Network Working Group
Request for Comments: 3175
Category: Standards Track

F. Baker
C. Iturralde
F. Le Faucheur
B. Davie
Cisco Systems

September 2001

Aggregation of RSVP for IPv4 and IPv6 Reservations

Status of this Memo

Copyright Notice

Abstract

This document describes the use of a single RSVP (Resource ReSerVation Protocol) reservation to aggregate other RSVP reservations across a transit routing region, in a manner conceptually similar to the use of Virtual Paths in an ATM (Asynchronous Transfer Mode) network. It proposes a way to dynamically create the aggregate reservation, classify the traffic for which the aggregate reservation applies, determine how much bandwidth is needed to achieve the requirement, and recover the bandwidth when the sub-reservations are no longer required. It also contains recommendations concerning algorithms and policies for predictive reservations.

1. Introduction

A key problem in the design of RSVP version 1 [RSVP] is, as noted in its applicability statement, that it lacks facilities for aggregation of individual reserved sessions into a common class. The use of such aggregation is recommended in [CSZ], and required for scalability.

The problem of aggregation may be addressed in a variety of ways. For example, it may sometimes be sufficient simply to mark reserved traffic with a suitable DSCP (e.g., EF), thus enabling aggregation of scheduling and classification state. It may also be desirable to install one or more aggregate reservations from ingress to egress of

an "aggregation region" (defined below) where each aggregate reservation carries similarly marked packets from a large number of flows. This is to provide high levels of assurance that the end-to-end requirements of reserved flows will be met, while at the same time enabling reservation state to be aggregated.

Throughout, we will talk about "Aggregator" and "Deaggregator", referring to the routers at the ingress and egress edges of an aggregation region. Exactly how a router determines whether it should perform the role of aggregator or deaggregator is described below.

We will refer to the individual reserved sessions (the sessions we are attempting to aggregate) as "end-to-end" reservations ("E2E" for short), and to their respective Path/Resv messages as E2E Path/Resv messages. We refer to the the larger reservation (that which represents many E2E reservations) as an "aggregate" reservation, and its respective Path/Resv messages as "aggregate Path/Resv messages".

1.1. Problem Statement: Aggregation Of E2E Reservations

The problem of many small reservations has been extensively discussed, and may be summarized in the observation that each reservation requires a non-trivial amount of message exchange, computation, and memory resources in each router along the way. It would be nice to reduce this to a more manageable level where the load is heaviest and aggregation is possible.

Aggregation, however, brings its own challenges. In particular, it reduces the level of isolation between individual flows, implying that one flow may suffer delay from the bursts of another. Synchronization of bursts from different flows may occur. However, there is evidence [CSZ] to suggest that aggregation of flows has no negative effect on the mean delay of the flows, and actually leads to a reduction of delay in the "tail" of the delay distribution (e.g., 99% percentile delay) for the flows. These benefits of aggregation to some extent offset the loss of strict isolation.

1.2. Proposed Solution

The solution we propose involves the aggregation of several E2E reservations that cross an "aggregation region" and share common ingress and egress routers into one larger reservation from ingress to egress. We define an "aggregation region" as a contiguous set of systems capable of performing RSVP aggregation (as defined following) along any possible route through this contiguous set.

Communication interfaces fall into two categories with respect to an aggregation region; they are "exterior" to an aggregation region, or they are "interior" to it. Routers that have at least one interface in the region fall into one of three categories with respect to a given RSVP session; they aggregate, they deaggregate, or they are between an aggregator and a deaggregator.

Aggregation depends on being able to hide E2E RSVP messages from RSVP-capable routers inside the aggregation region. To achieve this end, the IP Protocol Number in the E2E reservation's Path, PathTear, and ResvConf messages is changed from RSVP (46) to RSVP-E2E-IGNORE (134) upon entering the aggregation region, and restored to RSVP at the deaggregator point. These messages are ignored (no state is stored and the message is forwarded as a normal IP datagram) by each router within the aggregation region whenever they are forwarded to an interior interface. Since the deaggregating router perceives the previous RSVP hop on such messages to be the aggregating router, Resv and other messages do not require this modification; they are unicast from RSVP hop to RSVP hop anyway.

The token buckets (SENDER_TSPECs and FLOWSPECS) of E2E reservations are summed into the corresponding information elements in aggregate Path and Resv messages. Aggregate Path messages are sent from the aggregator to the deaggregator(s) using RSVP's normal IP Protocol Number. Aggregate Resv messages are sent back from the deaggregator to the aggregator, thus establishing an aggregate reservation on behalf of the set of E2E flows that use this aggregator and deaggregator.

Such establishment of a smaller number of aggregate reservations on behalf of a larger number of E2E reservations yields the corresponding reduction in the amount of state to be stored and amount of signalling messages exchanged in the aggregation region.

By using Differentiated Services mechanisms for classification and scheduling of traffic supported by aggregate reservations (rather than performing per aggregate reservation classification and scheduling), the amount of classification and scheduling state in the aggregation region is even further reduced. It is not only independent of the number of E2E reservations, it is also independent of the number of aggregate reservations in the aggregation region. One or more Diff-Serv DSCPs are used to identify traffic covered by aggregate reservations and one or more Diff-Serv PHBs are used to offer the required forwarding treatment to this traffic. There may be more than one aggregate reservation between the same pair of routers, each representing different classes of traffic and each using a different DSCP and a different PHB.

1.3. Definitions

We define an "aggregation region" as a set of RSVP-capable routers for which E2E RSVP messages arriving on an exterior interface of one router in the set would traverse one or more interior interfaces (of this and possibly of other routers in the set) before finally traversing an exterior interface.

Such an E2E RSVP message is said to have crossed the aggregation region.

We define the "aggregating" router for this E2E flow as the first router that processes the E2E Path message as it enters the aggregation region (i.e., the one which forwards the message from an exterior interface to an interior interface).

We define the "deaggregating" router for this E2E flow as the last router to process the E2E Path as it leaves the aggregation region (i.e., the one which forwards the message from an interior interface to an exterior interface).

We define an "interior" router for this E2E flow as any router in the aggregation region which receives this message on an interior interface and forwards it to another interior interface. Interior routers perform neither aggregation nor deaggregation for this flow.

Note that by these definitions a single router with a mix of interior and exterior interfaces may have the capability to act as an aggregator on some E2E flows, a deaggregator on other E2E flows, and an interior router on yet other flows.

1.4. Detailed Aspects of Proposed Solution

A number of issues jump to mind in considering this model.

1.4.1. Traffic Classification Within The Aggregation Region

One of the reasons that RSVP Version 1 did not identify a way to aggregate sessions was that there was not a clear way to classify the aggregate. With the development of the Differentiated Services architecture, this is at least partially resolved; traffic of a particular class can be marked with a given DSCP and so classified. We presume this model.

We presume that on each link en route, a queue, WDM color, or similar management component is set aside for all aggregated traffic of the same class, and that sufficient bandwidth is made available to carry

the traffic that has been assigned to it. This bandwidth may be adjusted based on the total amount of aggregated reservation traffic assigned to the same class.

There are numerous options for exactly which Diff-serv PHBs might be used for different classes of traffic as it crosses the aggregation region. This is the "service mapping" problem described in [RFC2998], and is applicable to situations broader than those described in this document. Arguments can be made for using either EF or one or more AF PHBs for aggregated traffic. For example, since controlled load requires non-TSpec-conformant (policed) traffic to be forwarded as best effort traffic rather than dropped, it may be appropriate to use an AF class for controlled load, using the higher drop preference for non-conformant packets.

In conventional (unaggregated) RSVP operation, a session is identified by a destination address and optionally a protocol port. Since data belonging to an aggregated reservation is identified by a DSCP, the session is defined by the destination address and DSCP. For those cases where two DSCPs are used (for conformant and non-conformant packets, as noted above), the session is identified by the DSCP of conformant packets. In general we will talk about mapping aggregated traffic onto a DSCP (even if a second DSCP may be used for non-conformant traffic).

Whichever PHB or PHBs are used to carry aggregated reservations, care needs to be take in an environment where provisioned Diff-Serv and aggregated RSVP are used in the same network, to ensure that the total admitted load for a single PHB does not exceed the link capacity allocated to that PHB. One solution to this is to reserve one PHB (or more) strictly for the aggregated reservation traffic (e.g., AF1 Class) while using other PHBs for provisioned Diff-Serv (e.g., AF2, AF3 and AF4 Classes).

Inside the aggregation region, some RSVP reservation state is maintained per aggregate reservation, while classification and scheduling state (e.g., DSCPs used for classifying traffic) is maintained on a per aggregate reservation class basis (rather than per aggregate reservation). For example, if Guaranteed Service reservations are mapped to the EF DSCP throughout the aggregation region, there may be a reservation for each aggregator/deaggregator pair in each router, but only the EF DSCP needs to be inspected at each interior interface, and only a single queue is used for all EF traffic.

1.4.2. Deaggregator Determination

The first question is "How do we determine the Aggregator/Deaggregator pair that are responsible for aggregating a particular E2E flow through the aggregation region?"

Determination of the aggregator is trivial: we know that an E2E flow has arrived at an aggregator when its Path message arrives at a router on an exterior interface and must be forwarded on an interior interface.

Determination of the deaggregator is more involved. If an SPF routing protocol, such as OSPF or IS-IS, is in use, and if it has been extended to advertise information on Deaggregation roles, it can tell us the set of routers from which the deaggregator will be chosen. In principle, if the aggregator and deaggregator are in the same area, then the identity of the deaggregator could be determined from the link state database. However, this approach would not work in multi-area environments or for distance vector protocols.

One method for Deaggregator determination is manual configuration. With this method the network operator would configure the Aggregator and the Deaggregator with the necessary information.

Another method allows automatic Deaggregator determination and corresponding Aggregator notification. When the E2E RSVP Path message transits from an interior interface to an exterior interface, the deaggregating router must advise the aggregating router of the correlation between itself and the flow. This has the nice attribute of not being specific to the routing protocol. It also has the property of automatically adjusting to route changes. For instance, if because of a topology change, another Deaggregator is now on the shortest path, this method will automatically identify the new Deaggregator and swap to it.

1.4.3. Mapping E2E Reservations Onto Aggregate Reservations

As discussed above, there may be multiple Aggregate Reservations between the same Aggregator/Deaggregator pair. The rules for mapping E2E reservations onto aggregate reservations are policy decisions which depend on the network environment and network administrator’s objectives. Such a policy is outside the scope of this specification and we simply assume that such a policy is defined by the network administrator. We also assume that such a policy is somehow accessible to the Aggregators/Deaggregators but the details of how this policy is made accessible to Aggregators/Deaggregators (Local Configuration, COPS, LDAP, etc.) is outside the scope of this specification.

An example of very simple policy would be that all the E2E reservations are mapped onto a single Aggregate Reservation (i.e., single DSCP) between a given pair of Aggregator/Deaggregator.

Another example of policy, which takes into account the Int-Serv service type requested by the receiver (and signalled in the E2E Resv), would be where Guaranteed Service E2E reservations are mapped onto one DSCP in the aggregation region and where Controlled Load E2E reservations are mapped onto another DSCP.

A third example of policy would be one where the mapping of E2E reservations onto Aggregate Reservations take into account Policy Objects (such as information authenticating the end user) which may be included by the sender in the E2E path and/or by the receiver in the E2E Resv.

Regardless of the actual policy, a range of options are conceivable for where the decision to map an E2E reservation onto an aggregate reservation is taken and how this decision is communicated between Aggregator and Deaggregator. Both Aggregator and Deaggregator could be assumed to make such a decision independently. However, this would either require definition of additional procedures to solve inconsistent mapping decisions (i.e., Aggregator and Deaggregator decide to map a given E2E reservation onto different Aggregate Reservations) or would result in possible undetected misbehavior in the case of inconsistent decisions.

For simplicity and reliability, we assign the responsibility of the mapping decision entirely to the Deaggregator. The Aggregator is notified of the selected mapping by the Deaggregator and follows this decision. The Deaggregator was chosen rather than the Aggregator because the Deaggregator is the first to have access to all the information required to make such a decision (in particular receipt of the E2E Resv which indicates the requested Int-Serv service type and includes information signalled by the receiver). This allows faster operations such as set-up or size adjustment of an Aggregate Reservation in a number of situations resulting in faster E2E reservation establishment.

1.4.4. Size of Aggregate Reservations

A range of options exist for determining the size of the aggregate reservation, presenting a tradeoff between simplicity and scalability. Simplistically, the size of the aggregate reservation needs to be greater than or equal to the sum of the bandwidth of the E2E reservations it aggregates, and its burst capacity must be greater than or equal to the sum of their burst capacities. However, if followed religiously, this leads us to change the bandwidth of the

aggregate reservation each time an underlying E2E reservation changes, which loses one of the key benefits of aggregation, the reduction of message processing cost in the aggregation region.

We assume, therefore, that there is some policy, not defined in this specification (although sample policies are suggested which have the necessary characteristics). This policy maintains the amount of bandwidth required on a given aggregate reservation by taking account of the sum of the bandwidths of its underlying E2E reservations, while endeavoring to change it infrequently. This may require some level of trend analysis. If there is a significant probability that in the next interval of time the current aggregate reservation will be exhausted, the router must predict the necessary bandwidth and request it. If the router has a significant amount of bandwidth reserved but has very little probability of using it, the policy may be to predict the amount of bandwidth required and release the excess.

This policy is likely to benefit from introduction of some hysteresis (i.e., ensure that the trigger condition for aggregate reservation size increase is sufficiently different from the trigger condition for aggregate reservation size decrease) to avoid oscillation in stable conditions.

Clearly, the definition and operation of such policies are as much business issues as they are technical, and are out of the scope of this document.

1.4.5. E2E Path ADSPEC update

As described above, E2E RSVP messages are hidden from the Interior routers inside the aggregation region. Consequently, the ADSPECs of E2E Path messages are not updated as they travel through the aggregation region. Therefore, the Deaggregator for a flow is responsible for updating the ADSPEC in the corresponding E2E Path to reflect the impact of the aggregation region on the QoS that may be achieved end-to-end. The Deaggregator should update the ADSPEC of the E2E Path as accurately as possible.

Since Aggregate Path messages are processed inside the aggregation region, their ADSPEC is updated by Interior routers to reflect the impact of the aggregation region on the QoS that may be achieved within the interior region. Consequently, the Deaggregator should make use of the information included in the ADSPEC from an Aggregate Path where available. The Deaggregator may elect to wait until such information is available before forwarding the E2E Path in order to accurately update its ADSPEC.

To maximize the information made available to the Deaggregator, whenever the Aggregator signals an Aggregate Path, the Aggregator should include an ADSPEC with fragments for all service types supported in the aggregation region (even if the Aggregate Path corresponds to an Aggregate Reservation that only supports a subset of those service types). Providing this information to the Deaggregator for every possible service type facilitates accurate and timely update of the E2E ADSPEC by the Deaggregator.

Depending on the environment and on the policy for mapping E2E reservations onto Aggregate Reservations, to accurately update the E2E Path ADSPEC, the Deaggregator may for example:

- update all the E2E Path ADSPEC segments (Default General Parameters Fragment, Guaranteed Service Fragment, Controlled-Load Service Fragment) based on the ADSPEC of a single Aggregate Path, or

- update the E2E Path ADSPEC by taking into account the ADSPEC from multiple Aggregate Path messages (e.g.,. update the Default General Parameters Fragment using the "worst" value for each parameter across all the Aggregate Paths' ADSPECs, update the Guaranteed Service Fragment using the Guaranteed Service Fragment from the ADSPEC of the Aggregate Path for the reservation used for Guaranteed Services).

By taking into account the information contained in the ADSPEC of Aggregate Path(s) as mentioned above, the Deaggregator should be able to accurately update the E2E Path ADSPEC in most situations.

However, we note that there may be particular situations where the E2E Path ADSPEC update cannot be made entirely accurately by the Deaggregator. This is most likely to happen when the path taken across the aggregation region depends on the service requested in the E2E Resv, which is yet to arrive. Such a situation could arise if, for example:

- The service mapping policy for the aggregation region is such that E2E reservations requesting Guaranteed Service are mapped to a different PHB that those requesting Controlled Load service.

- Diff-Serv aware routing is used in the aggregation region, so that packets with different DSCPs follow different paths (sending them over different MPLS label switched paths, for example).

As a result, the ADSPEC for the aggregate reservation that supports guaranteed service may differ from the ADSPEC for the aggregate reservation that supports controlled load.

Assume that the sender sends an E2E Path with an ADSPEC containing segments for both Guaranteed Services and Controlled Load. Then, at the time of updating the E2E ADSPEC, the Deaggregator does not know which service type will actually be requested by the receiver and therefore cannot know which PHB will be used to transport this E2E flow and, in turn, cannot pick the right parameter values to factor in when updating the Default General Parameters Fragment. As mentioned above, in this particular case, a conservative approach would be to always take into account the worst value for every parameter. Regardless of whether this conservative approach is followed or some simpler approach such as taking into account one of the two Aggregate Path ADSPEC, the E2E Path ADSPEC will be inaccurate (over-optimistic or over-pessimistic) for at least one service type actually requested by the destination.

Recognizing that entirely accurate update of E2E Path ADSPEC may not be possible in all situations, we recommend that a conservative approach be taken in such situations (over-pessimistic rather than over-optimistic) and that the E2E Path ADSPEC be corrected as soon as possible. In the example described above, this would mean that as soon as the Deaggregator receives the E2E Resv from the receiver, the Deaggregator should generate another E2E Path with an accurately updated ADSPEC based on the knowledge of which aggregate reservation will actually carry the E2E flow.

1.4.6. Intra-domain Routes

RSVP directly handles route changes, in that reservations follow the routes that their data follow. This follows from the property that Path messages contain the same IP source and destination address as the data flow for which a reservation is to be established. However, since we are now making aggregate reservations by sending a Path message from an aggregating to a deaggregating router, the reserved (E2E) data packets no longer carry the same IP addresses as the relevant (aggregate) Path message. The issue becomes one of making sure that data packets for reserved flows follow the same path as the Path message that established Path state for the aggregate reservation. Several approaches are viable.

First, the data may be tunneled from aggregator to deaggregator, using technologies such as IP-in-IP tunnels, GRE tunnels, MPLS label-switched paths, and so on. These each have particular advantages, especially MPLS, which allows traffic engineering. They each also have some cost in link overhead and configuration complexity.

If data is not tunneled, then we are depending on a characteristic of IP best metric routing , which is that if the route from A to Z includes the path from H to L, and the best metric route was chosen all along the way, then the best metric route was chosen from H to L. Therefore, an aggregate path message which crosses a given aggregator and deaggregator will of necessity use the best path between them.

If this is a single path, the problem is solved. If it is a multi-path route, and the paths are of equal cost, then we are forced to determine, perhaps by measurement, what proportion of the traffic for a given E2E reservation is passing along each of the paths, and assure ourselves of sufficient bandwidth for the present use. A simple, though wasteful, way of doing this is to reserve the total capacity of the aggregate route down each path.

For this reason, we believe it is advantageous to use one of the above-mentioned tunneling mechanisms in cases where multiple equal-cost paths may exist.

1.4.7. Inter-domain Routes

The case of inter-domain routes differs somewhat from the intra-domain case just described. Specifically, best-path considerations do not apply, as routing is by a combination of routing policy and shortest AS path rather than simple best metric.

In the case of inter-domain routes, data traffic belonging to different E2E sessions (but the same aggregate session) may not enter an aggregation region via the same aggregator interface, and/or may not leave via the same deaggregator interface. It is possible that we could identify this occurrence in some central system which sees the reservation information for both of the apparent sessions, but it is not clear that we could determine a priori how much traffic went one way or the other apart from measurement.

We simply note that this problem can occur and needs to be allowed for in the implementation. We recommend that each such E2E reservation be summed into its appropriate aggregate reservation, even though this involves over-reservation.

1.4.8. Reservations for Multicast Sessions

Aggregating reservations for multicast sessions is significantly more complex than for unicast sessions. The first challenge is to construct a multicast tree for distribution of the aggregate Path messages which follows the same path as will be followed by the data packets for which the aggregate reservation is to be made. This is complicated by the fact that the path taken by a data packet may

depend on many factors such as its source address, the choice of shared trees or source-specific trees, and the location of a rendezvous point for the tree.

Once the problem of distributing aggregate Path messages is solved, there are considerable problems in determining the correct amount of resources to reserve at each link along the multicast tree. Because of the amount of heterogeneity that may exist in an aggregate multicast reservation, it appears that it would be necessary to retain information about individual E2E reservations within the aggregation region to allocate resources correctly. Thus, we may end up with a complex set of procedures for forming aggregate reservations that do not actually reduce the amount of stored state significantly for multicast sessions.

As noted above, there are several aspects to RSVP state, and our approach for unicast aggregates all forms of state: classification, scheduling, and reservation state. One possible approach to multicast is to focus only on aggregation of classification and scheduling state, which are arguably the most important because of their impact on the forwarding path. That approach is the one described in the current draft.

1.4.9. Multi-level Aggregation

Ideally, an aggregation scheme should be able to accommodate recursive aggregation, with aggregate reservations being themselves aggregated. Multi-level aggregation can be accomplished using the procedures described here and a simple extension to the protocol number swapping process.

We can consider E2E RSVP reservations to be at aggregation level 0. When we aggregate these reservations, we produce reservations at aggregation level 1. In general, level n reservations may be aggregated to form reservations at level n+1.

When an aggregating router receives an E2E Path, it swaps the protocol number from RSVP to RSVP-E2E-IGNORE. In addition, it should write the aggregation level (1, in this case) in the 2 byte field that is present (and currently unused) in the router alert option. In general, a router which aggregates reservations at level n to create reservations at level n+1 will write the number n+1 in the router alert field. A router which deaggregates level n+1 reservations will examine all messages with IP protocol number RSVP-E2E-IGNORE but will process the message and swap the protocol number back to RSVP only in the case where the router alert field carries the number n+1. For any other value, the message is forwarded unchanged. Interior routers ignore all messages with IP protocol

number RSVP-E2E-IGNORE. Note that only a few bits of the 2 byte field in the option would be needed, given the likely number of levels of aggregation.

For IPv6, certain values of the router alert "value" field are reserved. This specification requires IANA assignment of a small number of consecutive values for the purpose of recording the aggregation level.

1.4.10. Reliability Issues

There are a variety of issues that arise in the context of aggregation that would benefit from some form of explicit acknowledgment mechanism for RSVP messages. For example, it is possible to configure a set of routers such that an E2E Path of protocol type RSVP-E2E-IGNORE would be effectively "black-holed", if it never reached a router which was appropriately configured to act as a deaggregator. It could then travel all the way to its destination where it would probably be ignored due to its non-standard protocol number. This situation is not easy to detect. The aggregator can be sure this problem has not occurred if an aggregate PathErr message is received from the deaggregator (as described in detail below). It can also be sure there is no problem if an E2E Resv is received. However, the fact that neither of these events has happened may only mean that no receiver wishes to reserve resources for this session, or that an RSVP message loss occurred, or it may mean that the Path was black-holed. However, if a neighbor-to-neighbor acknowledgment mechanism existed, the aggregator would expect to receive an acknowledgment of the E2E Path from the deaggregator, and would interpret the lack of a response as an indication that a problem of configuration existed. It could then refrain from aggregating this particular session. We note that such a reliability mechanism has been proposed for RSVP in [RFC291] and propose that it be used here.

1.4.11. Message Integrity and Node Authentication

[RSVP] defines a hop-by-hop authentication and integrity check. The present specification allows use of this check on Aggregate RSVP messages and also preserves this check on E2E RSVP messages for E2E RSVP messages.

Outside the Aggregation Region, any E2E RSVP message may contain an INTEGRITY object using a keyed cryptographic digest technique which assumes that RSVP neighbors share a secret. Because E2E RSVP messages are not processed by routers in the Aggregation Region, the Aggregator and Deaggregator appear as logical RSVP neighbors of each other. The Deaggregator is the Aggregator’s Next Hop for E2E RSVP

messages while the Aggregator is the Deaggregator's Previous Hop. Consequently, INTEGRITY objects which may appear in E2E RSVP messages traversing the Aggregation Region are exchanged directly between the Aggregator and Deaggregator in a manner which is entirely transparent to the Interior routers. Thus, hop-by-hop integrity checking for E2E messages over the Aggregation Region requires that the Aggregator and Deaggregator share a secret. Techniques for establishing that secret are described in [INTEGRITY].

Inside the Aggregation Region, any Aggregate RSVP message may contain an INTEGRITY object which assumes that the corresponding RSVP neighbors inside the Aggregation Region (e.g., Aggregator and Interior Router, two Interior Routers, Interior Router and Deaggregator) share a secret.

1.4.12. Aggregated reservations without E2E reservations

Up to this point we have assumed that the aggregate reservation is established as a result of the establishment of E2E reservations from outside the aggregation region. It should be clear that alternative triggers are possible. As discussed in [RFC2998], an aggregate RSVP reservation can be used to manage bandwidth in a diff-serv cloud even if RSVP is not used end-to-end.

The simplest example of an alternative configuration is the static configuration of an aggregated reservation for a certain amount for traffic from an ingress (aggregator) router to an egress (de-aggregator) router. This would have to be configured in at least the system originating the aggregate PATH message (the aggregator). The deaggregator could detect that the PATH message is directed to it, and could be configured to "turn around" such messages, i.e., it responds with a RESV back to the aggregator. Alternatively, configuration of the aggregate reservation could be performed at both the aggregator and the deaggregator. As before, an aggregate reservation is associated with a DSCP for the traffic that will use the reserved capacity.

In the absence of E2E microflow reservations, the aggregator can use a variety of policies to set the DSCP of packets passing into the aggregation region, thus determining whether they gain access to the resources reserved by the aggregate reservation. These policies are a matter of local configuration, as usual for a device at the edge of a diffserv cloud.

Note that the "aggregator" could even be a device such as a PSTN gateway which makes an aggregate reservation for the set of calls to another PSTN gateway (the deaggregator) across an intervening diff-serv region. In this case the reservation may be established in response to call signalling.

From the perspective of RSVP signalling and the handling of data packets in the aggregation region, these cases are equivalent to the case of aggregating E2E RSVP reservations. The only difference is that E2E RSVP signalling does not take place and cannot therefore be used as a trigger, so some additional knowledge is required in setting up the aggregate reservation.

2. Elements of Procedure

To implement aggregation, we define a number of elements of procedure.

2.1. Receipt of E2E Path Message By Aggregating Router

The very first event is the arrival of the E2E Path message at an exterior interface of an aggregator. Standard RSVP procedures [RSVP] are followed for this, including onto what set of interfaces the message should be forwarded. These interfaces comprise zero or more exterior interfaces and zero or more interior interfaces. (If the number of interior interfaces is zero, the router is not acting as an aggregator for this E2E flow.)

Service on exterior interfaces is handled as defined in [RSVP].

Service on interior interfaces is complicated by the fact that the message needs to be included in some aggregate reservation, but at this point it is not known which one, because the deaggregator is not known. Therefore, the E2E Path message is forwarded on the interior interface(s) using the IP Protocol number RSVP-E2E-IGNORE, but in every other respect identically to the way it would be sent by an RSVP router that was not performing aggregation.

2.2. Handling Of E2E Path Message By Interior Routers

At this point, the E2E Path message traverses zero or more interior routers. Interior routers receive the E2E Path message on an interior interface and forward it on another interior interface. The Router Alert IP Option alerts interior routers to check internally, but they find that the IP Protocol is RSVP-E2E-IGNORE and the next hop interface is interior. As such, they simply forward it as a normal IP datagram.

2.3. Receipt of E2E Path Message By Deaggregating Router

The E2E Path message finally arrives at a deaggregating router, which receives it on an interior interface and forwards it on an exterior interface. Again, the Router Alert IP Option alerts it to intercept the message, but this time the IP Protocol is RSVP-E2E-IGNORE and the next hop interface is an exterior interface.

Before forwarding the E2E Path towards the receiver, the Deaggregator should update its ADSPEC. This update is to reflect the impact of the aggregation region onto the QoS to be achieved E2E by the flow. Such information can be collected by the ADSPEC of Aggregate Path messages travelling from the Aggregator to the Deaggregator. Thus, to enable correct updating of the ADSPEC, a deaggregating router may wait as described below for the arrival of an aggregate Path before forwarding the E2E Path.

When receiving the E2E Path, depending on the policy for mapping E2E reservation onto Aggregate Reservations, the Deaggregator may or may not be in a position to decide which DSCP the E2E flow for the processed E2E Path is going to be mapped onto, as described above. If the Deaggregator is in a position to know the mapping at this point, then the Deaggregator first checks that there is an Aggregate Path in place for the corresponding DSCP. If so, then the Deaggregator uses the ADSPEC of this Aggregate Path to update the ADSPEC of the E2E Path and then forwards the E2E Path towards the receiver. If not, then the Deaggregator requests establishment of the corresponding Aggregate Path by sending an E2E PathErr message with an error code of NEW-AGGREGATE-NEEDED and the desired DSCP encoded in the DCLASS Object. The Deaggregator may also at the same time request establishment of an aggregate reservation for other DSCPs. When receiving the Aggregate Path for the desired DSCP, the Deaggregator then uses the ADSPEC of this Aggregate Path to update the ADSPEC of the E2E Path.

If the Deaggregator is not in a position to know the mapping at this point, then the Deaggregator uses the information contained in the ADSPEC of one Aggregate Path or of multiple Aggregate Paths to update the E2E Path ADSPEC. Similarly, if one or more of the necessary Aggregate Paths is not yet established, the Deaggregator requests establishment of the corresponding Aggregate Path by sending an E2E PathErr message with an error code of NEW-AGGREGATE-NEEDED and the desired DSCP encoded in the respective DCLASS Object. When receiving the Aggregate Path for the desired DSCP, the Deaggregator then uses the ADSPEC of this Aggregate Path to update the ADSPEC of the E2E Path.

Generating a E2E PathErr message with an error code of NEW-AGGREGATE-NEEDED should not result in any Path state being removed, but should result in the aggregating router initiating the necessary aggregate Path message, as described in the following section.

The deaggregating router changes the E2E Path message's IP Protocol from RSVP-E2E-IGNORE to RSVP and forwards the E2E Path message towards its intended destination.

2.4. Initiation of New Aggregate Path Message By Aggregating Router

The aggregating Router is responsible for generating a new Aggregate Path for a DSCP when receiving a E2E PathErr message with the error code NEW-AGGREGATE-NEEDED from the deaggregator. The DSCP value to include in the Aggregate Path Session is found in the DCLASS Object of the received E2E PathErr message. The identity of the deaggregator itself is found in the ERROR SPECIFICATION of the E2E PathErr message. The destination address of the aggregate Path message is the address of the deaggregating router, and the message is sent with IP protocol number RSVP.

Existing RSVP procedures specify that the size of a reservation established for a flow is set to the minimum of the Path SENDER_TSPEC and the Resv FLOW_SPEC. Consequently, the size of an Aggregate Reservation cannot be larger than the SENDER_TSPEC included in the Aggregate Path by the Aggregator. To ensure that Aggregate Reservations can be sized by the Deaggregator without undesired limitations, the Aggregating router should always attempt to include in the Aggregate Path a SENDER_TSPEC which is at least as large as the size that would actually be required as determined by the Deaggregator. One method to achieve this is to use a SENDER_TSPEC which is obviously larger than the highest load of E2E reservations that may be supported onto this network. Another method is for the Aggregator to keep track of which flows are mapped onto a DSCP and always add their E2E Path SENDER_TSPEC into the Aggregate Path SENDER_TSPEC (and possibly also add some additional bandwidth in anticipation of future E2E reservations).

The aggregating router is notified of the mapping from an E2E flow to a DSCP in two ways. First, when the aggregating router receives a E2E PathErr with error code NEW-AGGREGATE-NEEDED, the Aggregator is notified that the corresponding E2E flow is (at least temporarily) mapped onto a given DSCP. Secondly, when the aggregating router receives an E2E Resv containing a DCLASS Object (as described further below), the Aggregating Router is notified that the corresponding E2E flow is mapped onto a given DSCP.

2.5. Handling of E2E Resv Message by Deaggregating Router

Having sent the E2E Path message on toward the destination, the deaggregator must now expect to receive an E2E Resv for the session. On receipt, its responsibility is to ensure that there is sufficient bandwidth reserved within the aggregation region to support the new E2E reservation, and if there is, then to forward the E2E Resv to the aggregating router.

The Deaggregating router first makes the final decision of which Aggregate Reservation (and thus which DSCP) this E2E reservation is to be mapped onto. This decision is made according to the policy selected by the network administrator as described above.

If this final mapping decision is such that the Deaggregator can now make a more accurate update of the E2E Path ADSPEC than done when forwarding the initial E2E Path, the Deaggregator should do so and generate a new E2E Path immediately in order to provide the accurate ADSPEC information to the receiver as soon as possible. Otherwise, normal Refresh procedures should be followed for the E2E Path.

If no Aggregate Reservation currently exists from the corresponding aggregating router with the corresponding DSCP, the Deaggregating router will establish a new Aggregate Reservation as described in the next section.

If the corresponding Aggregate Reservation exists but has insufficient bandwidth reserved to accommodate the new E2E reservation (in addition to all the existing E2E reservations currently mapped onto it), it should follow the normal RSVP procedures [RSVP] for a reservation being placed with insufficient bandwidth to support the reservation. It may also first attempt to increase the aggregate reservation that is supplying bandwidth by increasing the size of the FLOW_SPEC that it includes in the aggregate Resv that it sends upstream. As discussed in the previous section, the Aggregating Router should ensure that the SENDER_TSPEC it includes in the Aggregate Path is always in excess of the FLOW_SPEC that may be requested in the Aggregate Resv by the Deaggregator, so that the Deaggregator is not unnecessarily prevented from effectively increasing the Aggregate Reservation bandwidth as required.

When sufficient bandwidth is available on the corresponding aggregate reservation, the Deaggregating Router may simply send the E2E Resv message with IP Protocol RSVP to the aggregating router. This message should include the DCLASS object to indicate which DSCP the aggregator must use for this E2E flow. The deaggregator will also

add the token bucket from the E2E Resv FLOWSPEC object into its internal understanding of how much of the Aggregate reservation is in use.

As discussed above, in order to minimize the occurrence of situations where insufficient bandwidth is reserved on the corresponding Aggregate Reservation at the time of processing an E2E Resv, and in turn to avoid the delay associated with the increase of this aggregate bandwidth, the Deaggregator MAY anticipate the current demand and increase the Aggregate Reservations size ahead of actual requirements by E2E reservations.

2.6. Initiation of New Aggregate Resv Message By Deaggregating Router

Upon receiving an E2E Resv message on an exterior interface, and having determined the appropriate DSCP for the session according to the mapping policy, the Deaggregator looks for the corresponding path state for a session with the chosen DSCP. If aggregate Path state exists, but no aggregate Resv state exists, the Deaggregator creates a new aggregate Resv.

If no aggregate Path state exists for the appropriate DSCP, this may be because the Deaggregator could not decide earlier the final mapping for this E2E flow and elected to not establish Aggregate Path state for all DSCPs. In that case, the Deaggregator should request establishment of the corresponding Aggregate Path by sending a E2E PathErr with error code of NEW-AGGREGATE-NEEDED and with a DCLASS containing the required DSCP. This will trigger the Aggregator to establish the corresponding Aggregate Path. Once the Deaggregator has determined that the aggregate Path state is established, it creates a new Aggregate Resv.

The FLOW_SPEC of the new Aggregate Resv is set to a value not smaller than the requirement of the E2E reservation it is supporting. The Aggregate Resv is sent toward the aggregator (i.e., to the previous hop), using the AGGREGATED-RSVP session and filter specifications defined below. Since the DSCP is in the SESSION object, no DCLASS object is necessary. The message should be reliably delivered using the mechanisms in [RFC2961] or, alternatively, the CONFIRM object may be used, to assure that the aggregate Resv does indeed arrive and is granted. This enables the deaggregator to determine that the requested bandwidth is available to allocate to the E2E flows it supports.

In order to minimize the occurrence of situations where no corresponding Aggregate Reservation is established at the time of processing an E2E Resv, and in turn to avoid the delay associated with the creation of this aggregate reservation, the Deaggregator MAY

anticipate the current demand and create the Aggregate Reservation before receiving E2E Resv messages requiring bandwidth on those aggregate reservations.

2.7. Handling of Aggregate Resv Message by Interior Routers

The aggregate Resv message is handled in essentially the same way as defined in [RSVP]. The Session object contains the address of the deaggregating router (or the group address for the session in the case of multicast) and the DSCP that has been chosen for the session. The Filterspec object identifies the aggregating router. These routers perform admission control and resource allocation as usual and send the aggregate Resv on towards the aggregator.

2.8. Handling of E2E Resv Message by Aggregating Router

The receipt of the E2E Resv message with a DCLASS Object is the final confirmation to the aggregating router of the mapping of the E2E reservation onto an Aggregate Reservation. Under normal circumstances, this is the only way it will be informed of this association. It should now forward the E2E Resv to its previous hop, following normal RSVP processing rules [RSVP].

2.9. Removal of E2E Reservation

E2E reservations are removed in the usual way via PathTear, ResvTear, timeout, or as the result of an error condition. When they are removed, their FLOWSPEC information must also be removed from the allocated portion of the aggregate reservation. This same bandwidth may be re-used for other traffic in the near future. When E2E Path messages are removed, their SENDER_TSPEC information must also be removed from the aggregate Path.

2.10. Removal of Aggregate Reservation

Should an aggregate reservation go away (presumably due to a configuration change, route change, or policy event), the E2E reservations it supports are no longer active. They must be treated accordingly.

2.11. Handling of Data On Reserved E2E Flow by Aggregating Router

Prior to establishment that a given E2E flow is part of a given aggregate, the flow’s data should be treated as traffic without a reservation by whatever policies prevail for such. Generally, this will mean being given the same forwarding behavior as best effort traffic. However, upon establishing that the flow belongs to a given aggregate, the aggregating router is responsible for marking any

related traffic with the correct DSCP and forwarding it in the manner appropriate to traffic on that reservation. This may imply forwarding it to a given IP next hop, or piping it down a given link layer circuit, tunnel, or MPLS label switched path.

The aggregator is responsible for performing per-reservation policing on the E2E flows that it is aggregating. The aggregator performs metering of traffic belonging to each reservation to assess compliance to the token bucket for the corresponding E2E reservation. Packets which are assessed in compliance are forwarded as mentioned above. Packets which are assessed out of compliance must be either dropped, reshaped or marked to a different DSCP. The detailed policing behavior is an aspect of the service mapping described in [RFC2998].

2.12. Procedures for Multicast Sessions

Because of the difficulties of aggregating multicast sessions described above, we focus on the aggregation of scheduling and classification state in the multicast case. The main difference between the multicast and unicast cases is that rather than sending an aggregate Path message to the unicast address of a single deaggregating router, in the multicast case we send the "aggregate" Path message to the same group address as the E2E session. This ensures that the aggregate Path message follows the same route as the E2E Path. This difference between unicast and multicast is reflected in the Session objects defined below. A consequence of this approach is that we continue to have reservation state per multicast session inside the aggregation region.

A further challenge arises in multicast sessions with heterogeneous receivers. Consider an interior router which must forward packets for a multicast session on two interfaces, but has only received a reservation request on one of those interfaces. It receives packets marked with the DSCP chosen for the aggregate reservation. When sending them out the interface which has no installed reservation, it has the following options:

a) remark those packets to best effort before sending them out the interface;

b) send the packets out the interface with the DSCP chosen for the aggregate reservation.

The first approach suffers from the drawback that it requires nMF classification at an interior router in order to recognize the flows whose packets must be demoted. The second approach requires over-reservation of resources on the interface on which no reservation was

received. In the absence of such over-reservation, the packets sent with the "wrong" DSCP would be able to degrade the service experienced by packets using that DSCP legitimately.

To make MF classification acceptable in an interior router, it may be possible to treat the case of heterogeneous flows as an exception. That is, an interior router only needs to be able to recognize those individual microflows that have heterogeneous resource needs on the outbound interfaces of this router.

3. Protocol Elements

3.1. IP Protocol RSVP-E2E-IGNORE

This specification requires the assignment of a protocol type RSVP-E2E-IGNORE, whose number is at this point 134. This is used only on E2E messages which require a router alert (Path, PathTear, and ResvConf), and signifies that the message must be treated one way when destined to an interior interface, and another way when destined to an exterior interface. The protocol type is swapped by the Aggregator from RSVP to RSVP-E2E-IGNORE in E2E Path, PathTear, and ResvConf messages when they enter the Aggregation Region. The protocol type is swapped back by the Deaggregator from RSVP-E2E-IGNORE to RSVP in such E2E messages when they exit the Aggregation Region.

3.2. Path Error Code

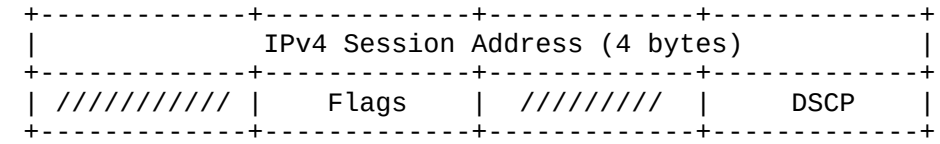
A PathErr code NEW-AGGREGATE-NEEDED is required. This value does not signify that a fatal error has occurred, but that an action is required of the aggregating router to avoid an error condition in the near future.

3.3. SESSION Object

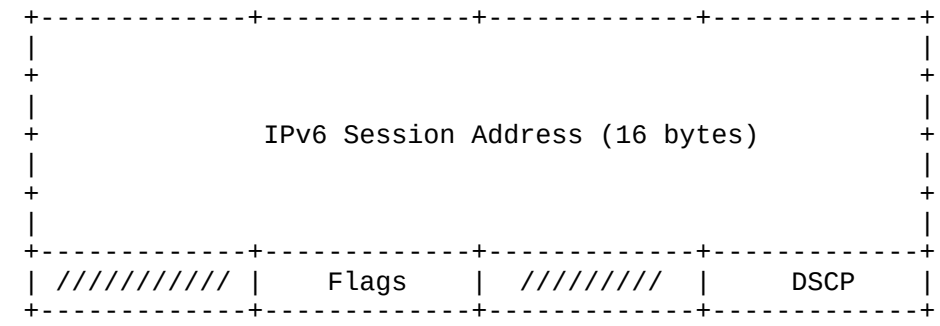
The SESSION object contains two values: the IP Address of the aggregate session destination, and the DSCP that it will use on the E2E data the reservation contains. For unicast sessions, the session destination address is the address of the deaggregating router. For multicast sessions, the session destination is the multicast address of the E2E session (or sessions) being aggregated. The inclusion of the DSCP in the session allows for multiple sessions toward the same address to be distinguished by their DSCP and queued separately. It also provides the means for aggregating scheduling and classification state. In the case where a session uses a pair of PHBs (e.g., AF11 and AF12), the DSCP used should represent the numerically smallest PHB (e.g., AF11). This follows the same naming convention described in [BRIM].

Session types are defined for IPv4 and IPv6 addresses.

- IP4 SESSION object: Class = SESSION,
  C-Type = RSVP-AGGREGATE-IP4



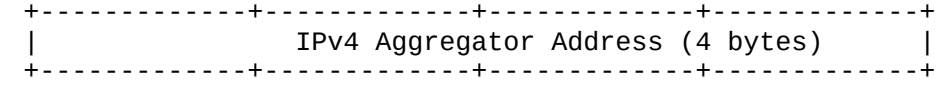
- IP6 SESSION object: Class = SESSION,
  C-Type = RSVP-AGGREGATE-IP6



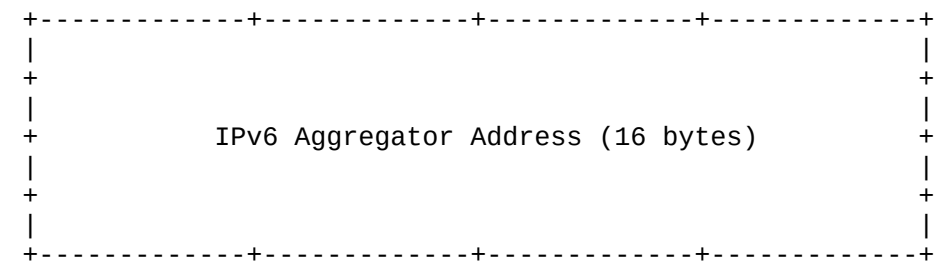
3.4. SENDER_TEMPLATE Object

The SENDER_TEMPLATE object identifies the aggregating router for the aggregate reservation.

- IP4 SENDER_TEMPLATE object: Class = SENDER_TEMPLATE,
  C-Type = RSVP-AGGREGATE-IP4



o IP6 SENDER_TEMPLATE object: Class = SENDER_TEMPLATE, C-Type = RSVP-AGGREGATE-IP6



3.5. FILTER_SPEC Object

The FILTER_SPEC object identifies the aggregating router for the aggregate reservation, and is syntactically identical to the SENDER_TEMPLATE object.

4. Policies and Algorithms For Predictive Management Of Blocks Of Bandwidth

The exact policies used in determining how much bandwidth should be allocated to an aggregate reservation at any given time are beyond the scope of this document, and may be proprietary to the service provider in question. However, here we explore some of the issues and suggest approaches.

In short, the ideal condition is that the aggregate reservation always has enough resources to allocate to any E2E reservation that requires its support, and never takes too much. Simply stated, but more difficult to achieve. Factors that come into account include significant times in the diurnal cycle: one may find that a large number of people start placing calls at 8:00 AM, even though the hour from 7:00 to 8:00 is dead calm. They also include recent history: if more people have been placing calls recently than have been finishing them, a prediction of the necessary bandwidth a few moments hence may call for more bandwidth than is currently allocated. Likewise, at the end of a busy period, we may find that the trend calls for declining reservation amounts.

We recommend a policy something along this line. At any given time, one should expect that the amount of bandwidth required for the aggregate reservation is the larger of the following:

(a) a requirement known a priori, such as from history of the diurnal cycle at a particular week day and time of day, and

(b) the trend line over recent history, with 90 or 99% statistical confidence.

We further expect that changes to that aggregate reservation would be made no more often than every few minutes, and ideally perhaps on larger granularity such as fifteen minute intervals or hourly. The finer the granularity, the greater the level of signaling required, while the coarser the granularity, the greater the chance for error, and the need to recover from that error.

In general, we expect that the aggregate reservation will not ever add up to exactly the sum of the reservations it supports, but rather will be an integer multiple of some block reservation size, which exceeds that value.

5. Security Considerations

Numerous security issues pertain to this document; for example, the loss of an aggregate reservation to an aggressor causes many calls to operate unreserved, and the reservation of a great excess of bandwidth may result in a denial of service. However, these issues are not confined to this extension: RSVP itself has them. We believe that the security mechanisms in RSVP address these issues as well.

One security issue specific to RSVP aggregation involves the modification of the IP protocol number in RSVP Path messages that traverse an aggregation region. If that field were maliciously modified in a Path message, it would cause the message to be ignored by all subsequent devices on its path, preventing reservations from being made. It could even be possible to correct the value before it reached the receiver, making it difficult to detect the attack. In theory, it might also be possible for a node to modify the IP protocol number for non-RSVP messages as well, thus interfering with the operation of other protocols.

One way to mitigate the risks of malicious modification of the IP protocol number is to use an IPSEC authentication header, which would ensure that malicious modification of the IP header is detected. This is a desirable approach but imposes some administrative burden in the form of key management for authentication purposes.

It is RECOMMENDED that implementations of this specification only support modification of the IP protocol number for RSVP Path, PathTear, and ResvConf messages. That is, a general facility for modification of the IP protocol number SHOULD NOT be made available.

Network operators deploying routers with RSVP aggregation capability should be aware of the risks of inappropriate modification of the IP protocol number and should take appropriate steps (physical security, password protection, etc.) to reduce the risk that a router could be configured by an attacker to perform malicious modification of the protocol number.

6. IANA Considerations

Section 1.2 proposes a new protocol type, RSVP-E2E-IGNORE, which is used to identify a message that routers in the network core will see; further processing of such messages may or may not be required, depending on the egress interface type, as described in Section 1.2. The IANA assigned IP protocol number 134, in accordance with [RFC2780], meeting the Standards Track publication criterion.

Section 1.4.9 describes the manner in which the Router Alert is used in the context of this specification, which is essentially a simple counter of the depth of nesting of aggregation. The IPv4 Router Alert [RFC2113] has the option simply to ask the router to look at the protocol type of the intercepted datagram and decide what to do with it; the parameter is additional information to that decision. The IPv6 Router Alert [RFC2711] turns the parameter into an option sub-type. As a result, the IPv6 router alert option may not be used algorithmically in the context of the protocol in question. The IANA assigned a block of 32 values (3-35, "Aggregated Reservation Nesting Level") which we may map to nesting depths 0..31, hoping that 32 levels is enough.

Section 3.2 discusses a new, required path error code. The IANA has assigned RSVP Parameters Error Code 26 to NEW-AGGREGATE-NEEDED.

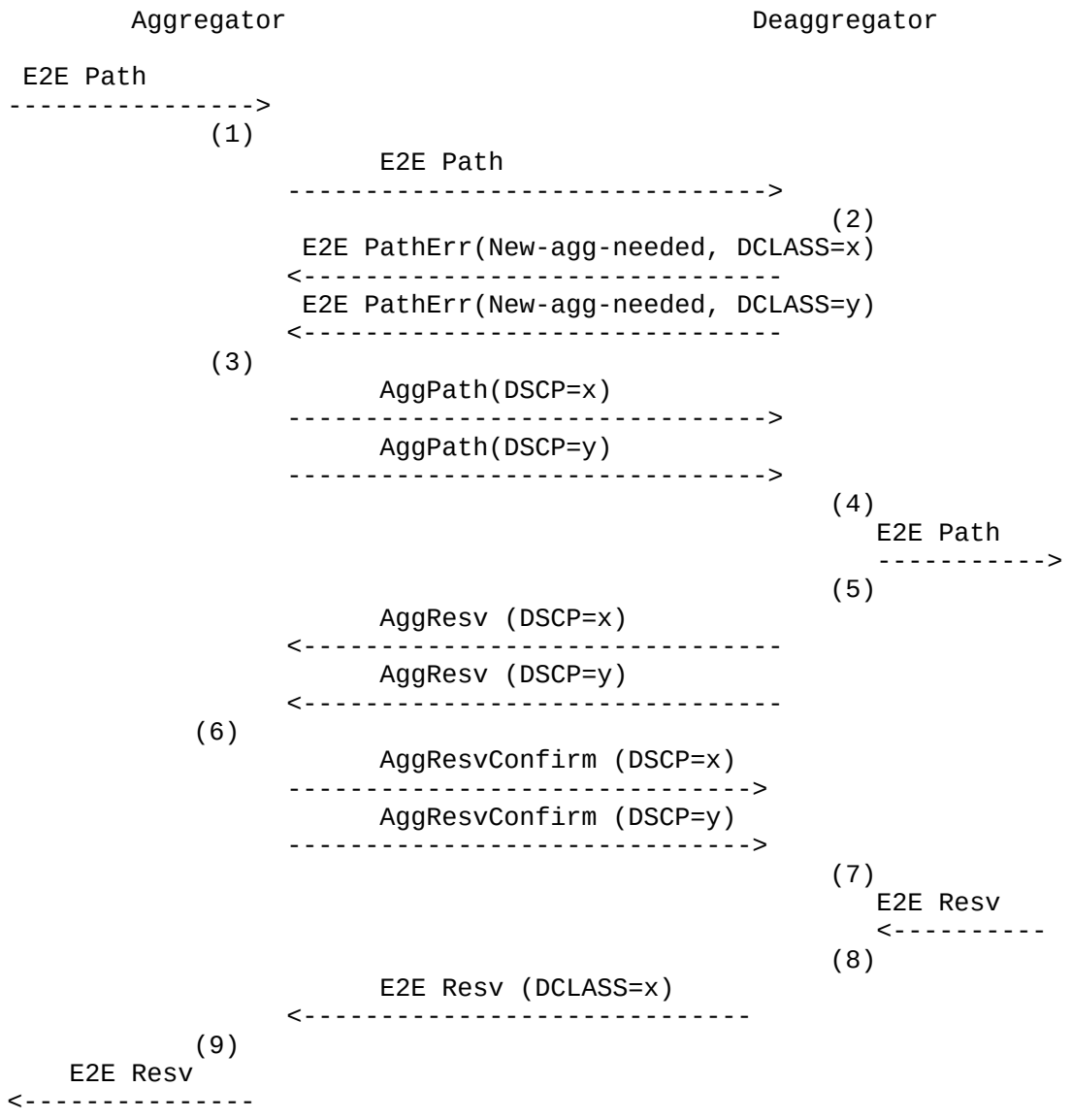
Sections 3.3, 3.4, and 3.5 describe extensions to three object classes: Session, Filter Specification, and Sender Template. The IANA has assigned two new common C-Types to be specified for the aggregator’s address. RSVP-AGGREGATE-IP4 is C-Type 9 and RSVP-AGGREGATE-IP6 is C-Type 10. In adding these C-types to IANA RSVP Class Names, Class Numbers and Class Types registry, the same numbering for them is used in all three Classes, as is done for IPv4 and IPv6 address tuples in [RSVP].

7. Acknowledgments

The authors acknowledge that published documents and discussion with several people, notably John Wroclawski, Steve Berson, and Andreas Terzis materially contributed to this document. The design is influenced by the RSVP tunnels document [TERZIS].

APPENDIX 1: Example Signalling Flow For First E2E Flow

This Appendix does not provide additional specification. It only illustrates the specification detailed above through a possible flow of RSVP signalling messages involved in the successful establishment of a unicast E2E reservation which is the first between a given pair of Aggregator/Deaggregator.



(1) Aggregator forwards E2E Path into aggregation region after modifying its IP Protocol Number to RSVP-E2E-IGNORE

(2) Let's assume no Aggregate Path exists. To be able to accurately update the ADSPEC of the E2E Path, the Deaggregator needs the ADSPEC of Aggregate PATH. In this example the Deaggregator elects to instruct the Aggregator to set up Aggregate Path states for the two supported DSCPs by sending a New-Agg-Needed PathErr code for each DSCP.

(3) The Aggregator follows the request from the Deaggregator and signals an Aggregate Path for both DSCPs.

(4) The Deaggregator takes into account the information contained in the ADSPEC from both Aggregate Path and updates the E2E Path ADSPEC accordingly. The Deaggregator also modifies the E2E Path IP Protocol Number to RSVP before forwarding it.

(5) In this example, the Deaggregator elects to immediately proceed with establishment of Aggregate Reservations for both DSCPs. In effect, the Deaggregator can be seen as anticipating the actual demand of E2E reservations so that resources are available on Aggregate Reservations when the E2E Resv requests arrive in order to speed up establishment of E2E reservations. Assume also that the Deaggregator includes the optional Resv Confirm Request in these Aggregate Resv.

(6) The Aggregator merely complies with the received ResvConfirm Request and returns the corresponding Aggregate ResvConfirm.

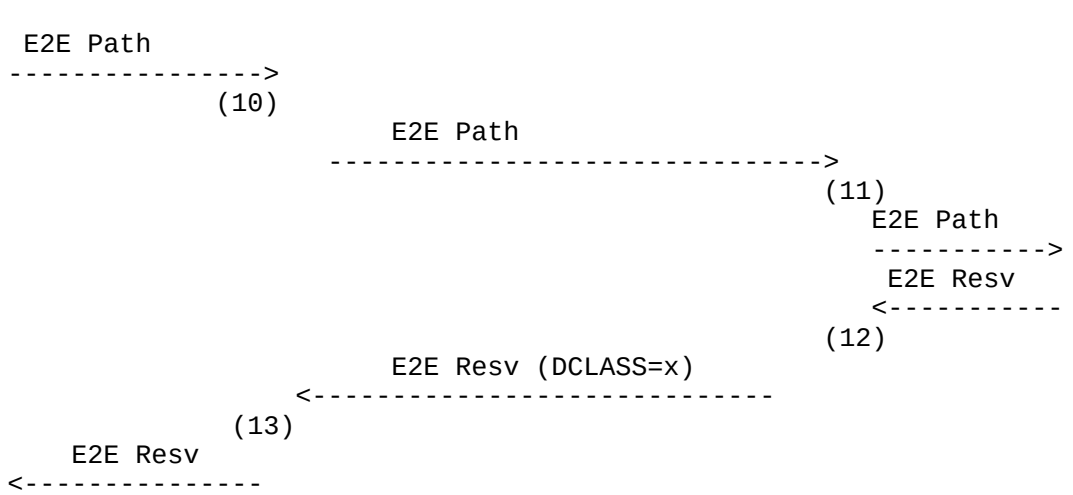
(7) The Deaggregator has explicit confirmation that both Aggregate Resv are established.

(8) On receipt of the E2E Resv, the Deaggregator applies the mapping policy defined by the network administrator to map the E2E Resv onto an Aggregate Reservation. Let's assume that this policy is such that the E2E reservation is to be mapped onto the Aggregate Reservation with DSCP=x. The Deaggregator knows that an Aggregate Reservation is in place for the corresponding DSCP since (7). The Deaggregator performs admission control of the E2E Resv onto the Aggregate Resv for DSCP=x. Assuming that the Aggregate Resv for DSCP=x had been established with sufficient bandwidth to support the E2E Resv, the Deaggregator adjusts its counter tracking the unused bandwidth on the Aggregate Reservation and forwards the E2E Resv to the Aggregator including a DCLASS object conveying the selected mapping onto DSCP=x.

(9) The Aggregator records the mapping of the E2E Resv onto DSCP=x. The Aggregator removes the DCLASS object and forwards the E2E Resv towards the sender.

APPENDIX 2: Example Signalling Flow For Subsequent E2E Flow Without Reservation Resizing

This Appendix does not provide additional specification. It only illustrates the specification detailed above through a possible flow of RSVP signalling messages involved in the successful establishment of a unicast E2E reservation which follows other E2E reservations between a given pair of Aggregator/Deaggregator. This flow could be imagined as following the flow of messages illustrated in Appendix 1.



(10) Aggregator forwards E2E Path into aggregation region after modifying its IP Protocol Number to RSVP-E2E-IGNORE

(11) Because previous E2E reservations have been established, let's assume that Aggregate Path exists for all supported DSCPs. The Deaggregator takes into account the information contained in the ADSPEC from the Aggregate Paths and updates the E2E Path ADSPEC accordingly. The Deaggregator also modifies the E2E Path IP Protocol Number to RSVP before forwarding it.

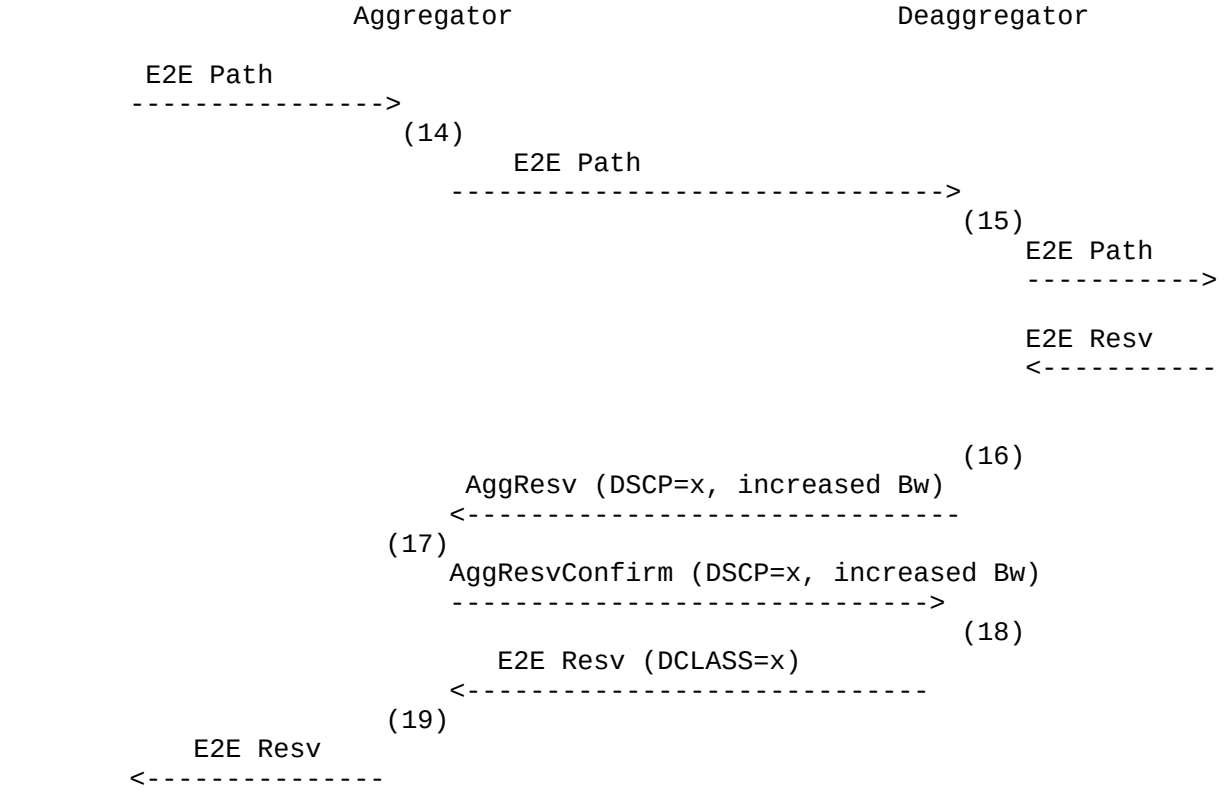
(12) On receipt of the E2E Resv, the Deaggregator applies the mapping policy defined by the network administrator to map the E2E Resv onto an Aggregate Reservation. Let's assume that this policy is such that the E2E reservation is to be mapped onto the Aggregate Reservation with DSCP=x. Because previous E2E reservations have

been established, let's assume that an Aggregate Reservation is in place for DSCP=x. The Deaggregator performs admission control of the E2E Resv onto the Aggregate Resv for DSCP=x. Assuming that the Aggregate Resv for DSCP=x has sufficient unused bandwidth to support the new E2E Resv, the Deaggregator then adjusts its counter tracking the unused bandwidth on the Aggregate Reservation and forwards the E2E Resv to the Aggregator including a DCLASS object conveying the selected mapping onto DSCP=x.

(13) The Aggregator records the mapping of the E2E Resv onto DSCP=x. The Aggregator removes the DCLASS object and forwards the E2E Resv towards the sender.

APPENDIX 3: Example Signalling Flow For Subsequent E2E Flow With Reservation Resizing

This Appendix does not provide additional specification. It only illustrates the specification detailed above through a possible flow of RSVP signalling messages involved in the successful establishment of a unicast E2E reservation which follows other E2E reservations between a given pair of Aggregator/Deaggregator. This flow could be imagined as following the flow of messages illustrated in Appendix 2.



(14) Aggregator forwards E2E Path into aggregation region after modifying its IP Protocol Number to RSVP-E2E-IGNORE

(15) Because previous E2E reservations have been established, let's assume that Aggregate Path exists for all supported DSCPs. The Deaggregator takes into account the information contained in the ADSPEC from the Aggregate Paths and updates the E2E Path ADSPEC accordingly. The Deaggregator also modifies the E2E Path IP Protocol Number to RSVP before forwarding it.

(16) On receipt of the E2E Resv, the Deaggregator applies the mapping policy defined by the network administrator to map the E2E Resv onto an Aggregate Reservation. Let's assume that this policy is such that the E2E reservation is to be mapped onto the Aggregate Reservation with DSCP=x. Because previous E2E reservations have been established, let's assume that an Aggregate Reservation is in place for DSCP=x. The Deaggregator performs admission control of the E2E Resv onto the Agg Resv for DSCP=x. Let's assume that the Aggregate Resv for DSCP=x does NOT have sufficient unused bandwidth to support the new E2E Resv. The

Deaggregator then attempts to increase the Aggregate Reservation bandwidth for DSCP=x by sending a new Aggregate Resv with an increased bandwidth sufficient to accommodate all the E2E reservations already mapped onto that Aggregate reservation plus the new E2E reservation plus possibly some additional spare bandwidth in anticipation of additional E2E reservations to come. Assume also that the Deaggregator includes the optional Resv Confirm Request in these Aggregate Resv.

(17) The Aggregator merely complies with the received ResvConfirm Request and returns the corresponding Aggregate ResvConfirm.

(18) The Deaggregator has explicit confirmation that the Aggregate Resv has been successfully increased. The Deaggregator performs again admission control of the E2E Resv onto the increased Aggregate Reservation for DSCP=x. Assuming that the increased Aggregate Reservation for DSCP=x now has sufficient unused bandwidth and resources to support the new E2E Resv, the Deaggregator then adjusts its counter tracking the unused bandwidth on the Aggregate Reservation and forwards the E2E Resv to the Aggregator including a DCLASS object conveying the selected mapping onto DSCP=x.

(19) The Aggregator records the mapping of the E2E Resv onto DSCP=x. The Aggregator removes the DCLASS object and forwards the E2E Resv towards the sender.

References

[CSZ] Clark, D., S. Shenker, and L. Zhang, "Supporting Real-Time Applications in an Integrated Services Packet Network: Architecture and Mechanism," in Proc. SIGCOMM'92, September 1992.

[IP] Postel, J., "Internet Protocol", STD 5, RFC 791, September 1981.

[HOSTREQ] Braden, R., "Requirements for Internet hosts - communication layers", STD 3, RFC 1122, October 1989.

[DSFIELD] Nichols, K., Blake, S., Baker, F. and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

[PRINCIPLES] Carpenter, B., "Architectural Principles of the Internet", RFC 1958, June 1996.

[ASSURED] Heinanen, J, Baker, F., Weiss, W. and J. Wroclawski, "Assured Forwarding PHB Group", RFC 2597, June 1999.

[BROKER] Jacobson, V., Nichols K. and L. Zhang, "A Two-bit Differentiated Services Architecture for the Internet", RFC 2638, June 1999.

[BRIM] Brim, S., Carpenter, B. and F. LeFaucheur, "Per Hop Behavior Identification Codes", RFC 2836, May 2000.

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

[TERZIS] Terzis, A., Krawczyk, J., Wroclawski, J. and L. Zhang, "RSVP Operation Over IP Tunnels", RFC 2746, January 2000.

[DCLASS] Bernet, Y., "Format of the RSVP DCLASS Object", RFC 2996, November 2000.

[INTEGRITY] Baker, F., Lindell, B. and M. Talwar, "RSVP Cryptographic Authentication", RFC 2747, January 2000.

[RFC2998] Bernet Y., Ford, P., Yavatkar, R., Baker, F., Zhang, L., Speer, M., Braden, R., Davie, B., Wroclawski, J. and E. Felstaine, "Integrated Services Operation Over Diffserv Networks", RFC 2998, November 2000.

[RFC2961] Berger, L., Gan, D., Swallow, G., Pan, P. and F. Tommasi, "RSVP Refresh Reduction Extensions", RFC 2961, April 2001.

[RFC2780] Bradner, S. and V. Paxson, "IANA Allocation Guidelines For Values In the Internet Protocol and Related Headers", RFC 2780, March 2000.

[RFC2711] Partridge, C. and A. Jackson, "IPv6 Router Alert Option", RFC 2711, October 1999.

[RFC2113] Katz, D. "IP Router Alert Option", RFC 2113, February 1997.

Authors' Addresses

Fred Baker
Cisco Systems
1121 Via Del Rey
Santa Barbara, CA, 93117 USA

Phone: (408) 526-4257
EMail: fred@cisco.com

Carol Iturralde
Cisco Systems
250 Apollo Drive
Chelmsford MA, 01824 USA

Phone: 978-244-8532
EMail: cei@cisco.com

Francois Le Faucheur
Cisco Systems
Domaine Green Side
400, Avenue de Roumanille
06410 Biot - Sophia Antipolis
France

Phone: +33.4.97.23.26.19
EMail: flefauch@cisco.com

Bruce Davie
Cisco Systems
250 Apollo Drive
Chelmsford MA,01824 USA

Phone: 978-244-8921
EMail: bdavie@cisco.com

Full Copyright Statement

Acknowledgement

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