Shortcutting IP Flows over Large ATM Networks

J. Schmitt[1], L. Wolf[1], M. Karsten[1], C. Siebel[2], Y.-O. Lorcy[2], R. Steinmetz[1,3]

[1] KOM, Darmstadt University of Technology, Merckstr. 25, D-64283 Darmstadt, Germany
Tel.: +49-6151-166150, Fax: +49-6151-166152
{Jens.Schmitt, Lars.Wolf, Ralf.Steinmetz}@kom.tu-darmstadt.de

[2] Deutsche Telekom AG, Technologiezentrum Darmstadt,Germany
{lorcy, siebel}@tzd.telekom.de

[3]GMD IPSI, German National Research Center for Information Technology,
Dolivostr. 15 • D-64293 Darmstadt • Germany

Abstract

In this paper we propose approaches for shortcutting of IP flows over large ATM networks. With large ATM networks we mean that the single physical ATM network is logically structured into multiple logical ATM subnetworks. Shortcutting across such large, logically structured ATM networks is a technique to avoid network layer processing as much as possible by maximizing the switched path across the ATM network. Existing schemes for shortcutting only provide mechanisms for constrained situations, as e.g. being solely applicable to unicast best-effort transmissions. Hence we try in particular to approach the currently unsolved respectively untreated cases of QoS and multicast transmissions.

Keywords: Shortcut, Heterogeneous IP/ATM networks, QoS, Multicast.

1. Introduction

1.1 Motivation

The amazing growth of interconnected networks such as the Internet is one of the most important developments in telecommunications in the last decade. Today, Internet traffic is to a very large extent carried over telecommunication networks. Thus the Internet benefits from the important and significant improvement in capacity of telecommunication networks by the use of fiber optic and ATM technology, those factors actually enabling the Internet's transition from a research network to a mass-scale information infrastructure. Technically, IP networks are often virtual networks over an ATM network infrastructure, at least partially. Such an overlaid network results in a separation of the control planes of the two networks, in particular for the routing of data respectively connections. In principle, the IP network can be operated without "knowledge" of the mechanisms of the underlying ATM network and that is how current production networks are usually operated. While this is certainly a simple approach, it lacks in efficiency. For efficiency reasons a more integrated relation between IP and ATM network layers is favorable. In particular, IP's awareness of the underlying ATM network can help in the domain of routing by, e.g., using PNNI's [6] QoS routing capabilities. Since PNNI knows about the topology and the dynamic state of the whole ATM network it can take much better decisions than a statically preconfigured routed path through the ATM network.

1.2 Assumptions and Terminology

For our discussions we assume a large heterogeneous IP/ATM network according to the overlay model. While there are other models of interaction between IP and ATM networks [17], this is the most simple one and more or less the only one playing a role in current networks. By a large ATM network we mean one that is logically structured into clusters or subnets (i.e. uses routers inside the large ATM cloud) for policy, administrative and/or scalability reasons. We call routers with connectivity to both the IP and the ATM network edge devices and with respect to the data flow ingress or egress devices. Alternatively we also call these devices subnet-sender and subnet-receiver or virtual source and destination.

Furthermore, we assume that RSVP ([11]) is used as the protocol to convey QoS information inside the IP network, i.e., as IP's signalling protocol.

1.3 Introductory Example

Let us take a look at the difference between hop-by-hop routing and shortcutting by regarding the case of a QoS unicast transmission.

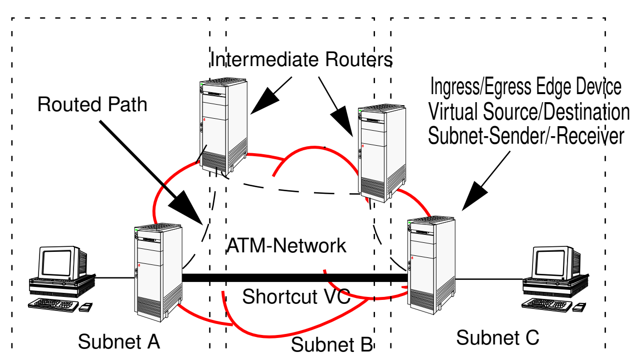

Figure 1: Hop-by-Hop vs. Shortcut.

In the scenario depicted in Figure 1, where three subnets are linked via routers, PATH messages would be delivered hop-by-hop (from router to router) to the final egress edge device. If the intermediate routers do standard RSVP processing then they will each "sign" the PATH messages as previous hops. According to the thus established PATH state, RESV messages will be transported back hop-by-hop in the reverse direction setting up a concatenation of VCs between the routers connecting different ATM subnets. Let us suppose now that all intermediate edge devices forward PATH messages without modifying the previous hop object. Then the RESV message of the final egress edge device would be sent straight to the ingress edge device and from there a shortcut VC to the egress edge device could be established. This is only an example of a modification in the router's behavior that would allow shortcuts for RSVP-signalled IP data flows. More detailed treatment of this and other cases will be presented in the rest of the paper.

1.4 Existing Approaches and Related Work

The first two standards allowing to transmit IP traffic over ATM networks were the IETF's Classical IP over ATM (CLIP [13]) and the ATM Forum's LANE [5]. While differing in many details (see [1] for an overview), they both follow the overlay model and have the concept of clustering the ATM network into LISes (Logical IP Subnets) respectively VLANs (Virtual LANs), where traffic between these has to pass routers, i.e. has to be transmitted hop-by-hop. While LANE allows for IP multicast transmissions, although in a simple and non-scalable manner, pure CLIP is restricted to unicast. Therefore, the IETF proposed the MARS (Multicast Address Resolution Server) architecture as an extension to CLIP [2].

In order to tackle the obvious inefficient use of the ATM network by using a routed path where a switched path is available, the IETF developed the Next Hop Resolution Protocol (NHRP [14]) in order to allow for unicast shortcuts. The ATM Forum in turn used NHRP in order to allow for its "successor of LANE", MPOA (Multi-Protocol over ATM [8]), the possibility of inter VLAN communication. Another solution developed by the ATM Forum to achieve shortcuts is currently proposed as PNNI Augmented Routing [7]. All of these shortcutting techniques however are currently only applicable to unicast transmissions. There are two proposals inside the IETF for best-effort multicast shortcuts, called VENUS [4] and EARTH([18]).

All of the above approaches only take into account best-effort transmissions. However if IP networks start to carry QoS-sensitive data flows as well, e.g. by applications using the RSVP protocol to convey their requirements, this area also has to be taken into account both for unicast as well as multicast transmissions. Early work with regard to this can be found in [10], which presents different alternatives for setting up shortcuts in response to RSVP-signalled information. In [9] the issue of using shortcuts across ATM networks when overlaying RSVP onto ATM is shortly touched, but more or less simply stated as shortcuts could be beneficial.

For the rest of the paper, the reader is assumed to be familiar with the concepts of CLIP, MARS, NHRP, and RSVP because discussions will be based on these protocols, although the proposed approaches should with minor changes also be applicable to other alternatives like LANE, MPOA, or PAR.

2. Basic Shortcutting Issues

2.1 IP vs. ATM Shortcuts

A fundamental issue of shortcutting is the question about which control plane does the routing of VCs through the ATM network. There are currently several approaches where shortcuts are provided by just splicing the concatenation of VCs at the routing hops, thus removing the IP processing inside the ATM cloud (see for example [15] or the work in the IETF MPLS (Multi-Protocol Label Switching) Work Group). The IP control plane virtually takes over ATM and does the routing itself. We call this IP shortcuts. Another approach is to let ATM's routing protocols like PNNI decide about which route to choose through the ATM network for establishing a shortcut VC from the ingress to the egress edge device. Here, the ATM control plane remains intact, that is why we call this approach ATM shortcuts. In [12] it was shown that ATM shortcuts lead to a better utilization of network resources depending however on the topology of the overlaid IP/ATM network. Furthermore, IP flows can benefit from ATM's advanced routing protocol, PNNI, by e.g. the use of its QoS routing capabilities. Especially for large and logically structured ATM networks it might well be possible that for a QoS transmission with certain delay and bandwidth requirements, as signalled by RSVP, there is no more capacity on the routed path but there is ample capacity on a different path through the ATM network. The blocking on the routed path can be due to two reasons: router resources shortage or bandwidth shortage on the routed path. While the first problem is adressed by IP shortcuts as well as ATM shortcuts the second problem can only be solved by using ATM shortcuts.

In this paper, we are dealing with ATM shortcuts. Besides the advantages mentioned above this is also due to our assumption of telecommunication networks, which are multi-service networks, carrying different types of traffic, not just data, although this is expected to become their most important "customer" in the future. For the operation of these networks it would be very burdensome to use two different control planes, assuming that the other applications like e.g. voice would keep on using the standard ATM control plane.

2.2 Pro's and Con's of Shortcutting

Before considering approaches to support shortcuts (for best-effort as well as QoS transmissions), the general merits and drawbacks of this technique should be stated clearly:

- Advantages are
    - lower delays and higher throughput can be achieved due to maximizing the switched path, i.e. eliminating layer 3 processing and segmentation/reassembly inside the ATM network;
    - ATM's PNNI and its QoS routing capabilities can be utilized over the whole ATM subnetwork and not just a LIS/Cluster;
    - routers are off-loaded, thereby avoiding them to become bottlenecks;
    - if there is a setup cost for ATM connections then a shortcut saves expenses when compared to a concatenation of VCs.
- Disadvantages are
    - the virtual source to the ATM network might become overloaded due to a so-called "VC implosion" problem if the ATM network becomes large, in the QoS case this is when there will be too many reservations to be managed and too many RESV messages to be processed;
    - shortcutting reduces the potential for aggregation of flows at the network layer, since less flows will share the same ingress edge devices the closer the ingress edge devices are located to the actual sources;
    - policy and administrative reasons might also constrain shortcuts, e.g., security mechanisms implemented on layer 3 and above might prohibit use of shortcuts or at least call for similar mechanisms on layer 2 (which is here the ATM or AAL layer);
    - in the case of an RSVP multicast session that uses shortcuts over a large ATM network there will be no sharing and no merging of reservations inside the ATM network, thereby losing scalability in the number of participants of a session.

Hence, shortcutting is not intrinsically good, but can be beneficial in at least some cases. We must thus determine when establishing a shortcut is really worthwhile. A prerequisite to establishing a shortcut is that the amount of data and the lifetime of the flow are large enough to justify the effort. Since a shortcut is always an exception where a new connection has to be build up, whereas for data that takes the "default" routed path through the ATM network there will usually be an open connection after a certain initialization period. The decision to establish a shortcut should also be based on the load of the intermediate routers. If these are already very loaded, then a shortcut might actually be the only possibility to establish a data flow with certain QoS requirements across the ATM network. The VC management scheme to support shortcuts should thus take into account state parameters of the ingress edge device and all the intermediate routers of the hop-by-hop path.

In the next sections we will analyze some existing approaches and propose new ones for shortcutting. This investigation will be made along different types of IP traffic.

2.3 IP Traffic Types

Since we assume that it is helpful to separate the comprehensive problem of shortcutting into smaller subproblems, we decided to do so by differentiating IP flows by two criteria:

- whether the data is best-effort or has QoS requirements signalled by RSVP, and
- whether it is a uni- or multicast transmission.

This is also the approach taken by other proposals, which however always treat only a subset of the four different cases, while we examine all of them. Other proposals mainly focus on the best-effort case although shortcutting is especially interesting for the transmission of QoS data.

3. Shortcutting IP Flows

3.1 Shortcut for Best-Effort Unicast Communications

In the case of best-effort unicast traffic, one could argue that shortcuts should not be necessary, as this kind of traffic has no strict timing requirements. This line of argument however misses the fact that shortcutting also offloads the routers inside the ATM network and that while best-effort data do not crucially depend on delay they still often profit very much from a reduced delay.

So, if a shortcut is desired, there are several existing approaches as already mentioned: NHRP, MPOA, or PAR. So there is no requirement for yet another approach in this area of the problem space.

3.2 Shortcut for QoS Unicast Communications

The situation for QoS unicast communications is quite different. There is currently no standard approach for setting up ATM shortcuts triggered by RSVP signalling. Yet, especially in the QoS case shortcuts could be very valuable due to the intrinsic delay requirements of the corresponding applications. Also, the use of PNNI's QoS routing capabilities to setup the shortcuts according to the traffic specifications contained in RSVP messages could be very valuable for QoS-dependent applications. Furthermore, as we will show, shortcuts are easier to be setup due to access to the information delivered by RSVP signaling.

In principle, both, the subnet-receiver and the subnet-sender, could setup the shortcut VC since point-to-point VCs are bidirectional and asymmetric. However, a subnet-sender approach seems more reasonable, since the ingress edge device certainly knows best about its current load due to processing of shortcuts. Furthermore, if the ATM equipment still uses UNI 3.1 the subnet-sender approach is the only alternative. So, we only regard subnet-sender-initiated shortcuts here. If shortcut is desired for this case and a reservation request has actually been issued by the receiver, then the RESV messages should only be processed by the ingress edge device, and not by any of the intermediate routers. There are different approaches to achieve this:

1. The virtual source could include an object into the PATH message that contains its IP or ATM address, so that the subnet-receiver who requests a reservation can send its RESV message right to it, as suggested in [10]. That would, of course, mean modifying the RSVP protocol by including this new object and adequate processing for it.

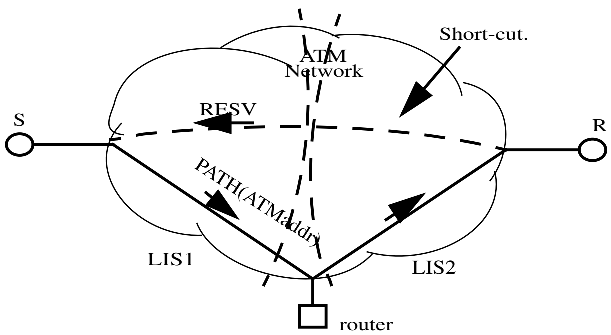

Figure 2: PATH message with source ATM address.

2. By means of some indication (e.g. by setting a flag in the RESV message) the receiver could tell the routers not to process the RESV message but forward it to the ingress edge device. This way the subnet-source would see the final egress edge device as next hop and could establish a shortcut to it.

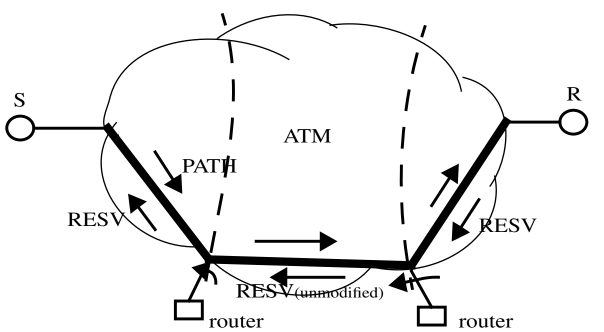

Figure 3: Forwarding "unmodified" RESV message upstream.

Another question is: how does the subnet-source know the ATM address of the subnet-receiver? Two possible solutions are:

1. Use NHRP to get the ATM address. This may take some time, however, since it is expected that the lifetime of the connection be significantly longer than this period, that may be acceptable and is a simple approach if NHRP is available.
2. Include a new object in the RESV message which carries the ATM address of the subnet-receiver to which the shortcut should be established [10]. This solution would also permit a non-RSVP capable egress edge device. The next RSVP-capable hop would be connected to this edge device and knows about the next hop being a non-RSVP capable ATM egress edge device. Therefore, it sends its RESV message including the ATM address of the egress edge device in the new object. This way, the source knows the ATM destination of the QoS VC.

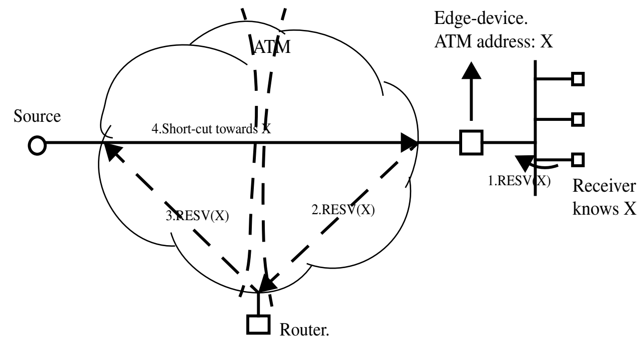

Figure 4: Non-RSVP capable edge device.

3.3 Shortcut for Best-Effort Multicast Communications

While the same arguments for the use of shortcuts as in the case of unicast traffic apply to the best-effort multicast case, it has to be observed that multicast shortcuts are significantly harder to achieve due to the potentially large size of IP multicast groups, their dynamics and the anonymity of the IP multicast model. On the other hand, multicast applications usually have longer durations and often are more delay-sensitive than unicast applications, and are thus likely to benefit from using shortcuts.

Since we are in the best-effort domain there are of course no RSVP messages or, at least, there are no reservations yet. Let us suppose that we are using MARS in conjunction with the VC-mesh approach. Then, for inter-cluster communications, one or several multicast routers will be used as illustrated in figure 5.

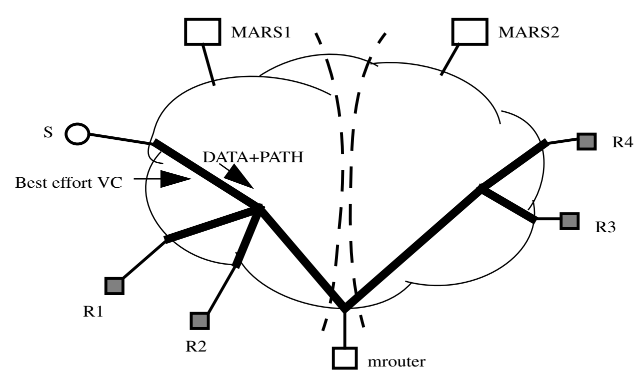

Figure 5: Multicast with multicast router (hop-by-hop).

If we desire to establish a multicast shortcut, MARS needs to be extended in a similar way as ATMARP had to be extended to NHRP in order to support shortcuts for the unicast case. There are however some serious problems when trying to establish shortcuts for the multicast case:

1. How does a source get to know the ATM addresses of receivers outside its own cluster and how should it keep track of membership changes outside the cluster. MARS should be modified to provide this information to the source. Therefore a form of coordination between MARSes is necessary.
2. It is possible that the number of receivers or members of a group exceeds the greatest point-to-multipoint VC a source or the ATM network is able to set up, as mentioned in [3]. In this case, either the number of group members must be limited, or a mixed scheme of using shortcuts and multicast routers could be designed. However, both options have their drawbacks. To limit the number of members in a group could certainly be very restricting for future large-scale multicast applications. The mixed scheme of using shortcut and hop-by-hop requires a complicated management due to the fact that some receivers receive data through the shortcut VC while others get them from the hop-by-hop path. This would also result in different QoS for those two kinds of receivers.

An alternative to alleviate at least the second problem would be to have some kind of multicast servers in order to aid the source, in case no more leaf nodes can be added to the point-to-multipoint VC. This scheme would result in a cascade of sources. Usually, a very small number of cascaded sources will suffice. In many cases, no more than one of these devices should be needed in any multicast communication. This is valid if the number of group members is less than twice the maximum number of nodes allowed in a point-to-multipoint VC. The case of one auxiliary source is depicted in figure 6.

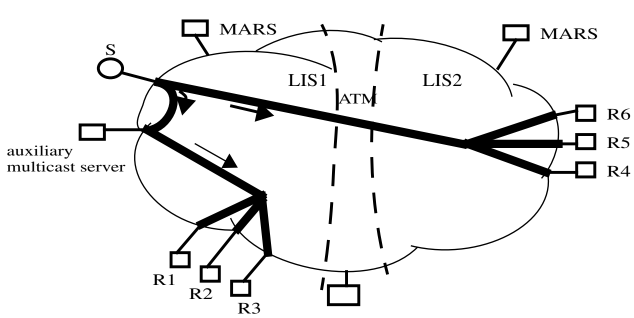

Figure 6: Cascaded Sources.

No IP processing and no segmentation and reassembly is needed in the auxiliary multicast server, because its only function is to extend the point-to-multipoint VC of the source. Thus, instead of a multicast server at the IP level it could be a device which only takes incoming cells and forwards them on a point-to-multipoint VC.

Extending MARS

In this section we propose extensions to MARS in order to tackle the first problem area of the preceding section. One of the problems that must be treated is how the source gets to know the ATM addresses of the receivers outside its own cluster. First of all, MARS_REQUEST messages should be modified in order to let the source specify if shortcutting instead of normal hop-by-hop routing is desired. This could be achieved by adding a new TLV(Type-Length-Value) field in the MARS_REQUEST message, which indicates to MARS that the source would like to establish a multicast shortcut.

In turn, the MARS should then answer in its MARS_MULTI message with all the ATM addresses of the ATM subnet-receivers of the group. To be able to do that, a scheme that allows MARS to solicit the addresses of receivers registered at other MARSes is required. Therefore some messages between MARSes from different clusters are necessary. In order to distinguish the extended MARS from the original MARS we call it cMARS (communicating MARS).

In the unicast case with NHRP, a request message is sent inside an IP packet, being forwarded to different NHSes (NHRP Servers) until one of them knows the ATM address requested. In case of cMARS, this request message should be addressed to other cMARSes. However, the requesting cMARS does not know which other cMARSes have members of the group, so two approaches are possible:

1. Send the request message, one by one, to all the cMARSes of the network. This certainly shows scalability problems if the number of cMARSes is becoming large. Furthermore, the requesting cMARS would need to know the ATM addresses of all cMARSes in the ATM network.
2. cMARSes should be IP nodes. This way, they could join a specific IP multicast group dedicated to the inter-cMARS communication (or possibly a hierarchically structured tree of multicast groups if the ATM network becomes extremely large and high scalability is required). Thus, requests for group members of specific IP multicast groups would be received by all cMARSes, and the ones that have members of that group in their cluster could answer with a list of the group members and their ATM addresses. The answer could be sent as an IP packet back to the source IP address of the multicast packet received, i.e., the requesting cMARS, or, alternatively, to the multicast group of all cMARSes. In general, the second option will result in more traffic than the first one and seems therefore inferior, but would have the advantage that group membership information could be cached by cMARSes even if they have not yet requested it.
3. cMARSes are in a higher level cluster with one dedicated MARS to which requests are sent and which sends back answers.

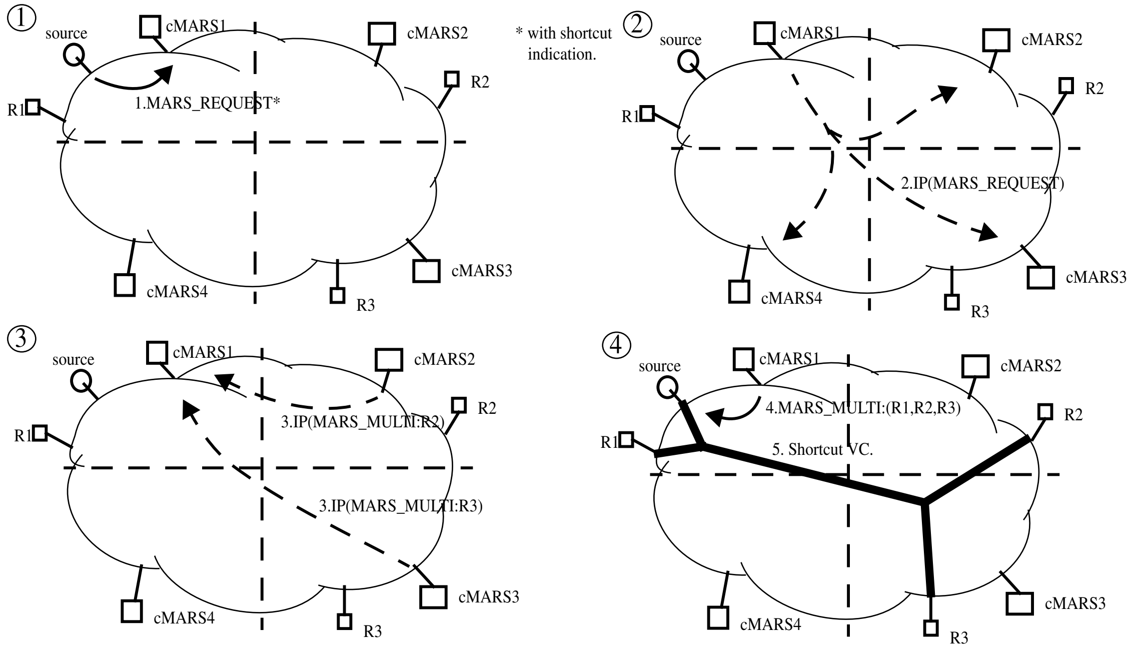

Figure 7: MARS extensions.

With one of theses approaches, it is now possible for the cMARS of the cluster in which the source is located to get to know all the receivers that currently belong to the group, as illustrated for the second approach in figure 7. However, IP multicast groups are dynamic, thus membership changes in other clusters must be tracked in some way.

One possible approach to tackle this problem is that each of the cMARSes adds the requesting cMARS to its Control Cluster VC, so that if changes in group membership occur, the requesting cMARS is aware of them. This solution is certainly not scalable since the requesting cMARS must be added to all the Control Cluster VCs and must process the information received on all of them. It should be noted that it would even need to be added to the Control Cluster VC of clusters that have no members for that group, because they might appear anytime.

A more scalable solution would be to indicate group changes to other cMARSes by encapsulating these messages in IP packets. These packets should only be sent in case that a requesting cMARS message has been received for the group that has changed, i.e., there is a source somewhere in the ATM network using multicast shortcut for that group. This way, group changes would be delivered by IP packets between different cMARSes. The question is how a cMARS knows where the shortcut sources are and whether they are still active. A MARS_REQUEST message with a shortcut indication from a source to its local cMARS can be seen as a way to register as a “shortcut source” within this cluster. Similarly, the request message sent to the IP multicast group of cMARSes can be a way to register within all other cMARSes as “cMARS with shortcut sources for that group”. With this information each cMARS whose cluster has changes for that group could notify them to the cMARSes which have a shortcut source for that group. When a subnet-source decides to finish its connection (may be due to inactivity), a message should be sent to the local cMARS to delete this source as a “shortcut source”. This could be done by introducing a new type of message in the MARS protocol, or simply using a MARS_REQUEST message with a TLV field indicating that the source does not need shortcuts any more. A similar message should be sent by the cMARS to the IP multicast group of cMARSes in order to be deleted as “cMARS with shortcut sources” for a particular group, if it has no “shortcut sources” for that group any more.

UNI 4.0 LIJ Facility

If UNI 4.0 Leaf Initiated Join is available, shortcut for multicast best-effort communications can be simplified to some extent. For best-effort multicast communications the LIJ facility may improve the scalability of the solution, since now the subnet-source does no longer need to add all the receivers of a group. With LIJ, it is the receiver who joins the point-to-multipoint VC if it desires to receive best-effort multicast data over a shortcut VC. Therefore, the problem now is to find out the identifiers (GCIDs) of an existing point-to-multipoint VCs for that group. MARS is currently designed to provide the ATM addresses of members of a group. Some extensions or a different protocol would be necessary to provide a receiver which wants to use LIJ with the GCIDs of the point-to-multipoint VCs of the group.

3.4 Shortcut for Multicast QoS Communications

One of the problems in implementing shortcut in the best-effort case is that, because of the anonymous IP multicast model, the source neither knows nor is informed about which members are in the group. Therefore, a procedure for the source to retrieve this information is required. MARS is an implementation of such a procedure, but its coverage is limited to the same cluster. The problem, thus, has been to extend MARS so that it works also for inter-cluster communications in a reasonably scalable manner even if dynamic membership is taken into account.

For the QoS case, where RSVP signalling is used, establishing shortcuts becomes actually easier than in the best-effort multicast case, though not as straightforward as for QoS unicast transmissions. When a receiver requests a reservation sending a RESV message to the previous hop, it is explicitly notifying its identity (by means of its IP address at least, if no extensions are being made). Therefore, the source knows which are the receivers of the group by means of the RESV messages. No additional mechanisms are necessary for finding the identity of the virtual destinations.

If shortcut is being used for best-effort multicast data and thus for PATH messages, the previous hop of the PATH message, i.e., where the receiver has to send its RESV message to, is the ingress edge device itself. If hop-by-hop is being used, the PATH message could be modified to contain an indication for the multicast routers to not modify the previous hop object of the PATH message. Hence, the receiver would send its RESV messages straight to the source. In both cases the ingress edge would know the IP addresses of the receivers to which a shortcut VC shall be established. However, what it needs to know is the ATM addresses of the subnet-receivers, the leaves of the shortcut point-to-multipoint VC. It is the same problem as in the QoS unicast case and thus the same approaches are principally applicable. The first option is to use NHRP to discover the ATM addresses of receivers outside the cluster. Besides the advantage of using a standardized mechanism, this has the following drawbacks:

1. The delay until a QoS VC is established could be too long, especially if the multicast group becomes larger and more dynamic.
2. There is a problem with non-RSVP capable egress edge devices, because for these next hop and ATM network egress will not be the same node.

A more efficient solution would be to include a new object into the RESV message, which contains the ATM address to which a shortcut should be established, as described for the unicast case. This way every member of the group which wants to receive data with QoS could send a RESV message containing additionally the ATM egress point. With this information, the ingress edge device could add this receiver to an existing shortcut point-to-multipoint VC, or could create a new shortcut VC, or could take any other decision depending on the VC management strategy being implemented.

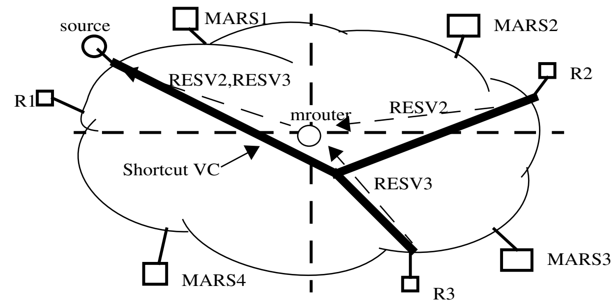

Figure 8: Forwarding RESV messages without merging.

UNI 4.0 LIJ Facility

On first glance, LIJ seems to be a good match with the receiver-oriented philosophy of RSVP. However, when a receiver requests a reservation, a RESV message is sent upstream, but the actual reservations are carried out in the downstream interfaces. Therefore, in the ATM context, it seems reasonable that the subnet-sender should be the one who sets up the branch of the point-to-multipoint VC.

With LIJ, however, it is possible that the branch is setup by the subnet-receiver when the RESV message arrives at the egress edge device. If LIJ is used, then a useful modification of RSVP would be to include the GCID of the shortcut point-to-multipoint VC into the PATH messages sent by the ingress edge device in order to be able to join that VC by the egress edge device. As an advantage of using LIJ the load in the ingress edge device would be lowered and thus a better scalability with the number of receivers could be achieved.

With a VC management strategy that permits the use of multiple VCs for a single RSVP session, e.g., in order to support some degree of heterogeneity [16], a receiver might either be offered a choice of different VCs which he could join or the source decides according to global criteria which VC is appropriate for a receiver to join and just sends one GCID in the PATH to the egress edge device. Here it becomes obvious that LIJ is not such an elegant solution as one would expect at first. While a choice of different VCs to join does not optimize the VC management according to global criteria, the other option of deciding which VC is appropriate for a receiver at the virtual source before a RESV message has been received is very restricting. The centralized nature of VC management strategies just does not fit very well to the decentralized concept of LIJ.

4. Location of the Shortcut Decision

The fact whether shortcuts are used for best-effort traffic or not, affects the way RSVP control messages are delivered over the ATM network. This, in turn, influences the way in which shortcuts for RSVP flows can be established and which instance decides about the establishment:

- If shortcut is being used for best effort traffic, establishing shortcut for the QoS case is straightforward, since the RSVP PATH messages travel from the ingress edge device straight to the egress edge device, without any intermediate IP nodes. Therefore, the previous hop is the ingress edge device. In fact, in this case there is no other choice than using shortcut for the QoS case.
- If hop-by-hop is being used for the best effort case, using shortcut for QoS traffic might be an initiative of:
    - the ingress edge device,
    - the egress edge device,
    - the intermediate routers.

If the virtual source decides to use shortcut, one of the methods explained above, as for example modifying the PATH message, should be utilized. Some changes are also needed in intermediate routers, in order to avoid them modifying the previous hop object of the PATH message.

If the receiver wants to use shortcut, RESV messages could be sent right to the source or ingress device, regardless of the previous hop of the PATH message. A possible way to achieve this is to include the ATM or IP address of the ingress edge device into the PATH message.

If hop-by-hop is being utilized for best-effort traffic and receiver-initiated shortcuts for the QoS traffic are desired, it requires some coordination between the receiver and the multicast router(s) the best effort traffic goes through. This is needed in order to delete the nodes that are using shortcut VCs from the concatenated best-effort VCs. In case QoS traffic would also be delivered hop-by-hop, deleting the receiver from the best-effort VC when it requests a reservation is also necessary, but here an intermediate router knows which receiver requested a reservation, what kind of reservation (i.e. style) and for which source(s). With this information, the multicast router can decide whether that node should be deleted from the best-effort VC and added to another, or whether it should be kept in the best-effort VC because there are other sources for which the receiver has made no reservation request (e.g. if FF is used). The problem in the shortcut case is now that an intermediate router does not have this information if the receiver sends its RESV messages directly to the source/ingress device. Therefore, RESV messages should be sent hop-by-hop but with an indication that they are for shortcut (this indication can be simply the presence of an ATM address object inside the RESV message).

If the decision of using shortcut is taken by intermediate router(s), this should be based on parameters like:

- current load of the router,
- TSpec contained in the PATH message,
- and/or FlowSpec of the RESV message sent by the receiver.

The aim of an intermediate router-initiated shortcut is to optimize its utilization, and at the same time, to avoid congestion (bottlenecks). In order to allow for the establishment of a shortcut the intermediate router should forward the RESV message to the previous hop without modifying the next hop object. This would enable the previous hop to set up a shortcut VC bypassing a possibly overloaded router. Note that the previous hop may be the ingress edge device or another router. In the first case, a complete shortcut will be established, while in the second case, two possibilities may occur. This router also takes the decision to be bypassed and also forwards the unmodified RESV message. The other choice is that the router decides to become the starting point of the shortcut. If this happens, a "partial shortcut" from that router to the receiver will be established and the RESV message sent to the previous hop will contain this router's address as next hop object.

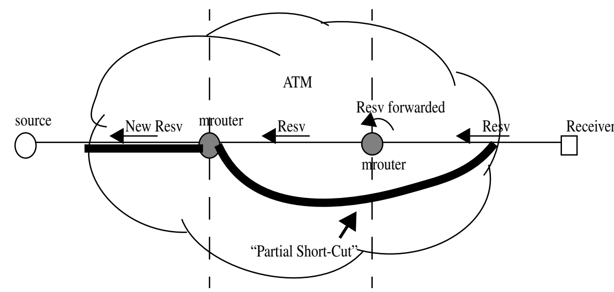

Figure 9: Partial Short-Cut.

5. Summary and Conclusion

We have shown which alternatives exist for shortcutting of IP flows over large ATM clouds. We identified the cases which are on the one hand most suitable for shortcutting and on the other hand could benefit most from it. We presented approaches to achieve shortcuts for all cases of IP flows: uni- and multicast, best-effort and QoS transmissions.

Most suited is certainly QoS unicast because of its supposedly long duration, stringent delay requirements and simplicity of handling when compared to multicast traffic. Nevertheless, QoS multicast traffic could also benefit because of its delay requirements and one may argue that for some multicast applications the dynamics of the group members are much less (even if it is only for the reason that members have to pay and thus really "think" about joining a group).

For best-effort, shortcuts might be arguable, however at least for unicast, they are not that difficult to setup and coordinate as for multicast.

We are currently in the process of implementing the proposed approaches in the OPNET™ network simulator in order to obtain more quantitative assessment of the proposed approaches.

References

[1] A. Alles. ATM Internetworking, May 1995. White Paper, Cisco Systems Inc.

[2] G. Armitage. Support for Multicast over UNI 3.1 based ATM Networks, November 1996. RFC 2022.

[3] G. Armitage. Issues affecting MARS Cluster Size, March 1997. RFC 2121.

[4] G. Armitage. VENUS - Very Extensive Non-Unicast Service, September 1997. RFC 2191.

[5] ATM Forum Technical Committee: LAN Emulation (LANE) over ATM 1.0, January 1995.

[6] ATM Forum Technical Committee: Private Network-Node Interface (PNNI) Signalling Specification, March 1996.

[7] ATM Forum Technical Committee: An Overview of PNNI Augmented Routing, April 1996.

[8] ATM Forum Technical Committee: Multi-Protocol over ATM v1.0, July 1997.

[9] L. Berger, E. Crawley, S. Berson, F. Baker, M. Borden, and J. Krawczyk. A Framework for Integrated Services with RSVP over ATM, August 1998. RFC 2382.

[10] A. Birman, V. Firoiu, R. Guerin, and D. Kandlur. Provisioning of RSVP-based Services over a Large ATM-Network. In *Proc. of IEEE Global Internet*. IEEE, November 1996.

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

[12] V. Firoiu, J. Kurose, and D. Towsley. Performance Evaluation of ATM Shortcut Connections in Overlaid IP/ATM Networks. Technical Report CMPSCI Technical Report TR97-40, University of Massachusetts, July 1997.

[13] M. Laubach and J. Halpern. Classical IP and ARP over ATM, April 1998. RFC 2225.

[14] J. Luciani, D. Katz, D. Piscitello, and B. Cole. NBMA Next Hop Resolution Protocol (NHRP), April 1998. RFC 2332.

[15] P. Newman, G. Minshall, and T. Lyon. IP Switching: ATM under IP. In *Proc. of IEEE Infocom*. IEEE, April 1996.

[16] J. Schmitt. Issues in Overlaying RSVP and IP Multicast on ATM Networks. Technical Report TR-KOM-1998-03, University of Technology Darmstadt, August 1998.

[17] J. Schmitt, L. Wolf, R. Steinmetz, Y.-O. Lorcy, and C. Siebel. Interaction Approaches for Internet and ATM QoS Architectures. In *Proceedings of the 1st IEEE International Conference on ATM (ICATM'98), Colmar, France*. IEEE, June 22–24 1998.

[18] M. Smirnov. EARTH- EAsy IP multicast Routing THrough ATM clouds, March 1997. Internet Draft, work in progress.