

BMP: AN EFFICIENT AND SCALABLE MULTICAST PROTOCOL

Siavash Samadian-Barzoki¹, Mozafar Bag-Mohammadi¹ and Nasser Yazdani²
¹{s.samadian, mozafarb}@ece.ut.ac.ir, ²yazdani@ut.ac.ir

Abstract. It is well known that IP multicast suffers from deployment problem. To solve the problem, many alternative group communication methods like application layer multicast and branching point (BP) based multicast have been proposed. Among them, BP based approaches have many invaluable features like incremental deployment, low memory requirement and hence high scalability. We propose a new BP based method called BMP (Branching based Multicast Protocol) which aimed to solve the existing problems with previous methods. Unlike other BP based proposals, BMP packet forwarding method has very little impact on unicast packet forwarding. In addition, BMP avoids packet duplications in other BP based approaches which occurs in the presence of network asymmetry.

1. INTRODUCTION

Many applications like video conferencing use multicast services to reduce the network load and data distribution delay. In several multicast routing protocols, the multicast tree is identified by its branching points where multicast data is delivered from one branching point to another using native unicast. A branching point (BP) in a multicast tree is a router which forwards the multicast data packets to multiple next-hop routers on the tree. We call these protocols BP based protocols. The basic motivation is that in typical sparse multicast distribution trees, the majority of routers are relay routers which forward incoming packets to an outgoing interface [6] [11]. In other words, the minority of routers are BPs. Traditional multicast routing protocols [4] [10] require group-related state maintenance at all on-tree routers, commonly known as Multicast Forwarding Table (MFT). In BP based protocols, only the BP keep MFT entries. All non-BPs forward multicast data packets using unicast forwarding engine. As a result, these protocols have low memory requirements compared to traditional multicast routing protocols.

The other benefits of BP based protocols are incremental deploy-ability, no need for domain-wide address allocation mechanism, and the tree construction in forward direction. Among them, the incremental deploy-ability is a vital feature as multicast suffers from deployment issues. Traditional multicast routing protocols like PIM-SM [10] and DVMRP [4] require every router in the network to implement the protocol. In contrast, BP based protocols like REUNITE [6] and HBH [7] have native support for incremental deployment. Since all packets have unicast destination addresses, routers that not implemented the protocol will forward the packets in unicast. Despite the fact that these routers can not act as BPs, they still can take part in multicast data distribution. The presence of such a router does not affect the correctness of the protocol but may result in some efficiency penalty [6].

Nevertheless, the main drawback of these approaches is duplicate lookup in both multicast and unicast forwarding

engines when handling data packets. In BP based protocols, when a multicast data packet arrives at an on-tree router, it looks up for a matching entry in MFT. If an entry is found, the packet is sent to the corresponding next hop routers. Otherwise, the packet is handled by unicast forwarding engine. Hence, multicast data packets in non-branching routers require two successive lookups in order to be forwarded. More importantly, these additional lookups are also needed for unicast data packets because the routers can not distinguish them from multicast ones.

However, each of the current BP based approaches has some flaws in their tree construction method which are briefly explained. For example, REUNITE [6] fails to construct SPT in the presence of unicast routing asymmetries. Asymmetries may also lead REUNITE to unnecessary packet duplications on certain links. Also, the departure of one receiver may change the route for another one. In SEM [1], when a new member joins the multicast session or one of the existing members leaves the session, the whole multicast tree must be constructed again. In the tree constructed by HBH [7], in addition to BPs, some non-BP nodes may also exist. Besides, HBH may also create duplicate packets on some links in asymmetrical networks.

BMP is proposed to solve the aforementioned problems. To avoid the excessive lookup problem in BP based protocols, BMP has a two-tire solution. First, it suggests the assignment of a special value to the *Protocol* field in the IP header of the multicast data packets to distinguish them from unicast ones. Second, BMP needs small modification in the parser code of the BMP-aware routers to direct the data packets correctly to their corresponding forwarding engine, i.e. unicast or multicast. Simulation results in [2] show that BMP reduces the overall number of required lookups in the constructed multicast tree at least by 45.89% compared with non-optimized BP based protocols. We do not explain parser modifications here and address interested readers to [2] for full explanation. In [2], we have proposed a general solution to eliminate excessive lookups which is applicable to every BP based protocols.

To solve the tree construction problems exist in other proposals, BMP uses a request and replay approach. The *Join* message of every new receiver reaches multicast senders. In response, the sender sends a *Join-ack* message toward the new receiver. This message is processed by every router in the path between the sender and the receiver. When a router finds that it is a new BP of tree, it sends a *Replace* message towards the previous BP (or source when the router is the first BP of the tree) and a modified *Join-ack* message toward the receiver. The Previous BP responds with a *Replace-ack* message which is sent toward the new BP. The *Join-ack* and *Replace-ack* messages in conjunction with another message named *Leave-ack* will install and delete appropriate states in on-tree routers.

In section two, we briefly introduce BP based protocols. We discuss BP concept in section three. Then, we present the BMP protocol in section four. Finally, section five concludes the paper.

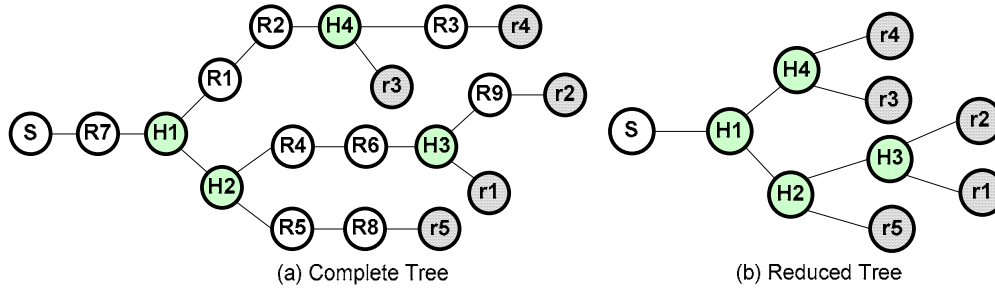


Figure 1. Ordinary protocols use complete trees where BP based protocols use reduced trees for multicast data distribution. The reduced trees do not contain the relay nodes.

2. RELATED WORK

REUNITE (REcursive UNicast TrEes) [6] implements multicast distribution based on the unicast routing infrastructure. They separate multicast routing information in two tables: a Multicast Control Table (MCT) that is stored in the control plane (slow path [13]) and a Multicast Forwarding Table (MFT) installed in the data plane (fast path [13]). Non-branching routers simply keep group information in their MCT and the branching nodes keep MFT entries which are used to forward multicast data packets. A multicast session is denoted by a $\langle S, P \rangle$ tuple, where S is the unicast address of the source and P is a 16-bit port number allocated by the source. Class D addresses are not used in their method. When the first receiver leaves the multicast session, the tree maintenance will be very complex in REUNITE.

Reference [5] proposes a scheme to achieve a same state reduction at non-branching nodes as REUNITE. However, it requires dynamically setting up tunnels between adjacent branching routers in a multicast tree. Using an additional layer of IP header introduces 20 more bytes overhead in each header and also may result in packet fragmentation. In addition, to support dynamic membership, a sophisticated and complex control protocol is needed to dynamically set up and tear down tunnels.

HBH [7] is proposed to solve deficiencies of REUNITE. In this protocol, MFT contains the IP address of next BPs instead of the receivers. Therefore, the tree is completely represented by its BP and the receivers. Furthermore, HBH identifies a multicast session using the channel concept existing in EXPRESS [12].

Simple Explicit Multicast (SEM) [1] is another BP based method with less tree construction complexity than REUNITE and HBH. SEM uses the receivers' list to construct the multicast distribution tree. The structure of MFTs in SEM is similar to HBH. When a new member joins the multicast session in this protocol or one of the existing members leaves the session, the whole multicast tree must be constructed again. However, this is an intolerable drawback which severely limits SEM application to semi-static and fully-static groups.

Originally proposed to reduce the forwarding cost in Xcast [9], Sender Initiated Multicast (SIM) [3] is a BP based protocol as well. Basically, SIM has two forwarding modes: list mode and preset mode. In the list mode, an SIM sender always attaches the receivers' list to multicast data packets. In contrast, the SIM sender periodically attaches the receivers' list to multicast data packets in the preset mode which is the prevalent SIM mode. SIM capable routers construct an MFT-like table to forward the packets in the preset mode. Since SIM uses receivers' list to construct the tree, it suffers from dynamicity of multicast groups same as SEM.

3. BRANCHING POINT IDEA

We classify on-tree nodes in a multicast tree into three distinct categories based on the number of their branches [11]:

1- **Member nodes:** Examples of these nodes are leaf receivers and occasionally the senders. These nodes have degree one on the multicast distribution tree graph. In Fig. 1a, nodes S and r1 to r5 are member nodes.

2- **Relay nodes:** These on-tree nodes have degree two and just relay the multicast data packets from an incoming interface to another outgoing interface. Their presence is ignored in BP based protocols. Traditional multicast schemes maintain multicast states in relay (or non-BP) nodes consuming invaluable memory space in their data paths [10] [4]. On the contrary, in BP based protocols, some protocols maintain these states in control plane [6] [7] and some do not require them at all [1] [3] [5]. These states are not used in BP based protocols when forwarding the multicast packets. They may only be stored for tree maintenance purposes. Relay nodes are shown with Ri in Fig. 1a which i varies between 1 and 9.

3- **Branching points:** These nodes have degree more than two and are responsible for making several copies from multicast data packets and sending them to next branching points. In BP based protocols, only these nodes are allowed to keep entries in their MFT in data path. H1 to H4 are example of BPs in Fig.1a.

Since the relay nodes are not used in multicast distribution tree for BP based protocols, we differentiate between the notions of multicast tree in these protocols from other multicast protocols. We use the terms "complete tree" and "reduced tree" when referring to the multicast tree in ordinary protocols and BP based protocols respectively [11]. These two kinds of tree are shown in figure 1. The complete tree may contain all three types of on-tree nodes, while the reduced tree only consists of member nodes and branching points.

Figure 2 depicts the differences between MFT structure in "complete tree" and "reduced tree". The MFT in complete tree consists of incoming link, outgoing links and group identifier (GI). Usually, the GI is (S, G) or $(*, G)$ pair where S is the source IP address, G is the group address and * is don't care. In a reduced tree, the MFT contains the GI and the IP addresses of next hop BPs and/or receivers. Here, GI may be (S, P) or (S, G) tuples, where P is the port number allocated by source. Even though, the size of MFT is less in a complete tree, the number of routers requiring MFT maintenance is smaller in a reduced tree. As a result, the total memory consumption in a reduced tree is less [6].

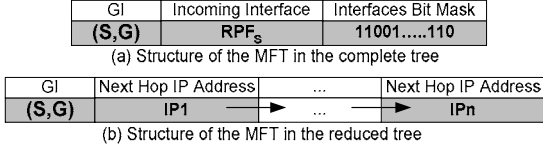


Figure 2. The MFT structure in complete and reduced trees

4. BMP

First, we discuss the data distribution mechanism of BMP, and then tree construction processes. Generally, BMP must hold following constraints in order to work correctly:

- 1- Every BP must be aware of its parent BP and children in reduced tree. The children can be other BP(s) and/or receiver(s).
- 2- Every non-BP must be aware of next BP in the tree. This information is stored in the MCT (Multicast Control Table) and helps non-BP routers to build MFT correctly when they become BP for new receiver.

4.1 Multicast Data Distribution

The data distribution mechanism of BMP is shown in figure 3. Here, S sends multicast data directly toward first BP of tree i.e. R1 using its MFT. R1 makes two copies of the incoming packet according to corresponding entry in its MFT and sends them toward r3 and R3. R2 forwards the received multicast packet same as a regular unicast packet. Using the appropriate entry in MFT, R3 sends three copies of incoming packets toward destinations r4, r1 and R5. The same is true for two other BPs of tree i.e. R5 and R6. It worth noting that, older routers U1 and U2 which have not deployed BMP yet, can forward the received multicast packet as unicast packet. This is possible because all multicast data packets have unicast destination.

4.2 Tree Construction

We describe BMP tree construction principles through an example network (see figure 4). First, suppose that r1 wants to receive multicast data of sender S. In BMP, all the receivers and hence r1 are designated routers that are directly attached to the

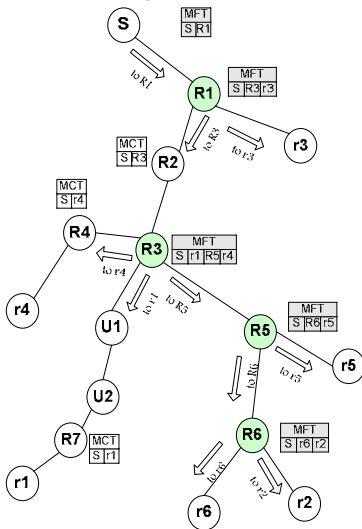


Figure 3. The multicast data distribution in BMP.

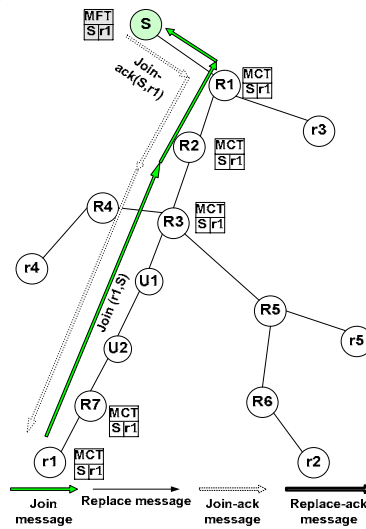


Figure 4. First receiver joins the multicast session.

interested host(s). The hosts use IGMP [14] to inform designated routers about their interest in a particular group. This part of BMP is same as traditional IP multicast model. However, r1 sends a *Join* message toward S. The *Join* message reaches the sender without any interception by intermediate routers along the way. S adds r1 to its MFT and sends a *Join-ack* message toward r1. The BMP-aware routers along the way between S and r1 add r1 to their MCT (see figure 4).

The older routers U1 and U2 which have not implemented BMP, simply forward the *Join-ack* message without any interception. The reason is that all BMP ack messages use the Router Alert option. As stated in [15], routers that do not recognize this option shall ignore it and routers that recognize the option shall examine packets more closely to determine whether or not further processing is necessary. Therefore, All BMP-aware routers on the path will examine the *Join-ack* message and other routers simply forward the message to its next hop towards the destination.

Now, r2 joins the multicast session and sends a *Join* message toward S. When S receives r2 *Join* message, it checks the outgoing interface toward r2. Since this interface is same as one previously computed for r1, S does not change content of its MFT. However, S sends a *Join-ack* message toward r2 as usual. The message is intercepted by every BMP-aware router along the way between S and r2 until it reaches a new BP of tree i.e. R3. R3 terminates the original *Join-ack* message and sends a new one down the tree. This helps downstream routers keep themselves up-to-date about their previous BP (in this example R3). R3 also must inform its previous BP about formation of new BP. Therefore, it sends a *Replace*(R3, r1, r2) toward S. S replaces r1 with R3 in its MFT when it receives the *Replace* message. Furthermore, S sends a *Replace-ack*(R3, r1, r2) toward R3. The *Replace-ack* message changes the MCT of router between S and R3. Figure 5 shows network status after r2 joined the tree.

Now, r3 sends a *Join* message to S and triggers it to send a *Join-ack* message toward r3. R1 recognize that it is a new BP of tree after processing the *Join-ack* message of S. Therefore, it terminates received *Join-ack* message and sends a

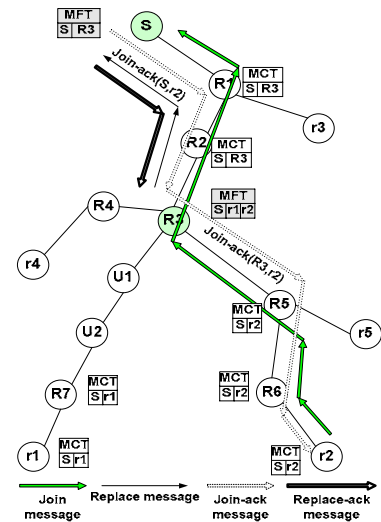


Figure 5. r2 joined the tree.

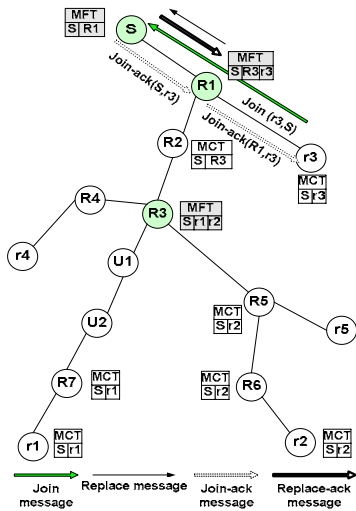


Figure 6. Multicast tree after r3 join.

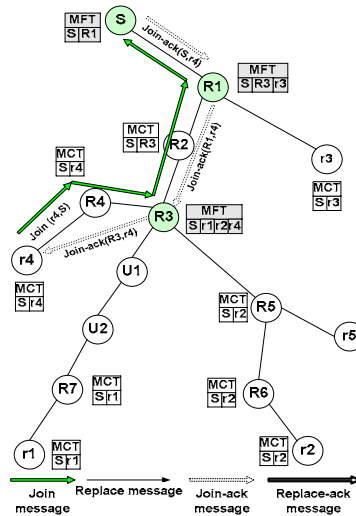


Figure 7. The tree after r4 join.

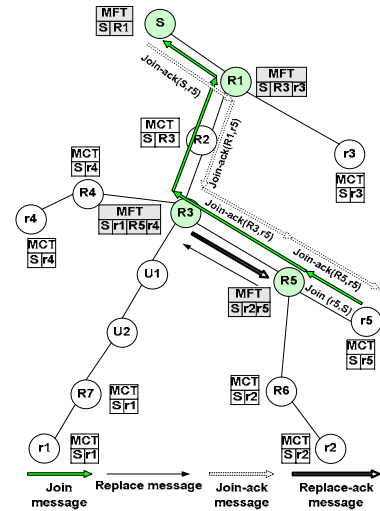


Figure 8. The tree after r5 join.

Replace(R1, r3, R3) to S and new *Join-ack*(R1, r3) to r3. It is worth noting that the MFT of R1 contains new receiver (r3) and next BP (R3) which is driven from R1 MCT (see figure 6).

The join process of r4 does not create new BP, but it adds r4 to the MFT of R3 as shown in figure 7. In this case, R3 does not send *Replace* message and only changes the received *Join-ack* message. Finally, when r5 joins the multicast tree, R6 become a new BP. The join process of r5 has a noticeable property. As can be seen in figure 8, the required *Replace* and *Replace-ack* messages are only sent in the affected part of tree. It means that the join process has some level of locality property and many join process can be done simultaneously.

5. CONCLUSION

We proposed a new BP based protocol which eliminates main inefficiencies that exist in the current proposals. The protocol uses a request and replay mechanism to insert proper states in the network nodes. Using idea in [2], BMP has minimum effect on the unicast packet forwarding. Also, BMP constructs so-called “reduced tree” with good quality in comparison with other approaches. These in conjunction with incremental deploy-ability feature make BMP as a promising candidate to implement IP multicast. We have implemented BMP protocol in NS-2 network simulator and verified its correctness in various circumstances such as network asymmetry. As a main drawback, the control overhead of BMP is relatively high compared to other BP based and conventional multicast protocols. As a future works, we will examine various aspects of incremental deployment property of BP based approaches and hence BMP such as quality of constructed tree. Also, we want to figure out the memory consumption of BMP and total reduction in the number of required MFT entries and size of MFT.

REFERENCES

- [1] Boudani, B. Cousin, "SEM: A New Small Group Multicast Routing Protocol", IEEE ICT2003, Tahiti, Feb. 2003.
- [2] M.Bag-Mohammadi, S.Samadian-Barzoki, N.Yazdani, "Improving Data Distribution in Branching Point Based Multicast Protocols", ICOIN2004, Posan, Korea, Feb. 2004.
- [3] V. Visoottiviseth, H. Kido, Y. Kadobayashi, S. Yamaguchi, "Sender-Initiated Multicast Forwarding Scheme", IEEE ICT2003, Tahiti, Feb. 2003.
- [4] D.Waitzman, C.Partridge, S.Deering, "Distance Vector Multicast Routing Protocol", RFC 1075, Nov.1988
- [5] J. Tian, G. Neufeld, "Forwarding state reduction for sparse mode multicast communication", IEEE INFOCOM'98, San Francisco, California, Mar. 1998.
- [6] I. Stoica, T. S. Eugene Ng, H. Zhang, "REUNITE: A Recursive Unicast Approach to Multicast", IEEE INFOCOM'2000, Mar. 2000.
- [7] L. H. M. K. Costa, S. Fdida, O. C. M. B. Duarte, "Hop By Hop Multicast Routing Protocol", ACM SIGCOMM'01, San Diego, USA, August 2001.
- [8] Paxson, "End-to-End Routing Behavior in the Internet", ACM SIGCOMM '96, Stanford, CA, August 1996.
- [9] R. Boivie, N. Feldman, Y. Imai, W. Livens, D. Ooms, O. Paridaens, "Explicit Multicast (Xcast) Basic Specification", IETF Internet Draft, 2003
- [10] S.Deering, et al., "The PIM architecture for wide-area multicast routing", IEEE/ACM Trans. on Networking, Vol.4, No.2, April 1996
- [11] J. Pansiot and D. Grad, "On routes and multicast trees in the Internet," *ACM Computer Communication Review*, vol. 28, no. 1, pp. 41–50, Jan.1998.
- [12] H.W.Holbrook, D.R.Cherton, "IP multicast channels: EXPRESS support for large-scale single-source applications", ACM SIGCOMM'99, Sept. 1999.
- [13] J. Aweya, "IP Router Architectures: An Overview", *Journal of Systems Architecture*, 46 (2000) pp.483-511, 1999.
- [14] B. Cain, S. Deering, I. Kouvelas, B. Fenner, A. Thyagarajan, "Internet Group Management Protocol, Version 3" RFC 3376, October 2002
- [15] D. Katz, "IP Router Alert Option", RFC 2113, Feb. 1997.