Distributed Core Multicast (DCM): a routing protocol for IP with application to host mobility

Ljubica Blazević Jean-Yves Le Boudec

Institute for computer Communications and Applications (ICA)
Swiss Federal Institute of Technology, Lausanne
email: {Ljubica.Blazevic, Leboudec}@epfl.ch

Abstract

We consider the problem of multicast routing in a large single domain network with a very large number of multicast groups with small number of receivers. Such a case occurs, for example, when multicast addresses are statically allocated to mobile hosts, as a mechanism to manage Internet host mobility [11]. For such networks, existing dense or sparse mode multicast routing algorithms do not scale well with the number of multicast groups.

We introduce an alternative solution called Distributed Core Multicast (DCM). DCM is based on an extension of the centre-based tree approach [2],[5]. It uses several core routers, called Distributed Core Routers (DCRs) and a special control protocol among them. The objectives are: (1) to avoid multicast group state information in backbone routers, (2) to avoid triangular routing across expensive backbone links, (3) to scale well with the number of multicast groups. We describe how our approach can be used to support mobile hosts. We argue that, when DCM is used to route packets to mobile hosts, good performance can be achieved during handoffs.

1 Introduction

We present a solution for providing low overhead delivery of multicast data in a large single domain network for a very large number of small groups. Such a case occurs when the number of multicast groups is very large (for example, greater than a million), the number of receivers per multicast group is very small (for example, less than five) and each host is a potential sender to a multicast group. We propose to apply this solution to support mobility in the Internet where a multicast address is statically assigned to a mobile host.

MSM-IP (Mobility Support using Multicasting in IP)[11] introduces the generic architecture to support host mobility in the Internet by using multicasting as the mechanism for routing packets to the mobile hosts. Every mobile host is statically assigned and addressed by a multicast address. A multicast router in a mobile host's current cell is responsible for joining the multicast distribution tree on behalf of a mobile host. This multicast router typically coexists with the base station in a cell. A base station that anticipates the arrival of a mobile host initiates a mobile host's group membership registration. Thus, a multicast group assigned to a mobile host has a few recipients. At the same time, a mobile host receives data only from a base station in its current cell. Hence, we have a form of unicast end-to-end communication which uses multicast routing. See [11] for a detailed description of the implications of using multicast addresses to support mobile hosts.

The benefits of using multicast addresses to support mobile IP are twofold:

- A fixed multicast address is assigned to a mobile host. This simplifies the task of the correspondent host and eliminates the need of explicit address translation (as in other proposals: IETF Mobile IP [12], SONY Scheme [14], IPv6 Mobility Proposal [3]).

- When a multicast address is assigned to a mobile host, base stations in neighbouring cells may have already joined a multicast group assigned to a mobile host through advance registration. This minimises packet losses and latency when a mobile host changes its location.

  In IETF Mobile IP, a response to a handoff of the mobile host happens only after the home agent becomes aware of the host's new location. Mobility support in IPv6 proposes router-assisted smooth handoffs [13] in order to reduce packet losses and latency during handoff. The mobile host can inform a previous router, which is the entity that served to deliver packets to the mobile host at its previous location, about its new location. Then the previous router must intercept any packets for the mobile host's previous care-of-address and tunnel them (using IPv6 encapsulation) to the mobile host's new care-of-address. As soon as the mobile host receives such encapsulated packet, it discovers the correspondent host that sent the packet and sends a binding update to the correspondent host. When the mobile hosts changes its point of attachment the previous router cannot immediately start tunnelling packets to the mobile host's new care-of-address. A certain interval of time passes, till the mobile host first configures its new care-of-address and then requests the previous router to act as its temporary home agent for the packets destined to the mobile host's previous care-of-address. During this time interval packets could be lost.

We propose an extension to an existing multicast routing protocol which aims to scale better than existing protocols when applied to support mobile hosts. Relevant aspects of existing multicast routing protocols are described in Section 2. Recent sparse multicast routing protocols, such as the protocol independent multicast (PIM-SM) [5] and the core-based trees (CBT) [2], build a single delivery tree per multicast group which is shared by all senders in the group. This tree is rooted at a single centre router called "core" in CBT, and "rendezvous point" (RP) in PIM-SM.

Those centre-based routing protocols have the following potential shortcomings:

- Traffic for the multicast group is concentrated on the links along the shared tree, mainly near the core router.

- Finding an optimal centre for a group is a NP-complete problem and requires the knowledge of the whole topology [20]. Current approaches typically use either administrative selection of centres or some simple heuristics [16]. Data distribution through a single core router could cause non optimal distribution of traffic in the case of a bad positioning of the core (or the RP) router with respect to senders and receivers. This problem is known as a triangular routing problem.

PIM-SM is not only a centre-based routing protocol, but it it also uses source-based trees. With PIM-SM, destinations can start building source-specific trees for sources with a high data rate. This partly addresses the shortcomings mentioned above, however, at the expense of having routers on the source-specific tree keep source-specific state. Keeping the state for each sender is undesirable when the number of senders is large.

We propose an alternative solution, called Distributed Core Multicast (DCM), for the efficient and scalable delivery of multicast data under the assumptions that are satisfied when multicast is used to support mobile IP (large number of multicast groups, a few receivers per group and a potentially large number of senders to a multicast group). Our solution is based on an extension of the centre-based tree approach.

We consider a network model that consists of several areas connected via the backbone area (see Figure 1). The objectives we want to achieve are: (1): to avoid multicast group state information in backbone routers, (2): to avoid triangular routing across expensive backbone links and (3) to scale well with the number of multicast groups. In this paper, we describe DCM for a large single domain network.

Here we summarize our proposal. DCM is based on several core routers per multicast group, called Distributed Core Routers (DCRs).

- The DCRs in each area are located at the edge of the backbone. The DCRs act as backbone access points for data sent by senders inside their area to receivers outside this area. A DCR also forwards the multicast data received from the backbone to receivers in the area it belongs to. When a host wants to join the multicast group M, it sends a join message. This join message is propagated hop-by-hop to the DCR inside its area that serves the multicast group. Conversely, when a sender has data to send to the multicast group, it will send the data encapsulated to the DCR assigned to the multicast group.

- The Membership Distribution Protocol (MDP) runs among the DCRs serving the same range of multicast addresses. It is fully distributed. MDP enables the DCRs to learn about other DCRs that have group members.

- Distribution of data uses a special mechanism among the DCRs in the backbone area, and the trees rooted at the DCRs towards members of the group in the other areas. We propose a special mechanism for data distribution among the DCRs that does not require that non-DCR backbone routers perform multicast routing.

We advocate the introduction of the DCRs close to any sender and receivers as solution for avoiding converging traffic to be sent to a single centre router in the network. Data sent from a sender to a group within the same area is not forwarded to the backbone. Our approach alleviates triangular routing problem common to all centre-based trees. Unlike PIM-SM, DCM is suitable for groups with many sporadic senders.

In this paper we examine the properties of DCM in a large single domain network. However, DCM is not constrained to a single domain network. Interoperability of DCM with other inter-domain routing protocols is object of ongoing work.

The structure of this paper is as follows. In the next section we give an overview of the existing multicast routing protocols. In Section 3, we give a detailed description of DCM for scalable delivery of multicast data. In Section 4 we show that DCM is suitable to support IP host mobility. We have implemented DCM using the Network Simulator (NS) tool [1]. In Section 5 we give a preliminary evaluation of our implementation.

2 Overview of Multicast Routing Protocols

There are two basic families of algorithms that construct multicast trees used for the distribution of IP multicast data: source specific trees and group shared trees. In the former case an implicit

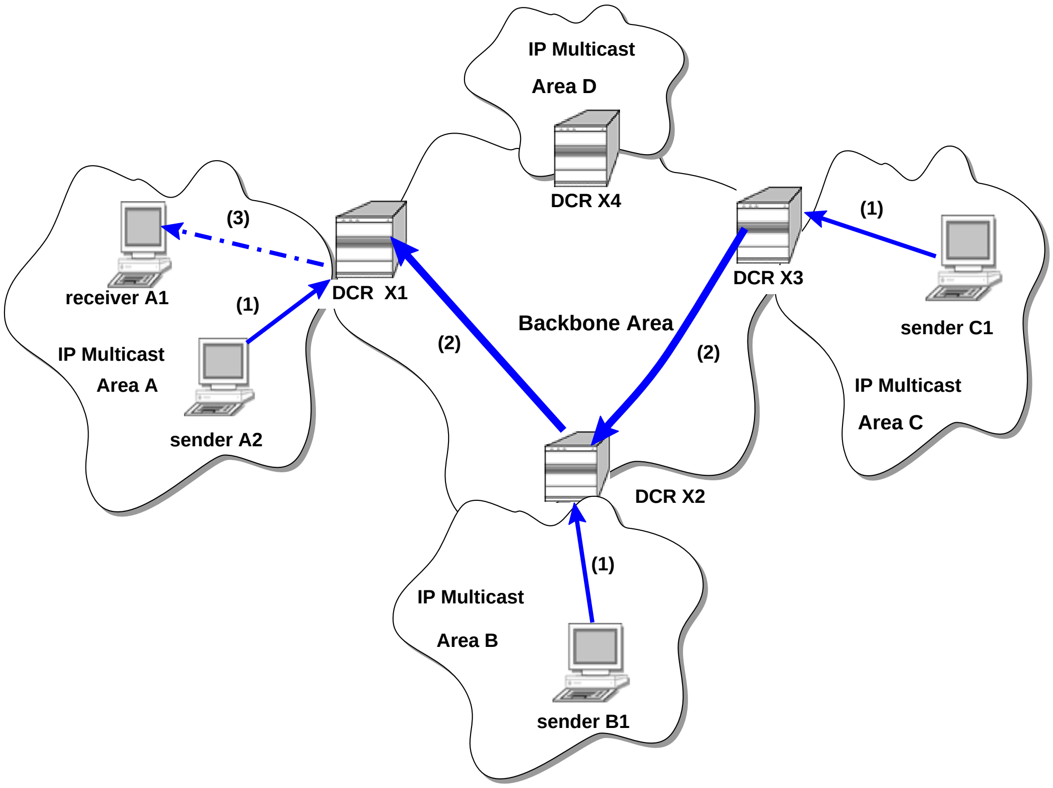


Figure 1: Model of a large single domain network and an overview of data distribution with DCM. We show one multicast group M and DCRs X1, X2, X3 and X4 that serve a range to which M belongs to. Step (1): Senders A2, B1 and C1 send data to the corresponding DCRs inside their areas. Step (2): DCRs distribute the multicast data across the backbone area to DCR X1 that needs it. Step (3): A local DCR sends data to the local receivers in its area.

spanning tree per source is calculated, which is minimal in terms of transit delay from a source to each of the receivers . In the latter case only one shared tree that is shared by all sources is built. There are two types of shared trees. One type is the Steiner minimal tree (SMT)[21]. The main objective is to build a tree that spans the group of members with the minimal cost and thus globally optimises the network resources. Since the Steiner minimal tree problem is NP-complete, numerous heuristics have been proposed [19]. No existing SMT algorithms can be easily applied in practical multicast protocols designed for large scale networks [20]. The other type of shared trees is a centre-based tree that builds the shortest path tree rooted "in the centre" of the networks and spans only receivers of the multicast group.

Below we describe briefly existing dense and sparse mode multicast routing protocols in the Internet.

Dense mode multicast routing protocols

Traditional multicast routing mechanisms, such as Distance-vector multicast routing protocol (DVMRP) [18] and Multicast open shortest path first (MOSPF) [9], are intended for use within regions where multicast groups are densely populated or bandwidth is plentiful. Both protocols use source specific shortest path trees. These routing schemes require that each multicast router in the network keeps per source per group information.

DVMRP is based on the Reverse Path Forwarding (RPF) algorithm that builds a shortest path sender-based multicast delivery tree. Several first multicast packets transmitted from a source are broadcasted across the network over links that may not lead to receivers of the multicast group. Then the tree branches that do not lead to group members are pruned by sending `prune` messages. After a period of time, the prune state for each (source, group) pair expires and reclaim stale prune state. Subsequent datagrams are flooded again until branches that do not lead to group members are pruned again. This scheme is currently used for Internet multicasting over the MBONE.

In MOSPF, together with the unicast routing information, group membership information is flooded so that all routers can determine whether they are on the distribution tree for a particular source and group pair. Like DVMRP, MOSPF has a high routing message overhead when groups are sparsely distributed.

Core Based Trees (CBT) sparse mode multicast routing architecture

Unlike DVMRP and MOSPF, a CBT [2] uses centre based shared trees: it builds and maintains a single shared bidirectional multicast distribution tree for every active multicast group in the network. This tree is rooted in a dedicated router for a multicast group that is called the *core* and it spans all group members. Here we give a short description of how a shared tree is built and how a host sends to the group.

A host starts joining a group by multicasting an IGMP[6] host membership report across its attached link. When a local CBT aware router receives this report it invokes the tree joining process (unless it has already joined the tree) by generating a `join` message. This message is then sent to the next hop on the path towards the group's core router (Figure 2). This join message must be explicitly acknowledged either by the core router itself or by another router that is on the path between the sending router and the core, which itself has already successfully joined the tree. Once the acknowledgement reaches the router that originated the join message, a new receiver can receive the multicast traffic sent to the group. The state of the shared tree is periodically verified by exchanging of `echo` messages between neighbouring CBT routers on the shared tree.

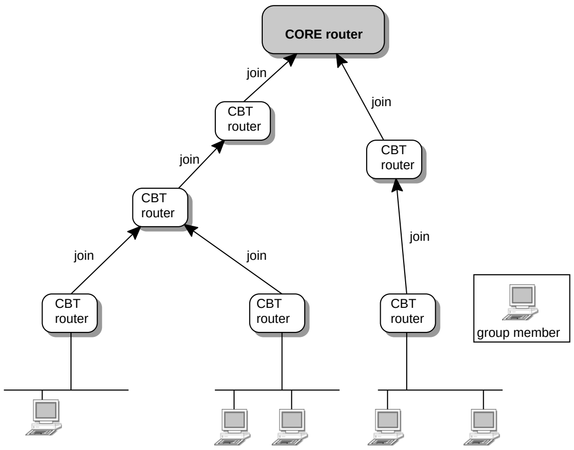


Figure 2: Construction of the shared tree with CBT

Data can be sent to a CBT tree by a sender that is not attached to the group tree. The sender originates native multicast data which is received by a local CBT router. This router finds out the relevant core router for the multicast group, and thus encapsulates the data packet(IP-in-IP) and unicasts it to the core router. After the core router decapsulates the packet it disseminates the multicast data over the group shared tree. When a multicast data packet arrives at the router on the tree, the router uses the group address as an index into the multicast forwarding cache. Then it sends a copy of the incoming multicast packet over each interface listed in the entry, except the incoming interface.

The main advantage of the CBT are that it is independent of the underlying unicast routing protocols and the routers keep forwarding information corresponding only to the multicast group and not depending on the source. This makes shared-based trees routing protocols more scalable than source-based trees routing protocols. The main disadvantage of CBT is that it has a potentially higher delay compared with DVMRP because multicast packets do not take the shortest path from the source to the destinations.

Protocol Independent Multicast Sparse Mode (PIM-SM)

PIM-SM [5] combines the source specific shortest path trees and centre based shared trees. On one hand, PIM-SM is conceptually similar to CBT: it builds a shared directed multicast distribution tree per multicast group centred at a special router called the Rendezvous Point (RP). However, unlike CBT, PIM-SM builds unidirectional trees. Sending of multicast data is similar to CBT. Initially, the sender encapsulates data in register messages and sends them directly to the RP where the data is distributed along the shared tree.

On the other hand, the unique feature of PIM-SM is that for those sources whose data rate justifies it, forwarding of multicast data from a particular source to the destination group can be shifted from the shared tree onto a source-based tree. However, the result is that the routers on the source-specific tree need to keep a source-specific state.

3 Distributed Core Multicast (DCM)

In this section, we describe the various elements of DCM. Those are: (1) addressing issues; (2) how hosts join the multicast group; (3) how membership information is distributed among DCRs; (4) how senders send to a multicast group; (4) how multicast data is distributed among DCRs; and (5) how multicast data is forwarded from DCR to members of the group inside its area.

In order to describe the DCR approach, we use the network model that is presented in Figure 1.

3.1 Addressing Issues

Within each area there are several routers that are configured to act as candidate DCRs. Candidate DCRs are known to all routers within an area by means of an intra-area bootstrap protocol [4]. This is similar to PIM-SM with the difference that the bootstrap protocol is constrained within an area. This entails periodic distribution of the set of reachable candidate DCRs to all routers within an area.

Routers use a common hash function to map multicast group address to one router from the set of candidate DCRs. For a particular group address M, we use the hash function to determine a DCR that serves[1] M.

The used hash function is $h(r(M), DCR_i)$. Function $r(M)$ takes as input multicast group address and returns the range of the multicast group, while DCR_i is the unicast IP address of the DCR. The target DCR_i is then chosen as the one of the candidate DCRs with the highest value of $h(r(M), DCR_j))$ among all $j \in \{1, .., J\}$ where J is the number of candidate DCRs in an area:

$$h(r(M), DCR_i) = \max\{h(r(M), DCR_j), j = 1, .., J\} \tag{1}$$

One possible example of the function that gives the range[2] of the multicast group address M is:

$$r(M) = M\&B \text{ , where } B \text{ is a bit mask.} \tag{2}$$

We do not present here hash function theory. For more information see [17], [4] and [15]. The benefits of using hashing to map a multicast group to DCR are the following:

- we achieve minimal disruption of groups when there is change in the candidate DCR set. This means that we have to do a small number of re-mappings of multicast groups when there is a change in the candidate DCR set. See [17] for more explanations.
- we apply the hash function *h(.,.)* as defined by Highest Random Weight (HRW)[15] algorithm. This function ensures load balancing among candidate DCRs. This is very important, because no single DCR is serving many more multicast groups than any other DCR inside the same area. We achieve, by this property, that when the number of candidate DCRs increases, the load on each DCR decreases.

All routers in all non-backbone areas should apply the same functions $h(.,.), r(.)$.

Each candidate DCR is aware of all ranges of multicast addresses for which it is elected to be a DCR. There is a function $m(r(M))$ that maps the range of the multicast group address M to another multicast address for control purposes. A DCR joins a control multicast address that corresponds to a range of multicast addresses that it serves. This multicast address is used by DCRs in different areas that serve the same range of multicast addresses to exchange control information. This is explained in more details in Section 3.3.

3.2 How hosts join the multicast group

When a host is interested in joining the multicast group M, it issues a join message via IGMP. A multicast router on its LAN, known as designated router (DR), receives the IGMP join message. DR determines the DCR inside its area that serves M as described in the Section 3.1.

The process of establishing the group shared tree is like in PIM-SM [5]. The DR sends a join message towards the determined DCR. Sending a join message forces any off-tree routers on the path to the DCR to forward a join message and join the tree. Each router on the way to the DCR keeps a forwarding state for M. When a join message reaches the DCR, this DCR becomes labelled with the multicast group M. In this way, the delivery subtree for receivers of the multicast group M in an area is established. The subtree is maintained by periodically

[1] A DCR is said to serve the multicast group address M when it is dynamically elected among all the candidate DCRs in the area to act as an access point for address M

[2] A range is the partition of the set of multicast addresses into group of addresses. A range to which a multicast group address belongs to is defined by Equation (2). e.g if the bit mask is (hex) 00000009 we get 4 possible ranges of IPv4 class-D addresses.

refreshing the state information for M in the routers (like in PIM-SM, this is done by periodically sending join messages).

Like in PIM-SM, when the DR discovers that there is no longer receivers for M, it sends a prune message towards the nearest DCR to disconnect from the shared distribution tree.

Figure 3 shows an example of joining the multicast group.

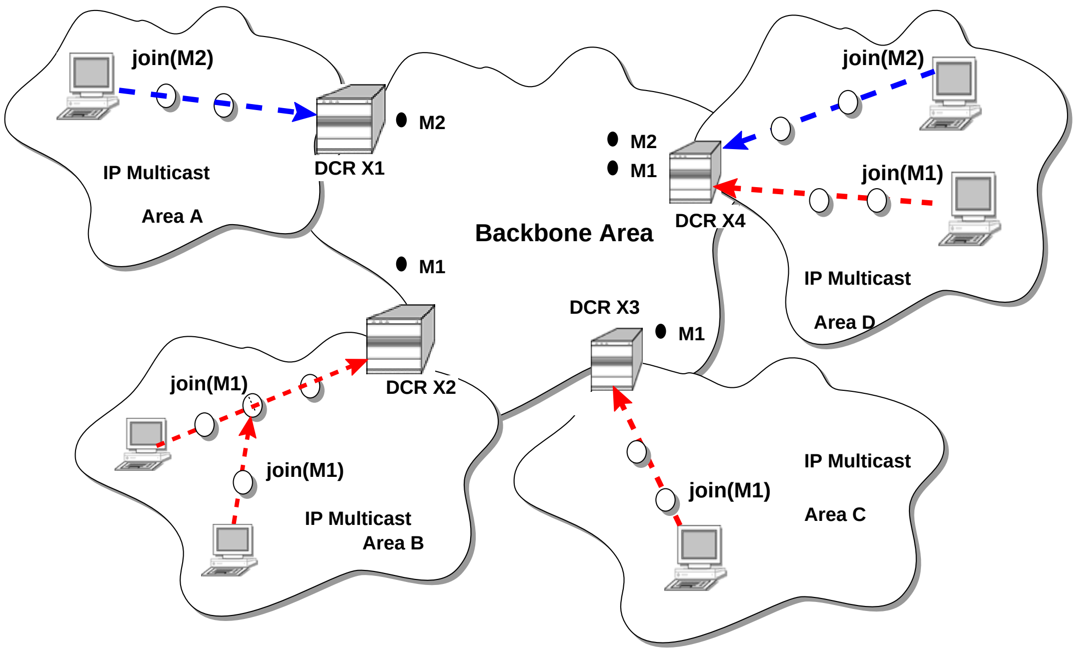


Figure 3: The figure shows hosts in four areas that join two multicast groups *M1* and *M2*. Four DCRs (X1,X2,X3 and X4) presented in the figure serve the range of multicast addresses where group addresses *M1* and *M2* belong to. A circle on the figure represents multicast routers in non-backbone areas that are involved in the construction of the DCR rooted subtree. These subtrees are showed with dashed lines. X2, X3 and X4 are now labelled with *M1*, while X1 and X4 are labelled with *M2*.

3.3 How membership information is distributed among DCRs

The Membership Distribution Protocol (MDP) is used by DCRs in different areas to exchange control information. As said above, within each non-backbone area, for each range of multicast addresses (as defined by Equation (2)) there is one DCR serving that range. DCRs in different areas that serve the same range of multicast addresses are members of the same MDP control multicast group. This group is defined by a MDP control multicast address used for exchanging control information. For example, in the network in Figure 3, DCRs X1, X2, X3 and X4 are members of the same MDP control multicast group. A DCR joins as many MDP control multicast group as the number of ranges of multicast addresses it serves. We do not propose a specific protocol for maintain the multicast tree for the MDP multicast group. This can be done by means of some existing multicast routing protocol (e.g CBT).

DCRs that are members of the same MDP control multicast group exchange the following control information:

- **periodical keep-alive messages.**

- **unicast distance information.** Each DCR sends to the corresponding MDP control multicast group, information about the unicast distance from itself to other DCRs that it learns to serve the same range of multicast addresses. This information comes from existing unicast routing tables and it is used for distribution of multicast data among the DCRs.
- **multicast group information.** A DCR, which is labelled with the multicast group M, informs DCRs in other areas responsible for M that it has receivers for M. In this way, every DCR keeps a record of every other DCR which has at least one member for a multicast address from the range that the DCR serves. A DCR should notify all other DCRs when it becomes labelled with a new multicast group or no longer labelled with some multicast group.

MDP uses its MDP control multicast addresses and performs flooding inside the groups defined by those addresses. An alternative approach would be to use MDP servers. This approach lead to more scalable, but also more complex solution. This approach is not studied in detail in this paper.

It is interesting to compare DCM to MOSPF[9] in the backbone. MOSPF is a multicast routing protocol designed atop a link-state unicast routing protocol called OSPF[10]. With OSPF, a large routing domain can be configured into areas which can be viewed as being organised in a two-level hierarchy. At the top level is a single backbone area to which all other areas connect. In MOSPF, all backbone routers have complete knowledge of all areas' group membership. Using this information together with the backbone topology information, the backbone routers calculate the multicast data distribution trees. With MOSPF, complexity in all backbone routers increases with the number of multicast groups. With DCM, DCRs are the only backbone routers that need to keep state information for the groups that they serve. As described in Section 3.1, the number of multicast groups that a DCR serves decreases as the number of candidate DCRs increases inside an area. Therefore, the load sharing among DCRs makes DCM more scalable than MOSPF.

3.4 How senders send to a multicast group

The sending host originates native multicast data for the multicast group M that is received by the designated router (DR) on its LAN. The DR determines the DCR within its area that serves M. We call this DCR the source DCR. The DR encapsulates the multicast data packet (IP-in-IP) and sends it with destination address equal to the address of the source DCR. The source DCR receives the encapsulated multicast data. This is similar to PIM-SM where the DR sends encapsulated multicast data to the RP corresponding to the multicast group.

3.5 How multicast data is distributed among DCRs

The multicast data for the group M is distributed from a source DCR to all DCRs that are labelled with M. Since we assume that the number of receivers per multicast group is not large, there are only a few labelled routers per multicast group. Our goal is to perform multicast data distribution in the backbone in such a way that backbone routers keep minimal state information while at the same time backbone bandwidth is used efficiently. We propose a solution, which can be applied in the Internet today. It uses point-to-point tunnels to perform data distribution among DCRs. With this solution, non-DCR backbone routers do not keep any state information related to the distribution of the multicast data in the backbone. In the Appendix A we propose two alternative solutions. With those solutions backbone bandwidth is used more efficiently, but at the expense of having the new routing mechanism that needs to be performed by backbone routers.

Point-to-Point Tunnels

The DCR that serve the multicast group M keeps the three following informations: (1) a set V of DCRs that serve the same range to which M belongs; (2) information about unicast distances between every pair of DCRs from V; (3) the set L of labelled DCRs for M. In this way, we present the virtual network of DCRs that serve the same range of multicast group addresses by means of an undirected complete graph $G = (V, E)$. V is defined above, while the set of edges E are tunnels between every pair of DCRs in V. Each edge is associated with a cost value that is equal to inter-DCR unicast distance.

The source DCR, called S, calculates the optimal tree that spans the labelled DCRs. In order words, S finds the subtree $T = (V_T, E_T)$ of G which spans the set of nodes L such that $cost(T) = \sum_{e \in E_T} cost(e)$ is minimised. We recognise this problem as the Steiner tree problem. Instead of finding the exact solution, we introduce a simple heuristic called Shortest Tunnel Heuristic (STH). STH consists of two phases. In the first phase a greedy tree is built by adding one by one the nodes that are closest to the tree under construction, and then removing unnecessary nodes. The second phase is further improving the tree established so far. STH is as follows:

Phase 1: Build a greedy tree

- **Step 1**: Begin with a subtree T of G consisting of the singe node S. $k = 1$.
- **Step 2**: if $k = n$ then goto **Step 4**. n is the number of nodes in set V.
- **Step 3**: Determine a node $z_{k+1} \in V$, $z_{k+1} \notin T$ closest to T (ties are broken arbitrarily). Add the node z_{k+1} to T. $k = k + 1$. Goto **Step 2**.
- **Step 4**: Remove from T non-labelled DCRs of degree[1] 1 and degree[2] 2 (one at a time).

Phase 2: Improve a greedy tree

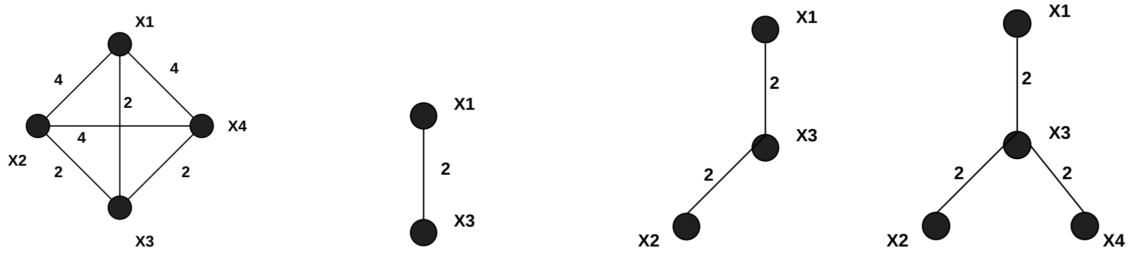
STH can be further improved by two additional steps:

- **Step 5**: Determine a minimum spanning tree for the subnetwork of G induced by the nodes in T (after the step 4).
- **Step 6**: Remove from the minimum spanning tree non-labelled DCRs of degree 1 and 2 (one at a time). The resulting tree is the (suboptimal) solution.

Figure 4, Figure 5 and Figure 6 illustrate three examples of the usage of STH. Nodes X1, X2, X3 and X4 present four DCRs that serve the multicast group M. In all examples the source DCR is X1, and the labelled DCRs for M are X2 and X4. For the first two examples, the tree that is obtained by the first phase cannot be further improved by steps 5 and 6. In the third example, steps 5 and 6 give improvements in terms of cost of the resulting tree.

[1] Degree of a node in a graph is the number of edges incident with a node

[2] A node of degree 2 is removed by its two edges being replaced by a single edge (tunnel) connecting the two nodes adjacent to the node being removed. The source DCR is never removed from a graph



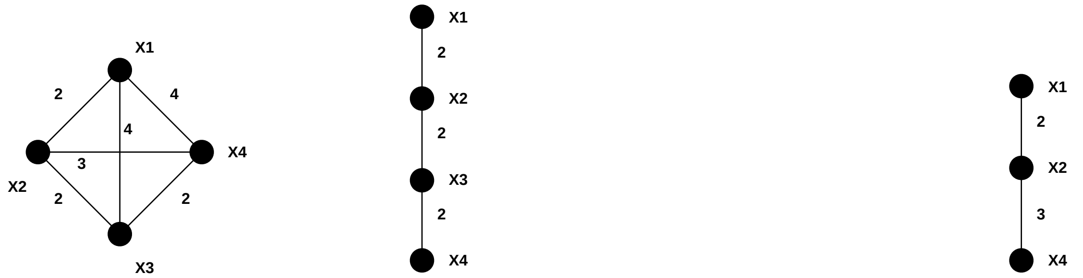
(a) the four node complete graph

(b) start with X1; add X3 to the tree because it is closer to X1 than X2 and X4

(c) X2 is added to the tree

(d) result of STH up to step 4; since X3 is of degree 3 it is not removed from the tree; this is STH solution since the tree cannot be improved by the steps 5 and 6.

Figure 4: The first example of the application of STH on the complete graph

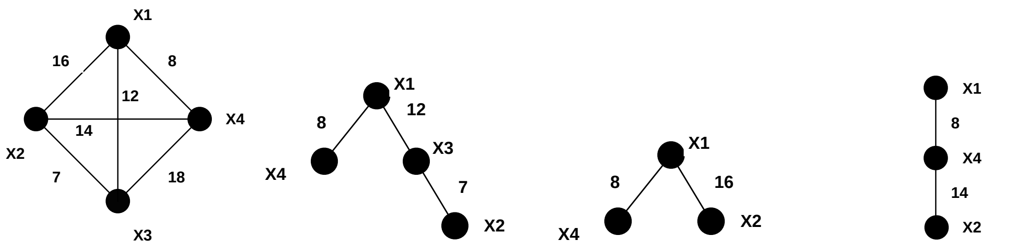


(a) the four node complete graph

(b) result of STH up to step 4

(c) X3 is of degree 2 and it is not the labelled DCR; it is removed from the tree; X2 and X4 are connected with the edge (a tunnel between them); this is STH solution since the tree cannot be improved by the steps 5 and 6.

Figure 5: The second example of the application of STH on the complete graph



(a) the four node complete graph

(b) result of STH up to step 4

(c) X3 is removed from the tree; result of STH up to step 5

(d) minimum spanning tree of X1, X2 and X4; this is STH solution

Figure 6: The third example of the application of STH on the complete graph

The source DCR applies STH to determine the distribution tunnel tree from itself to the list of labelled DCRs for the multicast group. The source DCRs puts inter-DCR distribution information in the form of an explicit distribution list in the end-to-end option field of the packet header. Under the assumption that there is a small number of receivers per multicast group, the number of labelled DCRs for a group is also small. Thus, an explicit distribution list that completely describes the distribution tunnel tree is not expected to be long.

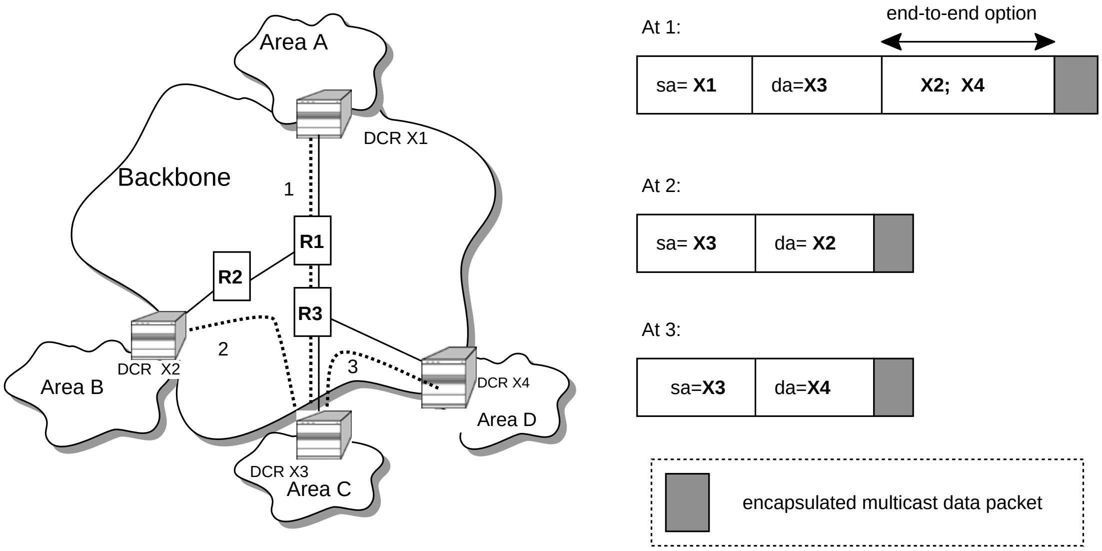


Figure 7: This figure presents an example of inter-DCR multicast data distribution by using point-to-point tunnels. The source DCR is X1 and labelled DCRs are X2 and X4. The source DCR router X1 calculates the tunnel distribution tree to X2 and X4 following the example in Figure 4. Then X1 sends the with encapsulated multicast data packet to X3. In the end-to-end option field of the packet, a distribution list is contained. X3 sends two copies of multicast data: one to X2 and the other to X4. On this figure are also presented packet formats in various points (points 1, 2 and 3) on the way from X1 to X2 and X4. A tunnels between two DCRs is shown with a dashed line.

When a DCR receives a packet from another DCR, it reads from the distribution list whether it should make copy of the multicast data and of the identities of the DCRs where it should send multicast data by tunneling. Labelled DCRs deliver data to local receivers in the corresponding area. An example that shows how multicast data is distributed among DCRs is presented in Figure 7.

We see that with DCM, only the source DCR calculates the multicast data distribution tree in the backbone. Other DCRs need only to forward the multicast data according to the already calculated distribution tree. This differs from MOSPF where calculations are performed at each backbone router that is on the distribution tree.

3.6 How multicast data is forwarded from DCR to members of the group inside its area

A DCR receives encapsulated multicast data packets either from a source that is within its area, or from a DCR in another area. A DCR checks if it is labelled with the multicast group that corresponds to the received packet, i.e whether there are members of the multicast group in its area. If this is the case, a DCR forwards the multicast packet along the distribution subtree that is already established for the multicast group (as is described in Section 3.2).

4 How to apply DCM to support host mobility

We claim the applicability of DCM as a mechanism for routing packets to mobile hosts.

We start this section with a short description of certain existing proposals for providing host mobility in the Internet and then illustrate how DCM can support mobility.

Overview of proposals for providing host mobility in the Internet

In the IETF Mobile IP proposal [12] each host has a permanent home IP address that does not change regardless of the mobile host's current location. When the mobile host visits a foreign network, it is associated with a care-of-address, that is IP address related with the mobile host current position in the Internet. When a host moves to visited network it registers its new location with its home agent. The home agent is a machine that acts as a proxy on behalf of the mobile host when it is absent. When some stationary host sends packets for the mobile host it addresses them to the mobile host's home address. When packets arrive on the mobile host's home network, the home agent intercepts them and sends by encapsulation packets towards the mobile host's current location. With this approach all datagrams addressed to a mobile host are always routed via its home agent. This causes the so-called triangle routing problem.

In IPv6 mobility proposal [3] when a handoff is performed, the mobile host is responsible for informing its home agent and correspondent hosts about its new location. In order to reduce packet losses during handoffs, IPv6 proposes router-assisted smooth handoffs [13]. Each time the mobile host moves its point of attachment from one IP subnet to the other, the mobile host configures its new care-of-address. Then a mobile host sends a so-called binding address option containing that care-of-address to its home agent, and to its correspondent hosts. For smooth handoffs, a mobile host should still accept packets at its previous care-of-address. The smooth handoffs in IPv6 consists in that the mobile host can inform a previous router, which is the entity that served to deliver packets to the mobile host at the previous location, about its new location. This is done by sending a binding update that associates the mobile host's previous care-of-address to the mobile host's new care-of-address. This means that the mobile host requests the previous router to serve as a temporary home agent for its own previous care-of-address. Thus, the previous router operates in the same way as when the mobile host's home agent (for its home address) receives a binding update from a mobile host. The previous router intercepts any packets destined for the mobile host's previous care-of-address and tunnels them (using IPv6 encapsulation) to the mobile host's new care-of-address. When the mobile host receives such encapsulated packet, it finds out the correspondent host that sent the packet and sends a binding update to the correspondent host. Router-assisted smooth handoffs reduce losing the packets during the handoffs. Still, when the mobile host changes its point of attachment the previous router cannot immediately start tunneling packets to the mobile host's new care-of-address. A certain interval of time passes because the mobile host first configures its new care-of-address and then requests the previous router to act as its temporary home agent for the packets destined to the mobile host's previous care-of-address. During this time interval packets could be lost.

The Columbia approach [7] was designed to support intracampus mobility. Each mobile host always retains one IP home address, regardless of where it is on the network. There is a number of dedicated Mobile Support Stations (MSSs) that are used to assure the mobile host's reachability. Each mobile host is always reachable via one of the MSSs. When a mobile host changes its location it has to register with a new MSS. A MSS is thus aware of all registered mobile hosts in its wireless cell. A source that wants to send a packet to a mobile host sends it to the MSS that is closest to the source host. This MSS is responsible for learning about the MSS that is closest to the mobile host and to deliver the packet. A special protocol is used to exchange information among MSSs.

MSM-IP (Mobility support using Multicasting in IP) [11] proposes a generic architecture to support host mobility in the Internet by using multicasting as a mechanism to route packets to the mobile hosts.

In this paper we do not consider a family of routing protocols designed for use in a Mobile Ad hoc NETwork (MANET) environment[8]. MANET is an autonomous system of mobile routers and hosts connected by wireless links. In MANET, mobility applies to the whole network: network topology changes rapidly and unpredictably, with the bandwidth and energy constraints in wireless links. This is different from the problem we consider in this paper: designing of a new multicast routing protocol that can support host mobility in a fixed network.

The DCM application to host mobility

The DCM is designed as a multicast routing protocol to support host mobility where each mobile host is statically assigned one class-D multicast address.

The routing of packets to the mobile host is performed with DCM. For the mobile host's assigned multicast address, within each area, there exists a DCR that serves that multicast address. Those DCRs are responsible for forwarding data to a mobile host. As was said before, the DCRs run MDP control protocol and are members of a MDP control multicast group for exchanging MDP control information.

A multicast router in the mobile host's cell initiates a joining to the multicast address assigned to the mobile host. Typically this router coexists with the base station in the cell. As described in Section 3.2 the join message is propagated to the DCR inside the area that serves the mobile host's multicast address. Then, the DCR sends to the MDP control multicast address a MDP control message that now the mobile host has registered . In the IETF Mobile IP proposal the home agent is the only place that knows the mobile host's current position and all communication with the mobile host is done via the home agent. In our approach this is avoided, because within each area there is a DCR aware of the mobile host.

In order to reduce packet latency and losses during a handoff, advance registration can be performed. The goal is that when a mobile host moves to a new cell, the base station in the new cell has already started receiving data for the mobile host. The mobile host continues receiving the data without disruption. There are several ways to perform this:

- A base station that anticipates[1] the arrival of a mobile host initiates joining to the multicast address assigned to the mobile host.
- In the case where a bandwidth is not expensive on the wired network, all neighbouring base stations can start receiving data destined to a mobile host. This guarantees that there would be no latency and packet losses during a handoff.

When a host wants to send data to a mobile host it sets the destination address to the multicast address assigned to the mobile host and sends the packet as a local multicast. This multicast packet is sent encapsulated to the DCR inside the area that serves the mobile host's

[1] The mechanism by which the base station anticipates the arrival of the mobile host is out of the scope of this paper

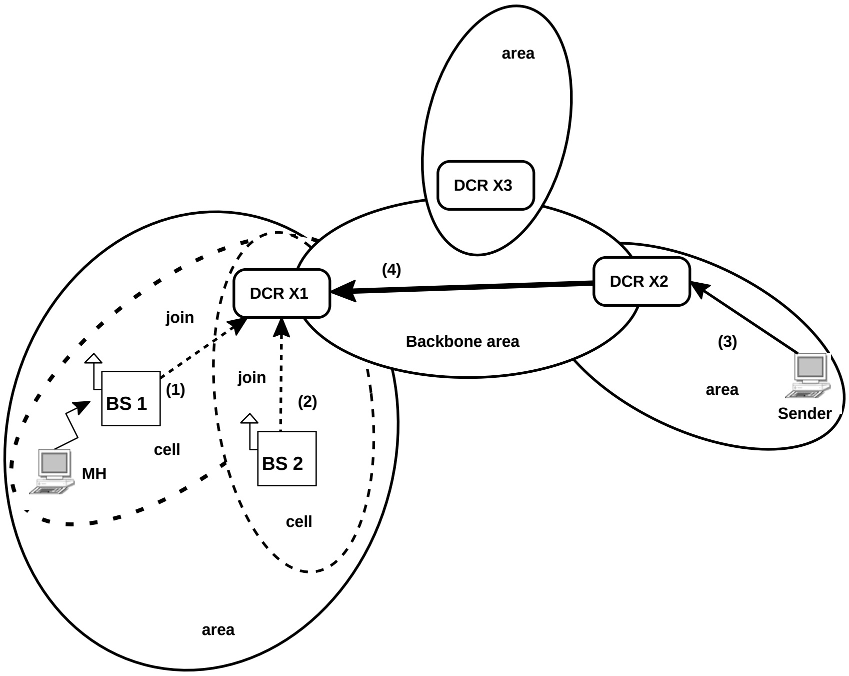


Figure 8: The mobile host (MH) is assigned multicast address M. Three DCRs, X1, X2 and X3 serve M. Step (1): Base station BS1 sends a join message for M towards X1. Step (2): Advance registration for M in a neighbouring cell is done by BS2. X1 informs X2 and X3 that it has a member for M. Step(3): The sender sends a packet to multicast group M. This packet gets delivered through the backbone to X1. Step (4): X1 receives encapsulated multicast data packet. From X1 data is forwarded to BS1 and BS2. MH receives data from BS1.

multicast address. From this DCR a multicast packet is delivered to the DCR(s) where a mobile host has registered, by using point-to-point tunnels mechanism described in Section 3.5. Upon receiving encapsulated multicast data for the mobile host, the DCRs forwards the data along established subtrees to base stations. A mobile host receives data only from a base station in its current cell. This is illustrated in one example in Figure 8.

Here we describe in more details how advance registration is performed. At its current cell, the mobile host receives data along the distribution subtree that is established for the mobile host's multicast address. This tree is rooted at the DCR and maintained with periodical sending of join messages. Now, suppose that the base station in the neighbouring cell anticipates arrival of the mobile host. It begins a joining process for the multicast group assigned to the mobile host. This process is terminated when a join message reaches a router that is already on the distribution tree. When the cells are close to each other, joining is terminated at the lowest branching point in the distribution tree. This ensures that the neighbouring base station quickly becomes a part of the multicast distribution tree with low overhead. This is illustrated in Figure 9. The neighbouring base station can start joining the multicast group assigned to the mobile host after the mobile host leaves its previous cell. Routers on the distribution tree keep forwarding information for a given time, even if the previous base station stops refreshing the tree because the mobile host leaves its cell. As before, if the base stations are close to each other, the multicast distribution tree for the new base station can be established in a short period of time that makes handoff efficient.

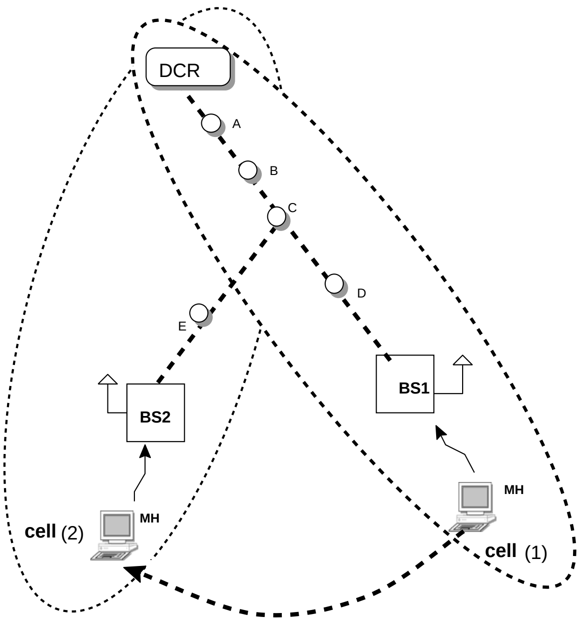


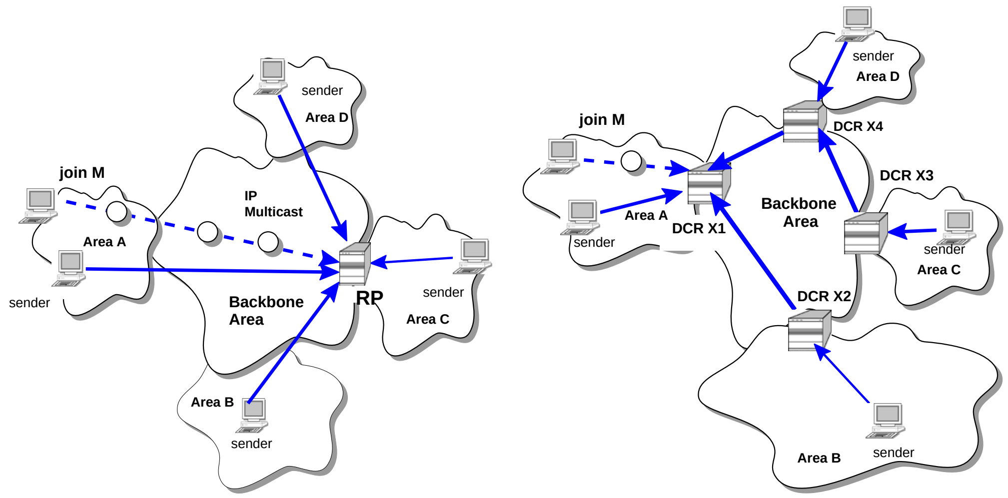
Figure 9: This figure presents an example of advance registration. At first, the mobile host (MH) is in cell 1. MH is assigned a multicast address M. Base station BS1 receives data for MH along the distribution subtree rooted at the DCR. On that subtree are routers A, B, C and D. Before host moves from cell 1 to cell 2, neighbouring base station BS2 initiates an advance joining for M. Joining at position 2 is terminated at router C.

In this paper we do not address the problems of using multicast routing to support end-to-end unicast communication. These problems are related to protocols such as: TCP, ICMP, IGMP, ARP. For these issues see [11]. A simple solution to this problem could be to have a special range of unicast addresses that are routed as multicast addresses. In this way, packets destined to the mobile host are routed by using a multicast mechanism. Conversely, at the end systems, these packets are considered as unicast packets and standard unicast mechanisms are applied.

5 Preliminary Evaluation of DCM

In this section, we evaluate the suitability of DCM as multicast routing protocol used to support host mobility in the Internet. We examine DCM performance under following assumptions: large number of multicast groups, a few receivers per group and a potentially large number of senders to a multicast groups. We show, in the cases we analyse, that DCM performs better than the PIM-SM shared-tree multicast routing protocol.

We implemented DCM using the Network Simulator (NS) tool [1]. To examine the performance of the DCM in a realistic manner, we performed simulations on a single-domain network model consisting of four areas connected via the backbone area. Figure 10 illustrates the network model used in simulations where areas A,B, C and D are connected via the backbone. The whole network contains 128 nodes. We examined the performance under realistic conditions: the links on the network were configured to run at 1.5Mb/s with a 10ms delay between hops. The link costs in the backbone area are higher than the costs in other areas.



(a) PIM-SM Shared-tree

(b) DCM Approach

Figure 10: The figure presents one member of the multicast group M in area A and four senders in areas A, B, C and D. Two different approaches for data distribution are illustrated: the PIM-SM shared-tree case and the DCM approach. In the DCM approach within each area there is one DCR that serves M. In PIM-SM one of the DCRs is chosen to be the centre router (RP) . With PIM-SM, all senders send encapsulated multicast data to the RP. In DCM each sender sends encapsulated multicast data to the DCR inside their area. With PIM-SM, multicast data is distributed from the RP along established distribution tree to the receiver (dashed line). With DCM, data is distributed from source DCRs X1, X2 and X3 by means of point-to-point tunnels (full lines in the backbone) and the established subtree in Area A (dashed line)

We analyse the following characteristics: size of the routing table, traffic concentration in the network, control traffic overhead and robustness. The evaluation of DCM in terms of delay and losses during the handoff is yet to be completed.

- **The amount of multicast router state information**

  DCM requires that each multicast router maintains a table of multicast routing information. In our simulations, we want to check the size of multicast router routing table. The routing table size becomes an especially important issue when the number of senders and groups grows, because router speed and memory requirements are affected.

  We performed a number of simulations. In all the simulations, we use the same network model presented in Figure 10, but with different numbers of multicast groups. For each multicast group there is only one receiver and 20 senders.

  Within each area, there is more than one candidate DCR. The hash function is used by routers within the network to map a multicast group to one DCR in the corresponding area. We randomly distributed membership among a number of active groups. For every multicast group, one receiver in the network is chosen randomly. In the same way, senders are chosen.

  The same scenarios were simulated with PIM-SM applied as the multicast routing protocol. In PIM-SM, candidate RP routers are placed at the same location as candidate DCRs in the DCM simulation.

We verified that among all routers in the network, routers with the largest routing table size are DRCs in the case of DCM. In the case of PIM-SM those are RPs and backbone routers. We define the most loaded router as the router with the largest routing table size. Figure 11 shows the routing table size in the most loaded router for the two different approaches. Figure 11 illustrates that the size of the routing table of the most loaded DCR is increasing linearly with the number of multicast groups. The most loaded router in PIM-SM is in the backbone. As the number of multicast groups increases, the size of the routing table in the most loaded DCR becomes considerably smaller than the size in the most loaded PIM-SM backbone router.

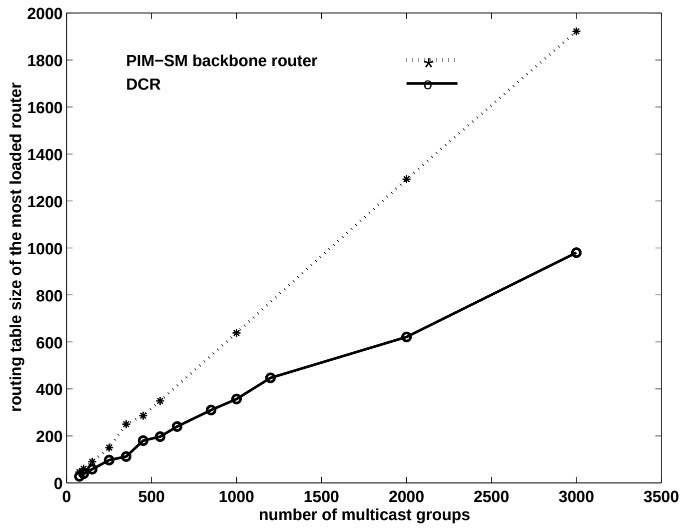


Figure 11: Routing table size for the most loaded routers

As it is expected, routing table size in RPs is larger than in DCRs. This can be explained by the fact that the RP router in case of PIM-SM is responsible for the receivers and senders in the whole domain, while DCRs are responsible for receivers and senders in the area where the DCR belongs.

For non-backbone routers, simulation results show that with the placement of RPs at the edge with the backbone there is not a big difference in their routing table sizes for two the approaches. Otherwise, if the location of RPs is elsewhere inside the area, non-backbone routers have smaller routing table size in case when DCM is applied as the multicast routing protocol than in case of PIM-SM.

Figure 12 illustrates the average routing table size in the backbone routers for the two routing protocols. In case of PIM-SM this size is increasing linearly with the number of multicast group. With DCM all join/prune messages from receivers in non-backbone areas are terminated at the corresponding DCRs situated at the edge with the backbone. Thus, in DCM, non-DCR backbone routers need not keep multicast group state information. And, this fulfils the DCM design objective to avoid multicast group state information in backbone routers.

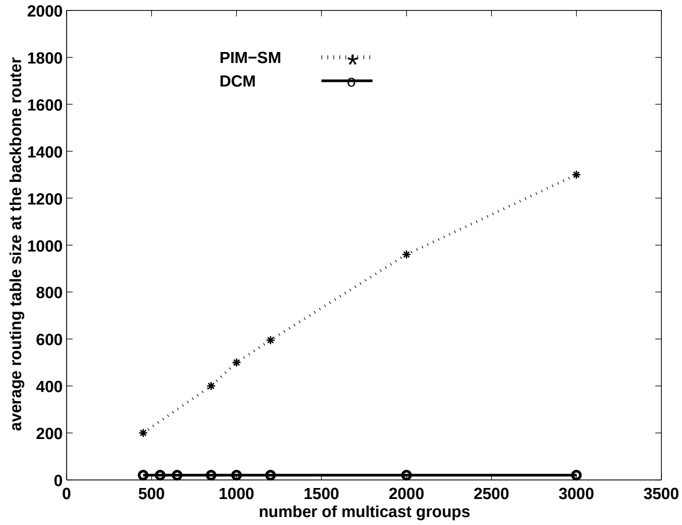


Figure 12: Average routing table size at the backbone router

- **Traffic concentration**

  In the shared-tree case of PIM-SM, every sender to a multicast group sends encapsulated data to the RP router uniquely assigned to that group within the whole domain. This is illustrated in Figure 10(a) where all four senders to a multicast group send data to a single point in the network. This increases traffic concentration on the links leading to the RP.

  With DCM, converging traffic is not sent to a single point in the network because each sender sends data to the DCR assigned to a multicast group within the corresponding area (as presented in Figure 10(b)).

  In DCM, if all senders and all receivers are in the same area, data is not forwarded to the backbone. In that way, backbone routers don't forward the local traffic generated inside an area. Consequently, triangular routing across expensive backbone links is avoided.

- **Control traffic overhead**

  Join/prune messages are overhead messages that are used for setting up, maintaining and tearing down the multicast data delivery subtrees. In our simulations we wanted to measure the number of such messages that are exchanged in two cases when DCM and PIM-SM are used as the multicast routing protocols. Simulations have shown that in DCM the number of join/prune messages is 20% smaller than in PIM-SM. This result can be explained by the fact that in DCM all join/prune messages from the receivers in the non-backbone areas are terminated at the corresponding DCRs inside the same area, close to the destinations. In PIM-SM join/prune messages must reach the RP that may be far away from the destinations.

  In DCM, DCRs exchange the MDP control messages. The evaluation of the overhead of these messages depends on the group joining/leaving dynamicity and updating frequency and is left for the future work.

- **Robustness feature in case of DCR failure**

DCM has a mechanism to insure robustness in case of the DCR failure similar to PIM-SM. If some of the candidate DCRs fail, this would be detected by all the routers with an area by means of the bootstrap protocol. A new DCR is elected from the set of reachable candidate DCRs within an area to replace the failed DCR.

6 Conclusions

We have considered the problem of multicast routing in a large single domain network with very large number of multicast groups with a small number of receivers. Such a case occurs, for example, when multicast addresses are statically allocated to mobile hosts, as a mechanism to manage Internet host mobility. Our proposal, called Distributed Core Multicast (DCM) is based on an extension of the centre-based tree approach. DCM uses several core routers, called Distributed Core Routers (DCRs) and a special protocol between them. The objectives achieved with DCM are: (1) to avoid state information in backbone routers, (2) to avoid triangular routing across expensive backbone links, (3) to scale well with the number of multicast groups. We have also described how our approach can be used to support mobile hosts. We have implemented DCM using the Network Simulator (NS) tool. We have presented a preliminary evaluation of our implementation. DCM is compared to PIM-SM multicast routing protocol and it is demonstrated that DCM performs better than PIM-SM when used to support mobile hosts in the Internet.

Appendix A

In section Section 3.5 we presented one solution called point-to-point tunnels for the distribution of multicast data between DCRs. Point-to-point tunnels avoid triangular routing across expensive links in the backbone, but does not completely optimise the use of backbone bandwidth. Here we present two alternative solutions called tree-based source routing and list-based source routing that use backbone bandwidth more optimally than point-to-point approach.

Tree-Based Source Routing

This solution assumes that the DCRs are aware of the backbone topology (e.g the backbone is one OSPF area) and backbone routers implement a special packet forwarding mechanism called tree source routing. This approach consists in that a source DCR for the multicast group computes a shortest path tree rooted at itself to a list of labelled DCRs for the multicast group. On a shortest path can be included DCRs in other areas that serve the multicast address, as well as non-DCR backbone routers. A description of a shortest path tree with destinations and branching points is included in the tree source routing header by the source DCR. Figure 13 shows one example of tree source routing approach.

This approach ensures that backbone bandwidth is used more optimally than if the "point-to-point tunnels" approach is used. This is achieved at the expense of introducing the new tree source routing mechanism that needs to be performed by backbone routers.

List-Based Source Routing

As the third solution we propose a new list-based multicast data distribution in the backbone. Here we give an initial description of this mechanism. The final solution is the object of ongoing

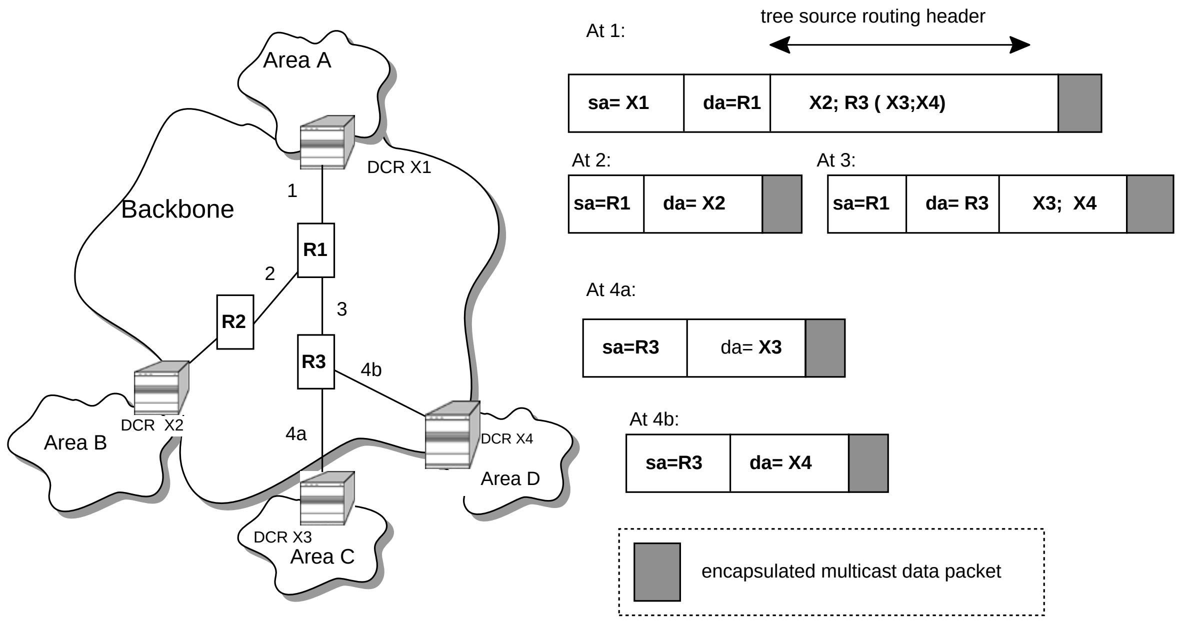


Figure 13: This figure shows how multicast data is distributed from source DCR X1 to labelled DCRs X2, X3 and X4 by using tree-based source routing approach. X1 puts distribution information in tree source routing header after computing a shortest-path tree to routers X2, X3 and X4. At first, the data should be delivered to backbone router R1 where two copies of the multicast data are made. One copy is sent encapsulated to X2, while the other is sent encapsulated to backbone router R3. As soon as router R3 receives a packet it reads from the tree source routing header that it should send two copies of the multicast data: one to X3 and the other to X4.

research.

As in the previous approach we assume that the DCRs are aware of the backbone topology. A special list-based source routing protocol is performed by the DCRs and backbone routers. This works as follows: as soon as a source DCR determines that it must forward a packet to a list of DCRs, it determines the next backbone router(s) to which it should send a copy of the packet to reach every listed DCR. The source DCR sends a copy of the packet to each determined router together with a sublist of the DCRs that should be reached from this router. This sublist is contained in a list source routing header. Unlike a tree-based source routing header, where in a tree source routing header can be included also non-DCR backbone routers, the list source routing header contains only the final DCR destinations.

Each backbone router performs the same steps until multicast data has reached every labelled DCR. Note that a DCR can also send a copy directly to another DCR.

On Figure 14 is presented one example of list-based source routing approach.

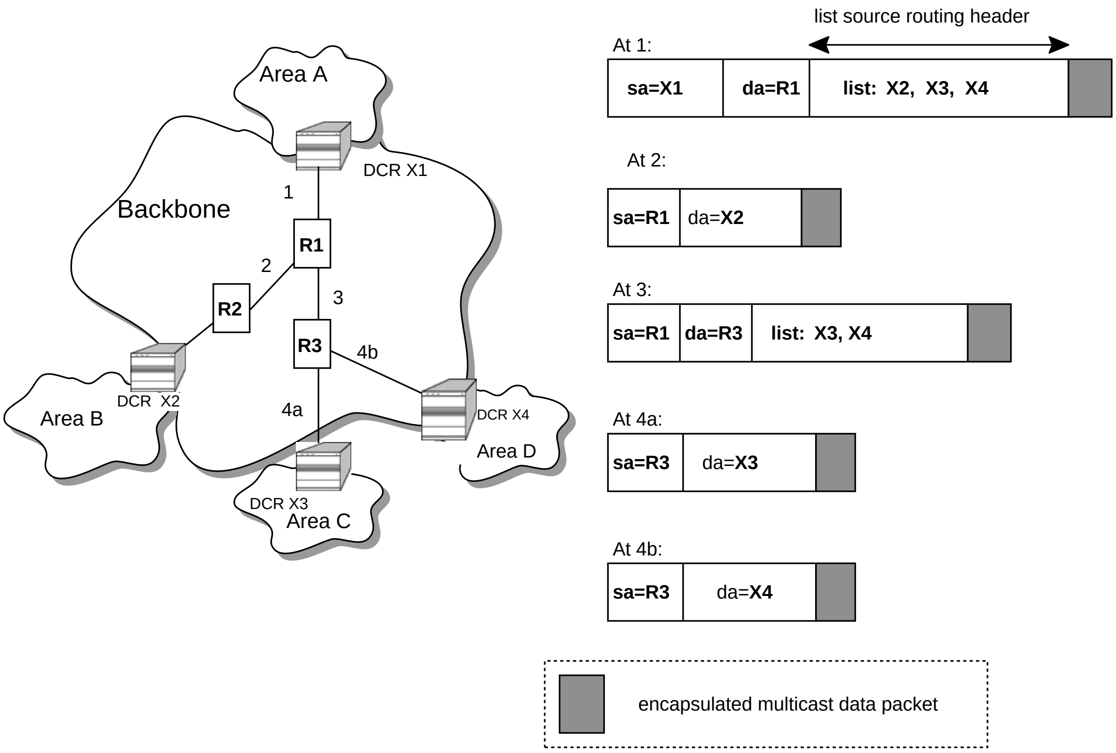


Figure 14: This figure shows how multicast data is distributed from the source DCR X1 to labelled DCRs X2, X3 and X4 by using the list-based source routing approach. X1 determines that it should send a copy of multicast data to backbone router R1. Router X1 puts in the the list source routing header information that X2, X3 and X4 should be reached from R1. As soon as R1 receives a packet it makes two copies of the multicast data. One copy of the multicast data is encapsulated is a packet that is sent to X2. Another copy of the multicast data is sent to R3. This packet contains in the list source routing header a list of DCRs that should be reached from R3. The list contains X3 and X4. As soon as R3 receives a packet from R1 it makes two copies of the multicast data. One copy is sent encapsulated to X3. Another copy is sent encapsulated to X4.

Appendix B: Glossary of Terms

Bellow is given a list of terms and definitions that are used throughout the paper.

- **DCR**. A DCR (Distributed Core Router) is an backbone access point for the data sent to multicast address by senders inside the same area to members outside the area. A DCR also forwards the multicast data received from the backbone to receivers in the area it belongs to.

- **DCR serves the multicast address m**. A DCR is said to serve the multicast address m when it is dynamically elected among all the candidate DCRs in the area to act as an access point for address m. (see Section 3.1)

- **Labelled DCR**. A DCR is labelled with the multicast address m if the DCR serves m and there are receivers for m in its area. A labelled DCR is root of a distribution subtree inside its area for m.(see Section 3.2)

- **Source DCR**. A source DCR for the multicast group m is the DCR that receives encapsulated multicast data for m by some source inside its area. (see Section 3.4)
- **Range**. A range is the partition of the set of multicast addresses into group of addresses. A DCR can serve several ranges of multicast addresses.(see Section 3.1)
- **MDP**(Membership Distribution Protocol). MDP is used for the source DCR in one area to learn about labelled DCRs in other areas. MDP is run between DCRs in different areas that serve the same range of multicast addresses. (see Section 3.3)
- **MDP control multicast address**. An MDP control multicast address is used for exchanging MDP control messages between DCRs in areas that serve the same range of multicast addresses. (see Section 3.3) There is one MDP control multicast address per range of multicast addresses.
- **STH**(Shortest Tunnel Heuristic). STH is used by the source DCR to compute the multicast data distribution tree in the backbone. The edges of this tree are tunnels between DCRs. (see Section 3.5)

Acknowledgements

The authors would like to thank Silvia Giordano for useful discussions and help during the editing of this paper.

References

[1] Network Simulator. Available from http://www-mash.cs.berkeley.edu/ns.

[2] A. Ballardie. Core Based Trees (CBT) Multicast Routing Architecture. RFC 2201, September 1997.

[3] S. Deering and R. Hinden. Internet Protocol, Version 6 (IPv6) Srecification. Technical report, RFC 1883, 1995.

[4] Deborah Estrin, Mark Handley, Ahmed Helmy, Polly Huang, and David Thaler. A Dynamic Mechanism for Rendezvous-based Multicast Routing. ACM/IEEE, May, 1997.

[5] D. Estrin et.all. Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification. RFC 2117, June 1997.

[6] W. Fenner. Internet Group Management Protocol, Version 2. RFC 2236, November 1997.

[7] John Ioannidis, Dan Duchamp, and Gerald Q. Maguire Jr. IP-based Protocols for Mobile Internetworking. In *Proc. of SIGCOMM'91*, Zurich, Switzerland, September 1991.

[8] J. Macker and M. S. Corson. Mobile As hoc Networking and the IETF. *ACM Mobile Computing and Communications Review*, 2(1), January 1998.

[9] J. Moy. Multicast Extensions to OSPF. RFC 1584, 1994.

[10] J. Moy. OSPF version 2. RFC 1583, 1994.

[11] Jayanth Mysore and Vaduvur Bharghavan. A New Multicasting-based Architecture for Internet Host Mobility. In *The Third Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom'97).*

[12] C. Perkins. IP Mobility Support, Network Working Group. RFC 2002, October 1996.

[13] Charles E. Perkins and David B. Johnson. Mobility Support in IPv6. In *Proc. of the Second Annual International Conference on Mobile Computing and Networking (MobiCom'96).*

[14] Fumio Teraoka, Yasuhiko Yokote, and Mario Tokoro. A Network Achitecture Providing Host Migration Transparency. In *Proc.of ACM SIGCOMM'91.*

[15] D. G. Thaler and C. V. Ravishankar. Using Name-Based Mappings to Increase Hit Rates. *IEEE/ACM Transactions on Networking*, 6(1), February 1998.

[16] David G. Thaler and Chinya V. Ravishankar. Distributed Center-Location Algorithms. *IEEE JSAC*, 15(3), April 1997.

[17] Vinod Valloppillil and Keith W. Ross. Cache Array Routing Protocol v1.0. INTERNET-DRAFT, 1998.

[18] D. Waitzman, S. Deering, and C. Partridge. Distance vector multicast routing protocol. RFC 1075, 1988.

[19] Bernard M. Waxman. Routing of Multipoint connections. *IEEE JSAC*, 6(9), December 1988.

[20] Liming Wei and Deborah Estrin. The Trade-offs of Multicast Trees and Algorithms. In *Proc.of the 1994 International Conference on Computer Communications and Networks*, San Francisco, CA, USA, September 1994.

[21] Pawei Winter. Steiner problem in networks: A survey. *Networks*, 17(2), 1987.