Internet Engineering Task Force
INTERNET DRAFT
22 February 1996

A. Birman
R. Guerin
D. Kandlur
IBM

Support for RSVP-based Service over an ATM Network
draft-birman-ipatm-rsvpatm-00.txt

Status of This Memo

Abstract

In this document we focus on RSVP-based resource reservations in a heterogeneous environment which includes ATM networks. We describe a method for establishing 'shortcuts' through an ATM network which avoids the performance penalty associated with level 3 processing in a 'classical RSVP over ATM' approach. For the case of guaranteed service we show how to map the RSVP flow characteristics to ATM call parameters, and thus enable end-to-end performance guarantees. We also discuss the extensions to RSVP and to ATM signaling required for the implementation of these solutions.

1. Introduction

We consider a heterogeneous environment in which legacy networks coexist with ATM networks. For applications that require performance guarantees the reservation of network resources is carried out using RSVP as the reservation setup protocol. The operation of RSVP over an ATM network is the focus of our paper.

Our starting point is the classical IP over ATM model [Lau94] in which an ATMARP server is used for address resolution within a LIS, while the inter-LIS traffic is routed through IP routers. For an application with QoS requirements the classical IP over ATM architecture does allow for QoS support over the VCCs between the routers. This solution is not altogether satisfactory, however, since it may not provide the QoS which would be possible if a direct VCC through the ATM network was established along the path through the ATM network. Such a direct connection, first proposed in [Mil95] and referred to as a 'shortcut', eliminates the level 3 processing at the routers and thus allows better performance. We describe below a method to establish such shortcuts over ATM networks, first for unicast and then multicast application flows (for additional details see [BFG+95]). We refer to this approach as 'classical RSVP over ATM with shortcuts'.

The approach described here has, without doubt, shortcomings. Some of these can be traced to the differences between RSVP and ATM signaling ([For94],[For95]), and their opposing design principles. This approach is offered as a possible first step in supporting QoS flows in a heterogeneous environment with ATM networks. If adopted and carried through to implementation, the experience thus gathered may be beneficial in the design of a better next scheme.

2. Reservation setup for unicast flows

We first consider unicast flows. The parameters necessary for setting up VCCs with QoS guarantees are obtained from RSVP messages Path and Resv.

2.1. Classical RSVP over ATM

Figure 1 shows an ATM network consisting of four LISs. A is the ingress router to the ATM network, B is the egress router. RSVP messages follow the IP route AEFGB. Thus, a Path message will travel downstream from A to B, while the corresponding Resv message will travel upstream from B to A. When the Resv message arrives at G the router has sufficient information to set up a VC from G to B.

Similarly, VCs will be set up from F to G, from E to F, and from A to E.

In particular, if the ATM network consists of a single LIS then the route from A to B has only one hop, although there could be multiple hops at the ATM level.

For the multi-hop case, while RSVP messages travel over best-effort VCs, data packets flow over QoS VCs and enjoy QoS support in the routers. Traversing the routers, however, entails IP-level processing and thus is less desirable than a shortcut VC from A to B. In the rest of this section we discuss several schemes for RSVP support using ATM shortcuts.

2.2. ATM shortcuts

In this scheme, we modify the RSVP operation in order to identify the appropriate egress router for the purpose of establishing a shortcut route through the ATM network. When the first Path message for a session arrives at A (Figure 2), the node determines that the message will be forwarded over an ATM link and thus node A is the ingress node into the ATM network. The Path message is routed along the default IP route, and is modified to carry both the ATM address and the IP address of A (the IP address of A is the `previous hop' or PHOP). At each node along the route an ATM connectivity check is performed to determine whether the current node is the egress point from the ATM network. This decision would be based on the ATM connectivity between the current router, the upstream router, and the downstream router as determined by the logical ATM network in which

```
        +-+       +-+       +-+       +-+       +-+
    --->|A|------>|E|------>|F|------>|G|------>|B|--->
        +-+       +-+       +-+       +-+       +-+

            LIS 1     LIS 2     LIS 3     LIS 4
          <-------> <-------> <-------> <------->

                        ATM network
         <---------------------------------------->
```

Figure 1: Reservation setup using ``classical'' RSVP support

they reside (the concept of the logical ATM network as described in [KP95]). If the current router is not an egress router, it forwards the Path message to the downstream router without updating the PHOP address field. If the current router is an egress router (e.g. B) it processes the Path message by creating a Path state for the session and by storing, along with other data, the IP address and the ATM address of A.

A Resv message is considered new when it represents either a reservation requests for a new flow or a request to modify the reservation of an existing flow. When such a Resv message arrives at B, B inserts its own ATM address as an object into this message, and forwards the message along the default routed path to A. Intermediate routers recognize the Resv message but do not create any session or reservation and simply forward the message upstream. When this Resv message arrives at A it carries in addition to the regular RSVP information, both the ATM address of the egress router B and QoS information necessary to determine the type of ATM VC that needs to be setup.

Since intermediate nodes do not need to process the Resv message, an alternative here is to encapsulate the Resv message into an IP datagram that is then tunneled from B to A. Tunneling provides the advantage that packet processing is expedited (along the fast-path through the router) since there is no special processing at intermediate nodes. On the other hand, the packet is not treated as a signaling packet and is susceptible to normal loss at intermediate nodes.

After the shortcut VC from A to B is set up, B needs to be able to associate the newly created VC with the RSVP flow. In order to achieve this, the flow identifier consisting of the tuple

(sourceIPaddress;destinationIPaddress; transportlayer)

is carried in the SETUP message in the Broadband High Layer Information (B-HLI) element (the length of this field would have to be extended from its current size of 8 bytes). The source and destination IP addresses cannot be inferred from the ATM addresses in the router--router case. The source and destination addresses themselves further consist of pairs of the form

(IPaddress;portnumber):

Note also that the receipt of the SETUP message provides an implicit acknowledgment that the Resv message was received at router A. This means that router A also has received all the information necessary to forward Resv messages upstream, i.e. the RSVP filter and service specifications that are not directly available from the ATM connection characteristics. As a result, the egress router B now suppresses the transmission of Resv refreshes towards router A, unless they carry a modified service specification.

Figure 2 shows a shortcut VC from A to B which bypasses nodes E, F and G. The shortcut VC is used for the RSVP data traffic, but Path messages continue to flow along the default routed path. It is noted that this scheme for creating shortcut routes is independent of the underlying routing mechanism and is oblivious to any IP routing domain boundaries. Moreover, RSVP state is required only in the edge routers A and B.

A possible variation to the method described above handles Path messages the same way, but differs from it in that it shifts the responsibility of establishing the ATM shortcut VC from the ingress router A to the egress router B (see Figure 2). This is possible because ATM unicast calls are always duplex, and resources can be reserved in both directions.

Specifically, when a Resv message arrives at the egress router B, B can generate a SETUP message towards A and specify the resources required in both directions. The SETUP message will specify QoS requirements in the direction A to B to accommodate the service specifications carried in the Resv message. Conversely, it will not request any QoS or bandwidth guarantees from B to A since there is no data flow in this direction. While the VC setup is now handled by the egress router, it is still necessary to forward the Resv message

```
      +-+       +-+       +-+       +-+       +-+
  --->|A|------>|E|------>|F|------>|G|------>|B|--->
      +-+       +-+       +-+       +-+       +-+
        :                                     :
        :                VC                   :
        .......................................


                      ATM network
       <-------------------------------------->
```

Figure 2: Reservation setup using ATM shortcuts

to the ingress router, so that it can propagate that information upstream (it cannot be accurately inferred for the traffic and QoS parameters carried in the SETUP message). In order to do that, Resv messages including refreshes for reliability purposes, will keep on being forwarded onto the IP route. However, as with the previous method, they are not acted upon at intermediate routers. Another alternative is to include the Resv message as higher layer information in the SETUP message.

The main advantage of this scheme is that it is consistent with the preferred solution for multicast flows when the LIJ capability of UNI 4.0 becomes available (see Section 3.2 for details).

3. Reservation setup for multicast flows

We consider two methods for establishing shortcuts through an ATM network. The ``root-initiated ATM shortcut'' is better suited to the present UNI 3.1 environment, while the ``leaf-initiated ATM shortcut'' would be preferred when the leaf-initiated join capability of UNI 4.0 becomes available.

3.1. Root-initiated ATM shortcuts

We start by extending the unicast scheme of Section 2.2 to single-sender multicast flows, as illustrated in Figure 3. As mentioned before, this is the approach best suited to a UNI 3.1 environment. The determination of the ATM shortcut follows the same steps as in Section 2.2. When a Path message for a session arrives at node A, the node determines that the message will be forwarded over an ATM link and thus node A is the ingress node into the ATM network (note that this step only needs to be performed upon receipt of the first Path message). The ATM address of A is inserted as an object into the Path message, which is fowarded onto the default IP route. In addition, a mechanism such as MARS [Arm95] is used for local multicast delivery on this path.

At each node along the route an ATM connectivity check is again performed to determine whether the current node is an egress point from the logical ATM network. If the current node, such as F in Figure 3, is not an egress point then the Path message is forwarded to the downstream nodes without updating the PHOP (previous hop) address field.

When the first Resv message arrives at an egress point, say B, B forwards the message along the reverse path to A. The ATM address of B is carried as an object in the Resv message. Intermediate

```
        +-+       +-+       +-+            +-+
    --->|A|------>|E|------>|F|----------->|B|--->
        +-+       +-+       +-+            +-+
                   |
                   |                       +-+
                   +---------------------->|C|--->
                                           +-+


                     ATM network
         <----------------------------------->
```

Figure 3: Reservation setup with maximum shortcut

routers, F and E in this case, simply forward the message upstream towards A. Specifically, they do not merge Resv messages and do not perform any reservation. As in the unicast case, an alternative is to tunnel the Resv message directly to A by encapsulating it into an IP message. When the first Resv message arrives at A, say from B, A has all the information necessary to create a shortcut point-to-multipoint VC with root A and leaf B. In order for B to associate the newly created VC with the RSVP flow, the flow identifier consisting of the pair

(sourceIPaddress; destinationIPaddress)

is carried in the SETUP message in the Broadband High Layer Information (B-HLI) element. Later, when the Resv message from C arrives at A, A adds C to the point-to-multipoint VC with an ADD PARTY signaling message. The ADD PARTY message will also carry the flow identifier in the B-HLI element.

In order to track route changes and changes in group membership, Path refresh messages keep flowing normally over the IP route. However, Resv refreshes from each router are suppressed as soon as the egress router receives the ATM setup message (ADD PARTY or SETUP for the first leaf). This is because the setup message indicates that the initial Resv message has been received by the ingress router, and that the reservation through the ATM network has been successfully performed. This suppression prevents the steady state implosion of refresh Resv messages at the ingress router.

However, the ingress router is still required to perform as many ATM connection SETUPs as there are leaves in the ATM network for the multicast address. This is because, the scheme always results in the use of a ``maximum'' ATM shortcut between the ingress and egress routers. The use of a maximum shortcut minimizes IP-level processing at intermediate nodes and thus shortens end-to-end packet delays, but the (signaling) load imposed on the ingress router may become a problem when dealing with large multicast groups.

Milliken [MillikenJul9501] proposed a scheme which is intended to alleviate the problem by distributing the (signaling) processing load among the routers. This load distribution is achieved by allowing some flexibility at each router on deciding whether or not to extend an ATM shortcut. A more promising and systematic approach to eliminate the possibility of signaling overload at the ingress router, it to use the Leaf-Initiated Join (LIJ) capability of of UNI 4.0. We discuss such a solution in the next section.

3.2. Leaf-initiated ATM shortcuts

Consider the ATM network in Figure 3 and assume the flow of Path messages is as described in the previous section. That is, Path messages continue to use the default IP routed path, and a mechanism such as MARS is used for local multicast delivery on this path. As before, the Path message processing at intermediate routers is changed, in that the PHOP is not modified, while the Path message carries the ATM address of A. In addition, A also chooses a ``global connection identifier'' (GCID) and inserts it into the Path message. This global connection identifier consists of a call identifier uniquely chosen by the root, which is paired with the root's ATM address for LIJ setup. For a given RSVP session, there may be multiple flows transiting through A and, for each flow, A would choose a distinct global connection identifier. This connection identifier will be used by egress routers when generating an ATM LIJ request to join the point-to-multipoint connection associated with the IP multicast address.

When the first Resv message reaches an egress router, say B, the router has all the information needed for generating an LIJ request to the GCID it received. The ATM point-to-multipoint connection is then created at this time, with the ingress router A as its root and B as the first leaf. As other egress routers, such as C in Figure 3, also receive their first Resv message, they signal their intention to join the connection in exactly the same manner, i.e. through a LIJ request to the specified GCID. They are then added as new leaves to the existing point-to-multipoint connection, but the ingress router A is not notified of this new join. This eliminates the potential processing overload at router A since it is only required to handle a single signaling request, i.e. when the first leaf joins.

However, note that as a result of not notifying the ingress router of new leaves joining, the information carried in the Resv messages arriving at the associated egress routers is not forwarded to the ingress router during the ATM setup process. This information is, however, necessary for the ingress router to further propagate Resv messages upstream, i.e. it needs information elements such as the RSVP service and filter specifications, which as mentioned before cannot always be directly inferred from the ATM traffic and QoS parameters. In order to achieve this, Resv messages, including refreshes, will continue to be propagated and merged on the IP path, but no reservation will be triggered at intermediate routers. The merging on the IP path ensures that the ingress router is not overwhelmed by the volume of refresh Resv messages it receives, while providing it with all the information it needs to forward Resv messages to its upstream neighbor. Note that even refreshes are sent in order to ensure reliable delivery of Resv messages to the ingress router.

4. Issues Related to Flow/Call Characteristics

The previous sections have dealt with many of the issues related to the mapping between RSVP and ATM control flows. In this section, we focus on similar problems but at the level of the data flows. Specifically, we consider issues related to the mapping of traffic parameters and QoS guarantees as well as connection types.

4.1. Flow mapping

RSVP defines a session as the set of data flows with the same (unicast or multicast) destination. As a result, at an endpoint of a flow (sender or receiver) the data flow is uniquely identified by the pair

(sourceIPaddress;destinationIPaddress):

The source and destination addresses themselves further consist of pairs of the form

(IPaddress;portnumber);

where the destination IP address can be a multicast IP address.

The ATM UNI identifies a connection through the Connection Identifier used in the SETUP, CONNECT, etc. messages. Connection Identifier is associated to an ATM flow from one sender to one or more receivers and is unique at the sender. A call can be uniquely identified in the ATM network by the pair (Connection Identifier, sender ATM address). Similarly, in ATM UNI 4.0 which introduces the Leaf Initiated Join (LIJ) capability, each LIJ capable call is uniquely identified by a Global Call Identifier (GCID). The GCID is formed by pairing the LIJ call identifier selected by the the ``ROOT'' of the call (point-to-multipoint connection) and the address of the ROOT itself. Network wide uniqueness is, therefore, ensured.

From the above discussion, we see that a node at the boundary between IP and ATM networks can map the quadruplet (source IP address, source port number, destination IP address, destination port number) that uniquely identifies an RSVP flow, to an ATM GCID consisting of (call identifier, sender ATM address).

4.2. Traffic and QoS handling

Traffic and QoS specifications are not defined in RSVP. They are deferred to the int-serv IETF draft documents. The Guaranteed Delay int-serv draft [SP95] defines the traffic specification (TSpec) as consisting of a token bucket with a given bucket depth b (in bytes) specifying the maximum allowed burst size for the flow, a bucket rate r (in bytes/second) giving the average rate of the flow, and a peak rate p (in bytes/second) giving the maximum transmission rate of the flow at the source. This traffic specification can be mapped into the corresponding ATM traffic parameters, which are specified in cell-based measures.

[SP95] defines the service specification (RSpec), and a procedure that describes how the RSpec is determined as a function of the delay requirements of the flow and the capabilities of the service elements (routers) on its path. The end-to-end delay d and the associated service specifications for the flow are not quantities that are initially provided explicitly. Rather, they are determined at the receiver upon receipt of the Path message carrying the values of the ``error terms'' Ctot= Sum Ci and Dtot = Sum Di, which have been accumulated on the connection's path. The terms Ci and Di correspond to the error contributed by router i when compared to a perfect fluid service model.

The RSVP and Int-Serv documents suggest that the resource reservation for a flow from S to D with guaranteed delay requirement is performed in the following way. The source S generates Path messages that contain the traffic characterization (TSpec)of the flow. The Path message, therefore, includes the parameters b, r, p, and two fields Ctot and Dtot which are both initialized to 0. At router i, these fields are incremented using the local values Ci and

Di:

Ctot:= Ctot+Ci; Dtot:= Dtot+ Di

At the receiver D, a desired end-to-end delay d is selected, and the required clearing rate R is computed as:

R = max(r, ((b + Ctot) / (d - Dtot)))

The clearing rate R is then loaded in the RSpec included in the Resv message sent towards S.

A key aspect of the above approach, that complicates the interactions with ATM is the decoupling between the advertising (accumulation of Ctot and Dtot as the Path message progresses) and the reservation phases (request for allocation of the clearing rate R). The main issue at the boundary of an ATM network is to determine which values to select for the terms CATM and DATM, when updating the Ctot and Dtot fields in the Path message. Also, delay guarantees based on the specification of a clearing rate may not always be supported by ATM switches and can not be readily expressed through ATM signaling. Hence, the ATM network has to be accounted for as a fixed delay component of the path. This requires information about (a) the ingress and egress points (routers) of the ATM network, (b) a delay bound, DATM, on the flow between the ingress and egress points.

The first item is available at the egress point as the Path message exits from the ATM network (as outlined in Section 2.2). The second item may be obtained by the egress router by querying the ATM network to find the best delay that can be guaranteed for a flow with the specific endpoints. While this information is not currently accessible over an ATM UNI, it is available as part of the ATM PNNI control information. The egress router would then be responsible for updating the Ctot and Dtot fields as the Path message exits from the ATM network (intermediate routers would leave these fields unmodified). This mechanism for updating the advertizement information at the egress points is consistent for both unicast and multicast flows.

5. Impact on RSVP and ATM signaling

In this document we proposed a method for supporting RSVP-based resource reservations in a heterogeneous environment which includes ATM networks. This method, classical RSVP over ATM with shortcuts, requires a number of modifications to RSVP and to ATM signaling. We review here these requirements.

5.1. Modifications to RSVP in the UNI 3.1 environment

In this environment, the general approach we take can be characterized as root oriented. This means that most of the interactions with the ATM signaling needed to extend RSVP flows across ATM networks, originate in the ingress router. Such extensions require a number of modifications to the processing of Path and Resv messages.

The first step at the ingress router is to identify that the flow is to cross an ATM network and should, therefore, be handled differently. Once this has been determined, subsequent modifications to the Path message handling vary somewhat as a function of the approach used. Typically, the Path message will be forwarded on the normal IP path, and extended to carry the ATM address of the ingress router. Path processing is also different at intermediate (non-egress) routers which do not update the PHOP field, so that it still points to the ingress router, and do not maintain state information. This helps lower the processing overhead for such messages. In addition, the Dtot field (and Ctot) is not updated until the Path message reaches the egress router(s), where it is incremented by an estimate of the maximum delay the ATM network would contribute. Path messages continue flowing on the IP route even after an ATM VC shortcut has been established for the flow.

The processing of Resv messages is also affected when crossing ATM networks. They are used to trigger the establishment of an ATM shortcut when received at an egress router(s). The connection request originates from the ingress router (ADD-PARTY for multicast flows, or SETUP for unicast flows) upon receipt of a new Resv message from an egress router. This Resv message carries the standard RSVP information, i.e. filter and service specifications, that are needed by the ingress router to forward Resv messages to its upstream neighbor. The Resv message also contains the ATM address of the egress router as well as the delay guarantees needed for the connection across the ATM network. Note that the receipt of the SETUP (or ADD-PARTY for multicast flows) at an egress router provides an implicit acknowledgment that the ingress router has received the Resv message and that the ATM reservation has been successful. Finally, refresh Resv messages are suppressed, i.e. not forwarded on the IP path, and connection liveness is guaranteed by ATM mechanisms.

5.2. Modifications to RSVP in the UNI 4.0 environment

The major enhancement in UNI 4.0, from the point-of-view of RSVP support, is the LIJ ability in point-to-multipoint connections. This allows us to use a leaf oriented approach to support RSVP flows (both unicast and multicast) which ensures better scalability.

The handling of Path messages remain essentially as for the UNI 3.1 case, in that they are forwarded on the normal IP path but not processed at intermediate routers, i.e. PHOP field and OPWA objects are not modified and no state is created. In addition to carrying the ATM address of the ingress router, the Path message also carries a global ATM call identifier (GCID) in the case of multicast flows. This GCID is then specified in the LIJ message generated by egress routers upon receipt of a new Resv message, when they want to join an existing point-to-multipoint connection associated with a given multicast flow. In the case of a unicast flow, the egress router simply initiates a SETUP to the ATM address of the ingress router.

Because in the leaf oriented approach the egress routers are responsible for the establishment of ATM connections, it is not necessary to forward Resv messages to the ingress router for that purpose. However, it is still necessary to transmit the RSVP information contained in the Resv message (filter and service specifications) to the ingress router, so that it can propagate it upstream. This is achieved by forwarding all Resv messages (including refreshes for reliability) on the IP route to the ingress router. Note that although Resv messages are processed at intermediate routers they are not acted upon, i.e. merging of Resv messages will take place when required but no reservations will be triggered and no state is maintained.

5.3. Modifications and extensions to ATM signaling

As stated above, it is clear that many of the extensions to be included in UNI 4.0 are key to an efficient support of RSVP flows across ATM networks. Foremost among them is the LIJ capability, which is critical to handle large multicast connections. This capability should, however, be such that different leaves are allowed to specify different service requirements. Other desirable extensions to be included in UNI 4.0 are the ability to renegotiate the characteristics of an established connection, and a B-HLI field larger than its current 8 bytes.

There are, however, other desirable extensions which may not be provided in UNI 4.0. One such example, is the ability for an RSVP router to query the ATM network for the best delay that can be guaranteed to a given destination. This can be achieved either by allowing ``soft'' requests, or by supporting both ``desired'' and ``acceptable'' QoS parameters. As a second example, the ability to let the root of a point-to-multipoint call assign a GCID even before any leaf has requested to join, could simplify some of the processing when establishing such calls.

6. References

[Arm95] G. Armitage, "Support for Multicast over UNI 3.1 based ATM Networks", Internet Draft, draft-ietf-ipatm-ipmc-10.txt, December 1995.

[BFG+95] A. Birman, V. Firoiu, R. Guerin and D. Kandlur, "Provisioning of RSVP-based Services over a Large ATM Network", IBM Research Report RC20250, October 1995.

[BZB+96] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1, Functional Specification", Internet Draft, draft-ietf-rsvp-spec-09.txt, February 1996.

[For94] ATM Forum, "ATM User-Network Interface Specifications, Version 3.1", September 1994.

[For95] ATM Forum, "ATM User-Network Interface Specifications, Version 4.0", ATM Forum/94-1018.

[KP95] D. Katz and D. Piscitello, "NBMA Next Hop Resolution Protocol (NHRP)", Internet Draft, draft-ietf-rolc-nhrp-07.txt, December 1995.

[Lau94] M. Laubach, "Classical IP and ARP over ATM", Internet Draft, Internet RFC1577, January 1994.

[Mil95] W. Milliken, "Integrated Services IP Multicasting over ATM", Internet Draft, draft-ietf-milliken-ipatm-services-00.txt, July 1995.

[SP95] S. Shenker and C. Partridge, "Specification of Guaranteed Quality of Service", Internet Draft, draft-ietf-intserv-guaranteed-svc-03.txt, December 1995.

7. Authors' Address

Alex Birman
T. J. Watson Research Center
IBM Corporation
30 Saw Mill River Rd.
Hawthorne, NY 10532

Phone: +1-914-784-7275

E-mail: birman@watson.ibm.com

Roch Guerin
T. J. Watson Research Center
IBM Corporation
30 Saw Mill River Rd.
Hawthorne, NY 10532

Phone: +1-914-784-7038
E-mail: guerin@watson.ibm.com

Dilip Kandlur
T. J. Watson Research Center
IBM Corporation
30 Saw Mill River Rd.
Hawthorne, NY 10532

Phone: +1-914-784-7722
E-mail: kandlur@watson.ibm.com