Internet-Draft
Expires: September 1997

Arun Viswanathan
Vijay Srinivasan
IBM Corp.

March 1997

Soft State Switching
A Proposal to Extend RSVP for Switching RSVP Flows

<draft-viswanathan-aris-rsvp-00.txt>

Status of This Memo

Abstract

This memo describes a mechanism for establishing a switched path with guaranteed Quality of Service for RSVP [1] flows in an integrated switch-router environment. It proposes an extension to the RSVP protocol that allows establishment of a sequence of virtual connections along the hop-by-hop routed path by enabling adjacent nodes to exchange Layer-2 labels. The labels correspond to information that identifies the virtual connections; for example, the VPI/VCI value in the case of an ATM-based Layer-2 infrastructure.

1. Introduction

An Integrated Switch-Router (ISR) is a switching node that has an IP Control Point (IP-CP) and implements a IP switching technology [2-4]. The ISRs form adjacencies through a well-known virtual circuit (VC), also called the default VC, that terminates at the adjacent ISR's IP-CP. This hop-by-hop VC connectivity gives a cloud of ISRs the same nature as any ubiquitous IP internet. The objective is to switch RSVP flows in such an environment.

This document proposes an extension to RSVP that introduces new objects to the existing RSVP messages. Using these objects, each downstream ISR provides its neighboring upstream ISR with the Layer-2 (L2) label on which it wishes to receive a RSVP flow. In an ATM-based ISR environment, this label would correspond to a VPI/VCI value for the ATM virtual circuit on which the ISR wishes to receive traffic from the RSVP flow. Then, using an approach similar to those outlined in [2], [3], and [4], the L2 labels are spliced hop-by-hop to form an end-to-end VC. The data from the RSVP flow then uses this this end-to-end VC, and the RSVP signalling messages are forwarded through the default VC. By moving RSVP flows from the default VCs to a dedicated end-to-end VC, it is possible to leverage the QoS capabilities of the underlying L2 technology to provide the type of service desired for the RSVP flow.

In this memo the term virtual circuit (VC) is used loosely to imply a switched data path under any switching technology that has the ability to isolate flows from each other, e.g. ATM, Frame Relay etc. The memo proposes a "one VC per sender" approach. Merging of RSVP flows into a single VC will be considered in a later revision of the draft.

It is assumed here that the ISRs on the edge of an ISR cloud can either auto-learn or are configured to indicate that they are edge ISRs (on a per interface basis).

2. Soft State Switching

In soft state switching, the goal is to switch traffic from an RSVP flow at L2 instead of having to forward them hop-by-hop as in conventional IP routers. By doing so, it is possible to leverage the high-performance switching and Quality of Service capabilities of the L2 technology. This is achieved when all neighboring ISRs along the routed path could exchange L2 labels for establishing the switched path for RSVP flows. Then, the L2 labels may be "spliced" hop-by-hop to setup an end-to-end (ingress-to-egress) VC along the preferred routed path. By splicing, we refer to the process by which an

incoming VC is associated with an outgoing VC at L2, without traffic from the incoming VC being processed at the network layer. For example, this can be achieved in ATM switches by establishing this association in the ATM switching tables. Once the splicing is complete, the default VC carrying best effort traffic between adjacent routers provides the IP forwarding path. The RSVP signalling messages are forwarded on the default VC.

The L2 labels are assumed to have only unidirectional significance. In other words, there exists a separate L2 label space for each direction of flow on a link. Moreover, the downstream ISR is chosen to be the L2 label space owner (allocator) on a link. The single owner approach keeps the L2 label usage simple and manageable. If a L2 label space had more than one owner, it would require that owners synchronize their use of the L2 labels or the space would have to be partitioned amongst the owners. For flexibility, the proposed extension to RSVP also supports the concept of "upstream on demand" allocation described in [3]. In this method, the upstream ISR allocates labels when demanded by the downstream ISR. This enables co-existence with other protocols that consume L2 labels.

3. Motivation

In this section, we discuss why the RSVP protocol is ideal for establishing a switched path for RSVP flows.

One motivating factor for using RSVP is that mapping the network layer QoS request to a L2 virtual connection is simple. The RESV message carries the QoS requested by the receiver(s) of the RSVP flow. For example, this could correspond to one of the Integrated Service classes described in [6-8]. This QoS information is needed when L2 labels are set up and spliced; i.e., when the resource reservations are made. Otherwise, the VC establishment protocol would have to carry its own QoS entity and/or map the VC setup to RSVP tables at each ISR hop.

Another motivating reason for extending RSVP is multicast support. RSVP is designed to scale well for multicast sessions requiring resource reservation. RSVP also allows receivers to join existing sessions with different QoS requirements. An independent VC establishment protocol should be able to handle such session "joins" equally well.

With the RSVP protocol the receivers can make sender selection through the provision of different filter styles. In this, multiple sender flows (as chosen by the receivers) in a RSVP session can be associated with a single reservation. In other words, sender flows

in a RSVP session can be merged into a single downstream reservation. A new VC establishment protocol would have to support a similar mechanism for seamless interoperability with the RSVP protocol.

Finally, any mechanism for set up the VCs would, in any case, require extensive interfacing with RSVP protocol and/or its state tables.

Due to these reasons, it is best if RSVP can be extended without changing its existing mechanics, to provide support for setting up the switched path for RSVP flows. This need not be viewed as "piggy-backing" another protocol on RSVP, but rather, a natural extension to RSVP to provide QoS in an integrated switch-router environment.

4. L2 Label Exchange Mechanism

The proposed extension to RSVP calls for adding a new object to carry the L2 label information in the RESV message. The egress ISR, say ISR A, (i.e. the "last" node in the ISR environment, or the ISR through which the RSVP flow exits the ISR environment) places this object in the RESV message and sends it to the PHOP ISR for the flow (as stored in the Path state for this flow) -- call this ISR B. The RESV message is sent to ISR B via the default routed path. ISR B will use the L2 label in the RESV message to setup a VC to ISR A (in this case, the egress ISR) on the outgoing interface. The QoS for this VC corresponds to a mapping of the Integrated Service class specified in the RESV message to an appropriate set of QoS values for the L2 technology.

ISR B then chooses a new L2 label on the incoming interface through which the RSVP flow enters the ISR, and sends this label to its own PHOP, ISR C, using the new object in the RESV message. When a VC is set up from ISR C to ISR B using this label, the L2 label on the incoming interface of ISR B is then mapped to the L2 label of the outgoing interface in ISR B's label swap table for L2 switching. This completes the splicing process at ISR B.

Repeating this process at each PHOP ISR, the RESV eventually reaches the ingress ISR (the ISR through which the RSVP flow enters the ISR environment). The ingress ISR will make necessary entries to forward packets for this flow through the VC identified by the L2 label in RESV message. All ingress ISRs will delete the L2 object before forwarding the RESV message to their PHOPs. The L2 labels used for an RSVP session are released whenever the RSVP session is torn down or is timed-out.

Using this process, an end-to-end switched path is established for an

RSVP flow through an ISR network. The data packets from the RSVP flow are forwarded via this switched path, while RSVP control messages continue to use the default VCs between ISRs.

The procedure described in this section does not describe how flow merging is performed in such an environment. Flow merging is a key feature of RSVP, and the ability to perform merging in an ISR environment is dependent on the capabilities of the ISRs. This topic is addressed in detail in the following section.

5. Merging

There are several switching technologies available today (ATM, Frame Relay etc.) and perhaps more in the future. Moreover, the capabilities of a switch of a certain technology vary from vendor to vendor. Three basic characteristics are identified that determine how the underlying L2 technology can be used in conjunction with this proposal to address merging of flows under appropriate environment. They are:

- Attribute A: Can correctly merge several upstream VCs into a single downstream VC. Frame switches are typically able to do this in a straightforward manner. However, for ATM switches without appropriate functionality built in, cells from different AAL SDUs may become interleaved on the outgoing VC, thus corrupting the higher layer information.

- Attribute B: Can treat a set of VCs as a single entity for QoS purposes. A switch with this property is able to treat all traffic from a set of VCs in a like manner for purposes of scheduling, fair queueing etc. For example, an ATM switch that performs per-class queueing would assign all the VCs from a given set to a particular class. Then, cells from all the VCs in the sets would receive the QoS corresponding to that class.

- Attribute C: Can demultiplex senders flows in a single VC into a separate VC for a sender. For example, using label stack for L2 tunneling ([3], [4]).

The current version of the document does not address the logical merging of sender flows in a RSVP session. The above attributes may be used to determine how RSVP flows are merged into a single VC.

6. Multicast Support

In order to support multicast sessions, at split points within the ISR network, where data from upstream ISRs splits into multiple downstream flows, the ISR can perform the required duplication (at L2 level) of flows through the hardware multicast capability (for example, point-to-multipoint VC) of the switch, if available. Otherwise, the flow has to be processed at the network layer and multicast in the normal manner. Note that the network layer forwarding is interoperable with all switch types.

7. Unreserved Receivers

When none of the receivers have reserved, the multicast session may flow through the default VC as best-effort traffic. But as soon as a receiver makes a reservation, the data flow may stop to receivers that haven't made any reservation yet. The receivers without reservation only get PATH messages but no data (even through best-effort). This problem can be addressed in several different ways determined by the switch architecture.

This problem does not exist for switches that support Attribute A. They can add the default forwarding VC as a branch in the point-to-multipoint VC. If the switch architecture allow adding the local IP-CP to the point-to-multipoint VC, then the IP-CP can multicast the packets only to those interface from which there is no reservation but are listed in the multicast table. This would be the most preferable approach for all switches.

Another way to alleviate this problem is to use the PATH message to do the VC establishment from the node downstream on which there are interfaces through which no reservation has been received [9]. This reservation uses a UBR-like QoS. This is done when there is at least one reservation in place for the RSVP session. This may not work in environments where upstream label allocation is not permitted.

8. TTL Decrement

When IP packets flow through a switched path, the TTL value in the IP header cannot be decremented. The decrementing of TTL value is used to delete packets in a routing loop to avoid/reduce congestion. For this purpose, the proposed PATH L2 Object carries a hop-count that counts the number of consecutive ISR hops. The ISRs increment the hop-count only if there is a switched path for that sender flow through that ISR. All ISRs maintain the hop count in the Path State. Only the egress ISR on which the VC terminates would use the count to

decrement the TTL on packets for that sender flow. The ISRs of a switching technology that have a TTL equivalent in the L2 header may not use the PATH L2 Object.

9. Upstream on Demand Label Allocation

The memo describes the RSVP extension when the downstream ISR is the label space owner. But a upstream label allocator can be supported. In this, the RESV message uses a NULL L2 Object to indicate a request for label allocation. The upstream ISR responds with a L2 label for the RSVP session in the PATH messages to the downstream neighbor. This flexibility allows co-existence with other IP switching protocols. This can also be useful in other environments such as [5].

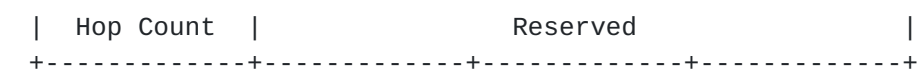
10. Adjacency

The ISR neighbors need some mechanism to establish adjacencies. This is required because the neighbors need to exchange the label range for correct label allocation. They also need to elect the label allocator. The current version of this memo does not propose any extension to RSVP protocol for this mechanism. It is assumed that adjacency would be established by another protocol (as proposed in [2], [3] or [4]) and such information would be made available to the RSVP module. In absence of such mechanism the ISRs would have to be configured with the required information to operate as described in this memo.

11. Object Formats

This section describes the object formats for the proposed extension. The L2 object for different scenarios are defined below:

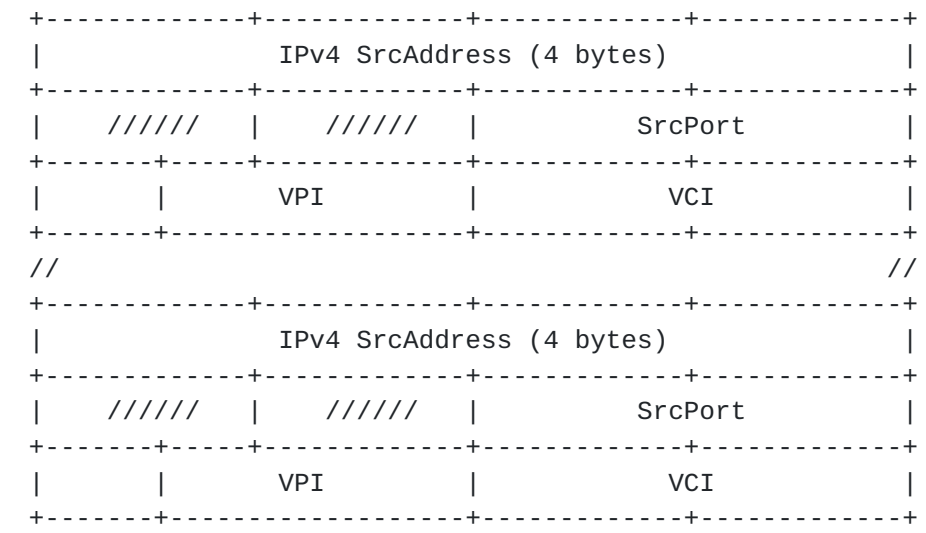
o L2 HOP COUNT object: Class = x, C-Type = 1



Hop Count
Counts the length (in ISR hops) of the switched path.

o NULL L2 Object: Class = y, C-Type = 1

o ATM RESV L2 object: Class = y, C-Type = 2

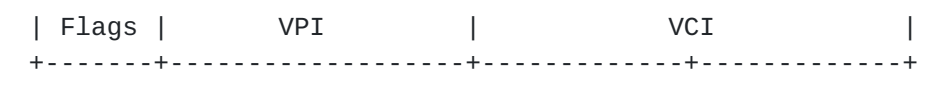


IPv4 SrcAddress
IPv4 address of the sender.

VPI - 12 bits
Virtual Path Identifier. If lesser than 12 bits then its right justified in this field.

VCI - 16 bits
Virtual Circuit Identifier. If lesser than 16 bits then its right justified in this field.

o ATM PATH L2 object: Class = y, C-Type = 3



Flags - 4 bits
0x01 - Implies that the PATH message is in response to an upstream on demand label allocation and may not be propogated any further.

VPI - 12 bits
Virtual Path Identifier. If lesser than 12 bits then its right justified in this field.

VCI - 16 bits
Virtual Circuit Identifier. If lesser than 16 bits then its right justified in this field.

The IPv6 extension and error codes will be defined in a later revision of the draft.

The reader may have noticed that the new RESV L2 object has duplicated information already present in the FILTER_SPEC object. Another approach could be to extend the FILTER_SPEC object definition to carry the link layer labels or insert the label object following the FILTER_SPEC object.

12. Security Considerations

Security considerations are not discussed in this document.

13. Acknowledgements

The authors wishes to acknowledge Nancy Feldman for her input.

14. References

[1] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification. Internet Draft, draft-ietf-rsvp-spec-14, November 1996.

[2] P. Newman, W. L. Edwards, R. Hinden, E. Hoffman, F. Ching Liaw, T. Lyon, G. Minshall, Ipsilon Flow Management Protocol Specification for IPv4, Version 1.0. Internet RFC 1953, May 1996.

[3] Y. Rekhter, B. Davie, D. Katz, E. Rosen, G. Swallow, D. Farinacci, Tag Switching Architecture - Overview. Internet Draft, draft-rekhter-tagswitch-arch-00.txt, January 1997.

[4] A. Viswanathan, N. Feldman, R. Boivie, R. Woundy, ARIS: Aggregated Route-Based IP Switching. Internet Draft, draft-viswanathan-aris-overview-00.txt, November 1996.

[5] D. Farinacci, Partitioning Tag Space among Multicast Routers on a Common Subnet. Internet Draft, draft-farinacci-multicast-tag-part-00.txt, December 1996.

[6] S. Shenker, C. Partridge, R. Guerin, Specification of Guaranteed Quality of Service. Internet Draft, draft-ietf-intserv-guaranteed-svc-06.txt, August 1996.

[7] J. Wroclawski, Specification of the Controlled-Load Network Element Service. Internet Draft, draft-ietf-intserv-ctrl-load-svc-03.txt, August 1996.

[8] F. Baker, R. Guerin, D. Kandlur, Specification of Committed Rate Quality of Service. Internet Draft, draft-ietf-intserv-commit-rate-svc-00.txt, June 1996.

Author's Address

Vijay Srinivasan
IBM Corporation
PO Box 12195
Research Triangle Park, NC 27709

Phone: +1 (919) 254-2730
Email: vijay@raleigh.ibm.com

Arun Viswanathan
IBM Corporation
17 Skyline Drive
Hawthorne, NY 10532

Phone: +1 (914) 784-3273
Email: arunv@vnet.ibm.com