a *Resv* message from a receiver Rj:

1. Wildcard-Filter

   WF(Rj; * {r})

   Here "* {r}" represents a Flow Descriptor with a "wildcard" filter (choosing all senders) and a flowspec of quantity r. For simplicity we assume here that flowspecs are one-dimensional, defining for example the average throughput, and state them as a multiple of some unspecified base resource quantity B.

2. Fixed-Filter

   FF(Rj; S1 {r1}, S2 {r2}, ...)

   This message carries a list of sender, flowspec pairs, i.e., Flow Descriptors.

3. Dynamic-Filter

   DF(nB, Rj;) or DF(nB, Rj; S1, ...)

   This message carries the number n of channels to be reserved, each channel having bandwidth B. It also has a list, perhaps empty, of senders. Since each channel must have the same flowspec, we omit flowspecs after the semicolon.

Figure 4 shows schematically a router with two previous hops (labeled a and b) and two outgoing interfaces (labeled c and d), which will be assumed in all the examples that follow. There are three upstream senders; packets from sender S1 (S2 and S3) arrive through previous hop a (b, respectively). There are also three downstream receivers; packets bound for R1 and R2 (R3) are routed via outgoing interface c (d, respectively).

```
                  ----------------
                a |                | c
( S1 ) ---------->|                |----------> ( R1, R2)
                  |     Router     |
                b |                | d
( S2,S3 ) ------->|                |----------> ( R3 )
                  |________________|
```

Figure 4: Router Configuration

```
                         |
         Send            |       Reserve              Receive
                         |
                         |        --------
  WF(R1; * {3B}) <- (a)  |  (c) | * {3B} |    (c) <- WF(R1; * {3B})
                         |      |________|
 [ WF(R2; * {B}) <- (a)] |                  [ (c) <- WF(R2; * {B}) ]
                         |
 [ WF(R3; * {B}) <- (a)] |
-------------------------|------------------------------------------
                         |        -------
  WF(R1; * {3B}) <- (b)  |  (d) | * {B} |     (d) <- WF(R3; * {B})
                         |      |_______|
 [ WF(R2; * {B}) <- (b)] |
                         |
 [ WF(R3; * {B}) <- (b)] |
```

Figure 5: Wildcard-Filter Reservation Example

Assuming this configuration, Figure 5 illustrates Wildcard-Filter reservations. The "Receive" column shows the *Resv* messages received over outgoing interfaces c and d, and the "Reserve" column shows the resulting reservation state for each interface. The "Send" column shows the *Resv* messages forwarded to previous hops a and b. In the "Reserve" column, each box represents one reservation "channel", with the corresponding filter.

As a result of merging, only the message with the largest flowspec is forwarded upstream to each previous hop. Merging will therefore purge the messages enclosed in square brackets. In the outgoing interface c, there is a common reserved channel shared by R1 and R2, while interface d has a reservation by R3.

Figure 6 shows Fixed-Filter style reservations. Merging takes place among the flow descriptors (i.e., filter spec, flowspec pairs); the unmerged messages are not shown. For example, the message forwarded to previous hop b, towards S2 and S3, contains flow descriptors received from outgoing interfaces c and d. Similarly, when FF(R2; S1 {B}) and FF(R3; S1 {3B}) are merged, the result is the single message FF(R3; S1{3B}) sent to previous hop a, towards S1.

For each outgoing interface, there is a private reservation for each source that has been requested, but this private reservation is shared among the receivers that made the request.

```
                        |
         Send           |       Reserve             Receive
                        |
                        |        --------
  FF(R3; S1{3B}) <- (a) |  (c) | S1{B}  |   (c) <- FF(R2; S1{B}, S2{5B})
                        |      |________|
                        |      | S2{5B} |
                        |      |________|
------------------------|---------------------------------------------
                        |        --------
                 <- (b) |  (d) | S1{3B} |   (d) <- FF(R3; S1{3B}, S3{B})
  FF(R3; S2{5B}, S3{B}) |      |________|
                        |      | S3{B}  |
                        |      |________|
```

Figure 6: Fixed-Filter Reservation Example

```
                        |
         Send           |      Reserve            Receive
                        |
                        |       -----------
  DF(2, R1; S1) <- (a)  |  (c) | R1: S1{B} |   (c) <- DF(2, R1; S1, S2)
                        |      |___________|
    DF(2, R2; ) <- (a)  |      | R1: S2{B} |   (c) <- DF(2, R2; S2)
                        |      |[R2: S2{B}]|
  DF(2, R3; S1) <- (a)  |      |___________|
                        |      | R2:  ?{B} |
                        |      |___________|
------------------------|--------------------------------------------
                        |       -----------
  DF(2, R1; S2) <- (b)  |  (d) | R3: S1{B} |    (d) <- DF(2, R3; S1)
                        |      |___________|
  DF(2, R2; S2) <- (b)  |      | R3:  ?{B} |
                        |      |___________|
    DF(2, R3; ) <- (b)  |
```

Figure 7: Dynamic-Filter Reservation Example

Figure 7 shows an example of Dynamic-Filter reservations. Within the reservation boxes, the receiver that owns the reservation is shown followed by a colon. R2 and R3 have made reservations for which there is currently no filter, as indicated by a filter of '?'.

Both R1 and R2 have requested reservations for S2. Since there is only one stream of packets from S2, only R1's reservation is actually made. However, the router keeps a record of R2's request as a "hidden reservation", shown in square brackets. If R1 removes its reservation for S2, the node must check for a hidden reservation for the same source, so that it can reassign the channel to R2 without waiting for R2's next refresh. This use of hidden reservations for Dynamic-Filter reservations is necessary to avoid new admission control decisions (which might fail) and to ensure continuity of service.

A router should not reserve more Dynamic-Filter channels than the number of upstream sources. Thus, in Figure 7, both R1 and R2 request that admission control admit two channels worth of bandwidth. However, there are only 3 sources upstream from this point, so these requests are satisfied with a total of 3 channel reservations.

Since there is only one source upstream from previous hop b, the first parameter of the DF message (the count of channels to be reserved) could be decreased to 1 in the forwarded reservations. However, this is unnecessary because the routers upstream will reserve only one channel, regardless.

We note that Dynamic-Filter reservation requests cannot be merged, since they tie each reservation to a particular owner. A node is permitted to include Flow Descriptors that are not relevant to the path (i.e., not upstream along that path). Thus, in Figure 7, the node could forward each of the three *Resv* messages exactly as received from c and d, to both previous hops a and b; the previous hop would ignore the irrelevant Flow Descriptors in each case.

As another example of these concepts, Figure 8 shows the previous hop router on incoming interface (a) in Figure 7. Here there are two hidden reservations.

```
        Send            |      Reserve          Receive
                        |       -----------
 DF(2, R1; S1) <- (a)   |  (c) | R1: S1{B} |   (c) <- DF(2, R1; S1)
                        |      |           |
   DF(2, R2; ) <- (a)   |      |[R2:  ?{B}]|   (c) <- DF(2, R2; )
                        |      |           |
 DF(2, R3; S1) <- (a)   |      |[R2: S1{B}]|   (c) <- DF(2, R3; S1)
                        |      |___________|
```

Figure 8: Previous Hop to Figure 7 interface (a).

2.5 Host Model

Before a session can be created, the session socket (comprised of DestAddress and DestSID) must be assigned and communicated to all the senders and receivers by some out-of-band mechanism. In order to join an RSVP session, the end systems perform the following actions.

H1 A receiver joins the multicast group specified by DestAddress.

H2 A potential sender starts sending RSVP *Path* messages to the DestAddress.

H3 A receiver listens for *Path* messages.

H4 A receiver starts sending appropriate *Resv* messages, specifying the desired Flow Descriptors.

There are several synchronization issues.

- Suppose that a new sender starts sending data but there are no receivers. There will be no multicast routes beyond the host (or beyond the first RSVP-capable router) along the path; the data will be dropped at the first hop until receivers(s) do appear (assuming a multicast routing protocol that "prunes off" or otherwise avoids unnecessary paths).
- Suppose that a new sender starts sending *Path* messages (H2) and immediately starts sending data, and there are receivers but no *Resv* messages have reached the sender yet (e.g., because

its *Path* messages have not yet propagated to the receiver(s)). Then the initial data may arrive at receivers without the desired QoS.

- If a receiver starts sending *Resv* messages (H4) before any *Path* messages have reached it, RSVP will return error messages.

  The receiver may simply choose to ignore such error messages, or it may avoid them by waiting for *Path* messages before sending *Resv* messages.

The interface (API) between RSVP and an application is not defined in this protocol spec, as it may be host system dependent. However, Section 3.6.2 discusses the general requirements and presents a generic API.

3 Functional Specification

BEGIN CHANGES
There are currently 6 types of RSVP messages: *Path*, *Resv*, *PathTear*, *ResvTear*, *PathErr*, and *ResvErr*.

END CHANGES

3.1 Message Formats

3.1.1 Path Message

```
       0             1              2             3
+-------------+-------------+-------------+-------------+
| Vers | Type |    Flags    |       RSVP Checksum       |
+-------------+-------------+-------------+-------------+
|                      DestAddress                      |
+-------------+-------------+-------------+-------------+
|                        DestSID                        |
+-------------+-------------+-------------+-------------+
|                     Refresh Period                    |
+-------------+-------------+-------------+-------------+
|                     State TTL Time                    |
+-------------+-------------+-------------+-------------+
|                 Previous Hop Address                  |
+-------------+-------------+-------------+-------------+
|       ///////////////     |         SD Count          |
+-------------+-------------+-------------+-------------+
|                Sender Descriptor List                 |
+-------------+-------------+-------------+-- ...
```

IP Fields:

Protocol
: 46

IP Source Address
: The IP address of the host or router sending this message.

IP Destination Address
: The IP address of the data destination (DestAddress).

RSVP Fields:

Vers

Version number. This is version 1.

Type

1 = Path Message

BEGIN CHANGES

Flags

8 = Drop

If this flag bit is on then data packets will be dropped when they are destined to this session but their sender is not currently selected by any filter. If this flag bit is off, such data packets will still be forwarded but without a reservation, i.e., using a best-effort class.

END CHANGES

RSVP Checksum

A standard TCP/UDP checksum, over the contents of the RSVP message with the checksum field replaced by zero.

DestAddress, DestSID

The IP address and stream Id identifying the session, i.e., the session socket.

Previous Hop Address

The IP address of the interface through which the host or router last forwarded this message.

The Previous Hop Address is used to support reverse-path forwarding of *Resv* messages. This field is initialized by a sender to its IP address (see IP Source Address above) and must be updated at each router hop as the *Path* message is forwarded.

Refresh Period

This field specifies the refresh timeout period in milliseconds. See Section 3.3 below.

State TTL Time

This field specifies the time-to-live for soft state, in milliseconds. It determines the cleanup timeout period; see Section 3.3 below.

SD Count

Count of Sender Descriptors that follow.

Sender Descriptor List

A list of Sender Descriptors (see below). The order of entries in this list is irrelevant.

Each sender must periodically send a *Path* message containing a single Sender Descriptor describing its own data stream. These messages are addressed to the uni-/multicast destination address for the session, and they are forwarded to all receivers, following the same paths as a data packet from the same sender. *Path* messages are received and processed locally to create path state at each intermediate router along the path.

If an error is encountered while processing a *Path* message, an RSVP error message is sent to all the sender hosts listed in the Sender Descriptor List.

Path messages are distributed from senders to receivers along the exact paths that the data will traverse, using uni-/multicast routing. This distribution actually takes place hop-by-hop, allowing RSVP in each router along the path to observe and modify the message. Routing of *Path* messages is based on the sender address(es) from the Sender Descriptor(s), not the IP source address. This is necessary to prevent loops; see Section 3.2.

Each Sender Descriptor consists of a sender template that defines the format of data packets and a corresponding Flowspec that describes the traffic characteristics:

```
+-------------+-------------+-------------+----------//-+
|  STempLen   |      Sender Template ...                |
+-------------+-------------+-------------+----------//-+
|  FlowSLen   |      Flow Spec ...                      |
+-------------+-------------+-------------+----------//-+
```

Here StempLen and FlowSLen are one-octet fields specifying the lengths in octets (including the length byte) of the sender template and flowspec, respectively. The sender template is composed of an explicit IP address plus a variable-length mask-and-value pair VF, in the following format:

```
+-----------+-----------+-----------+-----------+
|                Sender IP Address              |
+-----------+-----------+-----------+-----------+
|       VFlength        |      VFoffset         |
+-----------+-----------+-----------+-----------+  ---
|                 V: VF Value Part              | 4*VFlength
/                                               / octets
/                                               /
+-----------+-----------+-----------+-----------+  ---
|                 M: VF Mask Part               | 4*VFlength
/                                               / octets
/                                               /
+-----------+-----------+-----------+-----------+  ---
```

The value M and the mask V each have length 4∗VFlength octets. M and V define a filter using a simple mask-and-match algorithm applied to the packet at 4∗VFoffset octets from the beginning with the internet layer header. The VF part should not include the sender IP address or DestAddress; RSVP will effectively embed these explicit addresses into the template before using it to validate a filterspec.

The flowspec may be omitted, in which case FlowSLen will be zero. See Section 3.6.1 for information on the format of a flowspec.

3.1.2 Resv Message

Resv messages are sent from receivers to senders along reverse paths established by *Path* messages.

```
       0             1              2             3
+-------------+-------------+-------------+-------------+
| Vers | Type |    Flags    |       RSVP Checksum       |
+-------------+-------------+-------------+-------------+
|                       DestAddress                     |
+-------------+-------------+-------------+-------------+
|                         DestSID                       |
+-------------+-------------+-------------+-------------+
|                      Refresh Period                   |
+-------------+-------------+-------------+-------------+
|                      State TTL Time                   |
+-------------+-------------+-------------+-------------+
|                     Next Hop Address                  |
+-------------+-------------+-------------+-------------+
|                       RecvAddress                     |
+-------------+-------------+-------------+-------------+
| Dynamic Reservation Count |         FD Count          |
+-------------+-------------+-------------+-------------+
|                  Flow Descriptor List                 |
+-------------+-------------+-------------+-- ...
```

The fields are the same as defined earlier for a *Path* message, except for the following:

IP Fields:

IP Source Address

The IP address of the node sending this message.

IP Destination Address

The IP address of the next-hop router or host to which this message is being sent.

RSVP Fields:

Type

2 = Resv Message

Flags

The following flag bits indicate the reservation style.

2 = Fixed-Filter

4 = Dynamic-Filter

Next Hop Address

The IP address of the interface through which the last forwarded this message.

The Next Hop Address is used to support teardown. This field is initialized by a receiver to its IP address and must be updated at each router hop as the *Resv* message is forwarded.

RecvAddress

The IP address of (one of the) receiver(s) that originated this message, or one of the *Resv* messages that was merged to form this message.

Dynamic Reservation Count

The number of channels to be reserved, for a Dynamic-Filter style reservation.

If the ResvStyle is Dynamic-Filter, this integer value must be constant and equal or greater than (FD Count). For other ResvStyles, this field must be zero.

FD Count

Count of Filter Specs in the Flow Descriptor List.

Flow Descriptor List

A list of Flow Descriptors, i.e., (Filterspec, flowspec) pairs, to define individual reservations for Fixed-Filter and Dynamic-Filter styles. The first entry in the list may have special meaning (see below); the order of later entries is irrelevant.

Each Flow Descriptor has the following form:

```
+-------------+-------------+-------------+----------//-+
|  FiltSLen   |        Filter Spec ...                  |
+-------------+-------------+-------------+----------//-+
|  FlowSLen   |        Flow Spec ...                    |
+-------------+-------------+-------------+----------//-+
```

Here FiltSLen and FlowSLen are one-octet fields specifying the lengths in octets (including the length byte) of the filterspec and flowspec, respectively. See Section 3.6.1 for information on the format of filterspecs and flowspecs.

A Wildcard-Filter style reservation is encoded as a special case of a Fixed-Filter reservation: a single Fixed-Filter channel in which the Filterspec specifies the wild-card filter, i.e., it selects packets from all senders Si to the given session socket.

The following specific rules hold for different reservation styles.

- Wildcard-Filter

  To obtain Wildcard-Filter service, set FD Count = 1 and include a single Flow Descriptor whose Filterspec part is a *wild card*, i.e., selects all senders. and whose flowspec part defines the desired flow parameters.

- Fixed-Filter

  Include a list of FD Count >= 1 Flow Descriptors, each defining a sender Filterspec and a corresponding flowspec.

- Dynamic-Filter

  Include max(1, FD Count) Flow Descriptors in the message. Here the FD Count specifies the number of sender Filterspecs that are included. If DC is the Dynamic Reservation Count, then DC >= FD Count >= 0.

  The Flowspec part of the first Flow Descriptor defines the desired size of all the DC channels that are reserved. The Flowspec parts of later Flow Descriptors (if any) are ignored.

3.1.3 Error Messages

BEGIN CHANGES

There are two types of RSVP error messages: *PathErr* messages result from *Path* messages and travel towards senders, while *ResvErr* messages result from *Resv* messages and travel towards receivers. RSVP error messages are triggered only by processing of *Path* and *Resv* messages; errors encountered while processing error or teardown messages must not create error messages.

A *PathErr* message has the following form:

```
       0             1             2             3
+-------------+-------------+-------------+-------------+
| Vers | Type |    Flags    |       RSVP Checksum       |
+-------------+-------------+-------------+-------------+
|                      DestAddress                      |
+-------------+-------------+-------------+-------------+
|                        DestSID                        |
+-------------+-------------+-------------+-------------+
|  Error Code | Error Index |        Error Value        |
+-------------+-------------+-------------+-------------+
|       ////////////// (ignored) /////////////////      |
+-------------+-------------+-------------+-------------+
|       ////////////// (ignored) /////////////////      |
+-------------+-------------+-------------+-------------+
|     /// Reserved ///      |          SD Count         |
+-------------+-------------+-------------+-------------+
|                Sender Descriptor List                 |
+-------------+-------------+-------------+-- ...
|               Sender|Flow Descriptor List             |
+-------------+-------------+-------------+-- ...
```

The fields are the same as in a *Path* message, defined earlier, except for the following:

RSVP Fields:

RSVPType
  3 = *PathErr* message

Error Code
  A one-octet error description.
    01 = Insufficient memory
    02 = Count Wrong
      The SD Count field does not match length of message.

Error Index

Position of Sender Descriptor that caused the error within Sender Descriptor List. An integer between zero and SD Count - 1.

Error Value

(Unused)

A *ResvErr* message has the following form:

```
       0             1              2             3
+-------------+-------------+-------------+-------------+
| Vers | Type |    Flags    |        RSVP Checksum      |
+-------------+-------------+-------------+-------------+
|                       DestAddress                     |
+-------------+-------------+-------------+-------------+
|                         DestSID                       |
+-------------+-------------+-------------+-------------+
|  Error Code | Error Index |        Error Value        |
+-------------+-------------+-------------+-------------+
|      ////////////// (ignored) /////////////////     |
+-------------+-------------+-------------+-------------+
|      ////////////// (ignored) /////////////////     |
+-------------+-------------+-------------+-------------+
|                      RecvAddress                      |
+-------------+-------------+-------------+-------------+
| Dynamic Reservation Count |          FD Count         |
+-------------+-------------+-------------+-------------+
|                 Flow Descriptor List                  |
+-------------+-------------+-------------+-- ...
```

The fields are the same as in a *Path* message, defined earlier, except for the following:

RSVP Fields:

RSVPType

4 = *ResvErr* message

Error Code

A one-octet error description.

DEFINE THESE VALUES IN AN APPENDIX??

01 = Insufficient memory

02 = Count Wrong

The FD Count field does not match length of message.

03 = No path information for this Resv

04 = No Sender information for this Resv
There is path information, but it does not include the sender specified in any of the Filterspecs listed in the Resv messager.
05 = Incorrect Dynamic Reservation Count
Dynamic Reservation Count is zero or less than FD Count.
06 = Filterspec error
07 = Flowspec syntax error
08 = Flowspec value error
Internal inconsistency of values.
[What should be done with Flowspec Feature Not Supported?]
09 = Resources unavailable
[Sub-reasons? Depend upon traffic control and admission control algorithms?]

Error Index

Position of Flow Descriptor that caused the error within Flow Descriptor List. An integer between zero and FD Count - 1.

Error Value

Specific cause of the error described by the Error Code.
DEFINE THESE VALUES IN AN APPENDIX??

An error message may be duplicated and forwarded unchanged. Since *Path* and *Resv* messages may be merged, an error condition must be disseminated to all RSVP client applications whose requests contributed to the error situation. To accomplish this, RSVP must propagate and perhaps duplicate its error messages hop-by-hop. To accomplish this routing of an error message, it must include all the information used to route the original message that caused the error: the Sender Descriptor List, Flags, RecvAddress, and Flow Descriptor List fields, as appropriate. In particular, a *ResvErr* message carries the same style flags as the *Resv* message that caused the error.

To ease implementation, the error message formats are chosen to match the formats of the messages whose processing caused the error. In particular, a *Path* or *Resv* message that encounters an error can be simply converted to the corresponding error message by overwriting the Type and the Refresh Period fields.

A *PathErr* message is forwarded to all previous hops for all senders listed in the Sender Descriptor List. The routing of a *ResvErr* message is more complex.

- The router that creates a *ResvErr* message as the result of processing a *Resv* message should send the *ResvErr* message out the interface through which the *Resv* arrived.
- In succeeding hops, the routing of a *ResvErr* message depends upon its style.

  A DF-style *ResvErr* message is forwarded towards the RecvAddress.

A FF-style *ResvErr* message is forwarded on all outgoing interfaces for which there is already a reservation and for which a sender is listed in the Flow Descriptor List. That is, a *ResvErr* message is sent on a pruned version of the multicast distribution tree for the session; those branches that do not have reservations for any of the specified senders are pruned off.

An RSVP daemon delivers a copy of every relevant error message to its local application clients. It examines the set of RSVP requests that local clients have made through the API, and notifies every client that contributed to the error message. A match is required between the session, filters (senders), and reservation styles of the error message and the corresponding state in the latest API requests. A particular notification should include only the information (e.g., filters) relevant to that application.

3.1.4 Teardown Messages

There are two types of RSVP Teardown message, *PathTear* and *ResvTear*. A *PathTear* message tears down path state and travels towards all receivers downstream from its point of initiation. A *ResvTear* message tears down reservation state and travels towards all senders upstream from its point of initiation.

A *PathTear* message has the same format as a *Path* message, except that in a *PathTear* message:

- Type field = 5
- Refresh Period and State TTL Time fields are ignored.

A *ResvTear* message has the same format as a *Resv* message, except that in a *ResvTear* message:

- Type field = 6
- Refresh Period and State TTL Time fields are ignored.

Any Flowspec components of Flow Descriptors in a *ResvTear* or *PathTear* message are ignored.

Teardown messages are processed in the following way.

- *PathTear*

  Processing a *PathTear* message is straightforward. For each sender S in the message, the node removes path state for S and also deletes all related reservations. Finally, the node forwards the original *PathTear* message to all outgoing interfaces through which data packets from some S in the packet would be routed. That is, *PathTear* forwarding rules are the same as those for *Path* messages.

- *ResvTear*

  Processing an *ResvTear* message is more complex. Suppose a *ResvTear* message arrives through outgoing interface OI from next hop NH. For each sender S listed in the *ResvTear* message, the node checks the reservation, if any, for S on OI. If there is a reservation and if this reservation is shared among more than one next hop, then the only action is to remove NH from the list of next hops sharing this reservation. If there is only a single next hop, then the reservation is deleted. Finally, the node forwards the original *ResvTear* message to all incoming interfaces for senders listed in the message. That is, *ResvTear* forwarding rules are the same as those for *Resv* messages.

END CHANGES

3.2 Avoiding Message Loops

RSVP routes its control messages, and every routing procedure must avoid looping packets. The merging of RSVP messages delays forwarding at each node for up to one refresh period. This may avoid high-speed loop, but there can still be "slow" loops, clocked by the refresh period; the effect of such slow loops is to keep state active forever, even if the end nodes have ceased refreshing it. RSVP uses the following rules to prevent looping messages.

1. When an RSVP message is received through a particular incoming Interface F, the message must not be forwarded out F as an outgoing interface. This implies that RSVP must keep track of the interface through which each message is received, to avoid forwarding it out that interface. Note that, although RSVP distinguishes incoming from outgoing interfaces, in many cases the same physical interface will play both roles.
2. When a *Path* message is received, a route must be computed for each of its sender Flow Descriptors. These routes, obtained from the uni/multicast routing table, generally depend upon the (sender host address, destination address) pairs. Each route consists of a list of outgoing interfaces; these lists (with the incoming interfaces deleted by rule (1)) are used to create merged *Path* messages to be forwarded through the outgoing interfaces.

   Assuming that multicast routing is free of loops, *Path* messages cannot loop even in a topology with cycles.

   Since *Path* messages don't loop, they create path state defining a loop-free path to each sender. As a result, *Resv* messages directed to particular senders cannot loop. However, this cannot protect against looping *Resv* messages that are directed towards all senders ("wildcard" sender). The following three rules are needed for this purpose.
3. A *Resv* message whose RecvAddress matches one of the IP addresses of the local node must be discarded without processing.

```
                         ----------------
                      a |                | c
( R1, S1 ) <----->|     Router     |<-----> ( R2, S2 )
                        |________________|

    Send & Receive on (a)     |      Send & Receive on (c)
                              |
  WF(R2; *{3B}) <-- (a)       |   (c) <-- WF(R2; *{3B})
                              |
  WF(R1; *{4B}) --> (a)       |   (c) --> WF(R1; *{4B})
                              |
                              |
     Reserve on (a)           |         Reserve on (c)
        __________            |        __________
       |  * {4B}  |           |       |  * {3B} |
       |__________|           |       |__________|
                              |
```

Figure 9: Example: Rule (1) for Preventing Loops.

4. Each *Resv* message carries a receiver address. When the choice of address to place in a merged *Resv* message is otherwise arbitrary, RSVP must use the IP address that is numerically largest.
5. When a *Resv* message is received, the Reverse Path Forwarding rule is applied to the receiver address in the message: the message is discarded unless it arrived on the interface that is the preferred route to the receiver.

Figure 9 illustrates the effect of the rule (1) applied to *Resv* messages. It shows a transit router, with one sender and one receiver on each side; interfaces a and c therefore are both outgoing interfaces and physical previous hops. Both receivers are making a Wildcard-Filter style reservation, in which the *Resv* message is to be forwarded to all previous hops for senders in the group, with the exception of the interface through which it arrived.

BEGIN CHANGES
The loop-supression rules for *Resv* messages also prevent looping of *ResvTear* messages. Note that *ResvTear* messages are otherwise subject to fast loops, since they are not delayed by a refresh timeout period.

PathErr messages are routed upstream by the same rules used for FF and DF *Resv* messages (there is no equivalent of wildcard-filter for routing a *PathErr* message). Similarly, *ResvErr* messages are routed by the rules for *Path* messages. For reasons explained above, no special loop-suppression rules are required in either case.

END CHANGES

3.3 Soft State Management

The RSVP state associated with a session in a particular node is divided into atomic elements that are created, refreshed, and timed out independently. The atomicity is determined by the requirement that any sender or receiver may enter or leave the session at any time, and its state should be created and timed out independently.

Management of RSVP state is complex because there is not generally a one-to-one correspondence between state carried in RSVP control messages and the resulting state in nodes. Due to merging, a single message contain state referring to multiple stored elements. Conversely, due to reservation sharing, a single stored state element may depend upon (typically, the maximum of) state values received in multiple control messages.

For each element, there are two time parameters controlling the maintenance of soft state: the refresh period R and the TTL (time-to-live) value T. R specifies the period between successive refresh messages over the same link. T controls how long state will be retained after refreshes stop appearing.

Path and *Resv* messages specify both R and T. When messages are merged and forwarded to the next hop, R should be the minimum R that has been received, and T should be the maximum T that has been received. Thus, the largest T determines how long state is retained, and the smallest R determines the responsiveness of RSVP to route changes. In the first hop, they are expected to be equal. The RSVP API should set a configurable default value, which can be overridden by an application for a particular session.

To avoid gaps in user service due to lost RSVP messages, RSVP should be forgiving about missing refresh messages. A node should not discard an RSVP state element until K * Tmax has elapsed without a refresh message, where Tmax is the maximum of the T values it has received. K is some small integer; K-1 successive messages may be lost before state is deleted. Currently K = 3 is suggested.

Let X indicate a particular message type (either "Path" or "Resv") and a particular session. Then each X message from node a to node b carries refresh period Rab and TTL time Tab.

- As X messages arrive at node b, the node computes and saves both the min over the Rab values (min(Rab)) and the max over the Tab values (max(Tab)) from these messages.
- The node uses K * max(Tab) as its cleanup timeout interval.
- The node uses min(Rab's) as the refresh period.
- Each refresh message forwarded by node b to node c has Tbc = max(Tab) and Rbc = min(Rab)
- A node may impose an upper bound Tmax and a lower bound Rmin, set by configuration

information, and enforce: $Rmin <= R <= T <= Tmax$.

The receiver should be conservative about reacting to certain error messages. For example, during a route change a receiver may get back "No Path" error messages until Path messages have propagated along the new route.

BEGIN CHANGES
Teardown messages, like other RSVP messages, are sent as datagrams and may be lost (although a QoS is used that should minimize the chances of congestive loss of RSVP messages). To increase the reliability of teardown, Q copies of any given teardown message can be sent. Note that if the iteration count Q on initiating teardown messages is > 1, then the state cannot actually be deleted until Q teardowns have been sent. The state would be placed in a "moribund" status meanwhile.

The appropriate value of Q is an engineering issue. $Q = 1$ would be the simplest and may be adequate, since unrefreshed state will time out anyway; teardown is an optimization. Note that if one or more teardown hops are lost, the router that failed to receive a teardown message will time out its state and initiate a new teardown message beyond the loss point. Assuming that RSVP message loss probability is small (but non-zero), the longest time to delete state will seldom exceed one state timeout time K*Tab.

Here is an example. Here G1, G2, G3, and G4 are routers between a sender S and a receiver R. S initiates a *PathTear* message (denoted by "PT"), but this message is lost between routers G2 and G3. Since G2 has deleted its state for S, G2 will cease refreshing G3 (though G3 is still refreshing G4, etc.)

```
    PT       PT       PT
S ---> G1 ---> G2 -->x   G3         G4         R
```

After a time K*Tab, G3's state will time out, and G3 will initiate a teardown for S path state:

```
                                 PT        PT
                             G3 ----> G4 ---->  R
```

If some hop of this chain is lost, there will again be state timeout to continue the teardown. This process should terminate in a few timeout periods.

END CHANGES

3.4 Sending RSVP Messages

Under overload conditions, lost RSVP control messages could cause the loss of resource reservations. It recommended that routers be configured to give a preferred class of service to RSVP packets. RSVP should not use significant bandwidth, but the delay of RSVP packets needs to be controlled.

An RSVP *Path* or *Resv* message consists of a small root segment (24 or 28 bytes) followed by a list of descriptors. The descriptors are bulky and there could be a large number of them, resulting in potentially very large messages. IP fragmentation is inadvisable, since it has bad error characteristics. Instead, RSVP-level fragmentation should be used. That is, a message with a long list of descriptors will be divided into segments that will fit into individual datagrams, each carrying the same root fields. Each of these messages will be processed at the receiving node, with a cumulative effect on the local state. No explicit reassembly is needed.

3.5 Automatic Tunneling

It is impractical to deploy RSVP (or any protocol) at the same moment throughout the Internet, and RSVP may never be deployed everywhere. RSVP must therefore provide correct protocol operation even when two RSVP-capable routers are joined by an arbitrary "cloud" of non-RSVP routers.

RSVP will automatically tunnel through such a non-RSVP cloud. Both RSVP and non-RSVP routers forward *Path* messages towards the destination address using their local uni-/multicast routing table. Therefore, the routing of Path messages will be unaffected by non-RSVP routers in the path. When a *Path* message traverses a non-RSVP cloud, the copies that emerge will carry as a Previous Hop address the IP address of the last RSVP-capable router before entering the cloud. This will cause effectively construct a tunnel through the cloud for *Resv* messages, which will be forwarded directly to the next RSVP-capable router on the path(s) back towards the source.

This automatic tunneling capability of RSVP has a cost: a *Path* message must carry the session DestAddress as its IP destination address; it cannot be addressed hop-by-hop. As a result, each RSVP router must have a small change in its multicast forwarding path to recognize RSVP messages (by the IP protocol number) and intercept them for local processing. See Section 3.6.5 below.

(There is a potential defect in tunneling. Merged *Path* messages can carry information for a list of senders, and since multicast routing depends in general upon the sender, it is not possible to ensure that all the non-RSVP routers along the tunnel will be able to route the packet properly. The effect turns out to be that tunnels may distribute path information to RSVP routers where it should not go, which may in turn lead to unused reservations at these routers. This is hoped to be an acceptable defect.)

Of course, if an intermediate cloud does not support RSVP, it is unable to perform resource reservation. In this case, firm end-to-end service guarantees cannot be made. However, if there is sufficient excess capacity through such a cloud, acceptable and useful realtime service will still be possible.

3.6 Interfaces

3.6.1 Reservation Parameters

The Flowspec format is currently specific to the CSZ packet scheduler [CSZ92]. The parameters are:

- QoS Type (Guaranteed, Predictive, ...)
- Max end-to-end delay
- Average data rate (bits/ms)
- Token bucket depth (bits)
- Global share id

Although a filterspec may be in part opaque, RSVP must be able to extract the sender IP address (or wildcard) from it, and to compare it with a sender template field in the path state. For compactness and simplicity of processing, this version of the RSVP specification defines an RSVP Filterspec to be composed of an explicit IP address plus a variable-length mask-and-value pair VF, in the following format:

```
+-----------+-----------+-----------+-----------+
|                 Sender IP Address             |
+-----------+-----------+-----------+-----------+
|      VFlength         |      VFoffset         |
+-----------+-----------+-----------+-----------+  ---
|                 V: VF Value Part              | 4*VFlength
/                                               / octets
/                                               /
+-----------+-----------+-----------+-----------+  ---
|                 M: VF Mask Part               | 4*VFlength
/                                               / octets
/                                               /
+-----------+-----------+-----------+-----------+  ---
```

The value M and the mask V each have length 4∗VFlength octets. M and V define a filter using a simple mask-and-match algorithm applied to the packet at 4∗VFoffset octets from the beginning with the transport-layer header. VFlength can be zero, in which case VFoffset is ignored. A wildcard Filterspec, which will match any sender, has zero for Sender IP Address and VFlength.

The contents of M and V are opaque to RSVP. The VF part may or may not include the sender IP address or DestAddress; RSVP will embed these explicit addresses into the filterspec before handing it to traffic control. In general, RSVP cannot interpret the contents of the V and M fields, since they may depend upon protocols above the Internet layer.

There are many possible filters that cannot be expressed using a simple mask and value pair. A compact and general filter encoding is for further study.

3.6.2 Application/RSVP Interface

This section describes a generic API from an application to an RSVP control process. The details of a real interface may be operating-system dependent; the following can only suggest the basic functions to be performed. In particular, some of these calls cause information to be returned asynchronously.

An application could directly send and receive RSVP messages, just as an application can do file transfer using UDP. However, we envision that many applications will not want to know the details of RSVP operation, nor to provide the timing services necessary to keep the state refreshed, any more than an application wants to handle TCP retransmission timeouts. Therefore, a host using RSVP may be expected to have an RSVP control process to handle these functions. Using local IPC, applications will register or modify resource requests with this process and receive notifications of success or change of conditions.

1. Sender

   Call: SENDER(local socket, session socket, flowspec) $->$ sid

   This call is made by a sender host for each sender socket, to initiate RSVP processing. Wildcards may be inserted for the IP address and/or RSVP Id in the local socket, in which case the local system will choose appropriate values. The session socket consists of the uni-/multicast address of the receiving host(s) and the RSVP Id to use for this session.

   The SENDER call returns immediately with a local session identifier "sid", which may be used in subsequent calls. A local session control block is created and initialized, and the host begins sending periodic *Path* messages.

2. Receiver

   Call: RECEIVER(local socket) $->$ sid

   This call is made by a receiver host for each receiver socket, to initiate RSVP processing. The local socket consists of the uni-/multicast address for the desired session, and an RSVP Id.

   The RECEIVER call returns immediately with a local session identifier "sid", which may be used in subsequent calls. A local session control block is created and initialized, and the host begins listening for a *Path* message.

Following this call, data may be returned asynchronously, containing the flowspecs and associated foreign sockets for sender hosts. Error message(s) may also be returned asynchronously.

3. Reserve

   Call: RESERVE(sid, style, Flowspec-list, Filterspec-list)

   A receiver uses this call to make a resource reservation. Style is an integer index indicating the reservation style.

   The first RESERVE call following a RECEIVER call will initiate the periodic transmission of *Resv* messages. A later RESV call may be given to modify the parameters of the earlier call (but note that changing the reservations may result in admission control failure).

   The RESERVE call returns immediately. Following this call, an error message or a data reply from the earlier RECEIVER call may be returned asynchronously.

4. Close

   Call: CLOSE(sid)

   This call may be made by either sender or receiver to terminate RSVP state for the given session id. It will delete local state and cease sending refreshes, allowing distributed state to time out.

5. Teardown

   Call: TEARDOWN(sid)

   This call sends Teardown messages to actively remove state from the routers, and then issues a CLOSE.

3.6.3 RSVP/Traffic Control Interface

In each router and host, enhanced QoS is achieved by a group of inter-related functions: a packet Classifier, an admission control module, and a packet scheduler. We group these functions together under the heading *traffic control*. RSVP uses the interfaces in this section to invoke the traffic control functions.

((XXX Need some way to pass the Drop flag))

1. Make a Reservation

   Call: $Rhandle = TCAddFlow(Flowspec, DropFlag, [SessionFilterspec[, SenderFilterspec]])$

   Returns an internal handle Rhandle for subsequent references to this reservation.

   This call passes Flowspec to admission control and returns an error code if Flowspec is malformed or if the requested resources are unavailable. Otherwise, it establishes a new reservation channel corresponding to Rhandle, and if Filterspecs are supplied, installs a corresponding filter in the classifier.

For FF reservation requests, RSVP knows about sharing and calls AddFlow only for distinct source pipes.

For DF reservation requests: suppose that the *Resv* message specifies a Dynamic Reservation Count = D, and F flow descriptors, where $0 <= F <= D$. Then RSVP calls AddFlow D times, and D - F of those calls have null filterspecs.

2. Switch a Channel

   Call: $TCModFilter(Rhandle, [newFilterspec])$

   This call replaces the filter without calling admission control. It may replace an existing filter with no filter, modify an existing filter, or replace no filter by a filter.

3. Modify Flowspec

   Call: $TCModFlowspec(Rhandle, oldFlowspec, newFlowspec)$

   Here newFlowspec may be larger or smaller than oldFlowspec.

4. Delete Flow

   Call: TCDeleteFlow(Rhandle)

   This call kills the reservation and reduces the reference count of, and deletes if the count is zero, any filter associated with this handle.

5. Initialize

   Call: TCInitialize( )

   This call is used when RSVP initializes its state, to clear out all existing classifier and/or packet scheduler state.

3.6.4 RSVP/Routing Interface

An RSVP implementation needs the following support from the packet forwarding and routing mechanism of the node.

- Promiscuous receive mode for RSVP messages

  Any datagram received for IP protocol 46 is to be diverted to the RSVP program for processing, without being forwarded.

- Route discovery

  RSVP must be able to discover the route(s) that the routing algorithm would have used for forwarding a specific datagram.

$GetUCRoute(DestAddress)- > NextHop, Interface$

$GetMCRoute(SrcAddress, DestAddress)- > Interface$

- Outgoing Link Specification

  RSVP must be able to force a (multicast) datagram to be sent on a specific outgoing virtual link, bypassing the normal routing mechanism. A virtual link may be a real outgoing link or a multicast tunnel.

  This is necessary because RSVP sends different copies of outgoing path messages on different links, even though all of them use the same source and destination addresses.

- Discover (Virtual) Interface List

  RSVP must be able to learn what virtual interfaces exist.

4 ACKNOWLEDGMENTS

Lixia Zhang, Scott Shenker, Deborah Estrin, Dave Clark, Sugih Jamin, Shai Herzog, Steve Berson, Steve Deering, Bob Braden, and Daniel Zappala have all made contributions to the design of RSVP. We are grateful to Jamin, Herzog, and Berson for prototype implementations. The original protocol concepts for RSVP arose out of discussions in meetings of the End-to-End Research Group.

References

[CSZ92] Clark, D., Shenker, S., and L. Zhang, *Supporting Real-Time Applications in an Integrated Services Packet Network: Architecture and Mechanisms*, Proc. SIGCOMM '92, Baltimore, MD, August 1992.

[ISInt93] Braden, R., Clark, D., and S. Shenker, *Integrated Services in the Internet Architecture: an Overview*, working draft, October 1993.

[IServ93] Shenker, S., Clark, D., and L. Zhang, *A Service Model for an Integrated Services Internet*, working draft, October 1993.

[RSVP93] Zhang, L., Deering, S., Estrin, D., Shenker, S., and D. Zappala, *RSVP: A New Resource ReSerVation Protocol*, IEEE Network, September 1993.