Are Alterations Needed to the IP Multicast Service Model?

Flávio Alencar do Rêgo Barros[1] and Michael Anthony Stanton[2]

[1]Instituto de Computação
Universidade Federal Fluminense
R. Passo da Pátria, 156, bloco E, sala 350
24210-240, Niterói, RJ
Flavio Alencar [falencarrb@uol.com.br]

[2]Instituto de Computação
Universidade Federal Fluminense
R. Passo da Pátria, 156, bloco E, sala 350
24210-240, Niterói, RJ
Michael Stanton [michael@ic.uff.br]

Abstract

Deering's model of Internet group communication, in spite of its simplicity and elegance, imposes limits on the complete development of IP multicast, due either to the profusion of contradictory requirements of applications, or to the reduced support provided by the network core for technical or economic reasons. Even if some of these difficulties may be resolved at the transport or application layers, thus remaining outside the scope of the present model, there also appear defects within this model, mainly due to inter-domain routing, involving address management, to QoS insensitivity and to loss of functionality. The recognition of these defects has already led to the solution presented by the MASC/BGMP project, which involves a transition from the present model. However there are many who argue that the success of unicast applications in TCP/IP will only be repeated in multicast applications through the use of much simpler solutions than the existing ones, or by including greater intelligence in the network interior. With this in mind, we analyse here the proposals RAMA and XCAST, which may be seen to demonstrate greater simplicity. We also look at AIM, designed for organising receivers into subgroups with similar interests. Leveraging Internet group applications will certainly involve one or other, or even both, of these approaches.

Keywords: Internet protocols, services and applications; multicast; group communication.

1 Introduction

Group applications lead to new patterns of Internet usage and are quite demanding on its infrastructure. Such applications may be classified according to their communications patterns: *one-to-many* (programmed distribution of audio and video, file distributions, etc.), *many-to-one* (data collection, opinion polls, auctions, etc.) and *many-to-many* (real-time multimedia conferencing, group collaboration or conversation, distributed interactive systems, multi-user games, etc.), or a variant of this last class, *few-to-few* [27].

What is known as the current service model of IP Multicast, or simply as Deering multicast, is based on extensions to the Internet Protocol (IP), related to group communication using class D IP addresses. These extensions consist of two fundamental components: the *Group Model* and a *Multicast Routing Protocol* [9]. Whilst the Group Model is concerned with the organisation into groups of receivers of multicast traffic (a station transmitting to the group need not belong to the group, it *only* needs to know the IP address of the intended group of receivers), the Multicast Routing Protocol is used to build and maintain the distribution tree used to deliver messages, and is of interest primarily to Internet Service Providers. These two components are responsible for the very different nature of multicast transmission, compared with the more usual *unicas*t, and all studies, proposals and implementations of multicast transmission undertaken in the last ten years or so have been based on them. Naturally enough, we are also able to lay at their door the difficulties and deficiencies of multicast, which are not also shared by unicast.

Solving multicast issues is mainly a question of dealing with routing, especially inter-domain routing, due to the use of different policies and choices, both administrative and technical, and these choices may result in scalability problems or increased costs for the stations or for the network.

But the problems are not restricted to just routing: group communication applications in the Internet have raised questions of reliability, performance, flow and congestion control, security, group management and co-ordination, amongst others, which have caused support for multicast to extend beyond the network layer, and include also the transport and application layers, and even the link layer.

In section 2, we present the basis for the current service model for IP multicast. In section 3, we present an analysis of the transition currently in course to improve this model, performing in section 4 an evaluation of this service model and the promised extensions, with the objective of identifying which demands are still unsatisfied. In section 5, we analyse a number of new proposals which attempt to overcome these limitations. In section 6 we present our conclusions.

2 The Current Service Model

Deering's original proposal [8] is based on multicast transmission to a group of receiving stations characterised by a single class D IP address, where datagram delivery has the same "best effort" reliability as in unicast transmission. The group address is fixed, but the group membership may not be: members join and leave groups at their own convenience. The model relies on extensions to IP, through the introduction of the Internet Group Management Protocol (IGMP).

IGMP defines just two types of message: *requests* and *announcement*s. Local multicast-enabled routers periodically multicast request messages in order to discover which groups of receiver stations have members on the local area network. Receivers reply with multicast announcement messages, either immediately (for stations without timing capacity), or after a random delay, otherwise, in order to avoid the receipt of simultaneous announcements by the router. (Note that only one member of each group need reply, and additional members suppress their announcement messages after receiving the first message for the same group). A station may also spontaneously transmit an announcement message to join a new group, instead of waiting for the next request message.

Although the process of joining or leaving a multicast group depends on IGMP, further additional information is necessary for a potential member to decide to join the group, or not. An IETF working group called MMUSIC *(Multiparty Multimedia Session Contro*l) was charged with developing protocols to announce group transmission content. There are currently two ways used to deal with this point [22]. The simpler form is to use the broadcast model, through SAP *(Session Announcement Protoco*l), which continually distributes a list of the addresses of destination groups for potential receivers to join. SAP uses SDP *(Session Description Protoco*l) to describe the contact and access information for the multicast sessions being announced, and both SAP and SDP are widely used in distributive multimedia applications on the MBone[1]. This model, which resembles broadcast television, is weakly coupled, as session transmitters have no knowledge of who are the receivers of their transmissions.

A separate development of the MMUSIC working group is represented by SIP *(Session Initiation Protoco*l), which is used to invite others to participate in some kind of teleconference, which may be an audio- or videoconference, or even a simple telephone call.

As well as IGMP, used for managing stations' joining and leaving a group, two other protocols are necessary: an intra-domain routing protocol, also known as a MIGP *(Multicast Internal Gateway Protoco*l) [28], such as PIM-SM or DVMRP, and an inter-domain routing protocol used in border routers. A MIGP based on the current service model builds multicast distribution trees using one of several mechanisms which reflect the manner in which new receivers join the multicast group. In *sparse mode* protocols, the receiver explicitly joins the multicast group; in *dense mode* protocols, the mechanism used may be by broadcasting an association, or by a combination of flooding and pruning.

Practical proposals for building distribution trees are of two kinds. The first kind is the source-rooted tree, which is ideal for *one-to-many* applications or their variants, which involve high data rates and low delay with source access control, not unlike broadcast TV. The other kind, which uses a shared tree, is appropriate for applications with multiple, low data-rate sources, with little concern with delay re-

1 *Multicast Backbon*e. A virtual network providing support for IP multicast transmissions, in operation since 1992 [32][22].

quirements, as well as for applications with few sources and considerable superposition of paths, if separate source-rooted trees were to be used. The use of source-rooted trees implies that multicast routers need to store group state by source, in the form *(source-addres*s, *class D IP addres*s), whilst with shared trees routers need only store a single group state *(**, *class D IP addres*s), and the location of the group's core may be configured administratively, or discovered through a specific directory service.

So far as forwarding policies are concerned, there are two possibilities: once the distribution tree has been built, the traffic may be sent in *mono-directional* mode, from the source to the receivers, or in *bi-directional* mode, when traffic may be forwarded in any direction in the tree.

Among protocols which use source-based trees, DVMRP *(Distance Vector Multicast Routing Protoco*l) [30] computes its own routing table, uses mono-directional forwarding and periodically discards and rebuilds the distribution tree. PIM-DM *(Protocol Independent Multicast - Dense Mod*e) [10] is mono-directional, maintains soft state[2] in routers and utilises existing unicast routing tables. MOSPF *(Multicast Extensions to OSP*F) [23] complements the unicast link state advertisements of OSPF with group membership information, used to build separate multicast routing tables, is also mono-directional, and only alters the distribution tree in response to changes in topology or group membership.

Protocols using shared distribution trees based on a core include PIM-SM *(Protocol Independent Multicast - Sparse Mod*e) [11], which uses mono-directional forwarding and soft state in routers, and is based on unicast routing protocols, although information about the tree's core needs to be distributed previously. CBT *(Core-Based Tree*s) [2], which is bi-directional, has its distribution tree periodically updated, is based on unicast routing tables, but is not affected by changes in network topology.

For differing reasons, all these protocols suffer from scaling problems: flooding and pruning, used in DVMRP and PIM-DM, make inefficient use of network links, as well as requiring the maintenance of pruning state in routers; MOSPF requires that all routers know where all receivers are, which is not scalable in the case of dynamic association; PIM-SM requires that information about the group's cores be distributed previously, before any multicast traffic may flow [15]. Because of this lack of scalability, these protocols are appropriate only for use within a single routing domain, and not between different domains. The first practical inter-domain solution, MSDP *(Multicast Source-Discovery Protoco*l), was developed as an extension to the unicast inter-domain routing protocol, BGP *(Border Gateway Protoco*l), to be used to join together PIM-SM domains [13]. MSDP solves the problem of autonomous domains, but is not scalable. To deal with this limitation, one practical proposal has been to permit MSDP to work together with a multicast-capable inter-domain routing protocol called BGP+ or MBGP *(Multicast Border Gateway Protoco*l). The combination PIM-SM/MSDP/MBGP thus guarantees a complete, short-term solution for multicast routing.

MBGP consists of a set of multicast extensions to BGP version 4, separating unicast and multicast routing policies, but using the same ideas of routing aggregation for connecting autonomous domains. Domain border routers have to be capable of maintaining separate routing tables for unicast and multicast, and the aim of MBGP is to communicate to border routers the mapping to domains of intervals of class D IP addresses. With the use of MBGP, internal routers need only know the internal topology of their own domain and the path to a border router running MBGP. At the border, MBGP can announce multicast routes by adding to BGP+ messages the identifier of a family of consecutive addresses, specifying unicast or multicast forwarding information.

2 To maintain soft state in routers signifies keeping forwarding information for members of a multicast group for a limited time, after which it will be discarded, unless it has been confirmed or updated first. The cost of maintaining soft state is greatly influenced by the mechanisms used for scheduling and queue management in routers and packet re-ordering and dropping in receivers. Maintenance of soft state is preferred to hard state due to a number of factors, such as flexibility, management of algorithms, fault-tolerance, and so forth.

3 Transition from the current model: an improvement or just more complexity?

The transition being undergone by the current service model suggests the apparently natural solution for the problem of scalability in multicast communication: the use of *hierarchical routing*, where routers are organised in domains. External routing protocols, operating over the connections between domains, can ignore the internal details of other domains. The MASC/BGMP project [18] includes a set of three protocols[3] which deal dynamically with the addressing scheme, and, together with BGMP *(Border Gateway Multicast Protocol)* [29], represent this transition.

MASC *(Multicast Address-Set Claim)*, which operates at the inter-domain level, is designed to solve two basic problems: to find a globally unique multicast address, and to avoid collisions in the choice of this address. Each domain possesses a MAAS *(Multicast Address Allocation Server)* to co-ordinate the delivery of multicast addresses and to monitor address space usage. There are also one or more nodes (typically domain boundary routers) running MASC, which form a hierarchy reflecting the inter-domain topology. In this hierarchy there will exist backbone domains (without any parent domains), which exchange route utilisation messages amongst themselves, and reach bilateral agreements, which enable them to allocate address spaces for their child domains. The multicast address space is partitioned between regions, and in each region the MASC nodes advertise their addresses. MASC domains hear these advertisements and request portions of these address spaces, alert to possible collisions. Once a bilateral agreement is reached and the addresses of a domain are established, a child domain requests addresses from a parent domain using a *request-collide* scheme. A parent domain that had recently acquired MASC addresses uses a type of BGP route called a group route. Group routes include the band of multicast addresses allocated and injected into BGP by the MASC node. The part of the routing table which maintains group routes is called G-RIBs, and BGMP uses the G-RIBs to build a shared multicast tree. To achieve good utilisation, for a given domain an address allocation lifetime is associated with each address band. This hierarchical system favours scalability, but requires multicast applications to adapt to possible changes of address.

BGMP runs in domain border routers, and its function is to build shared bi-directional multicast distribution trees between neighbouring domains. Like BGP, BGMP uses TCP to transport its signalling messages. Tree building depends on two components: a MIGP, acting within a domain to keep the border router informed about group membership, and BGMP, used, together with neighbouring border routers, to build the shared tree, using a similar mechanism to PIM, except that the root of a BGMP tree is not a specific router but a whole domain. The choice of potential tree cores takes into account both administrative and performance considerations, which implies a fair probability of non-optimal selection.

This set of protocols is designed to provide a long term solution for multicast addressing and routing. Apart from the scalability which results from the reduced number of forwarding states, the objectives of the MASC/BGMP project include stability, due to the reduction of protocol overhead, the maintenance of the premises of the current multicast service model, and the ability to deal with policy difference between different domains. The project also presupposes independence of the intra-domain routing protocol, which thus reduces the impact on neighbouring domains of alterations or modifications of the intra-domain routing protocol in use in a given domain.

With the completion of this transition, a more complete generic solution to multicast communication would include the set formed by IGMP, for membership of receiver stations, PIM-SM, for routing in sparsely populated regions of the interior of a routing domain, the MAAA architecture, for relating domains, multicast stations and address servers, and BGMP for communication between domains.

4 Evaluation of the current model

One of the most significant characteristics of Deering's model is the anonymity of group participation, which simplifies the mechanisms used for multicast delivery. On the

3 **MASC** - *Multicast Address-Set Claim*, used between domains, **AAP** - *Address Allocation Protocol, used within a single domain, and* **MADCAP** - *Multicast Address Dynamic Client Allocation Protocol,* used by stations to request multicast addresses [25]. **MAAA** signifies *Multicast Address Allocation Architecture* [14][12][1][24].

other hand, a weakness of the IP multicast model is the lack of addressing information. A class D IP address is no more than a name, and implies nothing about the location of group members or of the distribution tree to which they are connected, and this restricts as much the delivery of data based on content, as it does the co-operation between the multicast delivery protocols and the applications, or between end-to-end protocols that use the multicast service [20]. The different needs of multicast delivery, in different levels, are met by a profusion of protocols, a situation very different to the simplicity of unicast. In **Figure 1** we present a comparison between multicast and unicast, after [12].

A minimal set of requirements for network support of multicast applications should include speed and consistency in address allocation, and adequate response to the dynamic behaviour of group membership and routes, supposing that these requirements are based on performance, and it should also be possible to deal with heterogeneity and information partition [19]. However, this set of requirements is not well supported by the current model of IP multicast, due to problems of this model which we will now discuss [4].

The Domain-dependence Problem - The existence of cores of multicast distribution trees in different domains from the respective transmission sources leads, necessarily, to conflicts of interests. It is undesirable that one domain depends on another, since:

1. Congestion control and transmission rate from a source in a given domain may become incompatible with local policies, when the traffic crosses a domain boundary;

2. If an Internet service provider (ISP) is based on a core located in another domain, certainly it will not be able easily to control the service received by its customers;

On the other hand, an ISP will normally not want to provide the core for a session for which its customers are neither sources or receivers. If this were to happen, the ISP would be expending its resources to provide service to customers of other ISPs, and, even worse, since the current service model provides no mechanism for estimating the size of a multicast group, it would be difficult even to measure the extent of such resource consumption.

The MSDP proposal solves the problem of domain-dependence, at the cost, however, of scalability and dynamic response [1]. Dynamic groups are characterised by frequent membership changes, or by sources transmitting in bursts. With MSDP, information about sources must be known before routing state can be created, and this is difficult to manage with dynamic groups. Thus, MSDP was launched as a solution for the short term, whilst another solution was sought for to solve the problem properly. Such a solution has now appeared in the form of the MASC/BGMP project, although its complexity has to be acknowledged, especially if we think in terms of the feasibility of setting up a global structure and stimulating ISPs to adopt it.

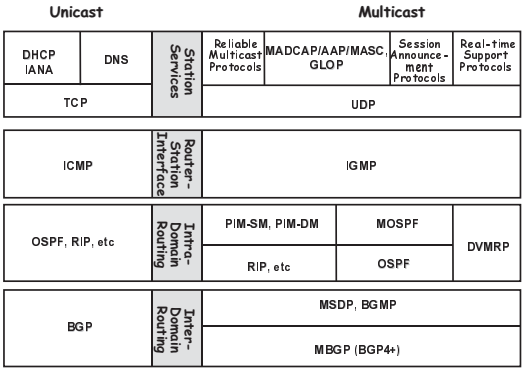

Figure 1 : Comparison between Unicast and Multicast (after [12])

The Non-resolved Functionality Problem - The elegance of the multicast model pays a price for its simplicity, offering few service options for group data delivery. Large scale applications will require much more from the network layer than anonymous packet delivery. In such applications, it is most probable that not all users will want to receive data from all sources. With such a demand profile, there will almost certainly occur some unnecessary network processing, as well as unnecessary occupation of queues in routers and of buffers in receiver stations. This negative scenario could be made even worse by the use of reliable protocols retransmitting lost packets, which might not be of any interest to some receivers. A more natural solution would be the use of some forms of addressing within the group, permitting their organisation in accordance with the receivers' interest in the data, thus reducing the problem at the network layer, and consequently reducing unnecessary traffic. It is not difficult for us to imagine such exam-

ples from distance learning, teleconferences, interactive simulation, and so on.

A multicast architecture should also suppose the existence of group management, either for charging for usage, for address recovery or for access control functions. The current model leaves this functionality unspecified, thus permitting that a multicast session be vulnerable to a number of threats, such as: denial of service attacks (through flooding), session collision, unauthorised reception and clandestine interference with authentic sources.

For the specific purpose of access control, there should exist some mechanism to control authorisations (for reception and transmission) and for group creation. A natural solution, which would prevent the dangers mentioned, would be to require explicit joining to receive transmissions from a list of known sources during the group session. Obviously, this approach is in conflict with the current paradigm, where receiver identity is not known by the source, and the source need not even belong to the group.

Protocol Cost, Group Latency and Scalability - Any of the solutions so far mentioned involve processing, time and signalling costs. The cost of association latency, for example, is particularly high with MSDP. Messages confirming association in the domain are sent periodically, and there may be a considerable delay between the moment the new candidate requests to join the group and when he receives confirmation. It is possible to reduce this problem in MSDP at the cost of increasing the number of states.

Another contribution to protocol cost is that multicast routing protocols periodically transmit information by flooding. MOSPF propagates link and group membership state packets to all routers so that they can build distribution trees. DVMRP and PIM-DM periodically flood data packets to the entire network. Even CBT and PIM-SM, which scale better, flood all routers which are candidates for the core with information about the mapping of the group.

Whether their distribution trees be source-based or shared, routing protocols suffer from new scaling problems caused by:

1. *Maintenance of connection state* - all routers which are part of a multicast tree maintain forwarding tables which can signify a significant cost in resources, either as memory or processing. Whilst shared trees imply costs for group maintenance, source-based trees involve group and source-related costs.

2. *Mechanism for advertising the source* - members of a multicast group become connected to sources without needing to know who these sources are. In sparse mode protocols, such as CBT and PIM-SM, this is because the core node requires complete knowledge and control of the entire domain. In dense mode protocols, flooding and pruning mechanisms guarantee the connection without the need for knowledge of the source. In both cases there are scaling problems, due either to the quantity of state to be maintained, or to the number of periodic update messages.

Address Space - Address allocation for IP multicast is not yet regulated. The implementations derived from the Deering model reveal the costs:

1. *Allocation of multicast addresses* - the creator of a multicast group should allocate a globally unique address. Since no means of doing this is defined in the current service model and there is no standard method for doing this, the IETF has experimented with static allocation of blocks of multicast addresses, although this is to be replaced by the MAAA set of protocols. If, on the one hand, the MAAA proposal will standardise address allocation, on the other there is concern at its complexity. More specifically, MASC, which was designed to solve inter-domain address allocations, has two controversial aspects. The first is the obviously compromise solution between address aggregation and expected demand for groups of addresses. The other is that group prefixes are not tied to domains, and are in fact leased. On the other hand, applications need to know the group address, and this would be easier if such addresses were fixed.

2. *Unknown receivers* - when a multicast packet arrives at a router, the router determines the output interfaces it will be forwarded through, but has no idea how many destinations will receive it. From the point of view of ISPs and carriers, this lack of knowledge affects security, accounting and policy requirements.

3. *Address collision* - with the growth of the Internet, it becomes more likely that more than one group creator will select the same multicast address. Obviously,

this imposes some hierarchical address allocation scheme, something absent from the Deering model, and which still does not exist.

Lack of Awareness of Quality of Service and Preferences - By the very nature of the Deering model, any group member will receive all data packets sent to the group with which he is associated. On the other hand, the modern tendency is to tailor data to the consumer. Applications will be created that will permit data to be selected by content. The supporting infrastructure should provide some alternative to merely discarding the unwanted packets at the destination, as, in this case, such applications would have to deal with unwanted network traffic, wasting resources both in the network and at the receiving stations.

For such a scenario, there is no support in the current service model, and, worse still, this model prevents a group from dealing with data based on its content. Customisation has no place in the current model of multicast, which was designed for a network offering best effort service. Clearly, this problem can be crudely worked around by creating multiple group address per session, or, alternatively, using QoS-sensitive routing, offering alternative routes using some metric. But this is not what happens in practice: QoS is not taken into account in route selection. CBT, PIM-SM and even BGMP are examples of this insensitivity [31][7]. All of these protocols build core-based trees which simplify routing, but introduce problems of core selection and of partitioning of addresses, affecting protocol performance.

5 New Multicast Paradigms

An initial movement in the direction of changing the current model, as we have seen, focussed on restrictions related to domain dependence, certain aspects of scalability, such as the reduction of router state, and a more efficient global scheme for address allocation and access control. The solution represented by MASC/BGMP is complex, whilst its practical implementations are based on static allocation and assignment of addresses, and are considered to be unsatisfactory. The only proposal which satisfies almost all address requirements is the future change to IPv6 addressing, which, as is known, will not happen sufficiently quickly enough, on account of all the required changes in Internet infrastructure.

Another, more radical, approach defends changes in the heart of the model, reasoning that the use of anonymous multicast allied to the lack of appropriate network support[4] is incapable of providing a real, long term solution which meets the needs of end users and service providers. Along these lines, there exist proposals to simplify the multicast paradigm, such as SM *(Simple Multicas*t) [26] and EXPRESS [16]. Yet another simplifying proposal, CLM *(ConnectionLess Multicas*t) [24], attempts to complement the current model, by removing the need for maintaining router state, even at the cost of restrictions on the size of groups. A further class of proposals, represented by AIM *(Addressable Internet Multicas*t) [20], concentrates on one of the main targets of criticism of the current model, the lack of addressing within groups, proposing a service which permits subgroups selectively to receive data based on content, and not merely on the multicast address. Another movement for change, which however remains beyond the scope of this paper, uses techniques which take into consideration QoS for selecting multicast routes.

5.1 AIM and Anycasting

AIM generalises the architecture of IP multicast by introducing addressing information within the multicast distribution trees. In this way, AIM seeks to provide a routing infrastructure which simultaneously supports low latency for address allocation, low overhead for group creation, a flexible anycasting[5] mechanism and a scheme for naming data. AIM is an instance of the protocol architecture ALF *(Application Level Framin*g)[6], designed better to express the requirements of applications to lower protocol levels. AIM will use any kind of multicast distribution tree, source-based or shared, extending IP multicast with the establishment of three kinds of label, to be associated with routers, each of which expresses a type of service.

4 In the traditional and predominant view of the network, its core is fast and simple, and complex processing is only performed at the network edge. However, in recent times, advocates have appeared who defend greater network support.

5 An anycasting service consists in the delivery of a packet to any one, and only one, of a group of group of possible receivers [33].

Positional outer Label - specifies the router's location relative to a fixed point of the tree, which is used a the addressing root, and is not necessarily identical to the source or core of the underlying distribution tree. Using positional labels, routers possess addresses within the multicast distribution tree, and, using this address, are able to select the reception of just a part of transmissions to the group. The positional labelling algorithm assigns the label "1 to the root, and each subsequent node has a prefix which corresponds to the parent node, and a unique suffix among the children of that parent. A new router prefix must always be passed on to its children, and so on successively to the leaves of the tree. With the routers so labelled, routing is implicit, and it is unnecessary to maintain any routing state in the routers. A simple comparison of the positional label of the receiver with the router label defines whether the receiver is an ancestor or a descendent within the tree, and forwarding is performed accordingly.

In order for a source to know to whom to send the packets, it must know the respective labels, or assign a corresponding mask, in the case that the receivers occupy an entire subtree. When receivers are sparse, only the GCP *(Greatest Common Prefi*x) of the list is processed by the corresponding router, and the list of further receivers is maintained in IPv6 extension headers (which implies a certain restriction of AIM). When the packet in question reaches the GCP router, this router decides the next GCP, by means of a flag (#), and so on, successively.

With positional labels it is possible for sources to address a selected part of the multicast group, based on information contained in the packet, without this generating additional traffic.

Distance Label - associated with each interface belonging to the multicast distribution tree, this specifies the distance, using some metric, to the next qualified router of the same group. This type of label also allows routers to address a specific subgroup within the multicast group. The distance label specifies a predefined type. Type 1 defines the *hop count* between the router and the nearest group member or receiver. Other types may be used to represent available resources or the average load at a station. A station which executes a given application advises its local router what is the kind of distance label desired. A router which receives an updated distance label from on of "its" stations (or from a neighbouring router) increments this label and associates it with the interface by which the update was received. Then the router announces to neighbouring router in the distribution tree its lowest distance label, which results in a labelled tree.

This scheme supports an anycasting service. To use it, packets are only forwarded by a given router using the interface with the lowest associated distance label. If two or more interfaces have the same distance label, the router selects one of them using some consistent criterion. Routes followed using anycast routing are not necessarily the same as those used in the underlying multicast routing. For example, in PIM-SM, all packets are first forwarded to the core, and then distributed, and this does not normally happen with anycasting.

Stream Label[7] - specifies the association of one or more receivers of the multicast group with the traffic generated by a subgroup of sources. This kind of label allows receivers to identify a subgroup of sources and rapidly to define receiver subgroups, without the need for new distribution trees. With stream labels, applications (or highest level protocols) may define the same contexts and meaning for different individual packets, labelling them with the same identifier, in such a way that multicast groups may be explicitly aggregated at the network layer.

AIM stream management is performed by each router maintaining a local stream table, containing for each stream its identifier and the forwarding state across each of the router's interfaces. A router is associated with each stream,

6 The main idea behind Application Level Framing is that the application's semantics should be reflected in the network protocol, thus optimising its performance both in the network and in the end systems. ALF integrates the transport and application layers, by proposing that the application manage the packaging of the application data, which become the only unit of processing, control and transmission throughout all network layers. This requires a new way to name data, placing application and data-flow entities in structures based on their relationships in the context of the applications [19][21].

7 In this context, *stream* signifies a logical grouping of packets within a multicast group. The use of such grouping is recursive within the main multicast tree. Such is the case of the grouping of audio and video data sent by a source to a multicast group.

and is responsible for advertising the stream, so that other routers may update their respective local stream table.

5.2 RAMA (*Root Addressed Multicast Architecture*) Proposals

The proposals of the RAMA model [1], which includes SM and EXPRESS, abandon some aspects of the current multicast model in the name of simplified routing. In this model, a channel is uniquely addressed through the tuple (N - address of core or source, G - group address). In the case of EXPRESS [16], two sources, S1 and S2, transmitting data to the same group, G, will only be received by receivers associated with both sources and thus with both channels. In the same way, in SM the advertisement of the core (N) of a shared tree will be associated with a specific group address, G.

SM and EXPRESS, however, have their differences. Whilst EXPRESS imposes a mono-directional tree per source, SM builds a bi-directional shared tree for various sources. SM requires changes to packet format, whereas EXPRESS does not.

From our point of view, SM is more representative of the RAMA model, because of its capacity to solve inter-domain routing [26][3]. The SM project supposes that the notion of a session can be determined at the level of a single shared tree (EXPRESS links the notion of session to the application layer), given that a great number of applications use multiple sources. In this way, the service model presented here is of the multi-partner type, using a bi-directional tree rooted in a core. Like EXPRESS, and different from the traditional model, this protocol separates group and core discovery from routing concerns. The address parameters of the group (N, G) may be obtained using some out-of-band application-level mechanism (WWW, e-mail, DNS, SAP, etc.), avoiding complexity and unifying intra- and inter-domain routing, without the need for the global allocation of a multicast address.

If a data source is already a group member, its SM data packet should carry an IP-encapsulated header, containing the group identifier (N, G), and in the destination address of the IP header there should be placed the code corresponding to ALL-SM-NODES, thus assuring that non-SM-enabled nodes will ignore the packet. A SM router which receives the packet does a forwarding table lookup on (N, G) to decide whether to forward or discard it. If the output port is a tunnel, the SM router substitutes the destination IP address by the address of the tunnel endpoint, which will certainly also be a SM router, and which will restore the ALL-SM-NODES destination value on packet arrival.

SM control messages, which are also encapsulated in IP headers, are sent in the direction of the core. A *join* message contains the IP option *Router Alert* [17], which involves the inspection and processing of this packet by all routers along the path to the core, with each SM router inserting its identifier in the appropriate header field, or tunnelling packets between non adjacent SM routers. On arrival and acceptance of a join request at the core, the acknowledgement follows the reverse path, confirming information on routing state all the way to the sender of the original join request. Tree maintenance is similar to CBT, with parent routers monitoring their children, and vice-versa, emitting and monitoring *liveness* and *heartbeat* messages.

5.3 XCAST: Explicit Multicast

Proposals have been presented at the IETF seeking to take advantage of the characteristics of real multicast applications without abandoning the facilities of unicast transmission, in recognition of the fact that practical applications are best classified as *few-to-fe*w. The different approaches used to handle the distribution of identical packets involve three separate mutually orthogonal costs, which will be described here, and which are illustrated in **Figure 2**. The pure unicast solution *(n* separate unicast transmissions) corresponds to point 1 of the figure. The traditional Deering solution (a single transmission to a multicast address) corresponds to point 2. The proposal represented by XCAST (the packet is sent only once, but contains a list of *n* receivers) corresponds to point 3. Solutions proposed within the scope of the current service model permit some intermediate point in the figure between points 1 and 2, since it is possible to exchange bandwidth for the cost of signalling and maintenance of state in the routers. The techniques proposed in XCAST (and mentioned in the **Figure 2**) allow us to exchange processing cost at a border router for bandwidth, or for state and signalling costs.

Boivie [5] asserts that there is a hyperbolic cost profile per group member, as a function of the number of members. This profile is evidence that the current service model is relatively expensive if groups are small and are many in number, as is the case in teleconferences, interactive games

and collaborative applications. XCAST is destined for such applications.

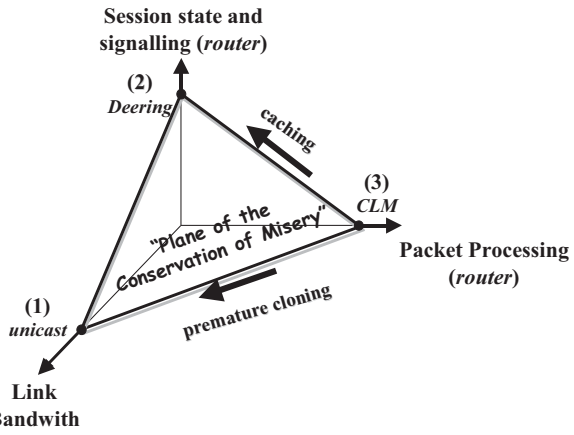

Figure 2: Plane of the Conservation of Misery

The idea behind XCAST is simple. To use unicast transmission implies a larger cost in bandwidth consumption, a scarce resource in access networks, with all the accompanying performance losses. To use Deering-style multicast generates costs due to state maintenance and signalling, requiring (memory and processing) resources which are relatively scarce in backbone routers, due to immense contention. The XCAST proposal seeks to transfer the brunt of costs to edge routers, not only in terms of bandwidth costs, but also in terms of per-packet processing costs. The implementation of this approach is similar to the MBone, and depends on an overlay virtual network of XCAST routers and stations, connected by tunnels, whereby XCAST packets are transported as common IP datagrams, taking advantage of the existing unicast support.

A XCAST data packet contains the IP address of all the members of a session. In each XCAST router the packet is processed, the output interfaces are identified, and for each such interface a new XCAST packet is built, containing only the IP addresses of members corresponding to that route. In the case that an output interface corresponds to just a single receiver, the remainder of the transmission uses pure unicast.

There are two ways to include a list of destination addresses. The first of these depends on creating a new IPv4 option; the other adds an additional header between the network and transport headers, known as the Xcast4 header. The Xcast header is processed only by Xcast-enabled routers, and contains control information to be interpreted at multicast tree branching nodes. The IPv4 and Xcast headers of a typical packet are described in [5], and briefly summarised below.

IPv4 Header
<Destination = IPv4 multicast address All_Xcast_Routers; value to be defined by IANA>
<Source = level 3 address of multicast source>
<Protocol = PROTO_Xcast>; value to be defined by IANA

Xcast4 Header
<Destination = list of IPv4 unicast address (and, possibly, also transport protocol port numbers) of multiclass group destinations>
<Protocol = UDP or TCP>
<bitmap = one bit for each destination, indicating whether it continues to be valid due to tree branching>

6 Conclusions

The alternatives which have been proposed to the traditional model of IP multicast and its known extensions concentrate their attention on overcoming insufficiencies and complexity which still exist. With AIM, Levine introduces group-relative addressing information in the multicast distribution trees, which allow new delivery and filtering services based on content preferences. This form of receiver group organisation simplifies important ISP requirements, such as security, session accounting and coexistence with established policies, whilst at the same time resulting in great savings through the reduction of unnecessary traffic and of router state. The anycast technique may prove useful for applications which need to select and locate the most appropriate within a set of possible servers, each of which replicates a given service, such as those which today are manually configured, like NTP *(Network Time Protoco*l) and DNS *(Domain Name Syste*m). With this kind of addressing, it is also simpler, as well as more economical in bandwidth, to perform retransmissions in reliable multicast protocols, eliminating unnecessary control traffic, and adding functionality to routers without increasing their state requirements.

Continuing in the addressing area, the RAMA proposal, represented here by its long term solution, SM, removes the distinction between intra- and inter-domain routing, separating group and core discovery from routing questions. With SM, some of the multicast problems we have alluded to are resolved as follows:

1. *Scalability* - implements a trivial allocation of multicast addresses, since, for each core, the whole class D address space is available. A receiving station can participate in a group, whether or not the adjacent router is SM-enabled.

2. Support for *group access control* - performed in the core, where lists are maintained of authorised/included and unauthorised/excluded nodes. Access rules defined in the core are propagated to the remaining routers in heartbeat messages, using an access control list for border routers.

3. *Scope of multicast transmission* - unlike in the current model, SM may use for multicast transmission the scope limits defined by unicast routing (subnet, area, autonomous system, federation of autonomous systems), provided that the border routers can process SM packets without requiring any special protocol to deal with the question. In addition, a single group identifier (N, G) may be used in multiple scopes.

4. *Domain independence* - when SM is used both intra- and inter-domain, it is necessary to guarantee that join messages from different internal receivers in one domain converge on a single point in the other domain.

5. *Adaptability to multicast dynamics* - several groups per session may be set up, due to the abundance of (N, G) pairs, which thus achieves lower delay in packet delivery or network load balancing.

A number of points in the SM proposal remain controversial or open questions:

1. *Filtering in the link layer* - as in a subnet, the mapping to the MAC address is made using the low order bits of the class D multicast IP address, which implies that different SM groups, using the same class D address, will receive packets unnecessarily.

2. *Performance questions* - SM packet forwarding involves searching a table based on the content of the SM header, and forwarding copies of the packet to the respective interfaces. Non-SM-enabled routers will suffer performance losses, due to software handling of SM packets. On the other hand it is usual that forwarding mechanisms be implemented directly in hardware.

3. *Group state aggregation* - no specific proposal has yet been made in SM for aggregating group routing information.

The last of the alternatives which were analysed, XCAST, explicitly abandons scalability for the simplified scenario of *few-to-few* applications. The strategy of abandoning class D address allocation for the coding of a list of group members in the data packet avoids the problems of maintaining router state, and facilitates adaptability to topological changes and knowledge of the group membership, which implies in total control of session management. With XCAST it is unnecessary to establish complex partnership agreements between domains, which thus renders XCAST a suitable inter-domain protocol to be used with SM or PIM-SM. These benefits are obtained at the price of packet overhead, and greater header processing, which is appropriate for small, sparse groups.

Apart from these questions, others remain which have not been considered here, both within and outside of the scope of the traditional IP multicast model. These include routing with restrictions, reliable multicast and secure multicast. A mature approach to multicast transmission should also include simpler, more general and more appropriate answers for these questions.

References

[1] K. C. Almeroth. The Evolution of Multicast: From the Mbone to Inter-Domain Multicast to Internet2 Deployment. In *IEEE Network*, September/1999

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

[3] T. Ballardie, R. Perlman, C. Lee, J. Crowcroft. Simple Scalable Internet Multicast. tech. rep., University College London, April/1999

[4] F. A. R. Barros. Difusão Seletiva: Confiabilidade, Escalabilidade e Qualidade de Serviço. MSc dissertation (in

Portuguese), Instituto de Computação, Universidade Federal Fluminense - UFF, 2001

[5] R. Boivie, et al. Explicit Multicast (Xcast) Basic Specification. IETF draft. draft-ooms-xcast-basics-spec-04.txt, *work in progress*, 2003

[6] D. E. Comer. Internetworking with TCP/IP. vol 1, 3rd ed., Prentice-Hall, 1995

[7] E. Crawley, R. Nair. B. Rajagopalan, H. Sandick. A Framework for QoS-based Routing in the Internet. RFC 2386, August/1998

[8] S. Deering. Host Extensions for IP Multicasting. RFC1112, 1989

[9] S. Deering. Multicast Routing in a Datagram Internetwork. PhD Thesis, Stanford University, December/1991

[10] A. Adams, W. Siadak, J. Nicholas. Protocol Independent Multicast Dense Mode (PIM-DM): Protocol Specification (Revised). IETF draft. draft-ietf-msdp-spec-14.txt, *work in progress*, 2002

[11] S. Deering et al. Protocol Independent Multicast Sparse-Mode (PIM-SM): Protocol Specification, RFC 2362, 1998

[12] C. Diot et al. Deployment Issues for the IP Multicast Service and Architecture. In *IEEE Network*, January/2000

[13] D. Meyer, B. Fenner. Multicast Source Discovery Protocol (MSDP). IETF draft. draft-ietf-msdp-spec-14.txt, *work in progress*, 2002

[14] D. Thaler, M. Handley, D. Estrin. The Internet Multicast Address Allocation Architecture, RFC 2908, September 2000

[15] M. Handley, J. Crowcroft. Internet Multicast Today. In *The Internet Protocol Journal*, Vol 2, No. 4, December/1999

[16] H. W. Holbrook, D. R.Cheriton. IP Multicast Channels: EXPRESS Support for Large-Scale Single-source Applications. ACM SIGCOMM'99, September/1999

[17] D. Katz. IP Router Alert Option. RFC 2113, February/1997

[18] S. Kumar et al. The MASC/BGMP Architecture for Inter-Domain Multicast Routing Protocol. ACM Sigcomm, 1998

[19] B. N. Levine, J. Crowcroft, C. Diot, J. J. Garcia-Luna-Aceves, J. F. Kurose. Consideration of Receiver Interest for IP Multicast Delivery. In *IEEE INFOCOM 2000*

[20] B. N. Levine. Network Support for Group Communication. PhD Thesis, University of California, June/1999

[21] S. McCanne. Scalable Multimedia Communication Using IP Multicast and Lightweight Sessions. In *IEEE Internet Computing*, 1999

[22] C. K. Miller. Reliable Multicast Protocols and Applications. In *The Internet Protocol Journal*, Vol 1, No. 2, September/1998

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

[24] O. Paridaens, D. Ooms, B. Sales. Security Framework for Explicit Multicast. IETF draft. draft-paridaens-xcast-sec-framework-02.txt, *work in progress*, 2002

[25] S. Hanna, B. Patel, M. Shah. Multicast Address Dynamic Client Allocation Protocol (MADCAP), RFC 2730, December 1999

[26] R. Perlman et al. SIMPLE Multicast: A Design for Simple, Low-Overhead Multicast. ietf draft, draft-perlman-simple-multicast-03.txt, *work in progress*, 1999

[27] B. Quinn, K. Almeroth. IP Multicast Applications: Challenges and Solutions, RFC 3170, September 2001

[28] L. Sahasrabuddhe, B. Mukherjee. Multicast Routing Algorithms and Protocols: A Tutorial. In *IEEE Network*, January/2000

[29] D. Thaler. Border Gateway Multicast Protocol (BGMP): Protocol Specification. IETF draft. draft-ietf-bgmp-spec-03.txt, *work in progress*, 2002

[30] D. Waitzman, C. Partridge, S. Deering. Distance Vector Multicast Routing Protocol. RFC 1075, 1988

[31] B. Wang, J. C. Wou. Multicast Routing and Its QoS Extensions. In *IEEE Network*, January/2000

[32] B. Wang, J. Hou. QoS-Based MCast Routing for Distributing Layered Video to Heterogeneous Receivers in Rate-based Networks. In *IEEE INFOCOM* 2000

[33] T. Wong, R. Katz, S. McCanne. An Evaluation of Preference Clustering in Large-Scale Multicast Applications. INFOCOM, 2000