The Border Gateway Multicast Protocol

BGMP

A Protocol specification

based on the

BGMP Working Group
Internet Engineering Task Force
INTERNET-DRAFT
Expiring July 2004

1. Introduction

The Border Gateway Multicast Protocol (BGMP) is a protocol used for inter-domain multicast routing. Run by the border-routers of a domain, inter-domain bidirectional shared trees are constructed by using BGP group routes. Further more the BGMP allows any existing multicast routing protocol to be used within individual domains. The resulting shared tree for a group is rooted at the domain who's address range covers the group's address; this domain is typically the group initiator's domain.
BGMP uses TCP as its transport protocol. Therefore, no implementation of message fragmentation, retransmission, acknowledgement or sequencing is required. For establishing its connections, BGMP uses port 264. This port is distinct from the port of the BGP, to provide protocol independence and to facilitate distinguishing between protocol packets, e.g. by packet classifiers, diagnostic utilities, etc.
The BGMP in general builds shared trees for active multicast groups, which consist of group specific bidirectional branches, but to also support any-source multicast (ASM), it is possible to set up source-specific, inter-domain, distribution branches, where needed.
Each shared tree is rooted at the domain whose address allocation includes the group's address.
Via BGMP, IP Multicast can be realized. This is an important mechanism to support applications such as multimedia teleconferencing, distance learning, data replication and network games, for example.

2. Tasks and rules of border routers

2.1 General

Messages are exchanged via two BGMP peers, which form a TCP connection between each other. Within these messages, connection parameters are opened and confirmed. After that, update messages, (join/prune) are sent, incremental as group membership changes.
Keep alive message are sent periodically to ensure the liveliness of the connection and also to confirm a received and accepted open message.
Once the open message is confirmed, update, keep alive and notification messages may be exchanged.
Notification messages are sent in response to errors or special conditions. If a connection encounters an error condition, a notification message is sent and the connection is closed if the error is a fatal one.
With BGMP, messages in general are only processed after entirely received and the maximum message size is set to 4096 octets.
All implementations are required to support this maximum message size.

2.2 Forwarding rules

Data packets are forwarded based on a combination of BGMP and MIGP rules (Fig. 2.2-1). If a packet arrives on an MIGP interface, it is accepted and forwarded according to the existing MIGP rules (Fig. 2.2-1).

If it arrives over a point-to-point BGMP interface (Fig. 2.2-2) and the packet got accepted, the following procedure is taken:

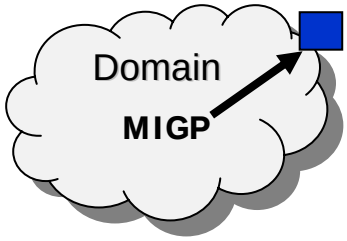

Fig. 2.2-1: A data packet arrives on an MIGP interface

If a (S,G) BGMP tree state entry exists, the router forwards according to these rules.
If not found, the router checks for a matching (*,G) BGMP tree state entry.
If neither is found, the packet is sent natively to the next-hop EGP peer for G.
If a matching entry was found, the packet is forwarded to all the targets in the target list. In this way, BGMP trees are able to forward data in a bidirectional manner.
If a target is an MIGP component, then the forwarding is subject to the rules of the MIGP protocol.

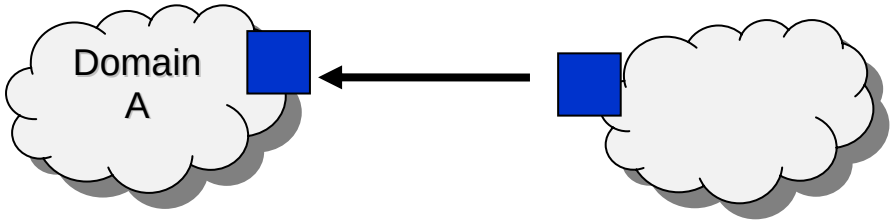

Fig. 2.2-2: A data packet arrives in the BGMP case

If a packet was not received by the next-hop target towards the group or the source, it will be dropped. After that, no further actions are taken.

3. Bidirectional trees

3.1 Third party dependency

BGMP builds bidirectional group-shared trees to minimize third-party dependencies and improve performance.

For example, in figure 3.1-1, members in domains C and D can communicate with each other along the bidirectional tree without depending on the quality of their connectivity to the root domain, A. This is also more efficient because of the shorter paths taken.

In contrast to this, the data from senders of unidirectional shared trees for a group, has to travel up to the root and then down the shared tree to all the members. This approach would introduce third party dependencies and potentially poor performance if applied at the inter-domain level.

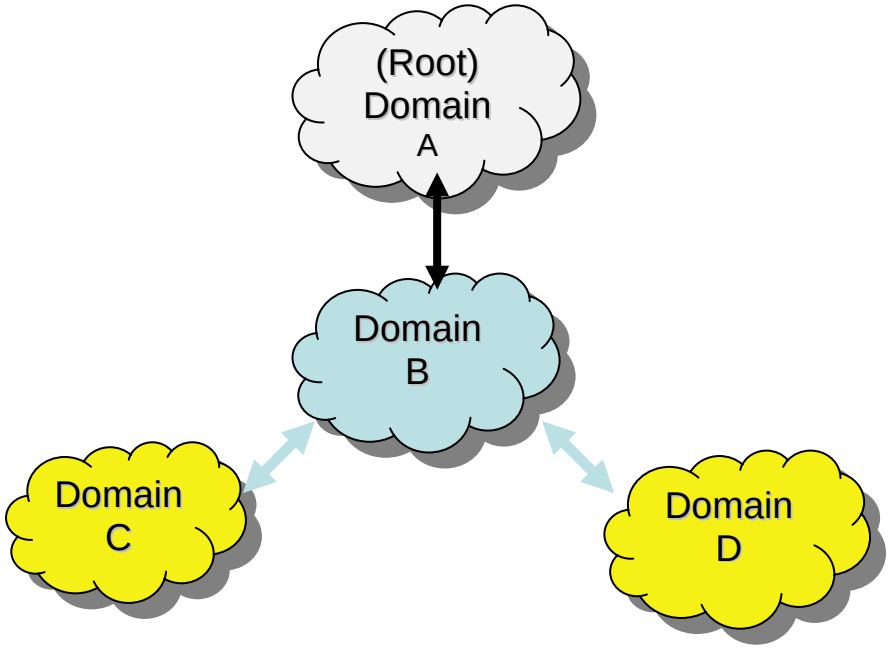

Fig. 3.1-1 Third party dependency is avoided by using bidirectional trees

3.2 Method of choosing the root

The choice of a group's shared-tree-root has implications for performance and policy. Within intra-domain shared tree protocols all routers are treated as equivalent candidates. It is more or less a random choice depending on load sharing and stability, for example.
In the inter-domain case (so in the BGMP) the choice of a group's root is subject to administrative control, depending on poor locality, e.g.
Usually a group gets rooted at the domain of the initiator of a group.

3.3 Establishing the bidirectional shared tree

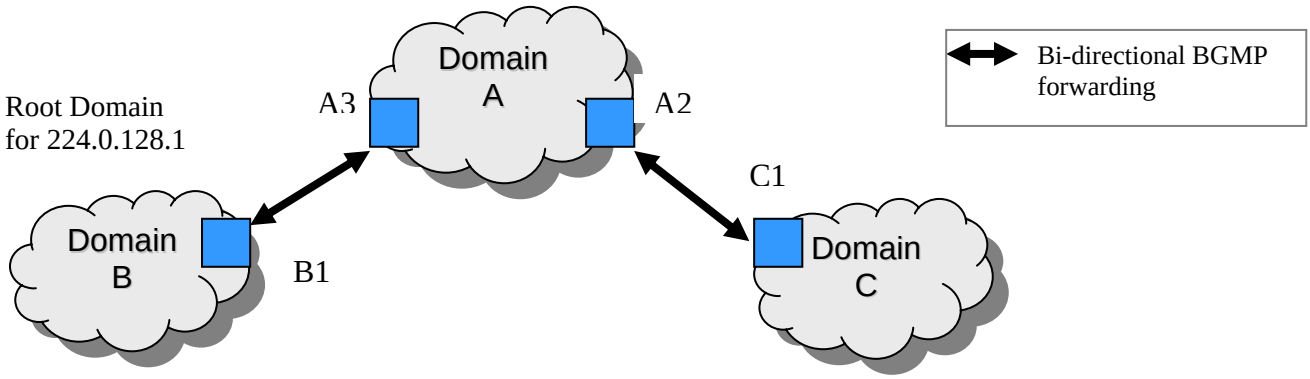

Fig. 3.3-1: A bidirectional shared tree

Consider a multicast group created by a host in domain B in Figure 3.3-1. Since the host acquires the address 224.0.128.1 from the address range, B will be the group's root domain. When a host in group C now joins this group, a join request is received from the MIGP by the BGMP component of the best exit router for 224.0.128.1, namely C1.
C1 looks up 224.0.128.1 in its G-RIB, finds (224.0.0.0/16,A2), and creates a multicast-group forwarding entry consisting of a parent target and a list of child targets.

The parent target identifies the BGMP peer that is the next hop towards the group's root domain. A child target identifies either a BGMP peer or an MIGP component from which a join request was received for this group. The parent and child targets together are called the target list (Fig. 3.3-2).

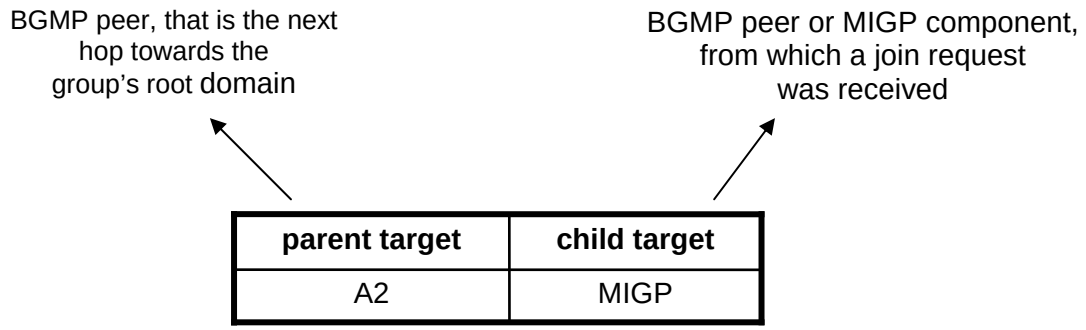

Fig. 3.3-2: A group specific target list / multicast-group forwarding entry

In the case of Cl, the parent target is A2 and the only child target is its MIGP component. This multicast forwarding entry, also known as a (*,G) entry, denotes that data packets sent to the group G, received at Cl from any source are to be forwarded to all the targets in the target list except the target from which the data packet came. BGMP border routers have persistent TCP peering sessions with each other for the exchange of BGMP control messages (in this case, group joins and prunes). After creating the (*,G) entry for 224.0.128.1, Cl sends a group join message over the connection to the parent target, A2. On receiving the group join message from Cl, router A2 looks up 224.0.128.1 in its G-RIB and finds the entry (224.0.128.0/24, A3) indicating that A3 is the next hop to reach the root domain for 224.0.128.1. It then sets up a (*,G) entry with the MIGP component to reach A3 as the parent target and Cl as the child target. A2 then transmits the join request to its MIGP component because A3 is an internal BGMP peer. The MIGP component of the border router performs the necessary actions to enable data packets to transit through the domain between A2 and A3.
On receiving the join request from its MIGP component, A3 creates a (*,G) entry with the MIGP component as the child target to enable the exchange of data packets with A2 through the MIGP. The parent target is Bl, since Bl is the next hop to reach the root domain according to its G-RIB entry, (224.0.128.0/24, Bl).
On receiving the join from A3, router Bl, which is in the root domain for the group, creates a (*,G) entry with its MIGP component as the parent target (since it has no BGP next hop) and A3 as the child target. A join request is sent to its MIGP component, which joins as a member of the group 224.0.128.1 within the domain using the MIGP rules.

3.4 Data from external domains

To illustrate how data reaches the shared tree from domains not on the tree, suppose a host in domain E (Fig. 3.4-1), that has no members of the group, sends data to the group 224.0.128.1. The data packets are transmitted through the MIGP to the best exit router El. Since El has no forwarding state for the group, it simply forwards the packets to the next hop towards the root domain (Al). Since Al also has no forwarding state for the group, it transmits the packet through the MIGP of A to reach the next hop border router to the root domain, A3. Since the border routers, A2, and A3, are on the shared tree for the group, they each forward the data packets they receive to all the targets in their (*,G) entry except the target from which the packet was received (i.e., their MIGP component). The data packets thus reach group members in domains B and C along the shared tree. When a BGMP router or an MIGP component no longer leads to any group members, it removes itself from the child target list of its parent target by sending a prune message or notification to its parent target. When the child target list becomes empty, the BGMP router removes the (*,G) entry and sends a prune message upstream towards the root domain. In this way, the multicast distribution tree is torn down as members leave the group.

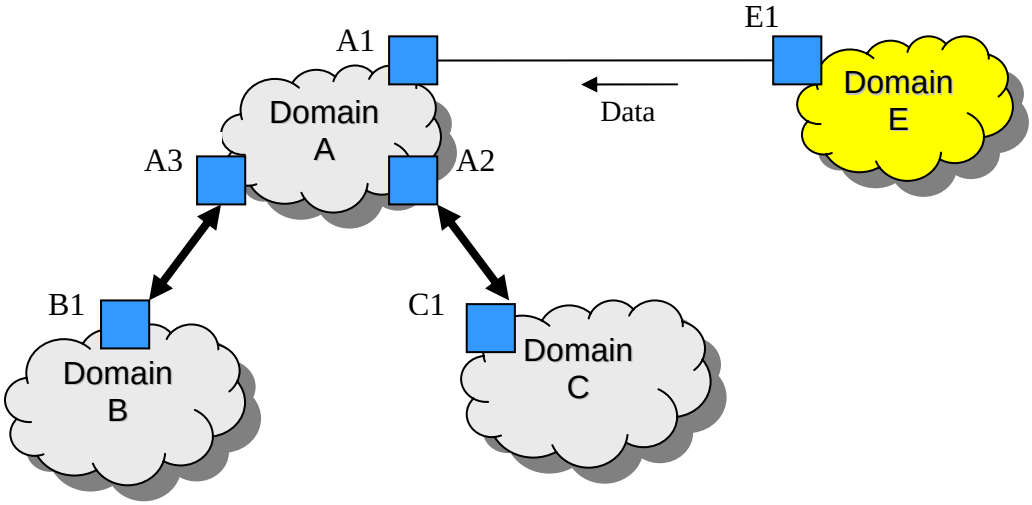

Fig. 3.4-1 Data from an external Domain (E) has to reach the tree

4. Source specific branches/trees

With the BGMP, source specific branches/trees are used to be compatible with source specific trees used by the MIGP or to construct trees for source specific groups.
A source specific branch is built only, if it is needed to pull traffic down to a BGMP router that has a source specific (S.G) state AND it is not yet in the shared tree AND the router does not want to receive encapsulated packets by a router in the shared tree.
Sometimes, data packets have to be transmitted encapsulated, to avoid them being dropped due to a RPF (Reverse Path Forwarding) check.

Data gets forwarded, if it arrives on a device which the router claims as a part of the shortest path to the source (*E2* in Fig. 4-1). Otherwise it is supposed to be duplicate data and gets dropped. Therefore data packets need to be sent encapsulated to be accepted by other routers. Due to this manipulation of the data packets, an overhead is created.

4.1. Establishing source specific branches/trees

BGMP can build source-specific branches in cases where the shortest path to a source from the domain does not coincide with the bidirectional tree from the domain (e.g. domain E in figure 4.1-1 for sources in domain D). In such domains, if the border router receiving packets on the shared tree is not on the shortest path to the source, it normally must send them encapsulated to the appropriate border router where they can be injected into the domain's MIGP. Otherwise the packets would be dropped by routers inside the domain due to failure of the RPF checks towards the source (Fig. 4-1).
If a source-specific branch is built, data can be brought into the domain from the source via the appropriate border router so that the data encapsulation overhead can be avoided. This is done by allowing the decapsulating border router the option of sending a source-specific join

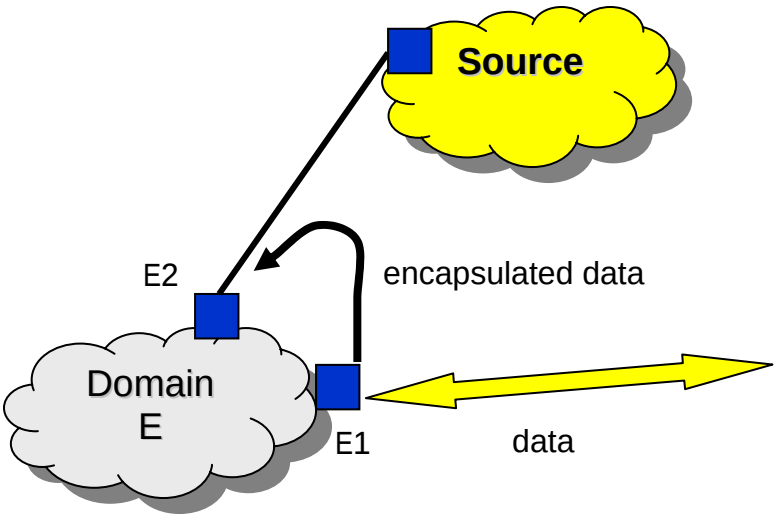

Fig. 4-1: To avoid dropping the data, it will be sent encapsulated

towards the relevant source, once data is flowing. The join messages then propagate until they hit either a branch of the bidirectional tree or the source domain itself. A source-specific prune is sent back to the encapsulating border router, which can then propagate it up the shared tree to prevent unnecessary copies of the packet.

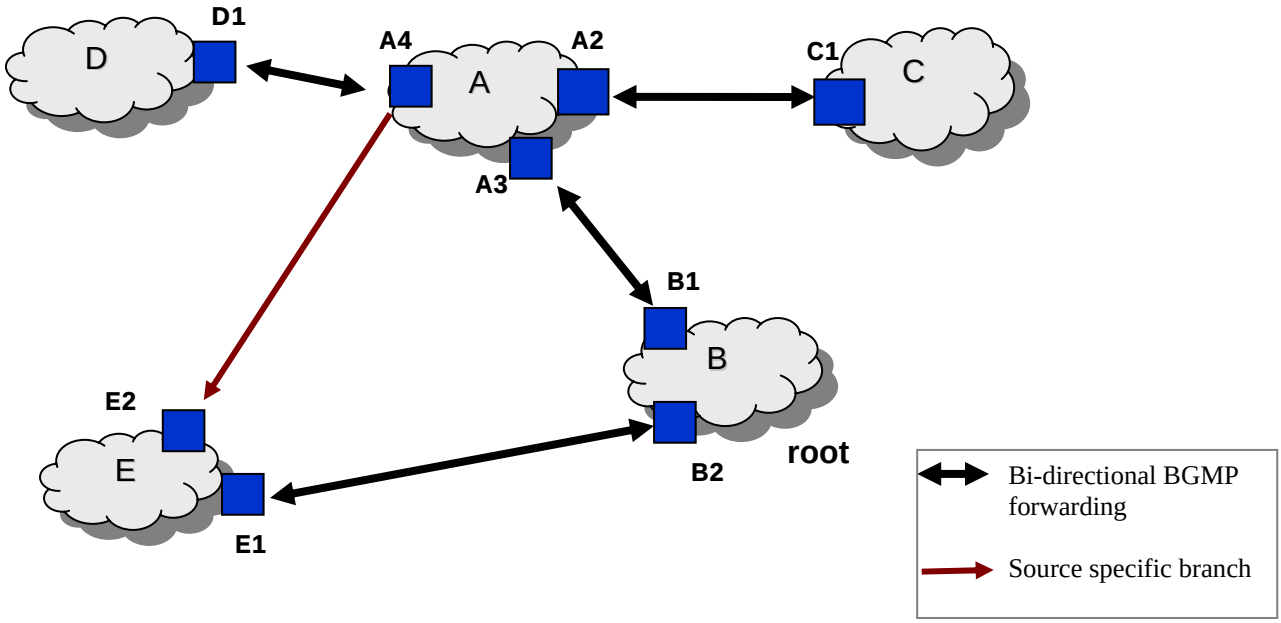

Fig. 4.1-1: The source specific branch

Suppose there are members of a group 224.0.128.1 in domains B, C, D and E, and B is the root domain for the group (Fig. 4.1-1). The bidirectional shared tree is set up as shown in the figure. Domain E has an inter-domain link to domain A via border router E2. Hence, the shortest path from domain E to hosts in D is through E2. E uses a MIGP, which implies that internal routers will only accept packets from a source which they receive from their neighbour towards that source. Since only El is on the bidirectional shared tree, data from a source S in domain D will be received by El. El must then encapsulate the data packets to E2 in order to avoid internal RPF check failures. E2 then sends the data to MIGP component, so that group members in E receive the data packets.

If E2 decides to stop the above encapsulations, it may send a source-specific join towards the source S. It also instantiates a multicast forwarding entry called a (S,G) entry, with the parent target being the next hop towards S (A4), and the child target list consisting of its MIGP component. Data packets that arrive from A4 will thus be accepted and forwarded to other targets listed in the (S,G) entry. The source-specific join from E propagates towards the source (in similar fashion to a shared-tree join propagating towards the root domain) setting up (S,G) state in the intermediate border routers until it reaches a border router that is on the shared tree for the group. In figure 4.1-1, A4 is on the shared tree for the group. A4, on receiving the source-specific join, creates an (S,G) entry, copies the existing target list from the (*,G) entry, and adds E2 to the child target list. The source-specific join is not propagated further by A4. Subsequent data packets sent by S and received by A4 are forwarded to all other targets in the (S,G) entry, including E2.

Once it begins receiving data from A4, E2 sends a source-specific prune to El, and starts dropping the encapsulated copies of the data packets from the source, coming from El. Since El has no other child targets for (S,G), it propagates the prune up the shared tree to B2, to stop receiving packets from S along the shared tree.

5. Security Considerations

If unauthorized or altered BMGP messages got accepted by BMGP components, a denial of service due to excess bandwidth consumption or lack of multicast connectivity can result.
To prevent this, it is possible to use an authentication of the BGMP messages.
In order to secure control messages, it is required, that a BGMP implementation is equipped with keyed MD5 (RFC2385).
Further on, it also has to be compatible with peers that do not support this.
But if one side of the connection is configured with keyed MD5 and the other side not, it is recommended not to establish the connection.

6. Terms and definitions

Domain:
A set of one or more contiguous links, and zero or more routers, surrounded by one or more multicast border routers. Note that this loose definition of domain also applies to an external link between two domains, as well as an exchange.

Root Domain:
When constructing a shared tree of domains for some group, one domain will be the "root" of the tree. The root domain receives data from each sender to the group, and functions as a rendezvous domain toward which member domains can send inter-domain joins, and to which sender domains can send data.

Multicast RIB:
The Routing Information Base, or routing table, used to calculate the "next-hop" towards a particular address for multicast traffic.

Multicast IGP (MIGP):
A generic term for any multicast routing protocol used for tree construction within a domain. Typical examples of MIGPs are: PIM-SM, PIM-DM, DVMRP, MOSPF, and CBT.

EGP:
A generic term for the inter-domain unicast routing protocol in use. Typically, this will be some version of BGP which can support a Multicast RIB, such as MBGP [MBGP], containing both unicast and multicast address prefixes.

Component:
The portion of a border router associated with (and logically inside) a particular domain that runs the multicast IGP (MIGP) for that domain, if any. Each border router thus has zero or more components inside routing domains. In addition, each border router with external links that do not fall inside any routing domain will have an inter-domain component that runs BGMP.

External peer:
A border router in another multicast AS (autonomous system, as used in BGP), to which a BGMP TCP-connection is open. If BGP is being used as the EGP, a separate "eBGP" TCP-connection will also be open to the same peer.

Internal peer:
Another border router of the same multicast AS. If BGP is being used as the EGP, the border router either speaks iBGP ("internal" BGP) directly to internal peers in a full mesh, or directly through a route reflector [REFLECT].

Next-hop peer:
The next-hop peer towards a given IP address is the next EGP router on the path to the given address, according to multicast RIB routes in the EGP's routing table (e.g., in MBGP, routes whose Subsequent Address Family Identifier field indicates that the route is valid for multicast traffic).

target:
Either an EGP peer, or an MIGP component.

Tree State Table:
This is a table of (S-prefix,G) and (*,G-prefix) entries that have been explicitly joined by a set of targets. Each entry has, in addition to the source and group addresses and masks, a list of targets that have explicitly requested data (on behalf of directly connected hosts or downstream routers). (S,G) entries also have an "SPT" bit.