THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學

Reliable Multicast Protocol (RMP)

for

Core-based Multicast Tree

Abstract

Abstract of dissertation entitled :

Reliable Multicast Protocol for Core-based Multicast Tree

submitted by CHAN, Kwok Wai

for the degree of MSc in Information Technology

at The Hong Kong Polytechnic University in November, 1999.

This dissertation is to describe the newly architecture of a Reliable Multicast Protocol (RMP) for providing reliable services at PIM-SM domain. In this approach, the protocol selected the appropriated PIM routers located at the branch for handling the NAK packets from the group members and providing the data re-transmission. RMP introduces the new term "NAK Responder" which actually is a PIM router selected for caching the sender's packets to provide the reliable services with the group members at the branch at which the NAK Responders is located. This protocol can also minimize the NAK implosion by increasing the group members because NAK packets sent from the group members are handled by the individual NAK Responder located at those branches which group members are connected.

It also describes the protocol details including the definition of packet type used at RMP, procedures processed by the entities of Sender, Receiver, Rendezvous Point and PIM Router as well as the packet formats. Regarding to the Simulation Result section, it demonstrates the performance metrics including Data Loss Recovery Time, Traffic Concentration and Scalability once the RMP is introduced into the PIM-SM domain for providing reliable service.

At the Conclusions Sections, this protocol can be further enhanced to reduce the overhead and made more effective once the existing PIM-SM protocol can be extended for adding some state information of the branch on it.

Acknowledgements

It is a pleasure and grateful to acknowledge the invaluable helps received from my academic supervisor Dr. Rocky Chang who fully supports in the form of technical papers, discussions, the review of various sections of my dissertation regularly.

Finally, I thank all my friends and colleagues who have contributed their time and efforts for supporting me to complete this dissertation.

Table of Contents

Page

List of Figures

Page

List of Tables

Page

Chapter 1 - Introductions

1.1. Motivation

In the last few years, the Internet has changed from a pure scientific network to the basis of the data communication in every-day life. The number of users grows still exponentially and has already reached the order of magnitude of tens of millions. The added spectrum and number of users introduce also new forms of communication into the Internet: Communication not just between two peers, but true group communication. Examples are a software house distributing the latest release of a software to its clients, and a financial institution disseminating market data to its clients. Information providers send data automatically to all their clients instead of serving requests individually.

The foundation for the group communication in the Internet is the IP-multicast service. However, this service provides no reliability for the data transmission. This may be tolerable for some real-time applications like video and audio transmission, but other services like information distribution require a guaranteed delivery of data.

Nowadays, there are several protocols for reliable multicast transmission available; however, they differ in the service or at topology of multicast tree they provide. And, those protocols support any number of sources within a multicast group, but since these sources operate entirely independently of each others.

In wide-area internetworks – like Internet, most members of sparsely distributed groups wish to participate in a multicast conference do not justify flooding the entire internetwork with periodic multicast traffic. PIM Sparse Mode (PIM-SM)[1] is developed to provide a multicast

[1] See Appendix I

routing protocol that provides efficient communication between those sparsely groups members. At the PIM RP-rooted tree, multi-senders joined to this share tree sending the multicast data to core of the tree (Rendezvous Point) for disseminate it towards the group members at the whole tree. However, there are no reliable multicast protocol working at this core-based share tree environment for providing reliable service today.

Most reliable multicast transport protocols – like AMPT[2] designed by Heinrichs B. and SRMT[3] developed by Marko Schuba, worked at the source specific multicast tree and considered one sender joined at the spanning tree also. For applying this kind of multicast protocols to the environment of multi-senders joining the PIM RP-rooted multicast tree, each sender is just considered entirely independently joining the RP-rooted tree for providing reliable services individually. Traffic concentration at the whole RP-rooted tree cannot be fully optimized. Because there are no centered control unit to co-ordinate the operability of reliable services each senders performed at the PIM RP-rooted tree, no optimization can be found at the whole performances by using those reliable protocols handling multi-senders performing the reliable services at the RP-rooted tree.

Therefore, Reliable Multicast Protocol, hereafter called as RMP, is proposed to be intended as a workable solution for giving the reliable service at root-based multicast share tree of PIM Sparse Mode (PIM-SM). RMP is specifically devised for multicast applications that requires ordered, duplicated-free, multicast data delivery from multiple sources to multiple receivers; and provides with basic reliability requirements rather than as a comprehensive solution for multicast applications with sophisticated ordering, agreement, and robustness requirement.

[2] See Appendix II
[3] See Appendix III

RMP makes use of the RP-rooted tree's properties and state information from PIM-SM protocol to provide the effective and efficient reliable services with the group members under the circumstance of minimizing the resource usage and traffic at PIM RP-rooted tree. At the initial state, RMP bases on the established RP-rooted tree to find out the appropriated PIM routers selected as NAK Responders (NR) from each branch, and those NRs will perform the multicast data storage and data retransmission process at the RMP operations. Senders at RMP operations deliver their multicast packet per clock tick from their independent transmit windows in the absence of negative acknowledgements from any receivers. Reliable delivery is provided within a source's transmit windows from the time a receiver joins the group until it departs. RMP guarantees that a receiver in the group either receives all data packets from transmissions and retransmissions from NRs, or is able to detect unrecoverable data packet loss. RMP supports any number of sources within the RP-rooted tree, each source fully identified by a globally unique Source Transport Identifier (STI).

In the following text, transport-layer originators of RMP data packets are referred to as sources, transport-layer consumers of RMP data packets are referred to as receivers or group members.

1.2. Summary of Operations

RMP runs over a datagram multicast protocol such as IP multicast. In normal process of data transfer, a source multicast sequenced data packet (OData), and receivers unicast selective negative acknowledgements (NAKs) to NAK Responders (NRs) for data packets detected to be missing from the expected sequence. NAK Responders (NRs) located at each branch of the RP-rooted tree provisionally store the multicast packets delivered from the RP, receiving the NAK from the receivers and confirming it by multicast a NAK confirmation (NCF) in response on the interface on which the NAK was received. Retransmission (RData) may be provided either by the NRs in response to a NAK packet, or by the source itself.

Upon detection of a missing data packet, a receiver repeatedly unicasts a NAK to the NR on its branch outgoing from the RP. A receiver repeats this NAK until it receives a NAK confirmation (NCF) multicast from that NAK Responder. That NR with an NCF to the first occurrence of the NAK and any further retransmissions of that same NAK from any receiver. If the data packets requested by NAK are stored at that NR, NR unicasts or multicasts the requested packets along the branch towards the receivers; otherwise, the NR will ask the other NRs located at another branches or sender within the same RP-root tree to get back the requested data packets firstly and then complete the data recovery process as usual.

In order to avoid a flurry of NAK packets, each receiver starts a randomly chosen NAK packet delay timer for each of its group memberships. If, during the delay period, another NAK packet is heard for the same group, the local receiver resets its timer to a new random value. Otherwise, the receiver transmits a NAK packet to the reported group address, causing all other members of the group to reset their NAK packet timers. This procedure guarantees that NAK packet are spread out over a period of time and that NAK packet traffic is minimized for each group with at least one member on the subnetwork.

At PIM-SM multicast protocol, Rendezvous point (RP) is acted as a root forwarding all sender's packets towards its group members from its outgoing interfaces or branches. For RMP protocol, PIM routers located at the PIM tree are selected to become the NAK Responders (NR) for accepting the NAKs and providing reliable service. The net effect is that unicast NAKs return from a receiver to a NR on the reverse of the branch on which OData was forwarded, that is, on the reverse of the distribution branch from the RP. The reasons for handling NAKs and allocating the NRs at the branch this way will become clear in the discussion of constraining retransmissions and storing the multicast packets at the branch, but first it is necessary to describe the mechanisms for establishing the NRs on the branch and announcing the NR availability to the receivers.

To select the PIM-routers as NRs in the RP-rooted tree, the basic operation is receiver periodically originating Search-NAK-Responder Message (SNRM) to its Designated Router (DR), which is the first PIM router connected to the subnet, prompting DR unicasts SNRM Message along the PIM tree to upstream PIM router to collect each PIM router's state including the number of outgoing interface corresponding to each RP-address. The SNRM Message come across the each PIM router and those state information from each PIM router are merged into the SNRM Message that eventually reached the RP. Each of SNRM Message has its branch identifier for identifying from where it came.

Obviously, RP obtains SNRMs from its outgoing interfaces and then selects the appropriated PIM routers as NRs for each branch. The outcome of the selection mechanism is multicast towards its outgoing interface such that PIM routers recognizes itself to be selected as NR and receivers acknowledge the availability of NRs located at its branch. Receivers use this information to address returning unicast NAKs directly to the upstream NR located at the branch on which the multicast packets are forwarded.

As a further efficiency, RMP specifies procedures for the constraint of retransmissions by NRs so that they reach those receivers that missed the original transmission. As NAKs are sent from receivers to NRs, they establish retransmit state in the NRs which is used in turn to constrain the forwarding of the corresponding RData. Packet retransmissions could be through using unicast or multicast to those receivers which missed the packets, however, some receivers have to discard the duplicated packets while using multicast retransmissions.

Finally, since PIM-SM protocol can switch the RP-rooted tree to source specified tree for providing better end-to-end performance, the reliable service of RMP protocol is still valid at this situation.

RMP defines six basic packet types : four of them flow downstream (OData, RData, BPM and NCF), and the others flow upstream (NAK and SNRM).

1.3. Design Goals and Constraints

RMP has been devised to provide the reliable data delivery service on core based multicast shared tree – PIM-SM, it carry out a efficiencies operations to serves those multicast applications that require relatively simple reliability requirements. The usual impediments to realize these efficiencies are the implosion of negative and positive acknowledgement from receivers to senders, retransmission latency from the source, and the propagation of retransmission to disinterested receivers.

1.3.1. Reliability

There is no any document addressing to the reliable data delivery across core based shared tree. In the tradition, at the source specified tree, reliable data delivery across an unreliable network is achieved through an end-to-end protocol in which a source (implicitly or explicitly) solicits receipt confirmation from a receiver, and the receiver responds positively or negatively. While the frequency of negative acknowledgements is a function of the reliability of the network, the frequency of positive acknowledgements is fixed at least the rate at which the transmit window is advanced, and usually more often.

At the PIM tree, no matter how many senders are joined the group, rendezvous point (RP) is to be became as a forwarder forwarding the multicast packets toward its outgoing interface within the multicast group. This situation is more similar to source specified tree at single sender. When these principles are extended without modification to multicast protocols, the result, at least for positive acknowledgements, is a burden of positive acknowledgments transmitted to the RP that quickly threatens to overwhelm it as the number of receivers grows. More succinctly, ACK implosion keeps ACK-based reliable multicast protocols from scaling well.

One of the goals of RMP is to make use of the characteristic of PIM tree to perform the reliability as possible. RMP sustains the negative acknowledgements for retransmissions and reliability. The approach taken in RMP is to the negative acknowledgements and resort instead to timeouts at the source to manage transmit resources. Re-transmit capability is effectively scattered over the whole PIM tree within the multicast group instead of relying on the original source. Obviously, it will be more scalable and load balancing while performing the reliable servicing.

The definition of reliability with RMP is a direct consequence of this design decision. RMP guarantees that a receiver either receives all data packets from transmissions or retransmissions.

1.3.2. Efficiency

While RMP avoids the implosion of position acknowledgements simply by abolishing the ACKs, the implosion of negative acknowledgements is addressed directly.

In order to avoid a flurry of NAK packet, each receiver starts a randomly chosen NAK packet delay timer for each of its group memberships. If, during the delay period, another NAK packet is heard for the same group, the local receiver resets its timer to a new random value. Otherwise, the receiver transmits a NAK packet to the reported group address, causing all other members of the group to reset their NAK packet timers. In addition, Receivers observe a random back-off before generating a NAK during which interval the NAK is suppressed by the receiver upon receipt of a matching NCF. The combination of these two strategies usually results in the NAK Responder receiving just a single NAK for any given lost data packet.

For RMP operation, NAK Responders located at each branch provide the fast reliable service to its downstream group member. More succinctly, receiver can directly get back the lost data from its near NAK Responders located at the branch on which the multicast packets is forwarded directly.

Once receiving first NAK packet, NAK Responder do not transmit the requested packets immediately because one or more receivers may miss the same packet. So, they wait a pre-defined delay time for receiving all NAKs corresponding to same packet and then determine whether the lost packet should be retransmitted using unicast or multicast transmission. At unicast retransmission, RMP specifies procedures for NAK Responders to store the NAK state, which will be used for performing the unicast retransmission later, so as to delivery those lost packet to reaches only those receivers that missed it. For multicast retransmission, it provide a better efficiencies data recovery process but few receivers need to discard the duplicated packets.

1.3.3. Simplicity

RMP is devised to achieve the greatest reliable multicast data service at PIM-SM with the least complexity. As a result, RMP does not address conference control, global ordering amongst multiple sources in the group, nor recovery from network partitions.

1.3.4. Scalability

Data retransmission is performed by the NAK Responder located at the branch on which the multicast data is directly delivered to receivers. Once the receiver joins the multicast tree, it builds the branch toward the RP; some intermediate PIM routers at the branch are selected as NAK Responder doing the reliable service. Obviously, establishment of the NAK Responder is totally depended on the receivers joining the tree so each receiver has its dedicated NAK Responder for providing data retransmission. Eventually, state information stored at the NAK Responder do not soar too much even number of receivers is increased.

Chapter 2 - Protocol Procedures

2.1. Overviews

At the PIM-SM domain, a Designated Router (DR) sends PIM's periodic Join/Prune messages toward a group-specific Rendezvous Point (RP) for each group for which it has active members. Each router along the path toward the RP builds a wildcard state[4] for the group. Reliable Multicast Protocol (RMP) is to provide sequenced, lossless delivery of bulk data from one or more senders to a group of receivers in PIM-SM domain. Actually, RMP is built on the top of PIM multicast protocol to provide reliable multicast service.

For RMP, each sender assigns data packet a sequence number, starting from zero; and PIM routers located at the branches are selected as NAK Responders for providing reliable services. NAK Responders are acted as a sender's packet reservoir, they provisionally stored up the sender's packets into its buffer. Group member received the packets from sender must be in order. When a receiver found the data received at out of sequence, it responds with NAK packet for each group to which it belongs. In order to avoid a flurry of NAK packets, each receiver starts a randomly chosen NAK packet delay timer for each of its group memberships. If, during the delay period, another NAK packet is heard for the same group, the local receiver resets its timer to a new random value. Otherwise, the receiver transmits a NAK packet to the reported group address, causing all other members of the group to reset their NAK packet timers. This procedure guarantees that NAK packet are spread out over a period of time and that NAK packet traffic is minimized for each group with at least one member on the subnetwork. Once the NAK Responder received the NAK packet, it examines

[4] State includes such field as the source address, the group address, the incoming interface from which packets are accepted, the list of outgoing interfaces to which packets are sent, timers, flag bits, etc.

its buffer whether having the requested packets firstly; if the requested packets are still alive at the buffer, NR sends the requested OData packets to its group members. Otherwise, NR conducts the data recovery process for getting back the requested packets from the sender or other NRs located at another branches so as to provide the reliable service with group members.

The following sections describe RMP operation in more detail.

2.2. SNRM Process

SNRM Message is periodically originated from group members to Rendezvous Point (RP). When passing through the PIM routers between DR and RP, SNRM Message collects those PIM router's state and merges it together. Those PIM router's state are eventually analyzed by RP to establish the Branch Path Message.

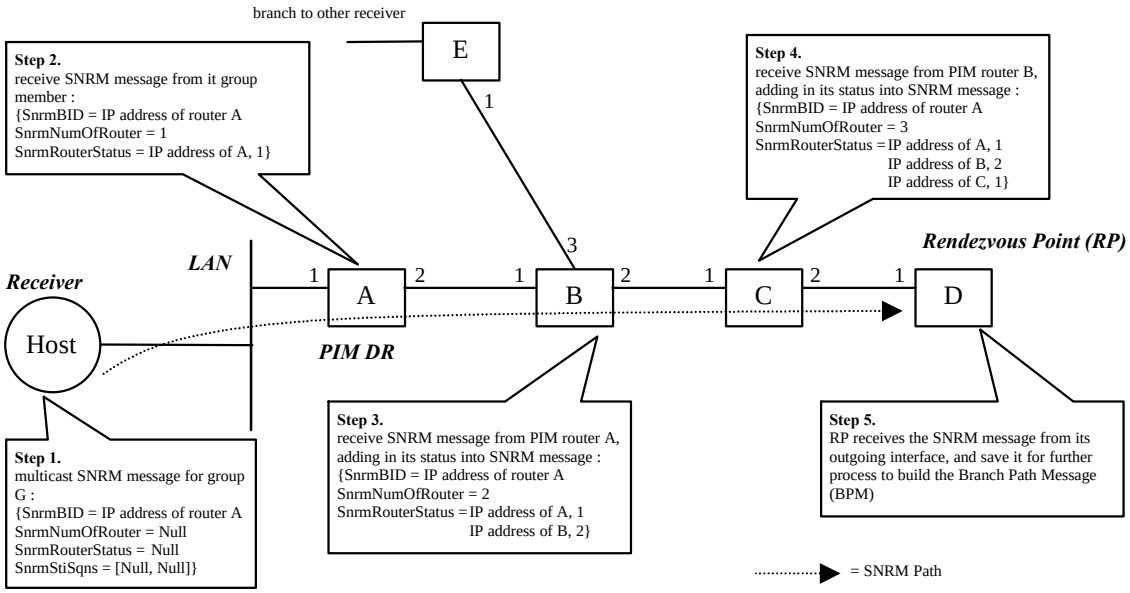


Figure 1. SNRM process at different elements of RP-rooted tree

2.2.1. Procedure for Receiver

Just building up the fledgling RP-rooted shared tree, receivers or group members multicasts Search-NAK-Responder Message (SNRM) at the subnetwork. Receivers suppress SNRM transmission for which a matching SNRM message is received during the SNRM transmit back-off interval. This mechanism can make only one SNRM message sent to the Designate Router for minimizing the traffic.

Receivers can obtain the IP address of the Designated Router from the periodical IGMP Membership Query such that parameter SnrmBID of SNRM message can be set to this IP address as Branch Identifier (BID). Other parameters, which consist of SnrmNumOfRouter,

SnrmRouterStatus and SnrmStiSqns, included in the SNRM are initialized to Null value. Receivers have to periodically originate SNRM for tracing the status of the branch.

2.2.2. Procedure for Intermediate PIM Router

Designated Router (DR) received the SNRM message from its group member, it need to examine the parameters included in the SNRM message. Once found the SnrmBID is null value, DR replaces the null value with its IP address and update the SnrmNumOfRouter parameter to 1 – it means that only one router is found. SnrmRouterStatus records down the PIM router status consisting of IP address and number of outgoing interface itself.

According to the PIM-SM protocol, Designated Router (DR) is periodically sent the PIM Join/Prune Message towards RP to adapt the change of the upstream router. SNRM message is then sent to RP as well.

Intermediate Routers adds in its status into SNRM message coming from the downstream PIM routers, and then forwards it to the upstream routers until reaching RP.

Those information from each PIM routers are merged into the SNRM that eventually reached the RP. Obviously, each SNRM has a unique SnrmBID for identifying from where it come.

2.2.3. Procedure for Rendezvous Point

RP stored up the SNRM received from its outgoing branches and then selects the appropriated PIM routers as NAK Responders (NR) for each branch on the basis of the receiving SNRM Message. The selection mechanism is that PIM router near the RP is selected as the NR firstly and is also set as the default NR corresponding to that outgoing interface or multicast path. The reason why the first PIM router leaving RP is selected as the default NR is the first PIM router must be able to receive all multicast packets came from RP if RP is able to receive all multicast packets and there are no link failure between RP and the first PIM router. Therefore, the first PIM router leaving the RP is a more reliable point for caching the source's packets. Secondly, PIM routers having more outgoing interfaces are chosen as the alternative NRs.

2.3. Establishment of Branch Path Message

Rendezvous Point (RP) receives the SNRM message coming from its outgoing interface to establish the Branch Path Message (BPM). BPM contains information for announcing which PIM routers are selected as NAK Responders. RP has to deliver the BPM to its group members within the RP-rooted tree periodically.

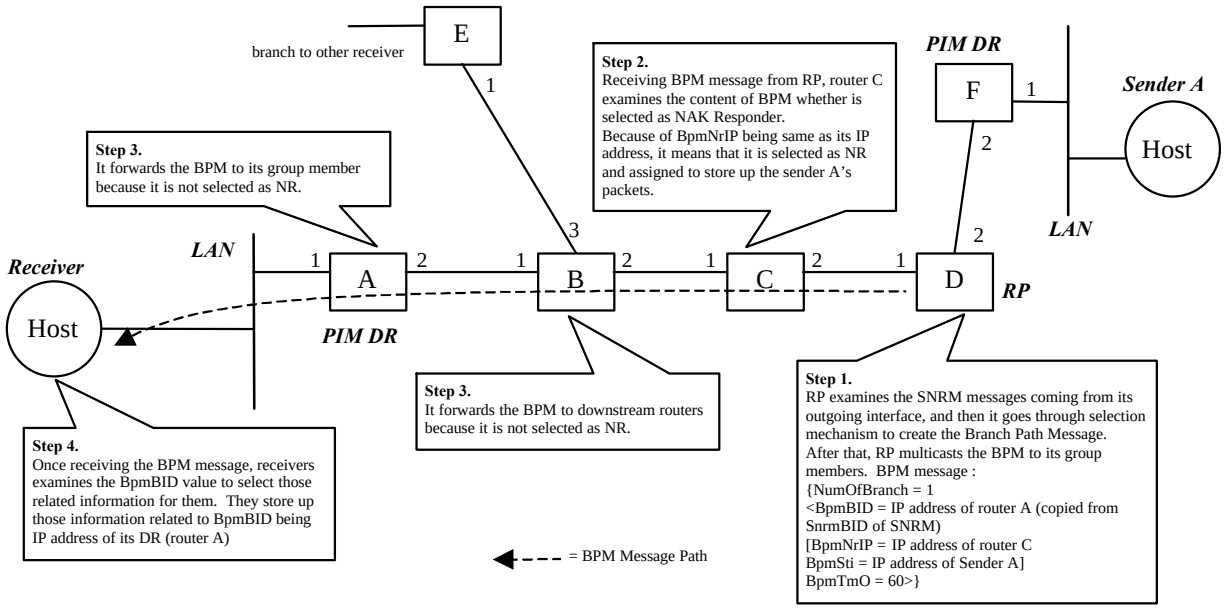


Figure 2. RP multicasts BPM message to its group members

2.3.1. Procedure for Rendezvous Point

Rendezvous Point processes the selection mechanism to single out the appropriated PIM routers as NAK Responders based on the received SNRM message.

Number of the NR chosen for each branch or each BpmBID is limited by the parameter MaxNrForBranch, which is set to 2 at initial state. Router C near the RP is firstly selected as NR and router B having more outgoing interface is chosen as the alternative NR. So, RP can construct the following state information (Table 1), and PIM routers marked with symbol '*' is the default NR for the corresponding branch. Each NR can be assigned to cache packets from one senders or more senders, it depends on the parameter MaxNrStoreForSender, which

is set to 1 at the initial state. However, RP can change this parameter to adapt the PIM tree change, which includes senders or receivers joining or leaving the group.

Branch ID (BpmBID)	IP Address of NR located at this branch (BpmNrIP)	NRs responsible for which Sender (BpmSti)	Timeout Value for this branch (sec) (BpmTmO)
BH1 (router A's IP)	router C's IP*	none	60
	router B's IP	none	
•	•	•	•
•	•	•	•
•	•	•	•

Table 1. BPM state stored at the RP

At the initial state or RP-rooted tree just built, no sender is joined to this group or RP so that parameter BpmSti may be null value – it means that the assigned NR does not need to store any multicast packets because of no sender being joined to RP-rooted tree. Parameter BpmBID and BpmNrIP shown at the BPM is to identify the branch at which the NRs are located and IP address of the selected PIM router as NAK Responder respectively. Those information related to the specific BpmBID has a time out value BpmTmO for keeping check the availability itself. When this value reached to zero, the corresponding NR and its related information is thrown out from BPM unless the new coming SNRM message updates the BPM to reset the BpmTmO value.

Once a sender A joins the RP or group, it sends the PIM Register Message[5] to RP. At this moment, from the state information, RP appoints the available NRs for performing the storage of the sender's packets such that it results in updating the state information shown at Table 2.

[5] At the PIM protocol, when sender first sends its packets to a group, its DR unicasts Register messages to the RP with the source's packets encapsulated within.

Obviously, PIM router C is assigned to provide reliable service for sender A's multicast packets.

Branch ID (BpmBID)	IP Address of NR located at this branch (BpmNrIP)	NRs responsible for which Sender (BpmSti)	Timeout Value for this branch (sec) (BpmTmO)
BH1 (router A's IP)	router C's IP*	Sender A's IP address	60
	router B's IP	none	
•	•	•	•
•	•	•	•
•	•	•	•

Table 2. State Information is updated after new sender joined

After updating the state information (Table 2), RP conveys those valid information from the state information passing through its outgoing interfaces towards all joined group members. This information hereafter is called as Branch Path Message (BPM) (Table 3).

Branch ID	IP Address of NR located at this branch	NR responsible for which Sender
BH1 (router A's IP)	router C's IP*	S1
•	•	•
•	•	•
•	•	•

Table 3. BPM Message sent to group members

The selection mechanism used by RP to appoint or allocate the NRs for saving the sender's packets is depended on the number of sender joined to the group or RP. At the above situation, only one sender A is joined to the RP so it needs one NR for each branch also. Later, there will mention another cases how effective the RP is to allocate the NRs for handling more senders.

2.3.2. Procedure for Intermediate PIM Router

When the intermediate PIM routers located at branch received this BPM message, each PIM router examines itself whether is selected as NR. If the PIM routers found their IP's address appeared at the BPM Message, it means that RP choose it as NR. Those PIM routers selected as NR have to store this BPM Message such that they know that which sender's packets they are be bound to store and know which sender's packets other NRs stored. For those PIM routers not to be selected as NR, they do not need to save this BPM message and forward it to next node along the branch.

2.3.3. Procedure for Receivers

When group members received the BPM message, they have to store this information also. By checking the branch ID (BpmBID), group members can find out who are its NRs, which are located at the local branch, for providing reliable service. Eventually, once group members found any discrepancy at the sequence number of the received packets, they can originate the NAK packet to its NRs for requesting packets retransmission.

2.3.4. Steady State Maintenance of BPM message

In the steady state group members send periodic SNRM to capture state, topology, and membership changes for RP establishing or updating the state information to create new BPM message. A SNRM message is also sent on an event-triggered basis each time a new branch is established for new group members joined. Once receiving the new SNRM messages from its branch, RP examines those messages whether having any new information. If those received SNRM messages are same as the last received one, RP can ignore those SNRM messages but it has to reset the parameter BpmTmO of BPM to 60 seconds; and then originates the updated BPM message to its joined group members. However, if the received SNRM messages demonstrates there are some new group members joined or topology change, RP necessarily reconstruct the state information and then delivers the updated BPM message to its joined group member as well. Apart from that, state information is also updated by the new senders joined.

2.4. Reliable Service for One Sender

The following diagram (Figure 3) demonstrates the sequence of packets recovery when receiver found packets lost.

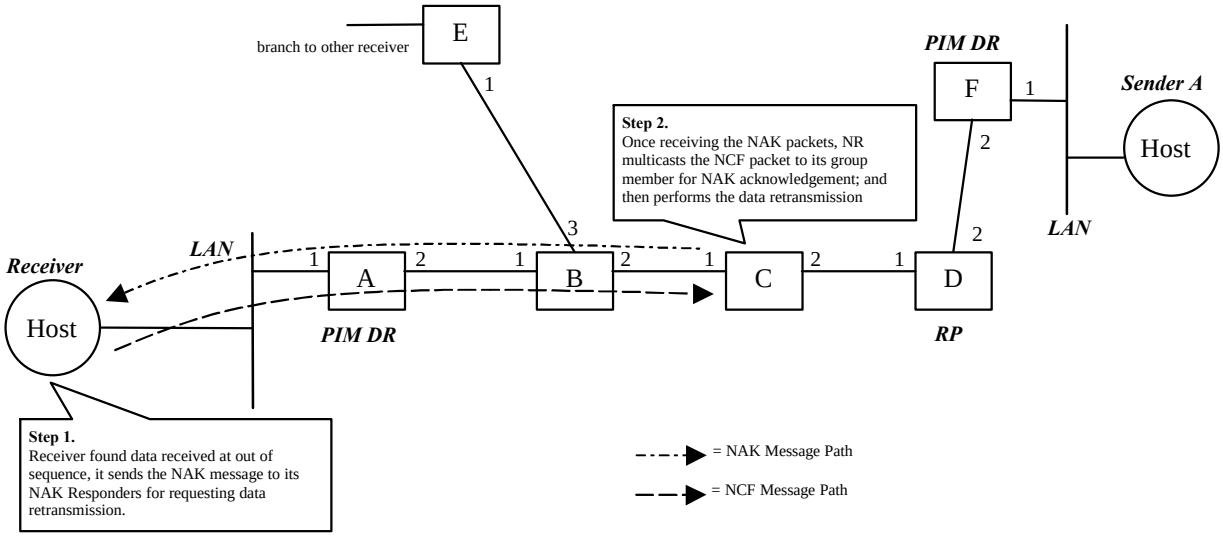


Figure 3. Receiver sends NAK packet to its NR

2.4.1. Procedure for Receiver

Data packets received by the receiver or group member have the sequence number encapsulated within. When group members found there are any discrepancy found at the sequence numbers of the received data packets, they responds with NAK packet for each group to which it belongs. Before sending the NAK packet, receiver has to find out its NR from the BPM message stored at its buffer first. Receiver periodically sends the NAK packet to its NR at interval T_{NAK} until they get the NCF packet for NAK acknowledgement from its NR. T_{NAK} is dynamically adjusted based on the time-stamp recorded at the NCF packet.

2.4.2. Procedure for NAK Responder

After receiving first NAK packet, NR do not transmit the requested packets immediately because one or more receivers may miss the same packet. So, they wait a pre-defined delay time (TimeRxSend) for receiving all NAKs corresponding to same packet and then determine

whether the lost packet should be retransmitted using unicast or multicast retransmission. Three parameters are included in the design for this purpose : TimeRxSend, MulticastThresh and a retransmission queue (ReTxQueue). Within TimeRxSend period, the NR processes NAKs from the group members in its local branch. The sequence numbers of the requested packets stored at the NAKs are added to the retransmission queue (ReTxQueue). A retransmission queue element contains the sequence number of a packet to be retransmitted. A counter (PacketCount) that counts the number of group members that have requested the packet, a table NakAddrTable that records the requesting group members' IP addresses. At the end of interval TimeRxSend, the NR checks the PacketCount value; if the value is larger than or equal to the MulticastThresh parameter, NR delivers the packet using multicast; otherwise, the NR delivers the packet to each receiver in NakAddrTable using unicast. The default setting of TimeRxSend and MulticastThresh is 20ms and 3 respectively.

When the requested packets are still alive at the NR's buffer, it is directly sent to the corresponding group members; otherwise, NR conducts one of two recovery mechanisms to provide the reliable service. There are two recovery mechanisms for handling those NRs which have not the requested lost packets stored. Selection of the recovery process from both is depended on the role of the NR.

Each branch must have one NR assigned as default NR. In fact, it is impossible or not guarantee that each senders have its corresponding NR for caching their data at each local branch, therefore, the function of the default NR is used for handling all NAK packets sent from those group members who does not find the corresponding NRs caching their requested packets at the local branch. At this situation, once default NR receives the NAK packet and found the NAK packet requested the sender's packet that NR is not appointed to cache. Default NR will conduct one recovery mechanisms finding another NRs, which locate at the

another branch and store the requested sender's packet also, from its BPM message firstly and then unicasts a NAK packet to that NRs for getting the requested packets.

Under another circumstance, the NR or the default NR is assigned for caching the sender's packet but they really have not the corresponding packets stored at its buffer due to some reasons. At this situation, those NRs will not look for another NRs from its BPM message to request the wanted packets; alternatively, they will unicast the NAK packets to the corresponding sender directly to request the packet they wanted. At this situation, the RMP protocol assumes the default NR is the reliable point for caching the packets came from the RP. Once this reliable point cannot receive the requested multicast packets it means that all other points cannot do also. So, the default NR requests the reliable service from the senders directly instead of from another NR located at the PIM tree.

As described before, senders have the opportunities to receive the NAK packets from the NRs, however, sender bases on the number of the received NAK packet to determine the packets delivery whether using unicast or multicast transmission. If there are too many NRs losing the requested packets, sender delivers it through multicast transmission; otherwise, it uses unicast service for sending the requested packets to those NRs. When receiving the response from the sender, the NRs will perform the recovery process as usual.

At RMP protocol, NRs cache the multicast data for providing the reliable services, however, size of memory or buffer is one of the crucial item affecting the performing of this protocol. There are two ways to clear up those obsolete packets. One of them helps from the receiving NAK packets. Once NR receives the NAK, all packets marked with the sequential number below the sequential number shown at the requested lost packet can be removed from the buffer for spacing more room to store the new coming packets. The second way is that

SNRM message has also included the recent sequential number of received packets for each STI, because SNRM message periodically sent by receivers not only provides node information at the local branch with RP establishing BPM message, but also could inform the NRs for removing some obsolete packets. Therefore, once receiving the SNRM message which includes the last continuous sequential number of the received packet that receiver had in order received all packets marked with the sequential number below, NR cleared those packets whose sequential number is below that last sequential number stored at the SNRM message for getting more room to store the new coming packets also.

2.5. Reliable Service for More Senders

Refer to the same PIM tree, there are two more senders – Sender B and Sender C joined the same group or RP (Figure 4).

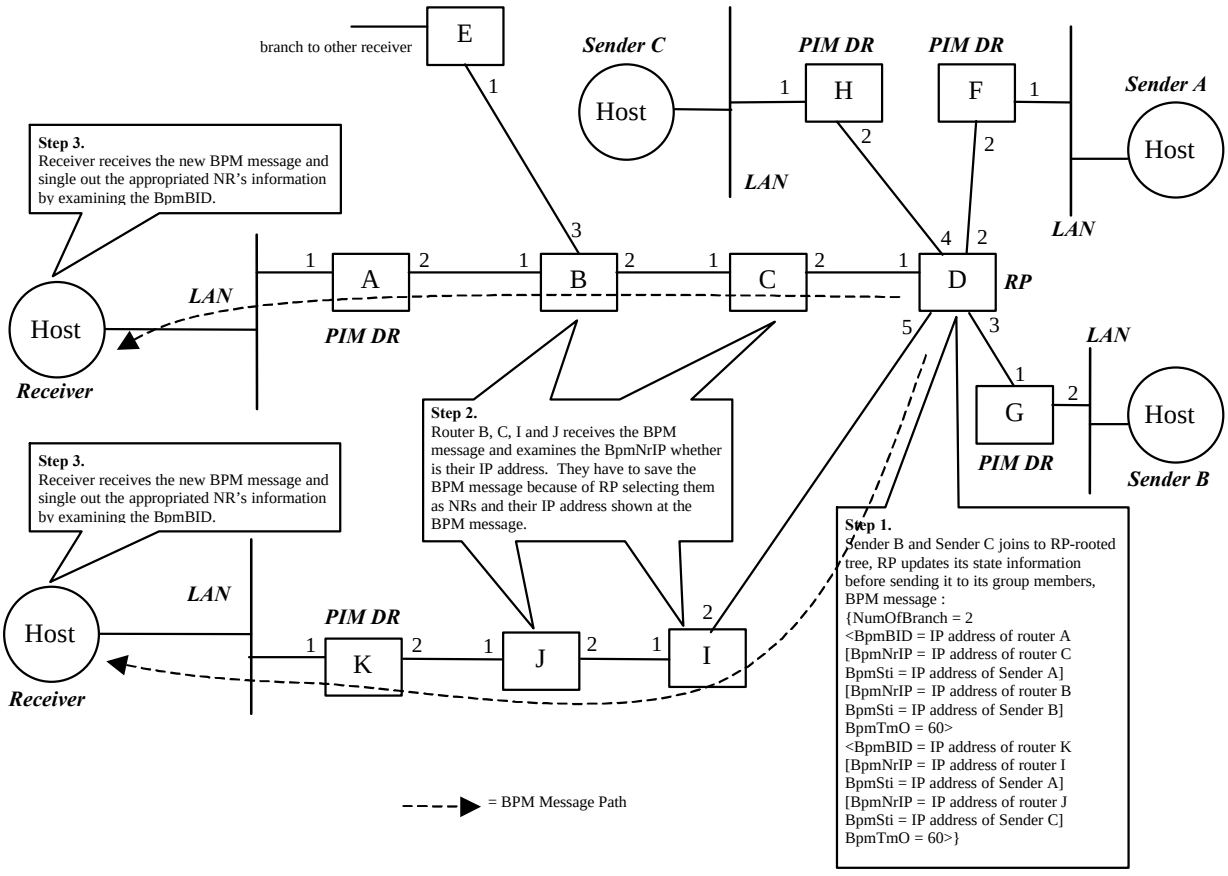


Figure 4. Two more senders join the RP-rooted tree

2.5.1. Procedure for Rendezvous Point

When RP received the PIM Register Messages from new senders, it updates the state information, which had already stored at its buffer. Before transmitting the BPM message to replace the old one stored at each group members and NRs, RP examines the number of joined senders whether is smaller than or equal to the number of the available NRs for each branch. If there are enough NRs at each local branch for handling all senders, it will be easy for RP to assign the NRs for saving all multicast packets at the same local branch; in other words, reliable service can be provided with group members at the local branch.

Otherwise, RP have to go through the selection mechanism to re-allocate the NRs, which are located at each branch, for selectively saving multicast OData packets.

At the above case, each local branch only have two NRs such that RP have to re-appoint the NRs for selectively handling the multicast packets of each sender. At the previous state information, there is only one NRs (PIM router C) for handling Sender A's multicast packets. At the new state information (Table 4), PIM router B is assigned for storing up the Sender B's packets; and PIM router I and J is appointed to backup the Sender A and C 's packets respectively.

Branch ID (BpmBID)	IP Address of NR located at this branch (BpmNrIP)	NRs responsible for which Sender (BpmSti)	Timeout Value for this branch (sec) (BpmTmO)
BH1 (router A's IP)	router C's IP*	Sender A's IP address	45
	router B's IP	Sender B's IP address	
BH2 (router K's IP)	router I's IP*	Sender A's IP address	45
	router J's IP	Sender C's IP address	
. . .	. . .	. . .	. . .

Table 4. New state information stored at the RP

However, the parameter BpmTmO is not updated in this case. This parameter is only updated by triggering from the SNRM messages, which are sent from group members when new group members joins to RP.

After updating the state information, RP sent out the new BPM message to its group members through multicast transmission for replacing the old BPM message stored at each group member and NRs.

2.5.2. Procedure for Intermediate Router

Intermediate PIM routers examines the BPM message originated from RP whether is selected itself as NAK Responders. They have to check the BmpNrIp parameter whether is same as its IP address. If matching, PIM routers are assigned as NAK Responders and need to take the responsibility to store up sender's multicast OData packets; otherwise, PIM routers just forward the BPM message to the downstream PIM routers.

2.5.3. Procedure for Receiver

Once receiving the new BPM message, receivers inspect the BpmBID of BPM message first to find out which NR information are for them; and then update the BPM message. When receivers found any discrepancy at the sequence number of the received multicast OData packets, they based on the stored BPM message to send the NAK message to the corresponding NAK Responder for requesting data recovery.

In case there are no NAK Responder for storing up the requested sender's multicast packets, receivers can send the NAK packets to the default NAK Responder for requesting the data recovery.

2.5.4. Procedure for NAK Responder

In case of PIM router selected as NAK Responder, they have to setup a finite queue for storing up the corresponding senders' multicast OData packets. When NAK Responder received the NAK packets from the group members, they multicast the NCF to its group members firstly and then examine themselves whether storing up the requested packets at their finite queue. If found, NAK Responder delivers the requested packets through unicast or multicast transmission depended on the data recovery mechanism. If the requested packets do not exist, NAK Responder can examine the stored BPM message to find out which NR having stored the requested packets; and then they can sends the NAK message to other NRs located at the another branch or sender to request the lost OData packets. Once the NAK Responder takes back the requested packets from another NR, they may perform the remaining recovery processes as usual (Figure 5). Basically, the action taken by the NRs is same as description at the before section – Reliable Services for One Sender.

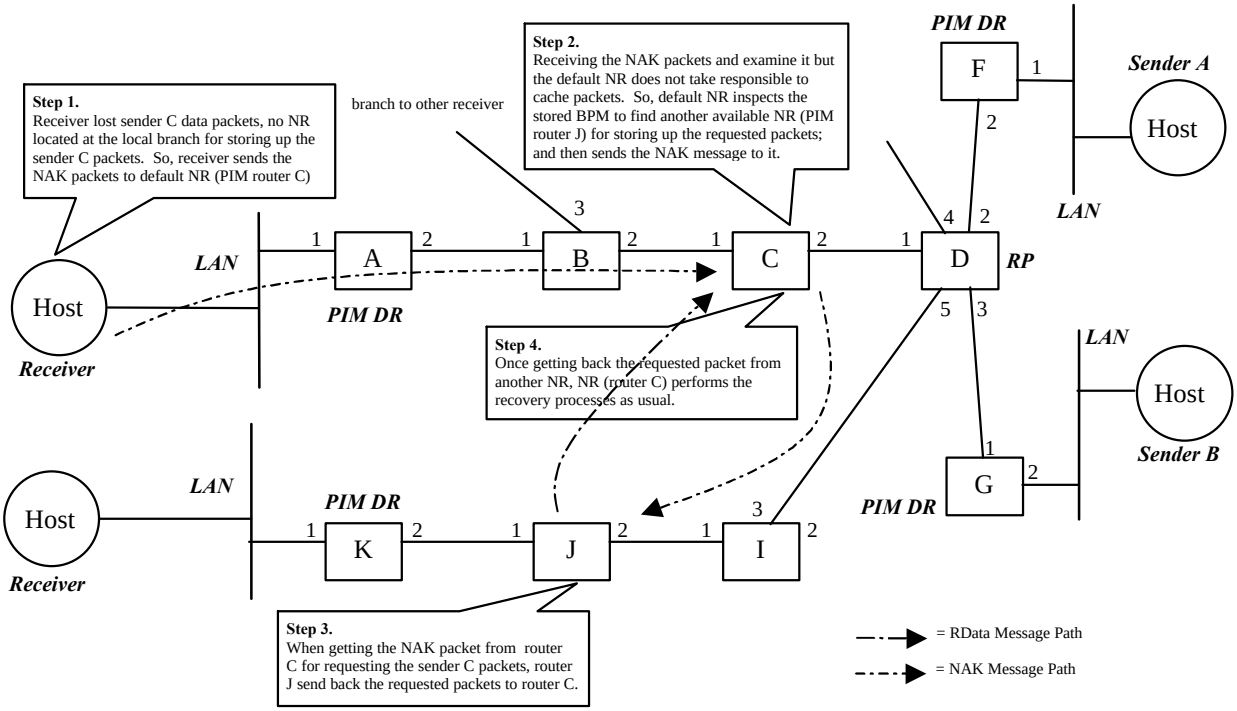


Figure 5. NAK Responder requests the multicast packet from another NR

2.6. New Group Members Join to RP

Apart from the state information stored at the RP is updated by triggering from the new senders joined or reaching to zero value of parameter BpmTmO corresponding to each branch. State information is updated in respond to the new group members joined the RP-rooted tree also.

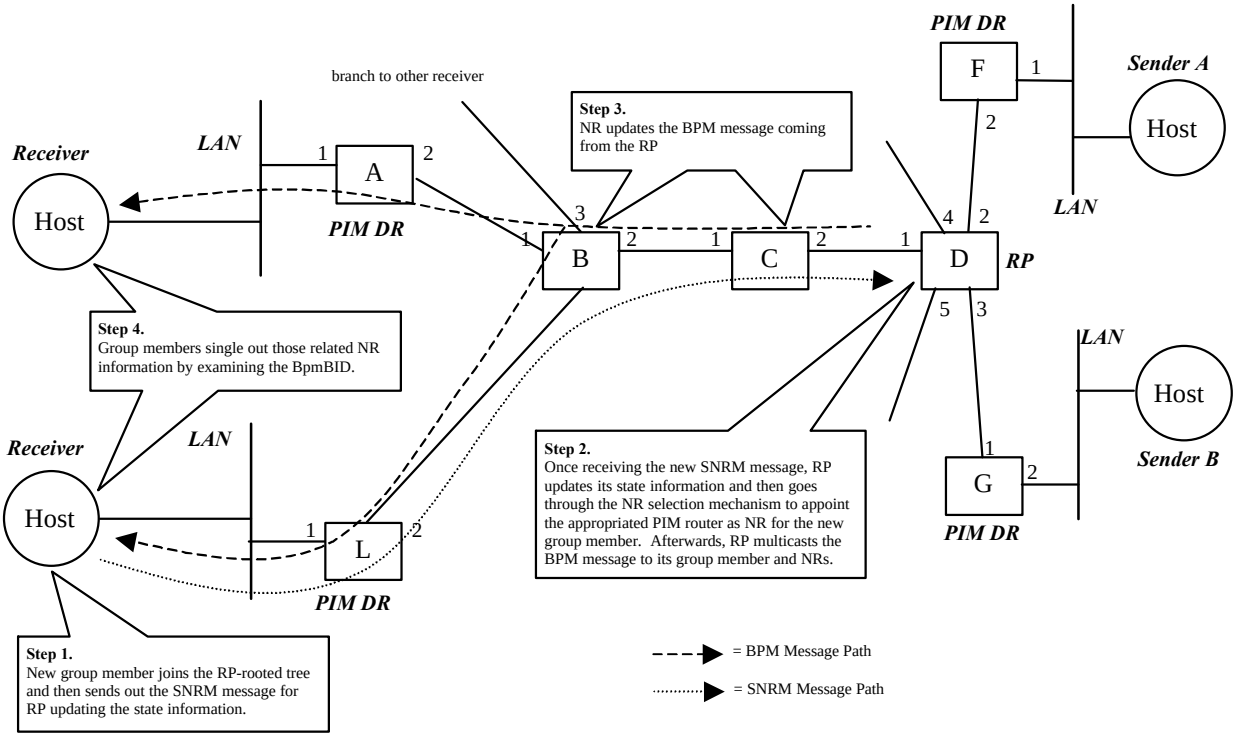


Figure 6. New group members joins the RP

2.6.1. Procedure for Receiver

Receivers joining the RP-rooted tree send the SNRM message along the branch towards RP to advertising new group members joined or branch structure change.

Later, receivers or new group members receive the BPM message and then select those related NR information by examining the BpmBID parameter of BPM message.

2.6.2. Procedure for Rendezvous Point

Rendezvous Point receives the new SNRM message coming from the new group member, it has to go through the NR selection mechanism to appoint the appropriated PIM router as NR for handling the data recovery process of the new group members. New state information is updated shown at the Table 5, obviously, PIM router C and B are assigned as NRs for the new group member or receiver.

Branch ID (BpmBID)	IP Address of NR located at this branch (BpmNrIP)	NRs responsible for which Sender (BpmSti)	Timeout Value for this branch (sec) (BpmTmO)
BH1 (router A's IP)	router C's IP*	Sender A's IP address	45
	router B's IP	Sender B's IP address	
BH2 (router K's IP)	router I's IP*	Sender A's IP address	45
	router J's IP	Sender C's IP address	
BH3 (router L's IP)	router C's IP*	Sender A's IP address	60
	router B's IP	Sender B's IP address	
•	•	•	•
•	•	•	•
•	•	•	•

Table 5. New state information stored at RP after new receiver joined RP-rooted tree

Parameter BpmTmO of new added branch is set to initial value – 60 seconds but those value for the existing branches are not necessarily to be updated until RP received their corresponding SNRM messages.

Obviously, PIM router C and B are already existing at the tree for providing reliable services of sender A and B respectively, and the new established branch (BH3) also contains both PIM routers so that RP uses existing available resource as NRs and then updated its state information to multicast the new BPM message to its outgoing interface.

If PIM routers located at the new branch are not shown at the current state information, RP goes through the selection mechanism to choose the appropriated PIM routers as NRs and assigns those selected NRs to provide reliable service for this new branches.

However, the late joined group members can get back all multicast OData packets from its NRs or default NR depended on the requested packets whether are saved at its NRs.

2.6.3. Procedure for Intermediate Router

Intermediate PIM routers inspect the new coming BPM message by checking their IP address whether exists the BPM message inside. If found, intermediate routers are selected as NAK Responder and they have to setup the finite queue to store up the corresponding sender's OData packets.

2.6.4. Procedure for NAK Responder

NAK Responders receive the new coming BPM message by examining the BpmNrIP parameter also. They need to know themselves whether getting rid of the NR listings or is assigned to store up another sender's multicast data packets.

2.7. Sender leaving the RP

When sender leaves the group or RP, RP updates its state information by re-allocating the NRs for each branch. After that, RP multicasts the new BPM message to group member and NRs for informing them there are senders leaving the group so that each group members or NRs based on the new received BPM message to update their information.

2.8. Designated Router leaving the RP

Once there are no group member attached to the Designed Router (DR), DR leaves the group or RP, such that there are no periodical SNRM message sent to RP for refreshing the state information. When the parameter BpmTmO reaches zero value, the information of this branch is thrown out from the state information. RP has to re-allocates the NRs for performing reliable services of the current senders and then multicasts the new BPM message derived from the new state information to group members or NRs for replacing their outdated information.

However, once found there are no any active group members joined, Designed Router (DR) sent the PIM Prune Message to RP for advertising to cut that branch. This Prune Message will also trigger RP to update its state information again such that the new BPM message is multicast to its group members also.

2.9. Reliable Service at Shortest Path Tree

According to the PIM-SM Protocol, a router with directly-connected members first joins the shared RP-tree, the router can switch to a source's shortest path tree (SP-tree) after receiving packets from that source over the shared RP-tree. It can see that, from Figure 7 below, DR (PIM router K) changed to the SP-tree from RP-rooted shared tree so that the branch towards

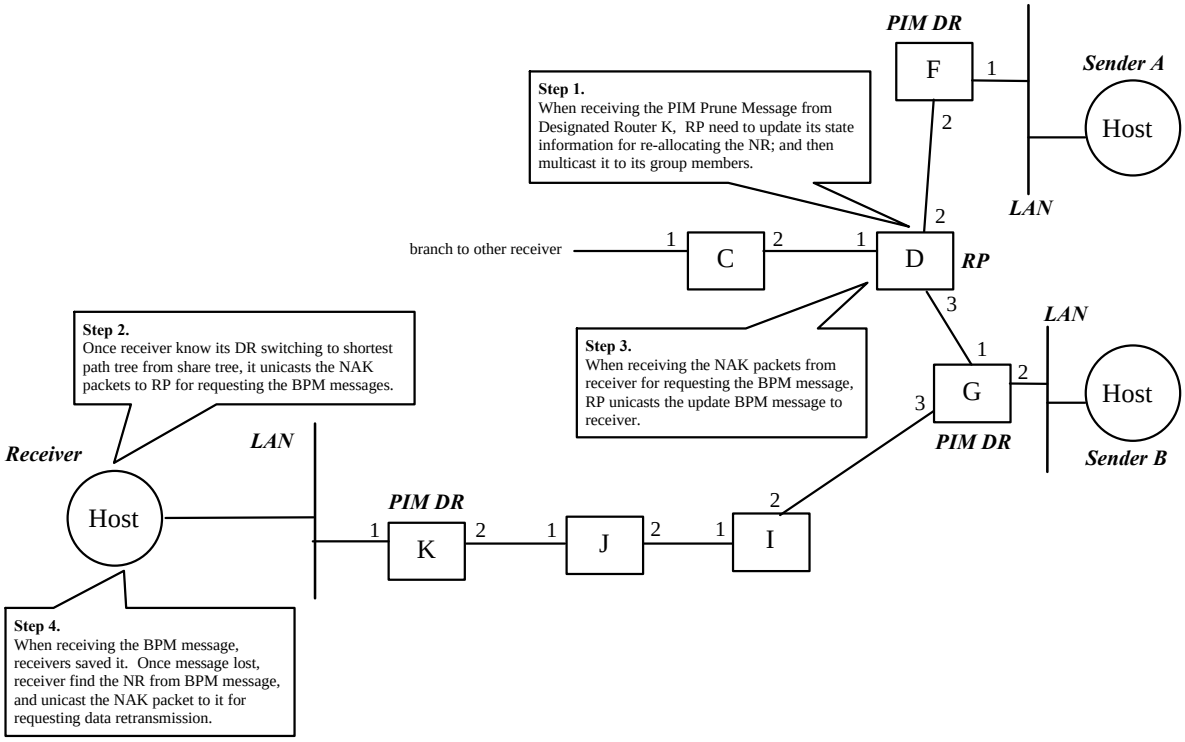


Figure 7. Reliable service at shortest path tree

RP is pruned and new shortest path connected to the sender B is established. Once Designated Router K switched to the shortest path tree, apart from the receiver attached to this Designated Router, all other group members receiving the packets are same as before but receiver attached to Designated Router K got the packets through the new established shortest path.

2.9.1. Procedure for Receiver

At shortest path tree, receiver does not have its NRs for providing the reliable service because the branch towards RP is pruned and the intermediate routers could not receive any multicast OData packets from RP.

For RMP, at this situation, this switching mechanism triggers receiver sending NAK packet through unicast transmission to RP for getting the BPM message, which is used to convey information about the NRs' location of the RP-rooted tree. When losing the OData packets, receiver can examine the BPM message to find out which NRs stored the corresponding sender's packets that it needed firstly and then unicasts the NAK packet to get back the requested packets from that NRs for data recovery. In case of receiver receiving null BPM message from RP, it means that there are no any NAK Responder distributed over the RP-rooted tree, such that receiver alternatively send the NAK packet to sender for getting the lost OData packets.

2.9.2. Procedure for Rendezvous Point

Once rendezvous point found PIM Prune Message with zero value of RPT-bit[6] for switching the branch to shortest path, it updates the state information and then multicasts the new BPM message to its group members.

[6] The RPT-bit is set to indicates that this join is being sent up the shared, RP-tree; otherwise, this join is being setup up through shortest path tree.

2.9.3. Procedure for Intermediate Router

Once the branch is pruned from the RP, the PIM routers located at that branch cannot receive the BPM message sent from the RP to construct the NR selection mechanism. In other words, group members does not know any NRs located at this shortest path towards sender. Designated Router (DR) at that branch will then inform its group member about that branch to be switched to the shortest path such that its group members can take appropriated actions to get back BMP message from RP.

Chapter 3 - Term and Entry Descriptions

This section will make more clear of some RMP's Messages and terms that have been mentioned at the previous sections. The functionality of entries of participants at the RMP will also be discussed.

3.1. Search NAK Responder Message (SNRM)

The principle function of SNRM is to provide the state information of PIM routers within RP-rooted tree for RP establishing the Branch Path Message required for group members to send the NAK packet to the appropriated NAK Responders for requesting reliable service.

Group members periodically multicast the SNRM message to request the NR information from RP. However, receivers suppress transmission for which a matching SNRM transmission is received during the transmit back-off interval.

The last-hop PIM router (DR) receiving the SNRM message from its downstream group members add the Branch Identifier (BID) in it, and forward it to next upstream PIM router. The SNRM message is upstream forwarded along the branch until reaching the Rendezvous Point (RP), state information of PIM router that the SNRM is passing through is merged into the SNRM message. RP based on the received SNRM message from its outgoing interface to select the appropriated PIM routers as NRs for performing reliable service.

SNRM messages passing through each PIM router at the branch grasp the state information consisting of number of outgoing interface of PIM router and PIM router's IP address.

Obviously, periodical SNRM messages sent to RP are used to update the knowledge of PIM tree change for RP building the new BPM message.

3.2. Branch Path Message (BPM)

BPM created by RP is used to advise some PIM routers of being selected as NAK Responder and inform group members about the location of NAK Responders at the branch.

Rendezvous Point (RP) creates the Branch Path Message (BPM) based on the received SNRM message sent from its group members. Once receiving new SNRM message, RP goes through the NR selection mechanism to update the BPM message and then multicast it to its group members within the RP-rooted tree.

BPM consists of Branch Identifier (BID), IP address of NAK Responder, STI and other protocol parameters. BPM message is multicast along each outgoing branch of RP, PIM router located at each branch must examine the BPM message by checking the parameter 'IP address of NR' whether matching with its IP address. If PIM router found its IP address appearing at the BPM message inside, it means that PIM router is selected as NR for performing reliable service at that branch; otherwise, PIM router does not need to do anything.

BPM along the branch towards the downstream group members, group members examine the BPM to single out those information that they concerned. Group members based on those information to assist itself to locate the appropriated NR for requesting reliable service when they wanted.

BPM is periodically updated by RP in respond to the new SNRM message. In order to minimize the traffic of BPM, RP does not generate or multicast the BPM message if there are no any updated information found at the new coming periodical SNRM messages.

3.3. Source Transport Identifier (STI)

RP-rooted tree is specified for handling multiple sender data delivering service within the same multicast group. Each sender joining with the RP-rooted tree should have a unique Source Transport Identifier (STI), which is embedded into those packet that it sent out. OData packet delivered from the sender should include STI inside, even retransmission packet RData contains this STI also for identifying the source.

Since all NAK packets originated by receivers are in response to missing data packet originated by a source, receivers simply add the STI into the NAK packets so as to get back the corresponding source data.

NCFs packets originated by the NR are in response to NAK packets delivered by receivers, NR simply embeds the STI heard from the NAK packets into the NCFs that should be delivered to those receivers sending the corresponding NAK packets for acknowledgement.

Rendezvous Point (RP) is the core of the PIM tree, all multicast data from multiple senders are flowed into it first. RP needs to record down the STI of all joined senders because STI has to be encapsulated into the Branch Path Message (BPM), which is originated by RP to advise of the appropriated PIM routers caching the corresponding sender packets.

3.4. Branch Identifier (BID)

Branch between the joined receivers and the Rendezvous Point (RP) is the data path for carrying the multicast packets. Each branch has its unique Branch Identifier (BID), which is embedded into the Search-NAK-Responder Message (SNRM) originated by the last-hop PIM router (DR) toward RP. RP receives SNRM messages from its outgoing interface, and then establishes the Branch Path Message (BPM) based on those received SNRMs. The BPM is multicast to all group members within the RP-rooted tree.

Downstream group members under each branch, they know the BID from its last-hop PIM router (DR). Therefore, after receiving the BPM message, group members rule out all unrelated information by examining the BID stored at the BPM to hunt out those information that they need.

3.5. NAK Responder (NR)

Rendezvous Point based on the SNRM messages coming from its outgoing interface to select the appropriated PIM routers as the NAK Responders (NRs) at each branch. NR at each branch performs the data retransmission once it gets the NAK packets.

Rendezvous Point (RP) performs the NR selection mechanism to appoint the appropriated PIM routers as NR Responders. PIM router at each outgoing branch near the RP is selected as the first NR and is also set as a default NR corresponding to that branch. Secondly, more outgoing interfaces of the PIM router is chosen as the alternative NR. Total number of NR selected is depended on the protocol parameters and how many available PIM routers located at the branch also. However, RP can relocate the NR to adapt for the PIM tree change, which includes senders or receivers joining or leaving the group.

Each NAK Responder is assigned to store one or more sender multicast packets. STI embedded in the multicast packets sent along the branch are examined by the NR to investigate whether have to be stored at its buffer.

Default NR is used for handling all NAK packets sent from those group members who do not find the corresponding NRs caching their requested lost packets at the local branch. At this situation, default NR conducts another recovery mechanism to get back the requested packets for providing the reliable services.

NR executes the packet retransmissions in respond to NAK packets sent from its downstream group members. They can also provide the data packets in respond to NAK packets emerged from other NRs located at another branch.

3.6. Sequence Number

RMP uses a circular sequence number space from 0 through (2^{32} - 1) to identify and order OData packets. Sources must number OData packets in unit increments in the order in which the corresponding application data is submitted for transmission. Within a transmit (defined below), a sequence number x is "less" or "older" than sequence number y if it numbers an OData packet preceding OData packet y, and a sequence number y is "greater" or "more recent" than sequence number x if it numbers an OData packet subsequent to OData packet x.

3.7. Leaky Transmission at Source

The concept of leaky transmission at the source is used for congestion control to avoid some slow receivers or low bandwidth links to be overwhelmed. Each source maintains its leaky transmission rate for OData transmission to its group members within the RP-rooted tree.

Each source has its Source Transport Identifier (STI) for identifying its multicast packet within the RP-rooted tree. Conceptually, each source is connected to RP-rooted tree by an interface containing a finite internal queue for holding the data coming from the application. OData packets stored at the finite queue are sent to Rendezvous Point (RP) at a clock-tick rate. The clock-tick rate can be adjusted in respond to the congestion condition of the RP-rooted tree. Once the internal queue is full, it will advise the application of suspending data delivery until it has the empty space being available again.

The definition is derived directly from a finite buffer (in bytes) that a source retains for retransmission (BufBytes), and the clock-tick rate sustained by a source for originating one packet per one clock tick (TickRate), and the maximum transmit rate TxRate (in bytes/second) maintained by a source sending data at each clock tick. The internal queue in sequence number (TxSqns) is (BufBytes / bytes-per-packet).

In terms of sequence numbers, the internal queue is the range of sequence numbers consumed by the source for sequentially numbering and transmitting the last packet of OData packets. The trailing (or left) edge of the internal queue (TxTrail) is defined as the sequence number of the oldest data packet available for retransmission from a source. The leading (or right) edge of the internal queue (TxLead) is defined as the sequence number of the most recent data packet a source has transmitted.

The fraction of the internal queue size (in number of packet) by which the internal queue is advanced (TxPackAdv) at a window advanced time (TxAdvSec) is called the window increment. The window advanced time can be also adjusted to adapt to the congestion condition of the RP-rooted tree.

In terms of sequence numbers, the increment window is the range of sequence numbers that will be the first to be expired from the internal queue. The trailing (or left) edge of the increment window is just TxTrail, the trailing (or left) edge of the internal queue. The leading (or right) edge of the increment window (TxInc) is defined as one less than the sequence number of the first data packet pointed by the TxTrail after TxPackAdv advanced time.

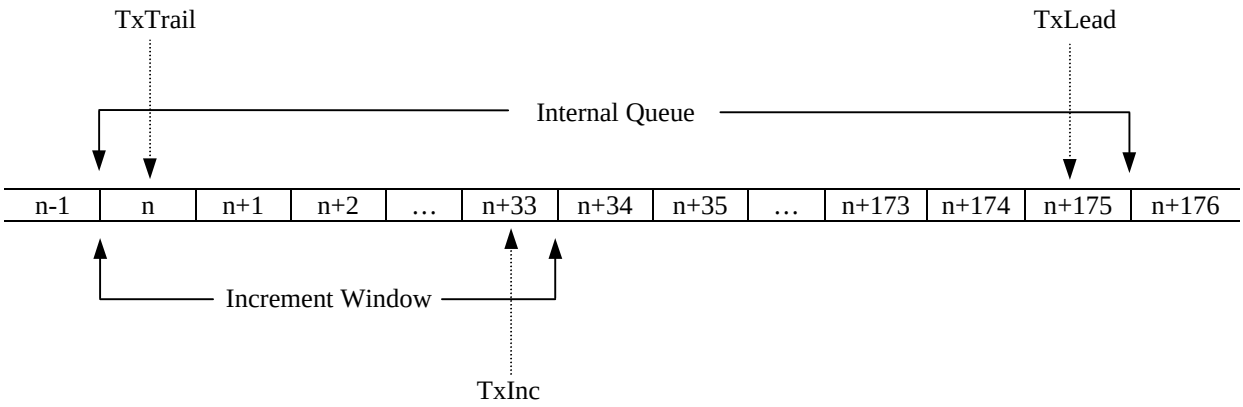
The internal queue is advanced across the increment window by the source at each TxPackAdv advanced time. When the internal queue is advanced across the increment window, the increment window is emptied (i.e., TxTrail is momentarily equal to TxInc), begins to refill immediately as transmission proceeds, is full again TxPackAdv later (i.e., TxTrail is separated from TxInc by TxPackAdv), at which point the internal queue is advanced again, and so on.

Consider the following example:

Assuming a constant internal queue 256 Kbytes and clock tick is 0.1 seconds, data packet size of 1500 bytes (includes 40bytes of IP and TCP header) and source advances 50 Kbytes once the last packets stored at the internal queue is sent, then

TxSqns = 256,000 / (1,500 – 40) = 176 (rounded up),
TxPackAdv = 50,000 / (1,500 – 40) = 34 (rounded down),
TickRate = 0.1 seconds,

so, the advanced time of Increment Window (TxAdvSec) should be (TxPackAdv * TickRate) 0.34 seconds. To continue this example, the following is a diagram of the transmit window and the increment window therein in terms of sequence numbers.



So the values of the sequence numbers defining these windows are :

TxTrail = n
TxInc = n+33
TxLead = n+175

However, once source receives the NAK packets from its group members, the clock-tick rate becomes twice so as to relieve the burden of NAK Responders or those communication link under congestion. If the sequence number shown at the NAK packets is within the Increment Window, the expiry time of Increment Window should be doubled also. If there are no any NAK packets received before advancing the Increment Window, the clock-tick rate and the expiry time of Increment Windows is reduced half until those values become to the initial settings.

So, for a source identified by a STI, a source maintains :

TxTrail	the sequence number defining the trailing edge of the internal finite queue, the sequence number of the oldest data packet available for retransmission
TxLead	the sequence number defining the leading edge of the internal finite queue, the sequence number of the most recently transmitted OData packet
TxInc	the sequence number defining the leading edge of the internal finite queue, the sequence number of the most recently transmitted data packet amongst those that will expire upon the next increment of the transmit window
TickRate	the clock tick rate defining one packet stored at the internal queue to be sent at a regular instant
TxPackAdv	to define how much of packets stored at the internal finite queue are thrown away at each Increment Window advancing
TxAdvSec	to define time interval of when advancing Increment Window

3.8. Finite Queue at NAK Responder

For a given Source Transport Identifier (STI), NAK Responder (NR) reserves internal finite queue for caching up the multicast packets came from Rendezvous Point (RP), those packets are used for packet recovery to its downstream group members or another NRs located at other branches.

Basically, there is also a Increment Window which is advanced at NRAdvSec time to spare memory to store the new coming multicast OData packet. Once the sequence number shown by NAK packet is within the Increment Windows, advancing time NRAdvSec has to be set double. If there is no NAK packet received or the requested sequence number is outside of Increment Window within NRAdvSec interval, advancing time NRAdvSec is reduced half until returning to initial value.

When internal finite queue is full and Increment Window is not advanced at this moment, the coming multicast packets will not be stored at the internal queue and it just pass through the NR to downstream to the group members.

For a given Source Transport Identifier, a NAK Responder maintains NRTrail, NRLead and NRInc for defining trailing edge and lead edge of NR's internal finite queue, and leading edge of Increment Window respectively.

There is also a mechanism for determining the re-transmission through unicast or multicast retransmission. The sequence numbers stored at the NAK packets are added to the retransmission queue (ReTxQueue). NR does not transmit the requested packets immediately because one or more group members may miss the same packet. So, they wait a pre-defined delay time (TimeRxSend) for receiving all NAKs corresponding to same packet and then

determine whether the lost packet should be retransmitted using unicast or multicast retransmission. A counter (MemberCount) that counts the number of group members requesting that packet, a table NakAddrTable that records the requesting group member's IP addresses.

At RMP protocol, SNRM message periodically sent by receivers not only provides node information at the local branch with RP establishing BPM message, but also informs the NRs for removing some obsolete packets. Obviously, SNRM message must include the last continuous sequential number of the received packet that receiver had in order received all packets marked with the sequential number below. When receiving SNRM message, NR cleared those packets whose sequential number is below that last sequential number stored at the SNRM message for getting more room to store the new coming OData packets.

3.9. Circular Queue at Receiver

Group members immediately deliver the packets to application once they found the received packets are in order. If lost packets found, group members stored up those new coming OData packets whose sequence number is behind the sequence number of that lost packets. Once the lost packets are obtained, it should be forwarded to application as well as those sequential packets behind.

For a given Source Transport Identifier, the multicast packets are stored into the circular queue at receivers. Receiver maintains RxLead and RxTrail pointing to the last packet just stored at the circular queue and to the packet that is available to be sent to the application respectively.

3.10. Packet Content

3.10.1. SNRM Search NAK Responder Message

SNRM is used to hurt out the status information of the PIM routers located the branch, Rendezvous Point (RP) based on SNRM message coming from its outgoing branch to establish the Branch Path Message (BPM). SNRM is periodically originated by the group members to examine whether the PIM routers located at the branch having any changes. Parameters included in the SNRM are :

SnrmBID	branch identifier is a unique number for differentiating itself from the other branches
SnrmNumOfRouter	to record how many PIM routers located at this branch
SnrmRouterStatus	to store up the PIM router's IP address and number of outgoing interface
SnrmStiSqns	the latest sequential number of packet that the downstream group members have been received, this parameter is used to remove the obsolete packets stored at the NAK Responder

3.10.2. BPM Branch Path Message

Rendezvous Point (RP) based on the SNRM message to establish Branch Path Message (BPM), which is multicast to its group member within the RP-rooted tree. BPM conveys some information for initiating PIM routers located at the branch to be became as NAK Responder and assigning them to store up the corresponding sender's packets. Once receiving the BPM message, group members can recognize who is its NAK Responder for handling its data recovery process. Parameters included in the BPM are:

NumOfBranch	number of branch information included in this BPM
BpmBID	branch identifier is a unique number for differentiating itself from the other branches, actually, it is copied from SNRM packets
BpmNrIp	IP address of the selected NAK Responder for a given BpmBID
BpmSti	Source Transport Identifier for which NAK Responder takes responsibility to store up the multicast packets.
BpmTmO	time period for which those information related to the BpmBID message is valid

3.10.3. OData Original Data

OData packets are transmitted by sources to send application's data to receivers and are unicast to RP and then it is multicast to the group and contains :

ODTsi	the globally unique source-assigned TSI
ODSqn	a sequence number assigned sequentially by the source in unit increments and scoped by ODTsi
ODPacket	sender's data stored into this fixed size packet, which is sent from the RP towards all group members.

3.10.4. RData Retransmitted Data

RData packets are retransmitted data packets transmitted by NAK Responder or sources in response to NAK packet. It may use multicast or unicast to be sent to the group for data recovery and contains :

RDTsi	ODTsi of the OData packet of which this is a retransmission
RDSqn	ODSqn of the OData packet of which this is a retransmission
RDPacket	re-transmitted data stored into this fixed size packet, which is sent from the NAK Responder or Sender towards the requested members.

3.10.5. NAK Negative Acknowledgement

NAKs are transmitted by receivers or NAK Responder to request retransmission of missing data packets. NAKs are unicast to the NAK Responder or source and contains :

NAKTsi	ODTsi of the OData packet for which retransmission is requested
NAKSqn	ODSqn of the OData packet for which retransmission is requested

3.10.6. NCF NAK Confirmation

NCF are transmitted by NAK Responder or sources in response to NAK packet. NCFs are multicast to the group and contains :

NcfTsi	NakTsi of the NAK being confirmed
NcfSqn	NakSqn of the NAK being confirmed
NcfTimeStamp	instant time at which the packet is sent, which is used by receivers to calculate the period for re-transmitting the NAK packet

3.11. Functions of Entity

As an end-to-end transport protocol, RMP specifies packet formats and procedures for sources to transmit and for receivers to receive data. To enhance the efficiency of this data transfer, RMP also specifies packet formats and procedures for NAK Responder to improve its reliability of data retransmission and to constrain the propagation of retransmissions. The division of these functions are described as below.

Basically, there are five entities, which logically include sender, receiver, NAK Responder, intermediate PIM router and rendezvous point, working at RMP. Next page, it shows the packets flow among this five entities.

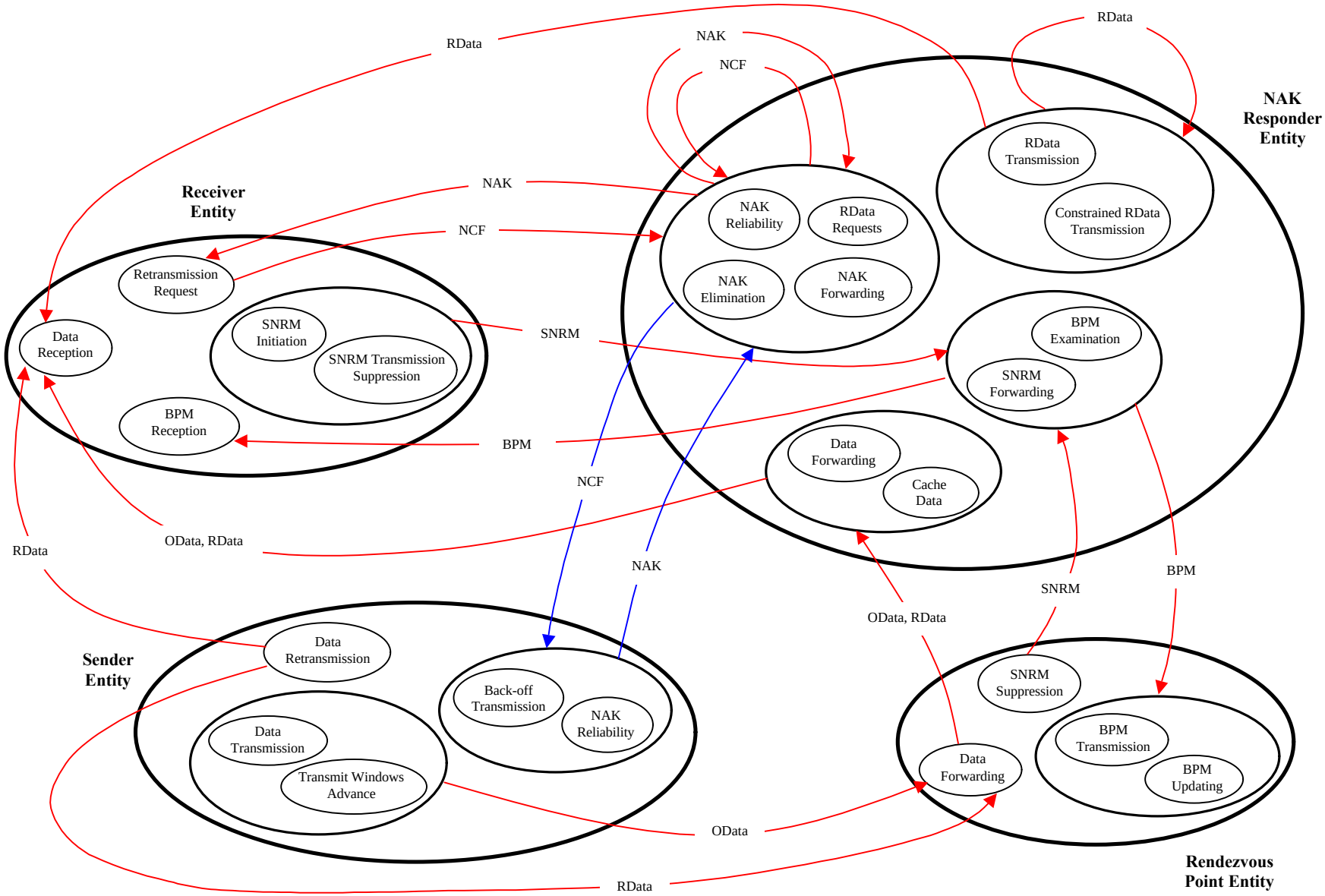


Figure 8. Packets flow among entities

3.11.1. Functions of Sender Entity

- ***Data Transmission***

Source multicast OData packets to the Rendezvous Point (RP) within the transmit windows at a leaky transfer rate.

- ***Back-off Transmission***

Once NAK packets appear, source decreasing the leaky transfer rate adapt for relieving the traffic and burden of buffer management at NAK Responder.

- ***NAK Reliability***

Sender multicasts NCF towards the RP-rooted tree in response to any NAK they received. For each NAK received, sender creates retransmit state recording the NAK information for making decision to perform the retransmissions whether go through the unicast or multicast.

- ***Data Retransmission***

Sources multicast RData packets to the RP in response to NAKs received for data packets within the transmit windows.

- ***Transmit Windows Advance***

Source has a per-defined transmit windows for holding the packets from applications. Packets within the transmit window are multicast packet by packet to the RP at a leaky transfer rate. In case of no NAK found within a timeout interval, packets within the increment window are expired and the transmit windows should be simply advanced for refilling new packets from application. However, once the NAKs appear, source may delay advancing the window until no NAKs for requesting the OData packets at the increment window are received within a timeout interval.

3.11.2. Functions of Receiver Entity

- ***SNRM Initiation***

Receivers multicast the Search-NAK-Responder Message (SNRM) to the last-hop PIM router (DR) for tracing the node state between receivers and RP; those state is used for RP establishing the Branch Path Message (BPM).

- ***SNRM Transmission Suppression***

Receivers suppress transmission of SNRM for which a same SNRM is received during the transmission back-off interval.

- ***Data Reception***

Receivers receive original data OData and re-transmitted data RData; and then delivers it to application immediately; any duplicated packets are eliminated.

- ***Retransmission Request***

Receivers unicast NAKs to the NAK Responder for data packets found to be missing from the expected sequence. Receivers must repeatedly transmit a given NAK until it receives a matching NCF.

- ***BPM Reception***

BPM Message is multicast from RP, and it convey information about the NAK Responders. Receivers base on those information to request the reliable service from NAK Responder when they need.

3.11.3. Functions of NAK Responder Entity

- ***RData Transmission***

NAK Responder retransmits the requested packets towards the group members through unicast or multicast transmission in respond to any NAK received. Retransmission is performed after a pre-defined delay time.

- ***Constrained RData Transmission***

RData Transmission could be performed through unicast or multicast transmission, it totally depends on the number of the received NAK corresponding to the same requested packets within a per-defined period.

- ***NAK Reliability***

NAK Responder (NR) multicast NCFs along the branch towards the group members in response to any NAK they receive. For each NAK received, NR creates retransmit state recording the Branch Identifier (BID), the sequence number of NAK, and the IP address of the group member that requested retransmission.

- ***NAK Forwarding***

NAK Responders (NRs) forwarding the received NAK to other NR located at another branch for getting back the requested packet when they takes response to cache it but the requested packets are not alive at its buffer. In addition, NRs forward the NAK packets to the corresponding sender when they are the default NRs and do not take response to store the request packets. NR repeatedly makes NAK forwarding until receiving the matching NCFs.

- ***NAK Elimination***

NAK Responder (NR) discards exact duplicates of any NAK for which it has already in retransmit state, and respond with a matching NCF.

- ***RData Requests***

NAK Responder (NR) can be received any NAK packets from another NAK Responder for requesting the lost packets. At that moment, NR sends NCF to that NAK Responder for acknowledgement and then perform the RData transmission in respond to that request.

- ***BPM Examination***

BPM Message is multicast along the branch so NR can receive that message. NAK Responder need to inspect the BPM message for checking itself whether is ruled out from the listings of NAK Responder.

3.11.4. Functions of Intermediate PIM Router Entity

- ***SNRM Forwarding***

Last-hop PIM router (DR) adding the unique Branch Identifier (BID) into SNRM which sent from its group member, it forwards this SNRM message to next upstream PIM router one by one until reaching the Rendezvous Point (RP). SNRM message records the state information of each passing intermediate router for RP building the Branch Path Message.

- ***BPM Examination***

BPM Message is multicast along the branch so each intermediate PIM router can receive that message. Each intermediate router need to examine BPM also for checking itself whether is selected as a NAK Responder.

3.11.5. Functions of Rendezvous Point Entity

- ***BPM Transmission***

Rendezvous Point (RP) receives SNRM Messages from its branches and bases on those information to create the Branch Path Message (BPM), which holds the information about the NAK Responders (NR) for each branch. BPM Message is updated by triggering the new SNRM from each outgoing branch, and is multicast toward RP's outgoing interface after updating.

- ***SNRM Suppression***

RP periodically receives the SNRM from its outgoing interface. RP examines each new received SNRM Messages whether having new state information of the PIM routers located at the branch. If there are no updated information, RP suppresses SNRM Messages and does not update the BPM.

- ***BPM Updating***

BPM is updated not only in respond to the new coming SNRM Messages, but also periodically updating to eliminate some parts of state information without being renew for a pre-defined expiry time.

Chapter 4 - Simulation Results

For evaluation of the RMP performance at PIM-SM tree, it modeled the PIM tree into four different scenarios; for each scenario, it also includes different configurations. Each scenario or configuration is static throughout the simulation and packet lost occurred at the receiver's node only.

Simulation will examine the performance metrics including data lost recovery time of group member, traffic concentration at NAK Responder and scalability in the following four scenarios while the PIM routers located at the branch are selected as NAK Responder (NR).

By comparing the performance of RMP, simulation will also make another evaluation of SRMT protocol, which is a multicast protocol working at the source specified tree, applied to the RP-rooted tree for providing reliable service. The operation of SRMT at multi-sender's situation, each sender joining to the RP-rooted tree is acted as a individual spanning tree structure and the Rendezvous Point (RP) to be became as a source. According to the SRMT, ACKs are eventually sent back to source for making reception acknowledge; therefore, the Rendezvous Point (RP) at SRMT's operation can be acted as NAK Responder. In the following section, it will compare the performance measured from the RMP against SRMT operated at RP-rooted tree.

For those scenarios, each branch contains fifteen PIM routers and connects towards one receivers eventually. Each PIM router at the branch is coded as unique number. The proper PIM routers located at each branch will be selected as NR by the Rendezvous Point for performing the data retransmission. Multicast data along the branch towards the group members will be stored at those nodes which have been appointed as NR. How many PIM routers selected as NR is depended on the number of senders joining the PIM tree.

4.1. Testing Scenarios

4.1.1. Scenario I One Sender with Three Members

For scenarios I, it only contains one sender but there are three configurations, which include one sender with one group member, one sender with two group members and one sender with three group members. Actually, each configuration at this scenario can be considered as independent PIM tree so that simulation is to simulate the performance metrics independently at each configuration; and those configuration at this scenario having the same property is only one sender.

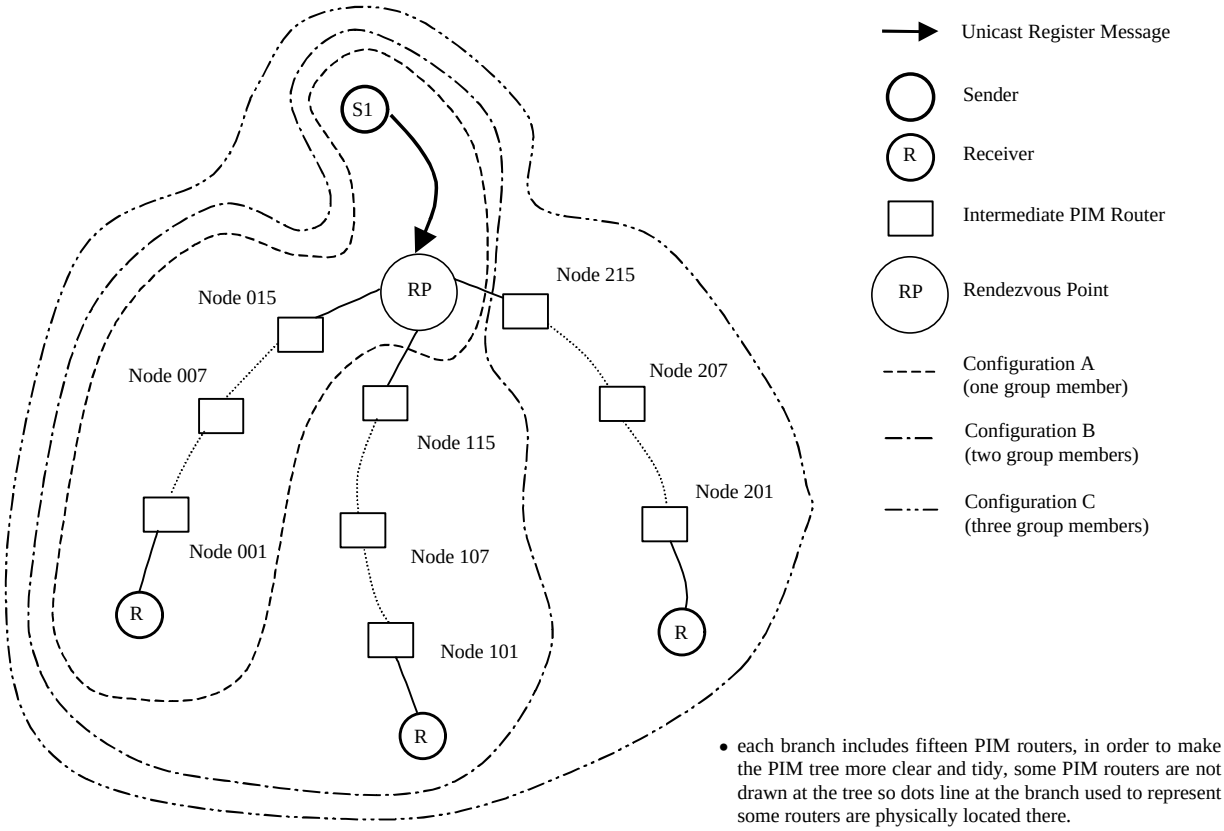


Figure 9. Scenario I Single Sender with different number of Group Member

At the next section - Experiential Results section, it will show out the simulation result for one sender against increasing the group member; and investigate the performance parameters at each configuration.

4.1.2. Scenario II Three Senders with One Member

This scenario also includes three independent configuration but those have the same property – only one branch with one group member.

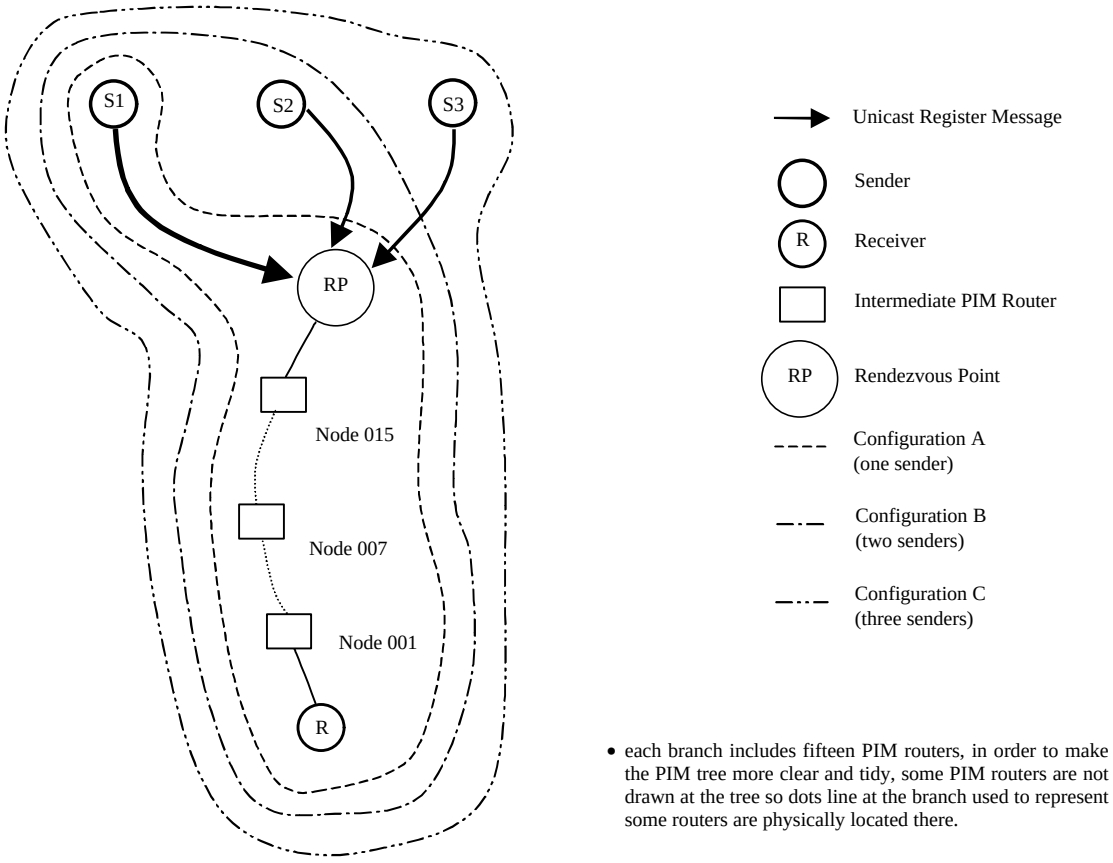


Figure 10. Scenario II One group member with different number of Senders

Scenario II's environment is used to investigate the performance metrics at one group member against more senders. Simulation will examine the performance metrics at each configuration, which include one sender with one group member, two sender with two group members and three senders with one group member independently.

4.1.3. Scenario III Three Senders with Three Members

This scenario is used to investigate the RMP's performance metrics at the situation of more senders and more group members joined the tree at which include three senders and three group members altogether.

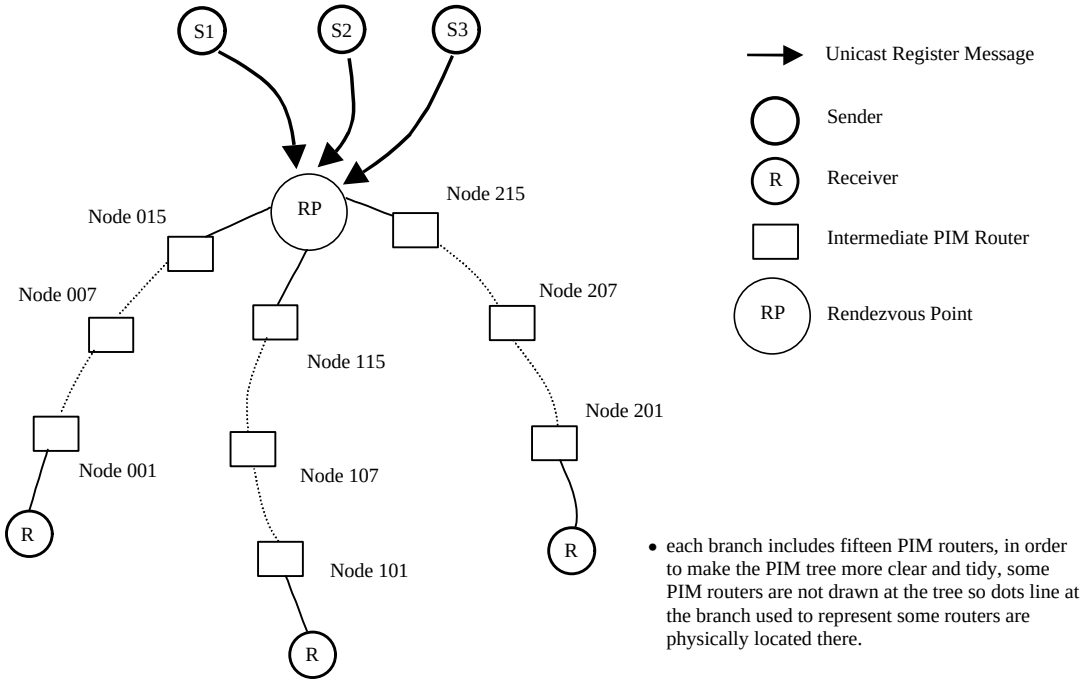


Figure 11. Scenario III - Three senders with Three Group Members

At the above scenario, it assumes those senders joining at the PIM tree at the same time; and senders and group members all stay at the RP-rooted tree during simulation. Each branch only contains one group member.

4.1.4. Scenario IV Single Branch with Three Senders and Members

This scenario is used to investigate the RMP's performance metrics at the situation of more senders joining the RP-rooted tree with only one branch having three group members.

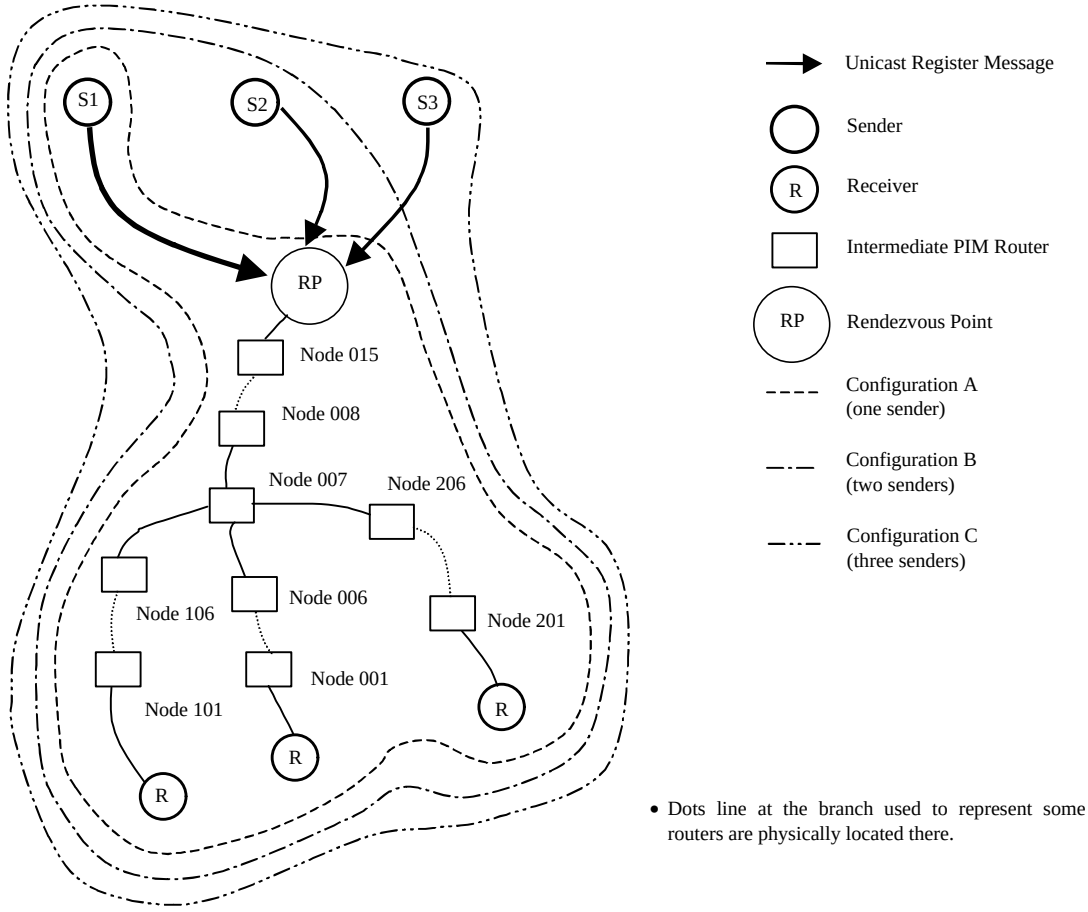


Figure 12. Scenario IV Single Branch with Three Senders and Members

Three members joining the Rendezvous Point through one branch, and PIM router (Node 007) has three outgoing interfaces connected to each group member. Multicast data delivered from the RP along the outgoing branch towards its group members; once the PIM router (Node 007) got the data, it will duplicate the multicast data and send it through its three outgoing interfaces toward the downstream group members.

4.2. Experimental Results

The RMP based on the NAK Responders distributed over the whole RP-rooted tree to provide the reliable services with the group members at PIM sparse mode domain. Rendezvous Point goes through the selection mechanism to appoint the appropriated PIM routers as NAK Responders for providing data recovery service.

This section will evaluate the performances metrics by comparing intermediate PIM routers as NAK Responders versus core Rendezvous Point (SMRT) performing the reliable service respectively. There are four performance metrics chosen to be compared, which includes Data Lost Recovery Time of group member, Traffic Concentration at NAK and Scalability.

4.2.1. Data Lost Recovery Delay Time Comparisons

Data recovery delay time is the time elapsed between when a host found the data lost and when that host got back the lost data. Group members send the NAK packet to its NAK Responder to request the reliable service once they found the received packet is not their expected sequential number. Therefore, this performance metric can express the relationship of the data lost recovery time being directly related to the location of the NAK Responder.

For this performance metric, it only considers the PIM tree structure having one sender with one group member (Scenario I – Configuration A) because this performance metric do not have any relationship at increment of group member and sender. And, this performance parameter only concerns about the recovery packets travelling time between group members and the location of the appointed NAK Responder. The data recovery time is represented in normalized value, which is the measured data recovery time divided by the normalized factor; and the normalized factor is the minimum travelling time for packet successfully delivered from sender to group members at whole simulation.

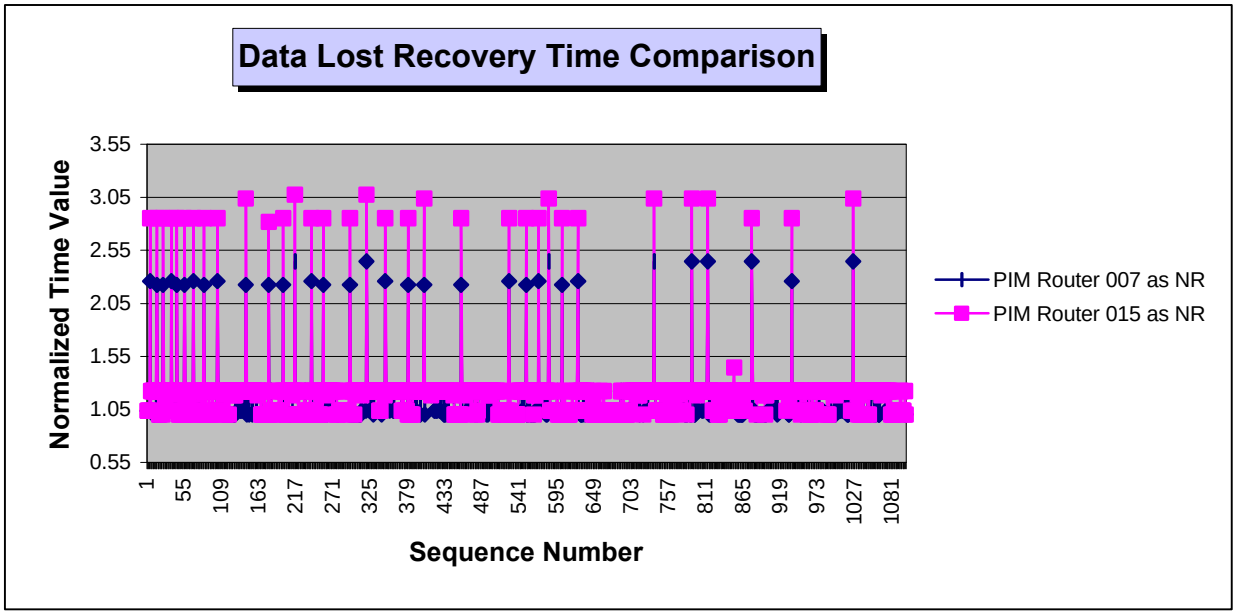


Figure 13. Recovery Time Comparison between two different node at same branch

Figure 13 shows out the normalized value of data lost recovery time elapsed for two NAK Responders located at the same branch but difference location. PIM Node 015 and 007 are selected as NR for lost data retransmission. Such that,

normalized average Data Lost Recovery time against PIM Node 007 = 2.29, and
normalized average Data Lost Recovery time against PIM Node 015 = 2.90

PIM Node 007 is located near the group member so the data lost recovery time is more faster compared with the PIM Node 015, which is located far away from the group member.

The above Figure 13 demonstrates the data lost recovery time comparison while the PIM routers located at the branch are appointed as NR. The following Figure shows out the time comparison for new configuration of RP selected as NR against the result shown at the above Figure.

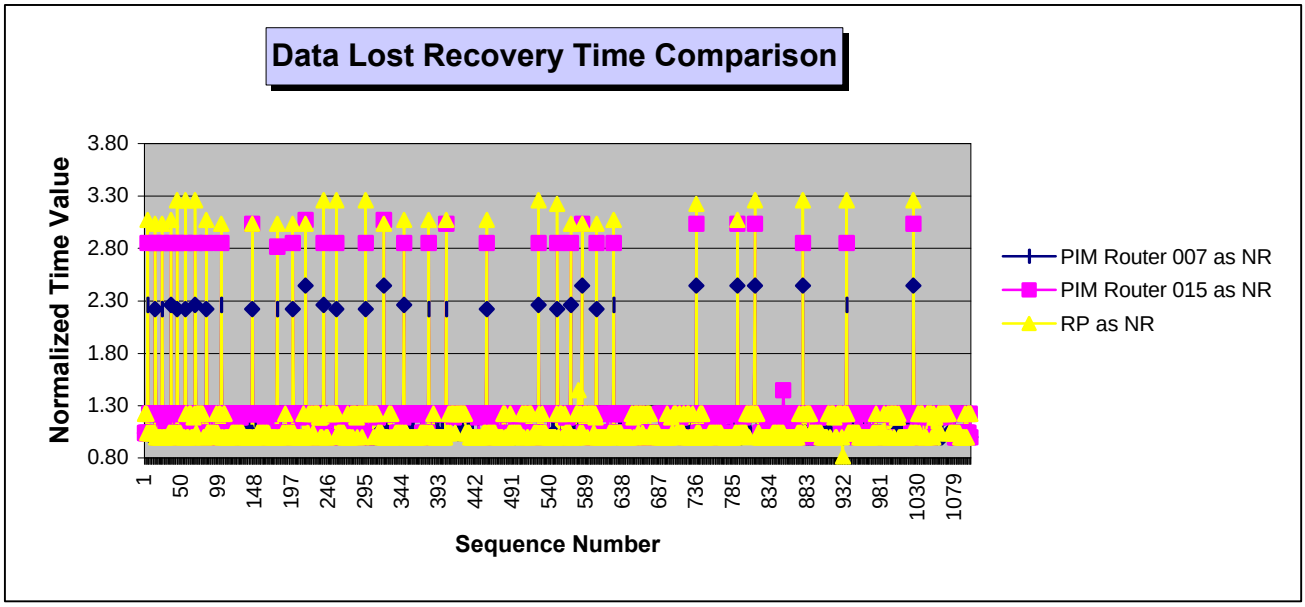


Figure 14. Data Lost Recovery Time Comparison against RP and PIM Router as NR

According to the Figure 13's result, normalized average Data Lost Recovery time for PIM Node 007 and 015 is 2.29, and 2.90 respectively. At new configuration, RP is appointed as NR instead of PIM routers selected as NR, the result shown at Figure 14, the new normalized average Data Lost Recovery time against RP is equal 3.13, which is more bigger than the result got from the Figure 13. Actually, RP is the core of the PIM tree, and is much far away from the group member so that the recovery time at this configuration being longer is reasonable.

Obviously, the appointed NR located near the group member can provide more faster reliable services for group members; in other words, group members could get the re-transmitted data from those appointed NR located near them more faster.

4.2.2. Traffic Concentration Comparisons

This section will examine the variance of the distribution of the traffic on the branch outgoing from the Rendezvous Point (RP) in the above four scenarios. In each scenarios, it will compare the traffic concentration metric between PIM routers assigned as NR and RP appointed as NR (SRMT approach).

Traffic Concentration measured in this section only concerns the traffic introduced from data lost recovery mechanism performed by RMP protocol, the traffic packet including NAK, NCF and RData packets and normal PIM's multicast traffic is not considered.

On the other hand, PIM tree is to provide one multicast data delivered environment for multi-sender joining the RP-rooted tree. Therefore, at the following investigation, it will pay more attention to examine the traffic influence caused by increasing the senders joined the RP-rooted tree.

In scenario I, RMP protocol appointed two PIM routers at each outgoing branch from RP to become NAK Responder, however, there is only one sender joined the PIM tree so that both NAK Responders at each branch are assigned to cache the same sender' packets. PIM Node 007, 107, 207, 015, 115 and 215 are appointed as NRs for data lost retransmission at each branch respectively. At this case, PIM node 015, 115 and 215 are set to default NAK Responders also.

The first investigation at scenario I, which is only one sender, is to compare the traffic concentration by increasing the group member joining the tree. The Traffic Concentration is also represented in normalized value which is average number of packets across the PIM routers at the tree due to the process preformed by the data lost recovery mechanism; and the normalized factor is the total number of packet lost at the simulation.

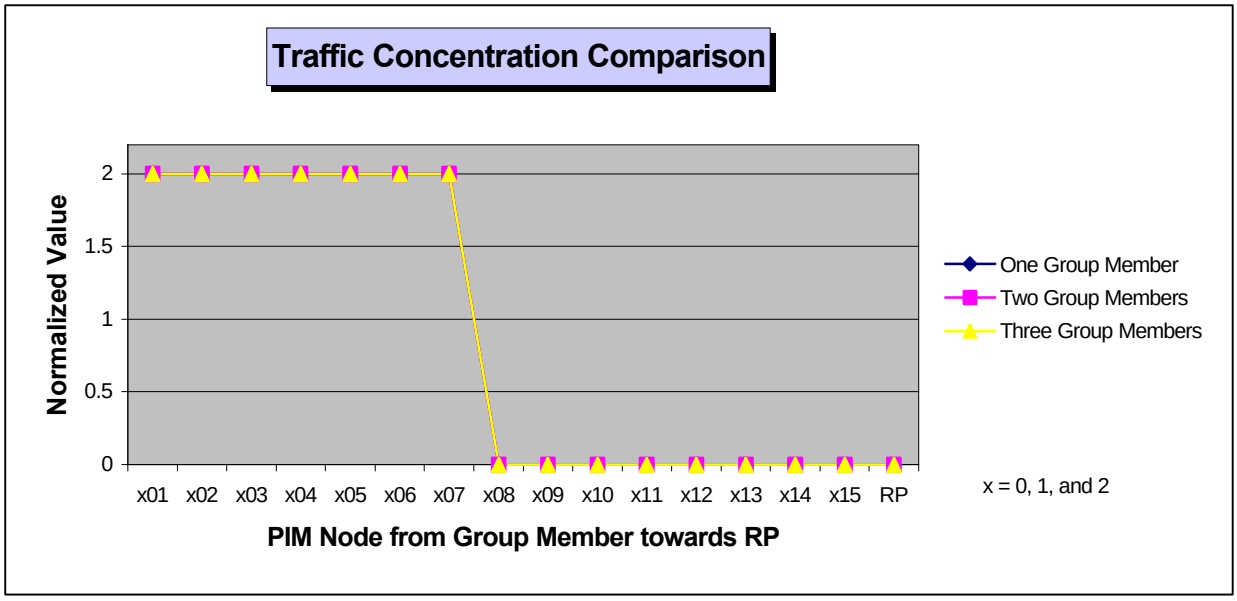


Figure 15. One Sender against increasing group member

From this above Figure, the curve of the three configurations at Scenario I demonstrate same result. Once group member located at each branch found the data lost, they send the NAK packet to its near NAK Responder to request reliable services. Therefore, node 007, 107 and 207 received the NAK packets from its group members at each branch respectively. Most traffic are concentrated at lower half of the branch starting from Node x01 to x07 at individual branch because those nodes are selected as NAK Responders and also are the nearest NRs at individual branch, they received the NAK packet from group members and sent the re-transmitted packets towards its requesters.

Obviously, from the above result, there are no any particular difference even increasing the group members. Actually, from the scenario I, each branch has its corresponding NR for

providing reliable services such that the traffic concentration at each branch has the same performance.

There are no any distinct influence at traffic concentration as number of group member is increased. Therefore, at the configuration of PIM Node selected as NR for providing reliable services, the traffic concentration is independent of increasing the group members joined.

The second investigation is to examine the above same scenario but Rendezvous Point is appointed as NAK Responder (SRMT approach) instead of PIM routers do. At this simulation, RP will provide the reliable services for caching the sender's packets and each group member will send the NAK packet to it for requesting the data recovery. All PIM router located at each branch is only to perform the package forwarding.

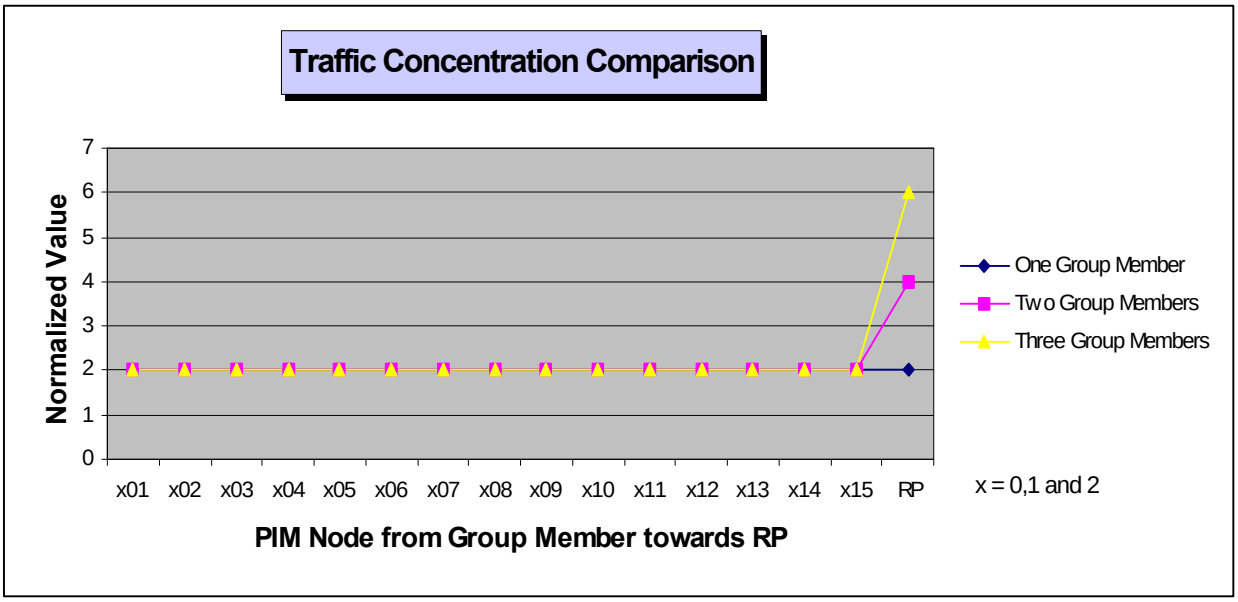


Figure 16. One Sender against increasing group members at SRMT approach

From the above Figure 16, it can see that all traffic is concentrated at the Rendevzous Point RP; and the more the group member joins the PIM tree, the higher the traffic rate appears at the RP. When group member lost the data, they send NAK packets to RP for requesting reliable services, therefore, RP need to handle more NAK packets by increasing the group

members. The relationship between traffic concenteration rate ($TrafficRate_{RP}$) and number of group members (NoRx) joined can be represnted as :

$$TrafficRate_{RP} = N \mathbf{x} NoRx \text{ ----------(1)} \qquad \text{where N is a constant}$$

therefore, traffic rate is linear proportional to the group members joined.

On the other hand, traffic distribution leaving the RP at the PIM routers along the branch is same even by increaing the group members. Because PIM routers at this simulation situiaon are only to handle the packet forwarding, it is obviously to show out the even traffic distribution at branch leaving the RP.

By comparing the result shown at the Figure 15 and 16, at Figure 15, it can see that the traffic contribution behind the NR – from PIM Node x08 to RP is to reach to zero. Group member at the individual branch found the data lost sends the NAK packet to its near NR for requesting the data lost recovery service so that all re-transmission traffic is only concentrated thoes PIM Node below the NR (Node x07).

In contrast, RP is assigned as NR (SRMT approach) at Figure 16, traffic is even distributed along the branch because the re-transmitted packets are sent from the RP towards the group members. Most traffic are concentrated at the Rendezvous Point.

The following figure shows the another experimental result of traffic concentration at increasing the number of sender instead of group member. It bases on the three different configurations at scenario II to investigate this performance metric against increasing the number of senders.

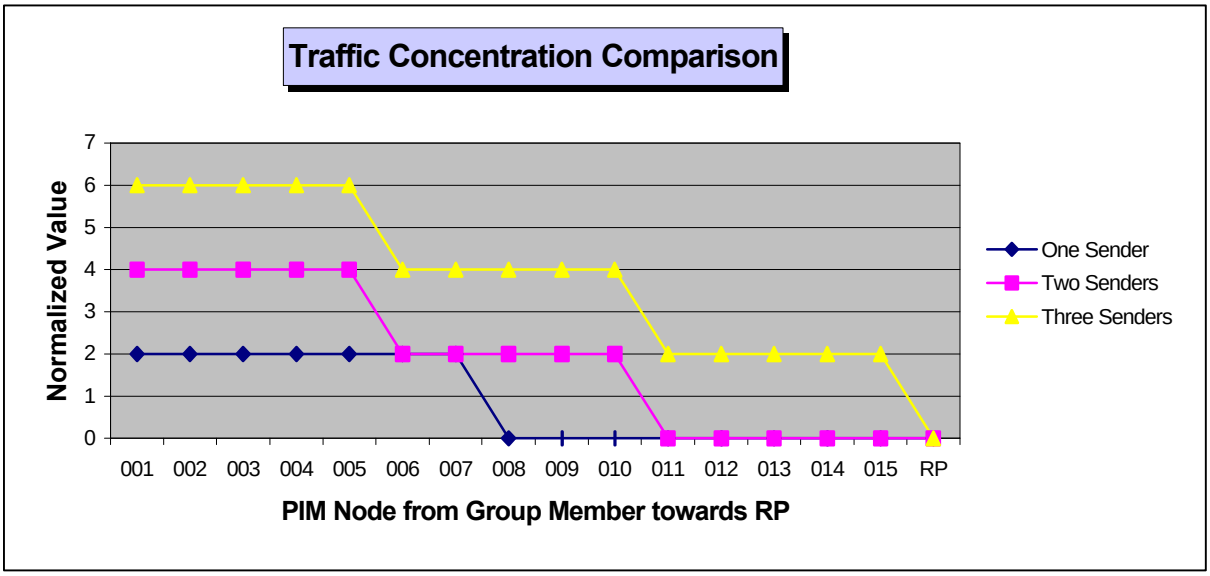


Figure 17. One Group Member against increasing the joined sender at Scenario II

The above Figure 17 contains three distribution curves of traffic concentration. At one sender's configuration - curve "--◆--", Node 007 and 015 are appointed as NRs for caching the sender' packets, however, group members find the nearest NR (Node 007) for requesting the reliable service. At two sender's configuration - curve "--■--", Node 005, 010 and 015 are assigned as NRs for providing reliable service, however, node 005 and 015 take responsible for caching the sender 1's packets; and Node 010 is to cache the sender 2's packets respectively. When group member found data lost from sender 1, they send the NAK packet to the nearest NR – Node 005 to request the reliable service. At three sender's configuration - curve "--▲--", Node 005, 010 and 015 are appointed as NRs for caching the sender 1, 2 and 3's packets respectively. Group members found data lost send the NAK packet to the corresponding NRs to get back the wanted packets.

Obviously, for one sender's situation, group member request the reliable service from its nearest NR (Node 007) so that more re-transmitted traffic are concentrated those nodes located at the NAK Responders (Node 007) and below. For two senders and three senders' situation, it can also see that the trend of the curve is the traffic distribution across those downstream PIM routers below the appointed PIM Nodes for providing data lost recovery service.

Number of peak traffic concentration point shown at each curve depends on how many senders joined the PIM tree. One PIM Node is reserved for providing reliable services of one sender at each branch such that more senders joining the PIM tree requests more PIM Node as NRs for providing service at each branch.

One the other hand, the location of the selected PIM Node at each branch is also related to how many senders joining the PIM tree. From the above Figure shown, the selection mechanism of NR is evenly to select the PIM Node across the branch.

From the above result, it know that traffic is concentrated at the NRs and the PIM Nodes below the NR. It is because that NR received the NAK packets from the group member for requesting the reliable service and re-transmitted packets are sent from the NR towards the group member along the branch. Therefore, this traffic distribution has the accumulative effect; once there are more PIM Node selected as NR at the branch, the traffic introduced from the NR located at the higher position is also to be became the traffic of the NR located at the lower position and the PIM Node below them as well. This effect can been obviously seen from the curve "--▲--" shown at the above figure.

Making the above simulation again but RP is assigned as NR (SRMT approach) instead of PIM Node is done. At this situation, RP will take the responsibility for providing reliable service once receiving the NAK packets from group members.

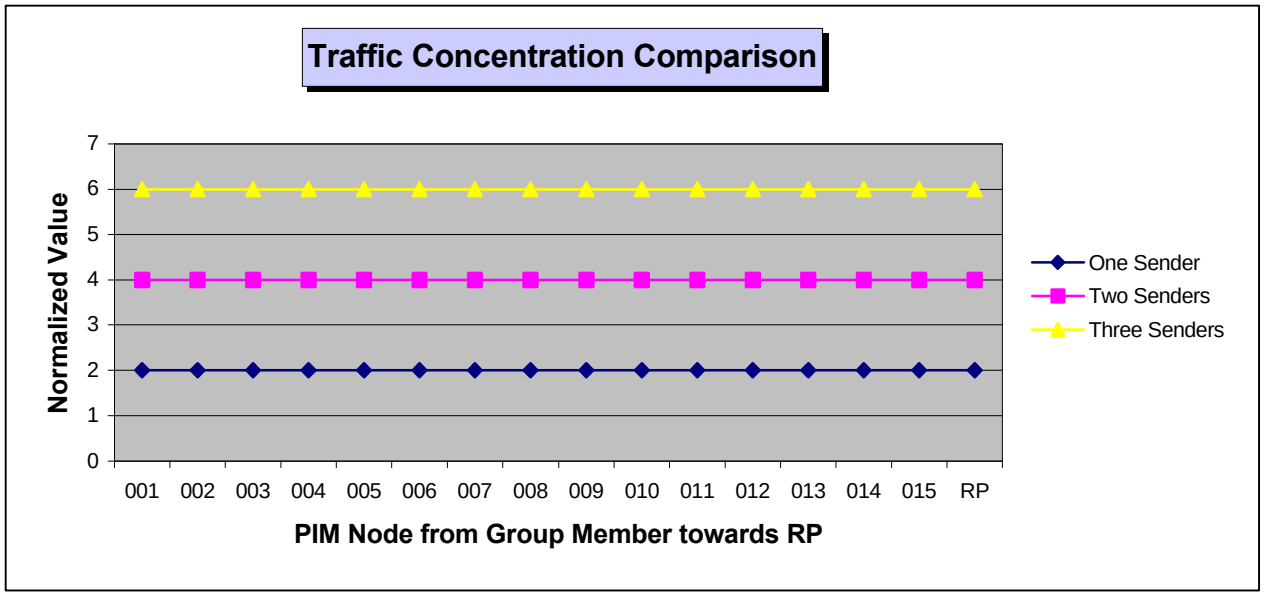


Figure 18. One Group Member against increasing the joined sender when RP as NR

From the above Figure 18, all re-transmitted traffic is evenly distributed at the whole branch; the more sender joins the tree, the higher the traffic rate is. Actually, RP takes the responsibility for providing the reliable services so they received the NAK packets and re-transmitted the requested packet also, and intermediate PIM routers receives same traffic from group members and Rendezvous Point. The relationship between traffic concenteration rate ($TrafficRate_{RP}$) and number of senders (NoTx) joined can be represnted as :

$$TrafficRate_{RP} = M \times NoTx \text{ ----------(2)} \qquad \text{where M is a constant}$$

therefore, traffic rate at RP is a linear function of senders joined.

Getting more analysis from the Figure 17 and Figure 18, it retrieves both "--▲--" curve's data, which represent the traffic concentration at three senders joining the tree, for comparison again.

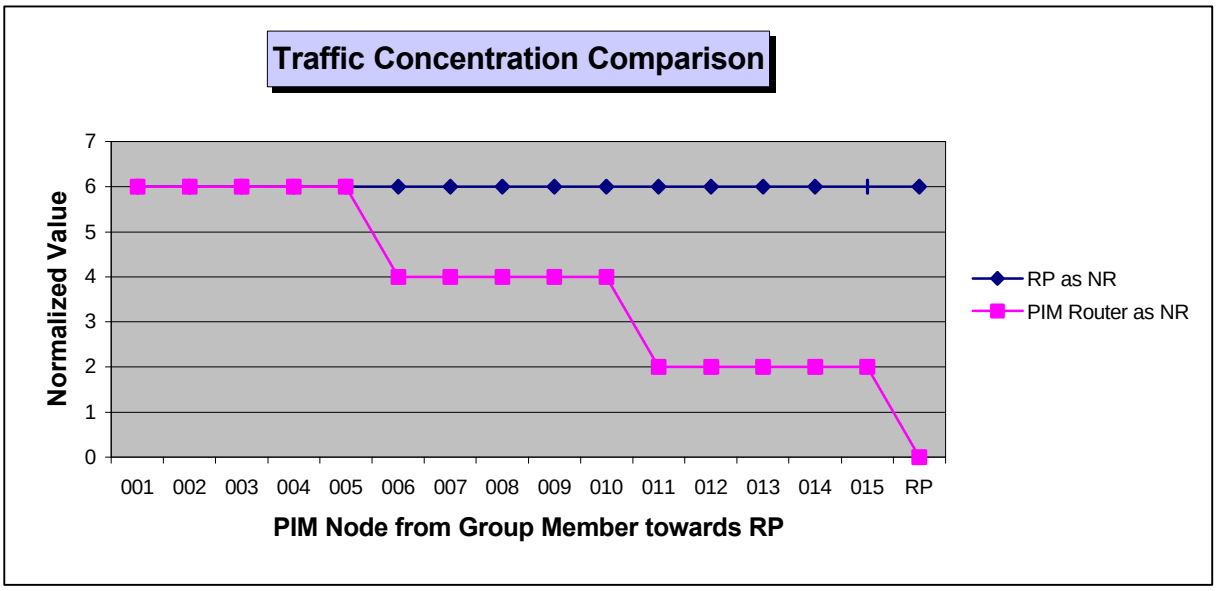


Figure 19. PIM Node as NR against RP as NR (SRMT approach) at three senders joined

The above Figure 19, it shows out the traffic concentration for three senders joining tree and compares the situation of PIM Node 005, 010 and 015 (curve "--■--") as NRs caching the packets of sender 1, 2 and 3 respectively and RP (curve "--◆--") as NR caching all multicast packets (SRMT approach).

It can see the accumulative effect appeared at the "--■--" curve, which represents the traffic distribution at configuration of PIM Nodes as NRs. In contrast, there are even traffic distribution appeared at the whole branch at the "--◆--" curve. Obviously, both configurations are configured to handle three senders joining the tree and provide the reliable services but the average reading of the traffic concentration at configuration of PIM Nodes as NRs is more less than that of configuration at RP as NR. The average traffic rate at RP as NR (SRMT) and at PIM Nodes as NRs is 6 and 3.75 respectively, which is less 37.5% traffic rate at the later configuration.

It can be also reflected from the trend of the both curves. The traffic rate of configuration at PIM Node as NR is more less from PIM Node 006 to RP compared with the another curve.

Actually, at the configuration of RP selected as NR (SRMT approach), all re-transmitted packets must go along the PIM Nodes below RP, as a result, all PIM Nodes along the branch bear the same burden of re-transmitted packets. In contrast, the higher location of the PIM Nodes at the configuration of PIM Node selected as NR, they only handle the re-transmitted packets came from those NRs whose location are higher than them such that it will relieve the overall traffic concentration of the branch. This merit is mainly introduced from the NRs at this configuration widely distributed along the branch instead of only one PIM node as RP at multi-senders configuration.

At the scenario III, it contains three senders joining the PIM tree at which there are also three group members joined. The following figure will examine the traffic concentration of this scenario and compare the result got from the configuration of PIM Nodes as NRs and RP as NR respectively.

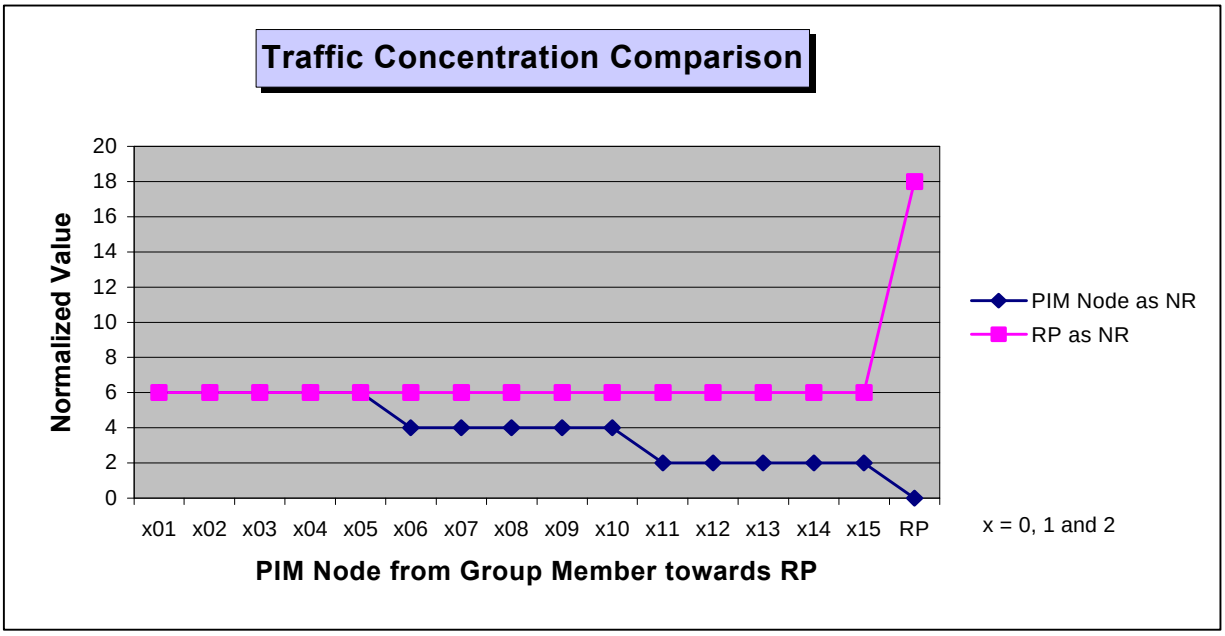


Figure 20. PIM Node as NR against RP as NR (SRMT approach) at Scenario III

At the Scenario III, three senders join the PIM tree so that three PIM nodes at each branch outgoing from the RP are selected as NRs. PIM node 005, 105 and 205 are selected for caching the sender 1's packets at each branch; and PIM node 010, 110 and 210 is appointed to cache the sender 2's packets at individual branch; and PIM node 015, 115 and 215 is assigned to store the sender 3's packets for providing reliable services.

From the above Figure 20, it can see that the traffic concentration for configuration of PIM Nodes selected as NRs is kept to be same performance at one group member with three senders joined. For this configuration, there are no particular influence by increase the group members joined.

On the other hand, at the configuration of RP selected as NR caching all multicast packets from three senders, the traffic rate at the RP has significant change. In fact, this result can be predicted from the equation (1) and (2) shown at the previous evaluation of one sender with three group members and three senders with one group member. The traffic rate at the RP can be expressed as

$$\text{TrafficRate}_{RP} = P \times \text{NoTx} \times \text{NoRx} \quad \text{----------(3)} \qquad \text{where P is a constant}$$

therefore, the traffic rate at the RP is directly proportional to the number of group members and senders joined the PIM tree.

RP selected as NR (SRMT approach) for providing reliable service will make the traffic rate concentrated at the RP is quite high; and even continuously grow once more and more senders and group members joined the tree.

PIM Node selected as NR for providing reliable service at the branch is more effective and efficient; and can relieve the traffic concentration along the branch. The influence against the traffic concentration at the branch by increasing group members and senders is quite small.

By comparing the traffic concentration for three group members joining the RP-rooted tree at Scenario IV, it evaluates the protocol's performance against increasing the senders at the approach of PIM routers selected as NR and RP as NR (SRMT approach) individually, and then compares both results.

For Scenario IV, at configuration A one sender joining the tree (curve "--◆--"), PIM node 007 and 015 are selected as NR caching the sender 1's packets; at configuration B two senders joining the tree (curve "--■--"), PIM node 007 and 015 are assigned as NRs caching the packets of sender 1 and PIM node 011 is appointed to cache the packet of sender 2 respectively; at configuration C three senders joining the tree (curve "--▲--"), PIM node 007, 011 and 015 are appointed to NRs taking response to cache the packets of sender 1, 2 and 3 respectively. Each traffic concentration at Scenario IV are exhibited as below figure.

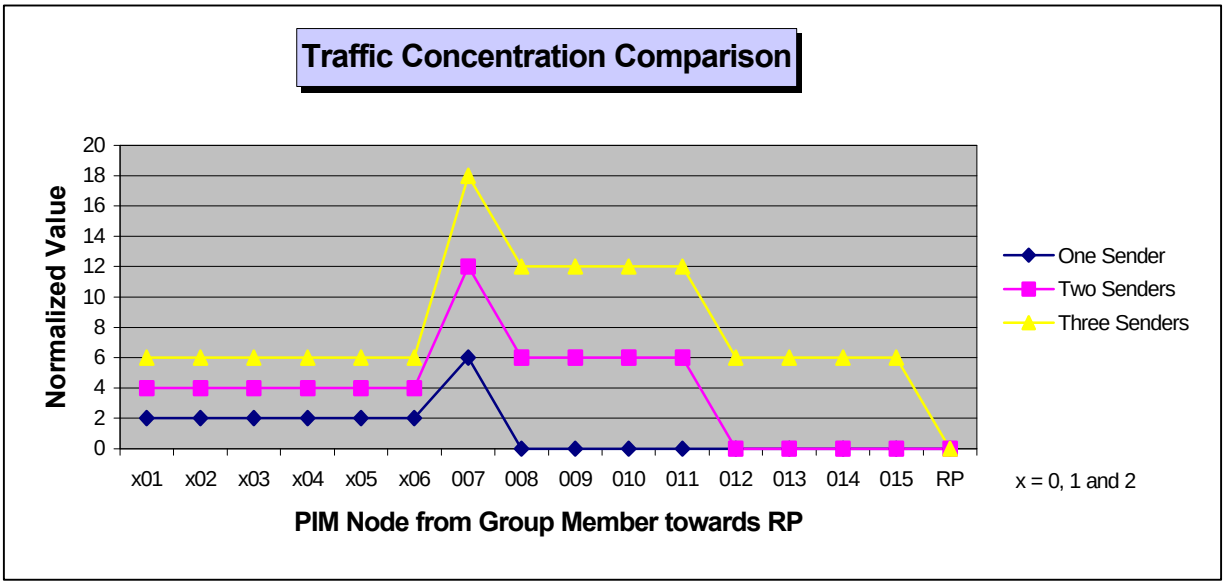


Figure 21. Three group members against increasing the joined sender at Scenario IV

At configuration A and B – one sender and two sender situation, although PIM node 015 is also assigned as NR caching the sender 1's packet, however, once the group members found the data lost of sender 1, they sent the NAK message to its nearest NR – PIM node 007 instead of PIM node 015, for requesting reliable service.

From this above Figure 21, it can see that the traffic concentration of sub-branch leaving from PIM router (Node 007) is same as that of Scenario III. At the Node 007, it is the conjunction of three group members joining the main branch and all NAK packets sent from group members and re-transmitted packets sent from NAK Responders must reach or pass through this node so that it is the peak traffic area of the branch. But for those PIM routers located at the upper position behind Node 007, traffic is less because there is only a part of NAK packets or re-transmitted packets reaching there.

When applying the SRMT approach to RP-rooted tree, Rendezvous Point receives all ACKs from its group members. The traffic concentration of this approach is shown as below.

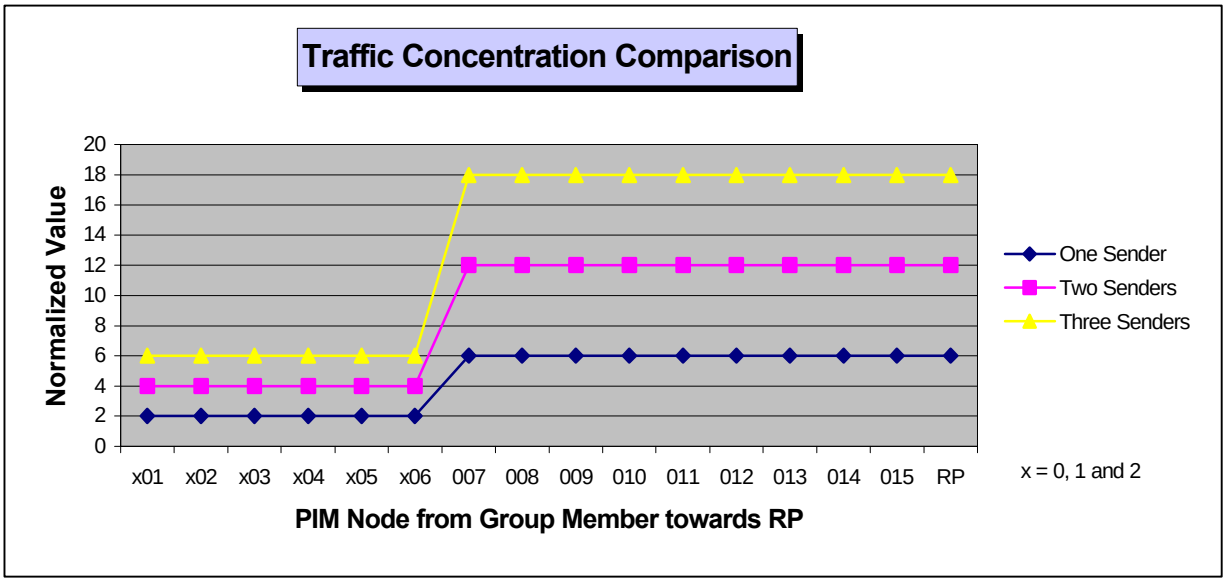


Figure 22. SRMT approach applied to Scenario IV

Three individual sub-branches joined together at the junction point located at the Node 007. For SRMT approach, ACKs sent from group members are going up along the branch towards

the RP for reception acknowledge. Therefore, the traffic concentration at sub-branch (Node x01 – x06) is same as the result at the approach of PIM routers as NR. However, those ACKs came from the sub-branches have also to be sent to the RP along the main branch (Node 007 – RP) for acknowledge. Obviously, more traffic appeared at the main branch at this approach compared with the approach of PIM routers selected as NR.

The main reason why the RMP protocol (PIM routers selected as NR) is less traffic at main branch (from PIM node 007 to RP) is the NAK Responder located at the main branch (Node 007, 011 and 015 at Figure 22) to take response for handling the data recovery process such that the NAK packets are not necessary to be forwarded again. Obviously, at the scenario IV, traffic introduced by RMP protocol at PIM-SM multicast tree is less than SRMT approach does.

4.2.3. Scalability

PIM Sparse Mode (PIM-SM) is designed to eliminate the scaling issues and limit the multicast traffic so that the multicast packets are only gone through those routers interested in receiving traffic for a particular group. RMP is introduced into the PIM-SM for providing the reliable services and it must sustain the scalability exhibited at the PIM-SM domain, in other words, RMP introduced into PIM-SM provides reliability but cannot give up the scalability property.

Actually, RMP needs state information stored into the different entities including the Rendezvous Point, those PIM routers selected as NAK Responders and group members for performance the reliable services. At the group member side, they only need to store part of related information retrieved from BPM message, which is sent from the Rendezvous Point; therefore, those stored message is independent of the number of group members. Regarding to the PIM routers selected as NAK Responders, they only cache the sender's packets travelling along the branch at which the NAK Responders are located; therefore, it is also independent to the increment of the group members. At the Rendezvous Points, it need to store up the SNRM message sent from the group members for building up the BPM message. Once the new group member joined the multicast group or exiting group members sent the periodical PIM Join Message to Rendezvous Point, the SNRM message is also sent to the Rendezvous Point for updating the BMP message. More group members joining the multicast group causes more state information stored into Rendezvous Point, however, only Rendezvous Point stores this state information in the whole multicast group and those related information will be also expired and erased once Rendezvous Point does not receiver the periodical PIM Join Message or RMP SNRM Message from the corresponding group members.

At the reliable data service, it must consider the NAK implosion's problem. This problem will cause the network congestion more serious and reduces the bandwidth of the network links. The overall performance of the reliable service and network will be totally degraded. Therefore, RMP already considers this problem and avoids it happening during providing reliable service at PIM-SM domain. At the configuration of PIM routers selected as NR, Rendezvous Point bases on the SNMR message to appoint the proper PIM routers as NAK responding to handling the NAK packets sent from the group members located at the same branch. Each NAK Responder located at the branch is only to receive the interested NAK packets sent from the group member. Actually, group members can know who is their NAK Responder from the received BPM message, which is sent from Rendezvous Point. As a result, once group members found the received packet out of sequence, they can send the NAK packet to the corresponding NAK Responder located at the branch for requesting the reliable services. Obviously, the NAK packets are not only handled by one responders, they are processed by the individual NAK Responders which are widely distributed at the multicast group such that the NAK implosion should not exist at the RMP introduced into the PIM-SM domain for providing reliable services.

Number of timers involved into the RMP protocol are used to set the expiry period of state information stored into the group members or Rendezvous Point. At the RMP, no matter how many group members joined the multicast group, there is only one timer used to keep track the validation of the BPM message stored at the Rendezvous Point, which is used to inform group members about its NAK Responders. Each joined group member also have a one count-down timer used to periodically send SNRM message to Rendezvous Point once this timer is reached to zero. However, the dedicated count-down timer is only related to the corresponding group member and independent of the other group members. Therefore, it can

see that timers used at the RMP do not affect the scalability property of PIM-SM mode once the RMP is introduced into it for providing the reliable services.

Chapter 5 - Conclusions and Future Works

5.1. Conclusions

RMP is mainly designed for providing the reliable service at the situation of multi-sender joined to the PIM-SM core-based multicast tree and it is suitable for non real-time applications.

RMP uses negative acknowledgment approach, which is based on NAK packets sent from the group member to provide the reliable services. NAK Responder (NR) at RMP is a crucial role for achieving this goal. NRs are widely distributed at the whole multicast group, and systematically conducts the reliable services towards those group members near them. The responsibilities of processing NAKs and performing retransmission is handled by the appointed widely distributed NRs located at the whole RP-rooted tree. Receivers in each local branch send their NAKs to its corresponding NRs, who is responsible for caching packets received from the sender and re-transmitting any packets lost in the local branch. Thus, RMP avoids the ACK implosion problem when the receiver population becomes large.

In addition, due to the widely distributed property of NAK Responders, it can optimize the traffic distribution of the multicast group also.

The state information maintained at each multicast receiver is independent of the number of group members. RMP uses a receiver-driven approach; it places the responsibility of ensuring sequenced, lossless data reception on each individual receiver. Thus, the sender is relieved of the burden of tracking the status of each receiver. And then, when a receiver joins or leaves a multicast group, it does not affect the state information of the sender or other receivers also.

Finally, from the whole analysis of the above sections, it demonstrate the RMP being a workable solution to construct a framework to provide the reliable services at the PM-SM domain.

5.2. Future Works

To improve the protocol further more, it notes two recommendations of RMP that could be altered to make the protocol more effective and efficient for providing reliable services in PIM-SM domain.

One issue it wish to address arise from the transmission of SNMR message, which is used by the group members for getting the status of PIM routers located at its branch. Actually, at the PIM-SM protocol, once group members joining the multicast groups they emitted the PIM Join Message towards the Rendezvous Point (RP) for joining the multicast group. If PIM-SM protocol could be extended for adding the status of PIM routers on it, Rendezvous Point (RP) can base on those information stored at PIM Join Message to build the RMP's BPM messages instead of getting that from SNRM message; such that RMP does not need to request group members to send the SNRM message for doing the same function. It will make the RMP more effective and minimize the overhead introduced from RMP protocol.

Secondly, the infant selection mechanism of NAK Responder used by the RMP is quite simple, it only bases on one simple algorithm to select the PIM routers, which are located at the branch, to be NAK Responder for providing reliable services. This mechanism does not consider the status including the traffic concentration, link bandwidth and congestion of the branch to select the NAK Responders. A proper NAK Responder can make the reliable service more effective and efficient. Therefore, one more intelligent selection mechanism is need to monitor the status of the branch at a periodic manner, it bases on the updated status of the branch to dynamically chose the PIM routers as NRs for providing the reliable service. However, this intelligent mechanism need much time to be developed and implemented. Ideally, the NAK Responders should be selected dynamically based on the status of the branch to make more efficient.

References

- Heinrichs B., ‘AMTP : Towards a High Performance and Configurable Multipeer Transfer Service’, in ‘Architecture and Protocols for High-Speed Networks’, Danthine, Effelsberg, Spaniol (eds.), Kluwer Academic Publishers, 1994.

- Marko Schuba, ‘SRMT – A Scalable and Reliable Multicast Transport Protocol’, Informatik 4, Aachen University of Technology, 52056 Aachen, Germany.

- Armstrong, S, Freier, A and Marzullo, K, `Multicast Transport Protocol', DARPA RFC 1301, February 1992.

- Braudes, R and Zabele S, `Requirements for Multicast Protocols' DARPA RFC 1458, May 1193.

- Crowcroft, J and Paliwoda, K, `A Multicast Transport Protocol', ACM SIGCOMM 1988: Commun. Arch. Protocols, 18(4), August 1988, p. 247-256.

- Deering, S, `Host Extensions for IP Multicasting', DARPA RFC 1112, August 1989

- Floyd, S., Jacobson, V., Liu, C., McCanne, S., and Zhang, L., ``A Reliable Multicast Framework for Light-weight Sessions and Application Level Framing", ACM SIGCOMM 95.
- Hofmann, M., A Generic Concept for Large Scale Multicast. Proceedings of International Zurich Seminar on Digital Communication (IZS'96), Springer Verlag, February 1996.

- Hofmann, M., Braun, T., Carle, G., Multicast communication in large scale networks, Proceedings of the IEEE Workshop on the Architecture and Implementation of High Performance Communication Subsystems (HPCS'95), Mystic, Connecticut, August 1995.

- Holbrook, H.W., Singhal, S.K., and Cheriton, D.R., Log-based Receiver-Reliable Multicast for Distributed Interactive Simulation. In Proceedings of SIGCOMM '95 , Cambridge, MA, August, 1995. ACM SIGCOMM.

- Koifman, A and Zabele,S, RAMP: A Reliable Adaptive Multicast Protocol, to be presented at IEEE INFOCOM '96, San Francisco, CA., March 1996.

- Lin, John C. and Paul, Sanjoy, RMTP: A Reliable Multicast Transport Protocol, IEEE INFOCOM '96.

- Montgomery, T, Reliable Multicast Transport.

- Stevens, W. Richard, TCP/IP Illustrated: the protocols/W. Richard Stevens. 1994.

- Thomas, Stephen A. IPng and the TCP/IP protocols : implementing the next generation internet / Stephen A. Thomas. 1996.

- Fluckiger, Francois. Understanding networked multimedia : applications and technology / Francois Fluckiger, 1995.

- APPENDIX -

Appendix I - PIM Sparse Mode

PIM Sparse Mode (PIM-SM) is being developed to provide a multicast routing protocol that provides efficient communication between members of sparsely distributed groups - the type of groups that are most common in wide-area internetworks. Its designers believe that a situation in which several hosts wish to participate in a multicast conference do not justify flooding the entire internetwork with periodic multicast traffic. They fear that existing multicast routing protocols will experience scaling problems if several thousand small conferences are in progress, creating large amounts of aggregate traffic that would potentially saturate most wide-area Internet connections. To eliminate these potential scaling issues, PIM-SM is designed to limit multicast traffic so that only those routers interested in receiving traffic for a particular group "see" it.

PIM-SM differs from existing dense-mode multicast algorithms in two essential ways:

- Routers with directly attached or downstream members are required to join a sparse-mode distribution tree by transmitting explicit join messages. If a router does not become part of the predefined distribution tree, it will not receive multicast traffic addressed to the group. In contrast, dense-mode multicast routing protocols assume downstream group membership and continue to forward multicast traffic on downstream links until explicit prune messages are received. The default forwarding action of the other dense-mode multicast routing protocols is to forward traffic, while the default action of a sparse-mode multicast routing protocol is to block traffic unless it is explicitly requested.

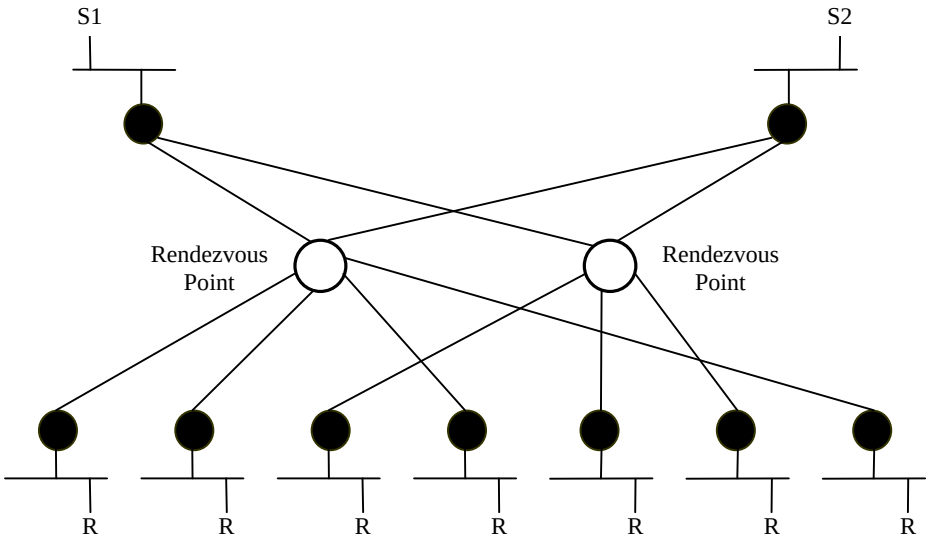


Figure I.1. Rendezvous Points

- PIM-SM is similar to the Core-Based Tree (CBT) approach in that it employs the concept of a rendezvous point (RP) where receivers "meet" sources. The initiator of each multicast group selects a primary RP and a small ordered set of alternative RPs, known as the RP-list. For each multicast group, there is only a single active RP. Each receiver wishing to join a multicast group contacts its directly attached router, which in turn joins the multicast distribution tree by sending an explicit join message to the group's primary RP. A source uses the RP to announce its presence and to find a path to members that have joined the group. This model requires sparse-mode routers to maintain some state (i.e., the RP-list) prior to the arrival of data packets. In contrast, dense-mode multicast routing protocols are data driven, since they do not define any state for a multicast group until the first data packet arrives.

I.1. Directly Attached Host Joins a Group

When there is more than one PIM router connected to a multi-access LAN, the router with the highest IP address is selected to function as the Designated Router (DR) for the LAN. The DR is responsible for the transmission of IGMP Host Query messages, for sending Join/Prune messages toward the RP, and for maintaining the status of the active RP for local senders to multicast groups (Figure I.2.).

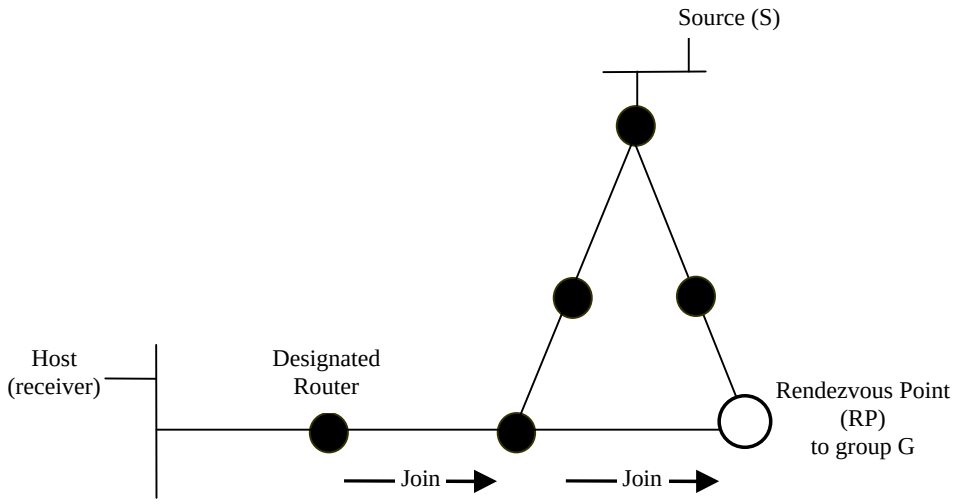


Figure I.2. Host Joins a Multicast Group

To facilitate the differentiation between DM and SM groups, a part of the Class D multicast address space is being reserved for use by SM groups. When the DR receives an IGMP Report message for a new group, the DR determines if the group is RP-based by examining the group address. If the address indicates a SM group, the DR performs a lookup in the associated group's RP-list to determine the primary RP for the group. The draft specification describes a procedure for the selection of the primary RP and the use of alternate RPs if the primary RP becomes unreachable.

After performing the lookup, the DR creates a multicast forwarding cache for the (*, group) pair and transmits a unicast PIM-Join message to the primary RP. The (*, group) notation indicates an (any source, group) pair. The intermediate routers forward the unicast PIM-Join message and create a forwarding cache entry for the (*, group) pair. Intermediate routers

create the forwarding cache entry so that they will know how to forward traffic addressed to the (*, group) pair downstream to the DR originating the PIM-Join message.

I.2. Directly Attached Source Sends to a Group

When a host first transmits a multicast packet to a group, its DR must forward the datagram to the primary RP for subsequent distribution across the group's delivery tree. The DR encapsulates the multicast packet in a PIM-SM-Register packet and unicasts it to the primary RP for the group. The PIM-SM-Register packet informs the RP of a new source, which causes the active RP to transmit PIM-Join messages back to the source station's DR. The routers lying between the source's DR and the RP maintain state from received PIM-Join messages so that they will know how to forward subsequent unencapsulated multicast packets from the source subnetwork to the RP.

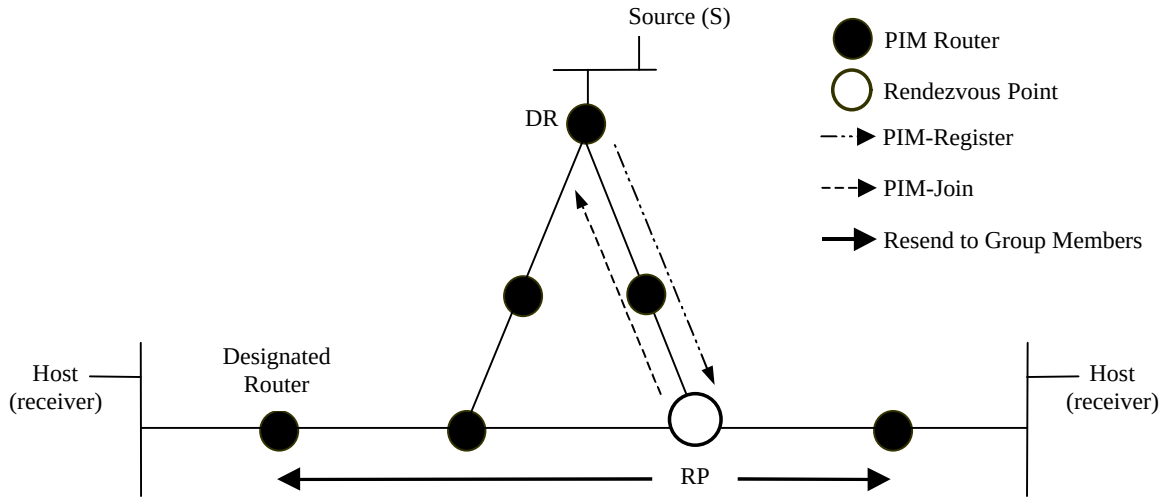


Figure I.3. Source Sends to a Multicast Group

The source's DR ceases to encapsulate data packets in PIM-SM- Registers when it receives Join/Prune messages from the RP. At this point, data traffic is forwarded by the DR in its native multicast format to the RP. When the RP receives multicast packets from the source station, it resends the datagram on the RP-shared tree to all downstream group members.

I.3. RP-Shared Tree or Shortest Path Tree (SPT)

The RP-shared tree provides connectivity for group members but does not optimize the delivery path through the internetwork. PIM-SM allows receivers to either continue to receive multicast traffic over the RP-shared tree or over a source-rooted shortest-path tree that a receiver subsequently creates. The shortest-path tree allows a group member to reduce the delay between itself and a particular source.

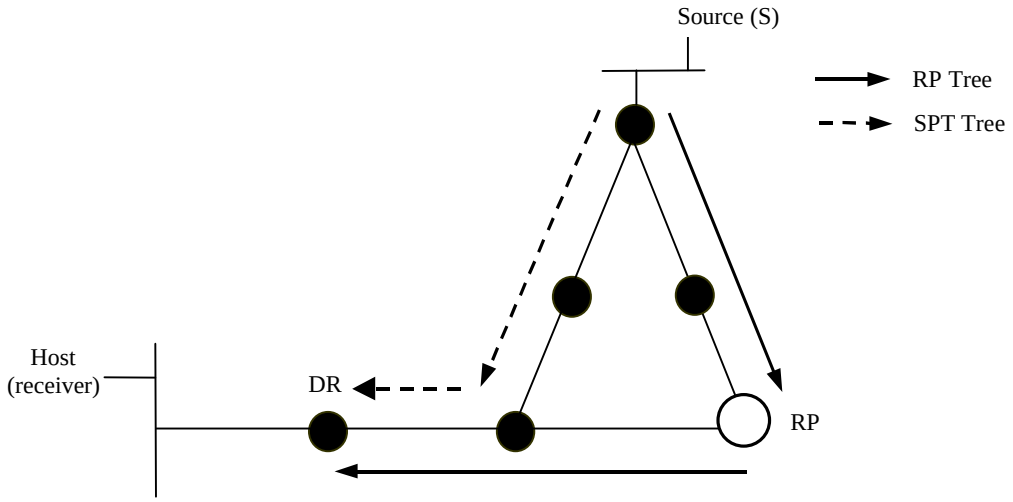


Figure I.4. RP-Shared Tree (RP Tree) and Shortest-Path Tree (SPT)

A PIM router with local receivers has the option of switching to the source's shortest-path tree as soon as it starts receiving data packets from the source station. The changeover may be triggered if the data rate from the source station exceeds a predefined threshold. The local receiver's DR does this by sending a Join message toward the active source. At the same time, protocol mechanisms guarantee that a Prune message for the same source is transmitted to the active RP. Alternatively, the DR may be configured to continue using the RP-based tree and never switch over to the source's shortest-path tree.

Appendix II Adaptive Multicast Transfer protocol (AMTP)

II.1. Introduction

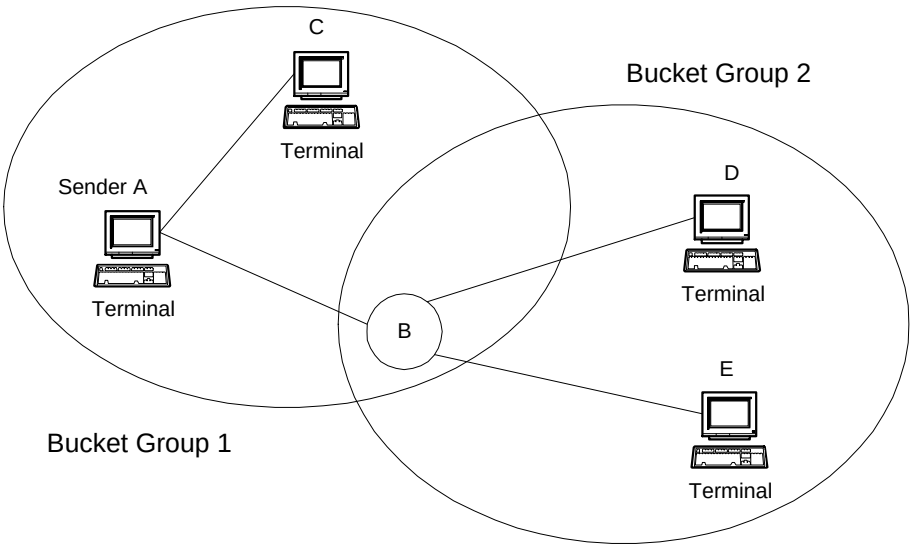
In AMTP the multicast data flow from the sender to the group is controlled by a bucket algorithm. A time window together with a set of control information for a certain time interval is called a bucket. After all packets of a time window have been forwarded, the sender (or extended intermediate nodes, so-called active routers) prompt their direct successor nodes (either active routers or receivers) to confirm the packets of the bucket. Following a timeout the collected ACKs are aggregated and sent to the next active router towards the source. Based on the information contained in these ACKs the sender either starts selective retransmissions or may clear the bucket's contents (if the transmission was successful). Since the number of buckets that may be used in parallel is limited, the multicast flow is controlled automatically.

Appendix III Scalable Reliable Multicast Transport Protocol (SRMT)

III.1. Introduction

SRMT is based on a distributed bucket algorithm similar to the one used in AMTP (Appendix II). In contrast to AMTP, SRMT allows intermediate nodes to store packets in buckets and to locally retransmit them if they are lost on the next hop. This significantly reduces the cost of retransmissions because each packet is retransmitted exactly where it has been lost.

SRMT distributes multiple copies of the (ATMP) bucket algorithm among the SRMT nodes of the multicast tree. A node group which performs the bucket algorithm consists of a parent node (acting as SRMT sender) and of child nodes (acting as SRMT receivers). Together these nodes form a so-called bucket group. The following Figure gives a simple example for two bucket groups.



In the first group sender A acts as SRMT sender whereas receiver C and router B act as SRMT receivers. The second bucket group consists of the nodes B, D and E, with B being the SRMT sender and D and E being SRMT receivers.

- END -