A Survey of Multicast Technologies

Vincent Roca, Luís Costa, Rolland Vida, Anca Dracinschi, and Serge Fdida

Université Pierre et Marie Curie (Paris 6)
LIP6 - CNRS, Network and Performance Group
8, rue du Capitaine Scott - 75015 - Paris, FRANCE
fax: (+33) 1.44.27.53.53
Firstname.Lastname@lip6.fr; http://www-rp.lip6.fr/

WORK IN PROGRESS, Version 1.3 - September 27th, 2000

Abstract

This document gives an overview of most of the directions taken by research in the multicast area. We first introduce the basic concepts. The following sections deal with high level services that can (or must) be provided on top of the underlying multicast routing infrastructure. Then we consider new evolutions in multicast routing: new protocols, their large scale deployment, and future trends. Finally we discuss multicast tools and applications.

Contents

1 Basic concepts

Many of the emerging applications in the Internet, such as Internet TV, tele-conferencing, distance-learning, and distributed games, fall in the category of group communications as opposed to the classical one-to-one conversations. These applications may have several sources and a huge number of receivers, up to millions in the case of Internet TV for example. These applications drove the development of the multicast service, as the use of several unicast channels is unfeasible in terms of network resources and processing power of end stations.

Multicast communications, i.e. the ability to send efficiently information to one or more receivers at the same time (i.e. in such a way that packets are not sent several times on a given link), tend to be more and more used. This is an efficient way:

- to distribute information to a set of receivers (e.g. for cooperative work applications like audio-conferences, white board, etc.), or
- to discover resources (e.g. at boot time an IPv6 host looks for a configuration router by sending a request to the "all routers" multicast address).

In this section we introduce the fundamentals of Internet multicast transmissions.

1.1 The two Levels of Multicast Transmissions

First of all we need to distinguish two levels of multicast transmissions: local and wide area. The reason is that these two situations are completely different as well as the solutions that are derived.

Local-area multicast transmissions

The local level *takes advantage of the possible multicast transmission capabilities of the physical layer.* Let's consider Ethernet: Ethernet supports point-to-point, broadcast and also multicast transmissions. This is made possible by the diffusion nature of the Ethernet technology (at least in the 10Base5, 10Base2, and non-switched 10BaseT versions). A block of MAC addresses has been reserved to multicast. They are identified by the Least Significant Bit of the Most Significant Byte of the address which is set to 1. The Ethernet multicast address is then created by copying the low 23 bits of the IP multicast address in `01.00.5E.00.00.00`. E.g. `224.0.2.2` leads to `01.00.5E.00.02.02`.

As a consequence, several IP multicast addresses that are only distinguished by the 9 Most Significant Bits lead to the same Ethernet address. It means that each time a multicast Ethernet frame is received, the receiver must check that the IP multicast address was the one expected.

It is also possible to emulate multicast transmissions using Ethernet broadcast and an incoming IP-level filtering. This is the solution used in old generation Ethernet cards that do not support the multicast mode.

Using multicast on a physical network that does not rely on diffusion (like ATM) is both more complex and less efficient. It often requires going through a server which performs distribution to each receiver either in many one-to-one connections, or in a one-to-many connection (if available, like in ATM).

The GARP (Generic Attribute Registration Protocol) and GMRP (GARP Multicast Registration Protocol) are the protocols defined in the IEEE 802.1p Standard [1] to coordinate multicast distribution between MAC bridges and switches. GARP provides the "routing" functionality at the link-layer level as GMRP provides the member registration services equivalent to IGMP at the network level. (We detail these network level functionalities later in this document.)

Wide-area multicast transmissions

Wide area multicast transmission is completely different. It requires the use of multicast routers, i.e. hosts that are capable of building and managing the multicast distribution tree[1]. According to whether

[1] Note that a multicast router is not necessarily a router attached to several different physical networks. A host (with one network card only) can also serve as a multicast router as it deals with "logical interfaces" rather than physical ones.

this multicast router is on a leaf network (i.e. with end-hosts connected) or not (i.e. on a transit network), it will have different functionalities:

- on a leaf network, the multicast router must discover the presence of local receivers (i.e. hosts willing to receive traffic destined to a multicast group). This is the purpose of IGMP (section 1.4).
- on a transit network, the multicast router participates in the distribution tree management and multicast packet forwarding (section 1.6.1).

1.2 The Group Model

The IP multicast model (or group model) is an open model where:

- anybody can belong to a multicast group, no authorization is required
- a host can belong to many different groups, there is no restriction
- a source can generate traffic to a multicast group no matter whether it belongs or not to the group
- the group is dynamic and one can subscribe to or leave a multicast group at any time
- the number and identity of group members is known neither to the source nor to the receivers

Each group is identified by a IPv(4 or 6) "multicast address" (section 1.3). Some addresses are also reserved for dedicated purposes (e.g. to identify all hosts, or all routers, etc.).

1.3 Multicast Addressing

Multicast IPv4 addresses are class D addresses that range from `224.0.0.0` to `239.255.255.255`. A multicast address, unlike other unicast addresses, does not identify a given host, but identifies a multicast group for its whole duration. A multicast address, unlike unicast addresses, is not assigned statically by an authority, but is dynamically assigned (in fact chosen by the source). Because there is no allocation scheme yet (but it will change, see section 4.4), there is a risk that several disjoint sources chose the same address. According to the scope of each traffic (i.e. TTL value), packets may be interleaved and applications may be corrupted.

1.4 The IGMP Protocol (Internet Group Management Protocol)

IGMP (Internet Group Management Protocol) is a level 3 protocol (like ICMP). Therefore an IGMP message is always encapsulated in an IP datagram. Its goal is to inform the local multicast router of the presence of hosts interested to receive traffic sent to a group. In that purpose the local multicast router periodically sends "QUERY" messages on the LAN, asking if anyone is currently listening some multicast group. Each receiver having subscribed to a group answers and informs the router of the group(s) identity. A feedback cancellation scheme (a receiver does not answers immediately but after some random time, listening to other announcements in the meantime) avoids an implosion of the router.

Several versions of IGMP exist:

- IGMP version 1: described in RFC_1112
- IGMP version 2: described in RFC_2236
  The major difference compared to IGMPv1 is the addition of the "fast leave detection" mechanism. It also enables a host to inform the local multicast router that it has left a group. It enables this router to update the multicast tree, dropping this branch immediately, if nobody else is interested in the group anymore.
- IGMP version 3: described in an Internet draft [12]
  This version essentially adds the possibility to do per-source filtering.

IGMP is a protocol which is restricted to the local dialog between receivers and their first-hop multicast router (local scope). This is completely different from the creation of the multicast distribution tree which is the responsibility of the multicast routers (wide-area scope) and the multicast routing protocols.

1.5 The Various Classes of Multicast Routing Algorithms

Several classes of algorithms for the creation of a multicast distribution tree exist. We can mention:

- flooding
- spanning tree
- Reverse Path Forwarding (RPF)
- Core-Based Tree (CBT)
- solution to the Traveling Salesman Problem (TSP)

Before introducing these approaches, we first focus on the underlying problem.

1.5.1 The Multicast Routing Problem

The multicast routing problem is the following. The interconnection network can be modeled as a *directed graph*, consisting of a set of nodes (vertices) V and a set of links (edges) E [58] [70]. Let $G = (V, E)$ be that directed graph. A directed link of G from node u to node v is represented by the ordered tuple (u, v). Let M be the multicast group including the sources. M is therefore a subset of set V.

The multicast routing problem consists in *finding one or more interconnection topologies, subset of G, that span all nodes included in M.* If a single topology is sufficient, no matter what the source is, this solution will be called a "shared interconnection topology" (e.g. see the CBT algorithm, section 1.5.5). If several topologies are required, one per source, this solution will be called a "source directed interconnection topology" (e.g. see the RPF algorithm, section 1.5.4).

Note that the above definition does not assume the presence of symmetric links. In practice links (u, v) and (v, u) will often be the same and have similar transmission features (bandwidth, propagation time, length, etc.). The directed graph G will often be composed of many undirected links.

In the general case, M is a subset of V, making the finding of an optimized solution complex. For instance the "Steiner tree problem" is an NP-complete problem [58]. In some cases (e.g. with host-based multicast, section 7.5), M is equal to V and the "Steiner tree problem" reduces to the "minimum spanning tree problem" for which polynomial-time algorithms exist.

1.5.2 The Flooding Approach

A node receiving a packet checks if it is the first reception of this packet. If this is the case the packet is forwarded on each outgoing interface except the one where it has been received. Otherwise the packet is dropped.

The difficulty relies in the "first reception" test. One solution is for a node to remember all the packets received so far. Another solution is that each packet contains the list of all the nodes crossed. For instance, `OSPF`, which relies on a flooding algorithm, compares the date of the (database update) packet received with the modification date of its own database.

If this algorithm is both simple and robust, it is also very memory and network resource consuming.

1.5.3 The Spanning Tree Approach

In this algorithm we select a subset of the physical links to create a loop-less tree of minimal cost, including all the nodes. This problem is known as the "Steiner tree problem in networks (SPN)" and is known in the general case a NP-complete problem. Besides the cost c_{uv} of each link (u, v) must be known.

If finding the minimum cost tree is not a requirement, then a simple algorithm exist: (1) select a core, and (2) keep only the links that are on the shortest path from this core and the other nodes (see the RPF algorithm, section 1.5.4).

Unlike the flooding algorithm, here some links may remain unused. This algorithm is simple, robust, requires little memory (a flag that indicates if a link belongs to the tree or not is sufficient). On the other hand, all the traffic is concentrated on a fixed subset of the physical layers, no matter the spreading of the receivers.

1.5.4 Reverse Path Forwarding (RPF)

This is the algorithm used by the DVMRP protocol (section 1.6.1). The principle is the following:

```
let P be the packet received from source S on the interface I;
if (I is on the shortest path to S)
        forward P on all the interfaces except I;
else
        drop P;
```

The algorithm efficiency can be increased by checking, before forwarding the packet to a neighbor, if the current node is on the shortest path between the source and this neighbor.

This algorithm has several advantages:

- it only relies on the point-to-point routing database (to know if I is on the shortest path to S),
- the multicast distribution trees of different sources are also different, which enables a better load balance on the various links,
- it guarantees the fastest possible delivery of packets as each step only keeps the shortest paths.

The “Flood and Prune” Variant

The RPF algorithm has the advantage of simplicity. A major drawback yet is the fact that it leads to distributing packets to all the possible nodes. A variation has been designed to enable the pruning of useless branches that are not on the path to a receiver. This is the “flood and prune” algorithm.

During the first step (“flood”), all the possible nodes receive all the packets. A leaf multicast router that receives a packet but has no local host interested by this multicast group, sends upwards a “prune” request. This is the second step (“prune”). Slowly, the intermediate routers are informed that some sub-trees do not lead to receivers and thus prevent themselves from forwarding the packets on these links.

This variant of RPF has the advantage of limiting the distribution tree only to the useful branches. But on the other hand it requires going through periodical flooding steps (to discover new receivers), and requires that each multicast router, no matter whether it is or not on a path to a receiver, keep state information for each multicast group (section 6.1).

1.5.5 Core Based Multicast

Here a core is first chosen for the multicast group. The receivers (in practice the last-hop multicast router) must send their subscribe request to this core. Each intermediate router remembers the interface from which this request has been received in order to include it into the distribution tree.

The created tree (called a “shared tree” for that reason) is the same for each source. This is both an advantage (less memory is required) and a drawback (the traffic is more concentrated). Another major asset of CBT is that transmissions are strictly limited to the routers that are on the path to a receiver.

1.5.6 Solution to the Traveling Salesman Problem (TSP)

If using a loop-less tree is usual, this is not the only possibility. A tour, solution to the Traveling Salesman Problem (TSP) [5] is also possible even if not optimal from a networking resource point of view.

1.6 The First Internet-Wide Multicast Deployment

Multicast has been deployed in the early 90s as an *experimental* world-wide service. The key protocol that forms the basis of this infrastructure is DVMRP. This section outlines its features and describes rapidly the MBONE multicast backbone.

1.6.1 The DVMRP Protocol

DVMRP (Distance Vector Multicast Routing Protocol) is the multicast routing protocol in use in the MBONE (Multicast Backbone). DVMRP relies on the exchange of "distance vector updates" like its unicast analog RIP. Each vector entry contains the destination (in this case a source), and a distance expressed in number of hops. These updates are sent on each multicast capable interface and on each multicast tunnel (see below). The multicast distribution tree is then created using the RPF algorithm and the distance database of DVMRP.

During the first DVMRP deployment, there was no pruning and transmissions were only limited by the TTL field (section 1.8). Since 1993, all the versions of DVMRP use the pruning version of RPF. Yet some packets regularly flood the whole MBONE, as far as made possible by the TTL value and the various thresholds (section 1.8). This periodic flooding is required in order to update the distribution tree, including new receivers and dropping those who left in the meantime.

1.6.2 The MBONE Infrastructure

The MBONE must be seen as an overlay of the Internet. As only a subset of the Internet routers can perform multicast packet forwarding, the MBONE consists in the *interconnection of multicast-aware areas through tunnels.* A tunnel is composed of two multicast routers, one at each extremity. Each multicast packet that needs to go through this tunnel is first encapsulated in a *point-to-point* packet (IP in IP encapsulation) addressed to the other side of the tunnel. This unicast packet is sent and can now be forwarded by unicast routers. When it reaches the other end of the tunnel, it is decapsulated and sent using native multicast in the second area.

If the presence of tunnels enabled a fast deployment of a world-wide multicast infrastructure, it also led to many inefficiencies. For instance, a packet will cross several times the same physical link if it goes successively through two tunnels that share this link.

1.7 Creating and deleting a multicast group

Creating a multicast group consists, for the source, in choosing a new multicast address and sending packets to that address. Upon receiving packets to a new address, multicast routers create new forwarding state and set up the distribution tree. Of course if there is no known receiver, this distribution tree is limited by the first-hop router (in case of DVMRP). But even in that case the tree exists and state is kept by multicast routers (e.g. to say that this group has been "pruned").

Deleting a multicast group consists, for the source, in avoiding to send packets to a multicast address. As the distribution tree consists of soft-state kept by multicast routers, this tree will slowly disappear (for instance `mrouted` uses a default 5 minute timer).

1.8 Limiting the Scope with the TTL Field

Controlling the scope of multicast packets relies on the TTL (Time To Live) field of the IP header. As with unicast transmissions, the TTL is decremented each time the packet is forwarded by a router, and dropped when it reaches zero.

This is generalized by IP multicast with the notion of *threshold.* These thresholds enable the creation of areas that a multicast packet can only cross if its TTL is superior to the threshold value. By convention:

- 32: defines the "organization boundary"
- 64: defines the "region boundary"

- 128: defines the “continent boundary”

Note that the notions of (organization, region, continent) are not strictly defined.

2 Providing a Reliable Transmission Service

2.1 Introduction

We have seen so far the routing aspects, i.e. how to send information to any number of receivers in a flexible and dynamic way. But this is a (little) part of the problem and most applications ask for additional services. The possibility of doing reliable transmissions is the most common requirement [21]. This is the goal of this section.

2.2 Classification of the Solutions

Two major tasks are required for a reliable transmission service: error detection and error recovery. In the following sections we describe the two principal solutions (FEC and ARQ) that are used by different reliable multicast protocols, as well as some hybrid solutions that have recently emerged.

2.2.1 ARQ Solutions (ACK and NAK-based)

The basic idea of the ARQ (Automatic Repeat reQuest) approach [50, 52] is to retransmit a packet only if it is lost by at least a receiver.

Depending on whether error detection is done by the sender or the receivers, reliable multicast protocols could use positive (ACK) or negative (NACK) acknowledgements.

When using *positive acknowledgements (ACKs)*, the sender retransmits messages until ACKs from all destinations are received. This approach does not scale well because ACKs sent by each receiver for each received packet may lead to serious network congestion (ACK implosion). In addition the source has to know the exact composition of the multicast group.

Using *negative acknowledgements (NACKs)* shifts the error detection load from the source to the destination. Receivers transmit negative acknowledgements only when a packet loss is detected. In order to reduce the implosion problem, different NACK suppression mechanisms could be applied, since the sender only needs to know that at least one receiver is missing data.

2.2.2 FEC-based Solutions

Forward error correction-based recovery [50] consists in sending redundant packets (also called parity packets) together with regular data packets. Reed-Solomon Erasure correcting code [44] [53] can be used to construct the parity packets. For every k data packets, n-k parity packets are constructed. All k the data packets can be reconstructed if at least k (original or parity) packets out of n are correctly received. This code has two drawbacks: there is a limited number of parity packets per data packet, and in practice the k parameter is limited to 30 or 60 for computational reasons. Other more powerful codes exist [40]: the Tornado code, Turbo code, LT code, etc. Some of them are patented.

The FEC-based approach reduces the end-to-end latency compared to the ARQ approach (a receiver does not have to wait for the retransmission of lost packets any more). But this is at the cost of bandwidth since additional packets are systematically sent.

2.2.3 Hybrid Solutions

In practice, using only FEC cannot guaranty a totally reliable service. Hybrid methods combining FEC with ARQ were thus proposed. There are two approaches to combine FEC and ARQ [50]: The *layered approach* considers FEC as an independent layer below the ARQ-based protocol [32]. The key advantage of such a solution is that FEC is transparent to the ARQ protocols and transparently improves ARQ performances. Besides, if an application does not require a complete reliability, the ARQ protocol may be skipped in order to only use the FEC layer.

The second approach proposes to *integrate FEC and ARQ in the same layer*, as part of the same protocol [28] [48]. The source uses the feedback from the receivers to learn how many packets were lost by the worst-case receiver. Only that number of FEC packets are then sent. This integrated solution outperforms the layered FEC-based approach, the benefits being more significant for large group sizes.

2.3 Examples of Reliable Multicast Transmission Protocols

2.3.1 The XTP Approach (for history)

XTP (Xpress Transport Protocol) relies on the bucket algorithm to provide statistically reliable transmissions. Time is divided into periods and to each period is associated a bucket. A bucket collects reception information sent regularly by receivers. There is a limited number of buckets. This number is a trade-off between the response time and the reliability level. If some information still misses from the oldest bucket at some receiver, the information contained are nonetheless sent to the application and the bucket reused.

XTP also introduced the "slotting and damping" techniques. Slotting consists in forcing receivers to multicast their repair requests in order to avoid NACK implosion. Receivers also apply damping, which means that they use timers to delay their NACKs. If in the meantime they receive a request for the same packet from an other receiver, they cancel their NACK. This scheme is now used in many other protocols.

2.3.2 The SRM Approach

SRM (Scalable Reliable Multicast) [25] is a NACK-based multicast protocol that guarantees out-of-order reliable delivery. The basic idea in SRM is that a lost packet need not be retransmitted by the original sender, but rather by the nearest receiver who has the segment. SRM uses the damping and slotting techniques of XTP, reducing the response time by estimating the delay from senders to hosts. Closer hosts will chose a smaller randomization interval than distant hosts, both for NACK and retransmission timers. If the timers cannot be estimated with reasonable accuracy, many of the apparent benefits of SRM may not be realized in practice. SRM is used in a well known MBONE application, the wb (section 9.4.

2.3.3 The Asynchronous Layered Coding (ALC) Approach

Asynchronous Layered Coding (ALC) [39] is a scalable reliable multicast protocol using several data streams in a layered fashion. These streams will often be sent to different multicast addresses. Therefore the number of streams delivered to each receiver is dictated by the local bandwidth availability and network conditions. In order to provide a good reliability level, ALC relies on a FEC codec. No feedback mechanism is used so as to promote a good scalability.

According to the application, several delivery modes are considered:

On demand mode: Receivers may join the ongoing object transmission session at their discretion, obtain the necessary encoding symbols to reproduce the object, and then leave the session. A typical example is a tool for the continuous transmission of popular files (e.g. a video-clip).

Streaming mode: Typically, receivers join and remain joined to a particular set of multicast groups to receive multiple related objects in consecutive object transmission sessions. A typical example is a streaming real-time MPEG video.

Push mode: Typically, receivers join a particular set of multicast groups to collect ALC packets for multiple unrelated objects they have interest in only. A separate multicast group may be devoted to the transmission of schedule information of object transmissions.

2.4 Other Approaches

As most of the reliable multicast protocols evolved out of the necessity for solving specific problems, their basic design criteria were different. We will present in the following some of the approaches.

2.4.1 Cycle Based Protocols

Cycle-based protocols divide a file into a sequence of fixed-size packets and transmit the entire file to all the receivers. After the transmission is terminated, the protocol gets into a series of retransmission cycles, receivers sending a list of missing packets. Senders retransmit lost packets until everything is received correctly. Starbust MFTP , RMTP* and MDP are examples of protocols that use this approach.

2.4.2 Group Communication Protocols

Group-communication protocols provide different ordering and delivery semantics to the application. In Reliable Broadcast Protocol (RBP) a token site broadcasts ACKs to the entire group, and answers the NACKs of other members by retransmitting the missing packets. The responsibility of the token-site rotates among group members. The Multicast Transport Protocol Version 2 (MTP2) uses the notion of a master node, responsible for message ordering and rate control. A sender has to obtain a token from the master in order to transmit. The Uniform Reliable Group Communication (URGC) protocol uses a co-ordinator for ordering, but the responsibility of the co-ordinator rotates among all the group members.

Finally, a special approach is used by the Reliable Adaptive Multicast Protocol (RAMP) that provides different modes of reliability to different receivers depending on their requirements. Senders and receivers can switch between reliable and unreliable modes.

2.4.3 Building ACK Aggregation Trees for Improved Scalability

Tree-based protocols group the receivers in local domains or subgroups, each with a representative (Designated Router or Group Controller). Local regions are organized in a tree hierarchy. NACKs are sent to the local representatives, which retransmit the missing packets, or if they do not have them, the request is sent to an upper level in the tree. Protocols such as RMTP TMTP LBRM LGMP or LORAX fall in this category.

To be simple to deploy, protocols that use ACK aggregation must be self-organizing, the receivers must be able to form the tree themselves using the local information in a scalable manner. Such mechanisms are possible, but are not trivial. The main scaling limitations come therefore from the tree formation and maintenance rather than from the use of ACKs.

3 Adding Congestion Control to Multicast Transmissions

3.1 Introduction

Multicast communication is by definition more greedy in bandwidth than unicast communications within the same number of receivers. The design of a multicast congestion control algorithm is then an important and useful task. There are two potential approaches for congestion control: within the network (it involves routers and distribution trees rather than simple paths as in the unicast case), and end-to-end.

A key unsolved problem for congestion control schemes that operate within the network for multicast traffic is how to retain the capability for heterogeneity. Other problems include defining fairness, the time scale for congestion control, scalability.

3.2 The Various Models

There are two main models:

- *sender oriented, and*
- *receiver oriented*

3.2.1 Sender Oriented Congestion Control

The sender is in charge of analyzing the congestion information from receivers. Some multicast applications require reliability, but the TCP-like congestion control approach (i.e. receiver loss detection, feedback, source rate adapting) is not obvious since it could raise scalability and aggregation problems. The implosion of these messages in case of large groups is avoided by combining probabilistic query / reply schemes, random delay responses and expanding scope search. The TCP-like approach has been proposed by [60], [42] [46] [8] [72]. [60] proposes a loss rate threshold at the sender to decide whether a receiver is congested and counts the number of congested receivers till a second threshold above which it reduces the sending rate; when the sending rate decreases after one congested receiver, this is equivalent to adapting to the slowest receiver. Multiple multicast groups based on data loss correlation are used by [42] and [68]. The former provides local recovery and loss feedback suppression. The latter uses smack messages (selective multicast acknowledgment) sent by receivers to compute loss statistics at the sender side. These proposals apply the TCP congestion control approach using the reliability support to infer the congestion information after this information has been scattered among receivers. Improvements concerning the source implosion and scalability have been done by reacting only to the most congested paths and ignoring other loss information [46] [8]. [72] achieves bounded fairness reacting randomly to receivers congestion signals when losses occur. Multicast transport protocols as RMP [73] and RMTP (Bell Labs) [37] are sender oriented and concentrate on organizing an explicit topology to provide repairs. These algorithms could eventually be used for congestion and flow control. PGM (Cisco) [19] proposes experimental congestion avoidance strategies based on received NAKs at the source host.

ECAM (Explicit Congestion Avoidance Mechanism) [16, 17] is a congestion avoidance mechanism that addresses unicast (TCP or UDP) as well as multicast (UDP) best effort flows. It originally combines Explicit Congestion Notification based on router active queue management to detect congestion, with ICMP source quench messages to inform the involved sources of the network congestion state. The fairness achieved is a trade-off between TCP-like and Max-Min fairness.

3.2.2 Receiver Oriented Congestion Control

This class includes those proposals using hierarchical coding for continuous streams as [66] and exponential layering $bandwidth\_subscribing\_N\_layers = bandwidth\_first\_layer * 2^{(N-1)}$ for continuous streams and bulk data [67]. A receiver belonging to more/less layer groups of the same session receives more/less data or at a faster/slower rate. The receivers compute the rate using packet loss rate and RTT with [24] [49] equation:

$$Bandwidth = 1.3 * MTU/(RTT * sqrt(Loss_Freq)) \qquad (1)$$

This rate is compared with the current rate at which they are receiving the data and the receivers will join/leave layers in order to efficiently adapt to this TCP-like computed bandwidth. [67] mimics the TCP behavior by special exponential layering (the receiver leaving a layer halves the bandwidth - as TCP does) and by use of synchronization points for receivers in order to react in a coordinated way. [35] uses a small number of video streams with the same data at different rates and combines both approaches: (1) receivers feedback their rate requirements to the source which adapts within defined limits the corresponding groups rate and (2) receivers may move among groups using for this decision packet loss rates. These congestion control approaches are more appropriate within multicast communications, especially for large groups.

3.3 Congestion Control for Layered Schemes (RLM, RLC)

These are receiver oriented mechanisms. RLM [45] assumes that the data stream can be divided into multiple layers of differing quality, and that receivers can subscribe to different distribution groups for the different layers. On detecting loss (through gaps in the send sequence number space), receivers dynamically adjust the number of groups they have subscribed to, using typical ”additive increase, multiplicative decrease” control algorithm.

4 Other Associated Services

4.1 Introduction

We have seen in the previous two sections the reliability and congestion control aspects. If they are the main two requirements of multicast enabled applications, additional services will often be useful. This section introduces some of them (not exhaustive list).

4.2 Heterogeneity Support

Addressing the problem of multicast transmissions to a set of heterogeneous sources and receivers is of the utmost importance. Indeed a multicast group may be composed of high-end receivers (workstations) attached to a high performance network (high-speed LAN), whereas other ones have limited processing power (PDA) or network access (low-speed modem, GSM modem). Three solutions are possible to address this heterogeneity:

- router-based filtering techniques
- packet scheduling techniques
- transcoding techniques

The first class of solutions relies on *filtering* [15] [41]. These techniques discard packets from flows whose bandwidth exceed their fair share. The TUF proposal [15] adds a tag to each packet sent to identify the "priority level" (or drop precedence). Adding information within the packet enables the sender to keep some control on the discard process. Therefore it is preferable to blind discard techniques like RED, RIO, etc. Besides no state is kept in routers which warrants a good scalability. [41] also advocates the use of filtering within routers but without any involvement by the source and the receivers. Here per-flow state must be kept in each router. Although promising, both approaches require extensions to multicast routers (or at least some key routers) for congestion reports ([41] only) and filtering. Therefore they cannot be deployed immediately and other solutions must be used in the meantime.

The second class of solutions relies on *packet scheduling within several transmission layers* [54]. Each layer can be associated to a different multicast group and the filtering is performed by the join/leave mechanism of multicast IP [45]. This approach is thus receiver oriented and has the advantage of being immediately deployed[2].

Finally, *transcoding* within some routers or proxy to dynamically adapt the data flow to the receiver features (e.g. to reduce an image complexity) can also be used . But these solutions are CPU intensive, can only be used with a limited number of data flows and require the availability of adequate routers or proxies.

Mechanisms that provide heterogeneous Quality of Service (QoS) support, like the Resource Reservation Protocol (RSVP)[75] may be used to provide heterogeneous resource reservations in QoS-enabled networks such as ATM. RSVP Switching [26, 56] combines RSVP and label switching technologies to provide heterogeneous QoS to multicast groups.

4.3 Group Management

A group is set of entities that respect at a certain moment a non ambiguous membership rule. We can define it also as a unique virtual entity, having a name and an address. This address usually points to a table containing the list of participants. A group can be static or dynamic, depending on whether its composition can change or not in time. The group is open if an entity can join without any control, and closed if not. The logical address of a group is in fact an index that will be translated, possibly in a distributed manner, into a list of individual addresses. For group members, being attached to a logical address means that the routing protocol will be charged of forwarding all data sent to that address

[2] Using a single multicast group and relying on router filtering (when available) is also possible with a cumulative scheduling. In that case each layer must be regarded as a "priority level" and the packets be marked accordingly.

towards all the members. One should mark the difference between the group in a social sense (those who want to communicate) and the group in a network sense (the abstraction of the communication vector that enables them to communicate). These two senses are closely related even if some think that they should be treated on different levels.

Multicast group management services deal with activities relating to group membership and group dynamics. They should provide the following functionalities [43]:

- creation and deletion of groups with specific characteristics,
- membership administration (joining and leaving) according to the authentication policy specified at group creation time,
- queries on group membership and characteristics,
- group event notification (i.e. informing interested parties of changes in group state),
- group property management, allowing changes in membership policies, group visibility or members' roles,
- floor and integrity conditions control.

Integrity Conditions

Integrity conditions are a set of rules on the composition and topology of the group that have to be satisfied before the creation and during the life of a group association. They are of two kinds:

AGI (Active Group Integrity): conditions on active group membership

ATI (Association Topology Integrity): conditions on the topology of the association

Integrity conditions can be classified as follows:

- *minimal conditions*: a minimal number of participants is required for group communication to take place
- *maximal conditions*: an upper bound is fixed on the number of participants
- *quorum conditions*: a given percentage of participants is required, for voting procedures for example
- *conditions on mandatory participants*: there are some special participants (i.e. the chairman of a video-conference) whose presence is mandatory for the communication
- *atomic conditions*: fixing the exact number and identity of members who should participate

If the integrity conditions are not satisfied, the group association can be released (hard AGI/ATI) or temporarily suspended until they will be satisfied again (soft AGI/ATI).

4.4 Multicast Address Allocation

The Multicast Address Allocation Architecture [64] elaborated by IETF's MALLOC Working Group is three layered, comprising a host-server protocol (MADCAP), an intra-domain protocol (Multicast AAP) and an inter-domain protocol (MASC).

The *Multicast Address Dynamic Client Allocation Protocol (MADCAP)* allows hosts to request multicast address allocation services from Multicast Address Allocation Servers (MAAS).

Multicast Address Allocation Protocol (Multicast AAP) is used by MAAS servers to co-ordinate allocations within a domain in order to ensure that they do not collide.

Multicast Address Set Claim (MASC) forms the top level of the hierarchical address allocation architecture. Routers use this protocol to claim address sets that satisfy the needs of MAAS servers within their allocation domain. Child domains listen to multicast address ranges acquired by their parents and select sub-ranges that will be used for their proper needs. When a MASC router discovers that there are not enough multicast address available, it claims a larger address set.

4.5 Administrative Scoping

Because the TTL (Time To Live) header field has limitations, an administrative scoping has been proposed . The MZAP (Multicast-Scope Zone Announcement Protocol) enables the definition of (hierarchical) multicast scope zones of any size to better control the dissemination of multicast traffic, which is another way to improve IP multicast scalability.

4.6 Session Announcement

Three session protocols are used for multimedia session management:

SDP (Session Description Protocol): RFC2327 describes multimedia sessions and gives scheduling information

SAP (Session Announcement Protocol): periodically multicasts to a well known address announcement packets containing the SDP description of the session

SIP (Session Initiation Protocol): RFC2543 is used to invite users to a session. It does not allocate multicast addresses, this is done by SAP

They are the underlying protocols used by the `sdr` session management tool (section 9.4).

4.7 The Building Block Approach

Since 1999 an activity has begun within the “Reliable Multicast Transport” group of the IETF in order to identify and standardize various “building blocks” [74]. These building blocks are in fact the components that are common to many different multicast protocols. Examples are FEC codecs, Congestion Control, Generic Router Support, etc. A protocol instantiation is thus composed of one or more building blocks linked together with highly protocol specific, tightly intertwined functions.

The following building-blocks are under progress (March 2000):

- Design Space/Building Blocks draft progress:

  ```
  draft-ietf-rmt-design-space-01.txt
  draft-ietf-rmt-buildingblocks-02.txt
  ```

- FEC Building Block (section 2.2.2):

  ```
  draft-ietf-rmt-bb-fec-00.txt
  ```

- ALC Protocol Instantiation (section 2.3.3):

  ```
  draft-ietf-rmt-pi-alc-00.txt
  ```

- Track Architecture:

  ```
  draft-ietf-rmt-track-arch-xx.txt
  ```

- Tree Building (section 7.5): no draft yet...

- NACK-Oriented Reliable multicast Building Block (section 2.2.1):

  ```
  draft-ietf-rmt-norm-bb-00.tx
  ```

5 New Multicast Routing Protocols and their Large Scale Deployment

5.1 Introduction

DVMRP, used in the first world-wide MBONE initiative, has too many limitations to be considered a viable solution. Therefore new multicast routing protocols appeared during the past few years: MOSPF, PIM-DM, PIM-SM, MSDP, MBGP, BGMP, etc. The trend is to create a *hierarchical multicast routing infrastructure*, as with unicast routing, with domains connected by inter-domain routing protocols. Some of the above protocols are limited to intra-domain multicast communications (MOSPF, PIM-DM, PIM-SM), while others are for inter-domain multicast (MSDP, MBGP, BGMP).

5.2 The MOSPF Routing Protocol (Multicast Open Shortest Path First)

Multicast Open Shortest Path First (MOSPF) is a multicast extension to the OSPF unicast link-state routing protocol. In OSPF each router periodically sends link-states to all other routers in the networks, so each router builds up a complete network map. With this information each router is able to compute the shortest-path to every destination in the network using the Dijkstra's algorithm. MOSPF extends these link-states to also carry information about group membership. Each router advertises the presence of multicast group receivers attached to it. Therefore MOSPF can construct a shortest-path multicast tree, i.e., the path from the source to each receiver in the multicast tree is the unicast shortest-path. Instead of relying on flooding/pruning information like DVMRP, in MOSPF each router keeps a database with the group members at all routers in the network. Yet a limitation is that this feature avoids MOSPF from scaling to large networks.

5.3 CBT Multicast Routing

Core Based Trees (CBT) is the early multicast routing protocol relying on center-based trees. CBT trees are bidirectional and shared by all the sources of the same group. It means that routers store per-group information instead of per-(source,group) information as in DVMRP and MOSPF. When a member wants to join the group, it sends a join message for the group towards the core router. This message instantiates forwarding state in the way to the core router, constructing the multicast tree. When a sender sends data to the group, the packet reaches a first on-tree router that then replicates this packet on all the on-tree interfaces except the one the packet came from. The good core placement is a difficult problem. Without it, multicast trees can be quite inefficient.

5.4 The PIM-DM/SM routing Protocols

Protocol Independent Multicast (PIM) consists of two multicast routing protocols:

Dense-Mode (DM) PIM: PIM-DM [18] is intended for intra-domain routing and assumes that group members are densely distributed in the network. PIM-DM is functionally very similar to DVMRP (i.e. relies on a RPF with flood-and-prune algorithm), but it differs on some details and on the fact that PIM assumes no specific unicast routing protocol.

Sparse-Mode (SM) PIM: PIM-SM [23] was firstly intended to wide-area multicast routing. It assumes that group members are sparsely distributed in the network, in which case source-based trees (i.e. trees built by dense-mode protocols) turn inadequate. PIM-SM constructs shared trees similarly to CBT, the difference being that PIM-SM trees are unidirectional. Rather than "core", PIM-SM uses the notion of "rendez-vous point" (RP). To each group is associated one (or more) RP. New group receivers send "join" messages to the RP. Like CBT, each intermediate router takes advantage of these "join" messages to update the distribution tree. Data issued by a source is first encapsulated in unicast and sent to the RP before being distributed in the multicast tree (the tree is unidirectional). For those very active sources, and because going through a RP before joining a receiver is moderately efficient, PIM-SM enables the creation of a RPF tree rooted at this source. It is thus less dependent on the center location than CBT.

5.5 Intra Versus Inter-Domain Routing: MSDP/MBGP

The operation of PIM-SM is difficult in the inter-domain level because routers are not all multicast capable. Since PIM-SM relies on the unicast routing protocol to construct multicast trees (assuming that the *reverse* unicast path is good to *forward* multicast traffic), join messages may reach non-multicast routers complicating PIM's operation. The use of PIM-SM in the inter-domain level still has two problems: designing a scalable mechanism for mapping multicast groups to RPs and the fact that ISPs do not desire to depend on other ISP's facilities - the RP location in other ISP will not be acceptable in many cases.

The near-term solution to these problems resides in the use of the Multiprotocol Extensions for BGP-4 (MBGP [7]), PIM-SM, and the Multicast Source Discovery Protocol (MSDP [20]). MBGP allows multiple routing tables to be maintained for different protocols. This way, routers may construct one routing table with unicast-capable routes and another with multicast-capable routes. PIM can then send join messages detouring non-multicast routes. MSDP provides a solution to the ISP interdependence problem. ISPs run PIM-SM within their own domain with their own set of RPs. RPs within one domain are interconnected and connected to RPs in other domains using MSDP to form a loose mesh. MSDP sets up a group-shared tree within each domain. When a source in a specific domain starts sending, the RP in this domain sends a *Source Active* message to RPs in other domains. Joining members in other domains send source-specific join messages following the MBGP routes in the inter-domain level. This solutions solves PIM-SM's problems only in the near-term because every RP in every domain must be told about every source, so MSDP does not scale with the number of senders.

5.6 Intra Versus Inter-Domain Routing: BGMP

The Border Gateway Multicast Protocol (BGMP [62]) is another solution proposed to inter-domain multicast routing. BGMP builds shared trees for active multicast groups and allows receiver domains to build source-specific inter-domain branches where needed. The default behavior of BGMP is to have shared trees in the inter-domain level because it assumes that intra-domain connectivity is richer than inter-domain, so inter-domain shared trees are likely to be efficient. Multicast trees are bidirectional to minimize third-party dependence.

5.7 The Internet2 Multicast Initiative

The MBONE (section 1.6.2) was the first experimental Internet-wide multicast deployment. The experience gained with the MBONE led to reconsidering the deployment of multicast services within the new Internet2 infrastructure. The guidelines requires to use sparse-mode protocols and native multicast. Therefore *no tunnel is allowed and all routers must support MSDP/MBGP for inter-domain multicast routing.*

Internet2 is in fact composed of two high-speed backbones: vBNS and Abilene. The vBNS initiative started in 1995 and supports inter-domain multicast since mid-1999. The Abilene backbone is newer and only recently (early 1999) become operational. The multicast service is therefore not as advanced as in vBNS. A link is provided to other multicast backbones (e.g. TEN-155).

More information can be found at URL: `http://www.internet2.edu/multicast/`

5.8 Multicast Deployment in Europe

A native multicast service has been deployed in Europe using the TEN-155 pan-European research network. IP Multicast routing/forwarding within the TEN-155 backbone and towards the European National Research Networks (NRNs) connected to it, make use of the PIM-SM/MBGP/MSDP protocols. It now completely replaces the DVRMP cloud of the previous MBONE.

More information can be found at URL: `http://www.dante.net/mbone/`

6 The Scalability Aspects of Multicast Routing

6.1 Some Costs Associated to a Multicast Group

Using a multicast group has a cost which is often underestimated. Of course the data packets sent to a group with no receiver associated (which can happen, especially with layered transmission schemes) will be dropped by the first-hop multicast router. But there are additional costs [55]:

- A dense mode protocol like DVMRP [69] or PIM-DM [18] relies on the periodical flooding of data as far as made possible by the RTT within the domain. It greatly increases the network traffic and the router load even on branches with no receiver, both during the flooding (downstream data packets) and pruning (upward control packets) stages.
- With dense-mode protocols, all the routers, even if they do not belong to the multicast distribution tree, keep information for this group [4]. For instance, a router running PIM-DM keeps a context with the source and group addresses, the state and a timer for each group in "prune state" (i.e. without any underneath receiver) [18]. Section 6.1 gives more details in case of DVMRP/`mrouted`.
- The MOSPF protocol [47] requires that each router of the domain keeps and exchanges state for each multicast group. This information is then used by multicast routers to create the local view of the distribution tree. An unused group will also require such a state.
- Using a sparse-mode protocol like PIM-SM [23] requires using MSDP (Multicast Source Discovery Protocol) [20] to inform remote domains of the presence of a source in the local domain. Whenever a new source becomes active, the local MSDP peer announces its presence (`SA` message) to all directly connected MSDP peers, who then forward the message to other peers. This information is periodically refreshed and may also be cached [4]. This Internet-wide periodic flooding (and state) takes place even if the group has no member!
- Finally a multicast address is reserved while class D addresses are a scarce resource with IPv4.

Therefore, minimizing the number of multicast groups used is all the more important as multicast applications are to become popular.

An Example: Cost of the Forwarding State Kept by DVMRP/mrouted

An analysis of the `mrouted` (version 3.8) distribution implementing DVMRP shows that a group leads to the allocation of (more than) 100 bytes of state information, no matter whether there are receivers beneath this router or not [55].

The multiplication of multi-layer multicast sessions (and more generally of multicast applications) and the common *wrong* idea that an unused multicast group has little associated cost may well make the situation become quickly serious.

Even if we only considered the DVMRP protocol here, per-group state is required with other protocols as well (section 6.1). In addition to the memory costs analyzed here, there is the cost of the flood-and-prune activity typical to dense-mode protocols.

6.2 Forwarding State Aggregation in Multicast Routers

One solution to improve IP multicast scalability is the aggregation of this multicast forwarding state. [63] promotes a novel data structure to better aggregate ranges of multicast addresses within a router's forwarding table. The authors also note that an appropriate address allocation scheme would greatly help improving this aggregability.

[65] introduces the idea of dynamically established tunnels to replace unbranched links when a sparse mode protocol is used.

6.3 Using the Exact Number of Layers

Minimizing the number of multicast groups used is another solution. This is the goal of the ODL (On Demand Layer addition) protocol [55] that is well suited to layered transmission schemes. Its goal is to enable the source to use only the layers that are actually required by the current pool of receivers.

ODL relies on both the source and the receiver sides: it is the *receiver*'s responsibility to ask for additional layers when appropriate (e.g. if his congestion control module says so) and to respond to `QUERY` messages. It is up to the *source* to create additional layers when asked so by a receiver. The source must also check periodically by sending `QUERY` messages that each multicast group is used, and if no receiver answers within a given delay, then this layer is dropped.

7 Future Trends in Multicast Routing

7.1 Introduction

As we saw before, multicast routing protocols for wide area networks have limitations. Recently, several new proposals appeared that question the traditional multicast approach. Some of them propose a simpler service, others remove the need for inter-domain multicast routing, or introduce QoS aspects in the routing decision.

7.2 Simple Multicast

The Simple Multicast proposal [6] tries to reduce or eliminate some of the complexity and overhead of traditional IP Multicast approaches. The basic idea is identifying the group by the pair (C,M), where C is the core router, while M is the multicast address. By routing on the destination and source address, there can be 2^{32} addresses per route/core/source, thus solving the address allocation problem.

Simple Multicast is scalable to the global Internet, this scalability being achieved by using a trivial multicast address allocation scheme, un-coupling core selection and discovery from the multicast protocol and using bi-directional trees. The inter-domain routing problem is solved by carrying the core IP address in the join message. Unicast forwarding tables can thus be used to deliver the join.

7.3 Express Multicast

Express (EXPlicitely REquested Single Source multicast) [30] is a recently proposed single-source protocol extending IP Multicast to support the channel model. A multicast channel is a datagram delivery service identified by a tuple (S,E), where S is the sender's source address and E is a channel destination address. Only the source host S may send to (S,E). When joining a channel, a new member receives only the packets sent by S to E. Two channels (S,E) and (S',E) are unrelated, despite the common destination address. Express is implemented using ECMP (Express Count Management Protocol), a management protocol that maintains the distribution tree and supports source-directed counting and voting.

The Express protocol is specifically designed for subscriber-based systems that use logical channels. Even if elaborated for single-source applications such as Internet TV or file distribution, multiple-source systems can also be built on top of it by using multiple channels (one per source) or by allowing several sources to share a channel, using higher level relaying through the channel's source host. Several references (section 7.4) explain how to implement the single-source model of EXPRESS using current protocols.

7.4 Using Single Source Variants of PIM-SM

The current trend is to use a simplified multicast service. The Channel model of Express has largely motivated this trend. The expected benefits are:

- a simplified addressing model (in particular that avoids cross-delivery of traffic when different sources send to the same multicast address),
- a simplified routing architecture.

The Single Source Multicast (SSM)[29] approach is essentially based on the Express model described above. The SSM service is composed of the PIM-SSM routing protocol[9] and the IGMPv3 group management protocol[12], used between the end stations and their attached edge routers. IGMPv3 already offers a per-source filtering capability.

PIM-SSM (Protocol Independent Multicast - Source Specific Multicast) is a modified version of PIM-SM able to treat IGMPv3 requests. It considers a single source per multicast channel and uses source-based trees for multicast distribution instead of a shared tree. There is no need thus for a rendezvous point (RP) to manage the shared tree, nor for inter-domain source discovery mechanisms such as MSDP. PIM-SSM assumes that receivers know the address of the source, being thus able to initiate IGMP source-specific requests using this address. The actual source discovery mechanism is out of the scope of PIM-SSM documents. It can be done using some services such as e-mail, web announcements, or the SAP protocol.

The exclusive address range of 232/8 was allocated by IANA for the SSM service, enabling the cohabitation of source-specific service (implemented through PIM-SSM in the exclusive range) and traditional IP multicast (using PIM-SM outside the reserved address range). Thus, PIM-SM has to be modified in edge and in core routers as well. In addition, IGMPv3 will have to be implemented in all the end stations and edge routers. However, the SSM model seems to become widely accepted by the research community, being much simpler and eliminating many of the limitations of the traditional PIM-SM/MBGP/MSDP multicast distribution model.

7.5 Host-based Multicast

Host-based multicast is a new hybrid scheme to offer a group communication service. In this approach end-hosts (or well identified hosts, for instance an edge router) auto-configure themselves to create a into a multi-node distribution topology, mixing multicast and unicast routing. For instance, where multicast is the most efficient technique (e.g. in case of several hosts on the same Ethernet segment), a multicast area is kept. In other situations, when unicast routing is the most efficient technique (e.g. in case of inter-domain connection), then a unicast tunnel is created. Several motivations exist:

- it offers a total control of the distribution tree created,
- additional properties can easily be offered (e.g. to offer reliable communications over a lossy link, in case of a path through frequently congested routers, etc.)
- to include mobile nodes that are not suited to standard multicast routing,
- to include very specific areas (e.g. an "ad-hoc" network where nodes communicate without the help of any fixed infrastructure) that rely on dedicated multicast routing protocols,
- to include nodes that do not have access to a multicast distribution service.

[38] and [11] introduce a host-based approach, AMRoute, for the particular case of ad'hoc networks. Even if this may not be the most efficient scheme they introduce new ideas (mesh and trees) that may be of interest even in the Internet case.

[27] introduces a host-based approach, Yoid, for the case of (non-mobile) hosts connected to the Internet. It introduces a rendez-vous point to provide information about the session and initiate several management signaling. The architecture is rather ambitious and other services (like adding reliability) are also considered.

REUNITE describes a recursive approach to build multicast distribution tree. Unlike the AMRoute and Yoid schemes, *REUNITE is not an end-host scheme* as it consists in building a tree among routers consisting in unicast tunnels.

7.6 Multicast Routing for Mobile hosts

Multicast routing for mobile host is still a hot research topic. Several situations must be identified:

- so-called *ad'hoc networks*, where a set of mobile hosts and routers are connected by wireless links, without the help of any fixed infrastructure nor any central administration. These routers/hosts are free to move randomly and to organize themselves arbitrarily. The network's wireless topology may thus change rapidly and unpredictably. This situation is addressed by the Mobile Ad-hoc Networks (manet) IETF working group:

  `http://www.ietf.org/html.charters/manet-charter.html`

- (wireless or wired) *mobile hosts*. This (more or less important) mobility can be hidden by the "mobile IP" protocol. More information can be found in the associated mobileIP IETF working group:

  `http://www.ietf.org/html.charters/mobileip-charter.html`

Doing multicast transmissions is a challenge in these two situations. The protocols required to support this mobility are totally different to that used in case of fixed hosts and routers.

Several proposals have been made in case of ad'hoc networks: AMRoute , ODMRP , and AMRIS . The most efficient ones are mesh-based and are more or less based on a flooding scheme [34].

The case of mobile hosts is completely different as the goal is to enable a seamless and transparent connection to the standard multicast infrastructure. Several inefficiencies exist with the classic triangular routing scheme of mobile-IP (traffic sent by a remote host is first captured by the mobile's home agent, and then tunneled to it's current location). With tunneling, the routing path may be far from optimal as all the traffic goes through the mobile's home agent. Besides, when several mobile hosts having subscribed to the same multicast group are visiting the same foreign network, a copy of each packet is sent to this foreign network! To improve the situation [36] introduces a new "multicast home agent" (MHA). This MHA remains the same as long as the mobile roams in the MHA's service range. If out of range, the MHA is moved to a location closer to the mobile's current position. This is a good solution to find a balance between efficiency and multicast tree updates.

7.7 QoS-based Multicast Routing

7.7.1 Minimum-cost delay-constrained multicast trees

Kompella et al. [33] treat the problem of constructing minimal-cost spanning trees constrained by delay. They propose an algorithm (KPP) that finds a minimal-cost tree with a bounded delay from the source to each destination. The algorithm is intended for source routing because it considers the complete knowledge of the network topology.

Salama et al. [59] evaluated several multicast routing algorithms that are both centralized and static. The routing objective is to construct trees with bounded maximum delay, while minimum cost to provide good network utilization.

7.7.2 Alternate path routing

MORF is a multicast *setup* protocol intended to be used in conjunction with the multicast routing protocol. The basic idea of MORF is to provide alternate paths for applications, in the case where the opportunistic default route is not able to support the QoS requirements; and to allow the possibility of route pinning to avoid service disruption when the opportunistic route changes. The work is extended in [76], the alternate path installation protocol functioning is explained and an alternate path computation heuristic is presented.

7.7.3 Routing with resource reservations

Rajagopalan and Nair [51] analyze the problem of multicast routing with resource reservation. In this work multicast trees are dynamic, unlike the majority of previous proposals on multicast tree construction based on heuristics. Route computations are made by receivers, in a decentralized fashion. The problem is that decentralized routing may lead to inefficient resource allocation. Three algorithms are proposed, based on a modified version of the dynamic MCP (Minimum-Cost Path) heuristic. In this Modified Dynamic Heuristic (MDH) routers keep track of the bandwidth allocated for each flow in each interface. The link cost reflects the amount of additional bandwidth to be allocated to the new flow. A delay policy may be added to avoid the addition of undesirable links, e.g. a satellite link.

Cavendish et al. [13] also consider the problem of routing multicast connections with bandwidth reservation. The difference in their approach is that their algorithm, BCST (Bandwidth Constrained Steiner Tree), constructs shared-trees instead of one tree per source. The algorithm also tries to minimize the tree cost. It considers a link-state unicast routing protocol so that nodes have a complete map of the network topology. Route computation is centralized. The work is extended in [14], where the maintenance of the multicast trees is analyzed. Group dynamics is added, and the impact on tree cost and feasibility is studied. The main problem is that when a member is added or pruned from a tree, the resulting tree may not be the minimum cost tree for the new configuration. It may also not be feasible.

7.7.4 Internet-related proposals

Hou et al. [31] propose some QoS extensions to the CBT protocol. The work concentrates on enhancements to the join/leave message processing on routers, aiming at the guarantee of different QoS requirements, one at a time. Each router keeps a database regarding the QoS level of its downstream members. When it receives a join message it checks if the QoS level required for the new member can be provided, and if the QoS level of the other members can still be maintained (in the case of a source member joining the tree). The join message has to travel to the core of the tree before the join to succeed (because routers keep QoS information about downstream members only). The delay and loss requirement of members may be heterogeneous, but the bandwidth requirement is homogeneous. The QoS requirements are not considered in the routing, they are just guaranteed on the tree construction and maintenance by restricting the success of join messages.

Rajagopalan et al. [10] propose a QoS routing framework that is mainly intended to work with PIM-SM. The hypotheses are that receiver requirements are heterogeneous and that a QoS unicast routing scheme exists that provides link and nodal resource availability information. Two decentralized schemes for incremental tree computation are proposed, TIQM (Tree Information Based QoS Multicast) and NUQM (Naive Unicast Based QoS Multicast). Both schemes are based on a heuristic algorithm to compute routes constrained on three QoS requirements: the bandwidth asked by the receiver, the delay from the source to the receiver and the packet loss probability. The amount of link-state information maintained by the routers is the main difference between TIQM and NUQM.

7.7.5 QoSMIC

QoSMIC (Quality of Service sensitive Multicast Internet protoCol) constructs shared trees and source specific trees for receivers that have stringent QoS requirements. QoSMIC uses a Manager router for a group, which administers the group and facilitate the joining of new members. The joining of a new router includes a local search and a multicast tree search procedures. The local search consists of searching the joining-router neighbors (normally within N hops), so as to candidates (routers on the tree) to send advertisement messages to the new router. The multicast tree search is conducted by the Manager router. The new router contacts the Manager which informs the on-tree routers of the new member. Some routers are selected as candidates and send advertisement messages to the new router. The algorithm for choosing between several candidates is not detailed, the work concentrating on the performance analysis of the several candidate-selection schemes.

7.7.6 Application-specific routing protocols

Rouskas and Baldine [57] treat the problem of constructing multicast trees constrained on the end-to-end delay but also on the delay variation between receivers. It is for the best of our knowledge the only routing algorithm that takes into account the delay variation between the receivers of the multicast group. The scheme is sender-oriented, i.e., the sender is responsible for route decisions. First, a feasible multicast tree is constructed by the DVMA (Delay Variation Multicast Algorithm) algorithm, and the multicast connection is setup. DVMA consists of firstly constructing a shortest-path tree. This tree is checked regarding the maximum delay and delay variation constraints. If it respects the maximum delay but not the delay variation, then the tree is reconstructed, beginning from the receiver with maximum delay.

Hou and Wang [71] propose a multicast routing algorithm focused on the distribution of layered video over rate-based networks. The algorithm aims the optimization of network resource utilization, being constrained by the delay and bandwidth requirements of heterogeneous receivers. With the assumption that networks use rate-based schedulers for packet transmission and that the sources are constrained by a traffic model (in this case a leaky bucket), the delay experimented by a connection is related to its reserved bandwidth. The problem is then to find a path from the source to each receiver that has enough bandwidth to support the QoS requirements of the receiver. The algorithm considers that a link-state unicast routing protocol provides the status of each link in terms of available bandwidth. Each link has an associated constant delay.

8 Tools to Help and Monitor the Multicast Deployment

8.1 Introduction

Using multicast services often means debugging problems as this service turns out to be complex and therefore instable. In this section we introduce some tools that can help this task, as well as a world-wide infrastructure to evaluate the performances of various multicast connections.

8.2 Tools for Multicast Configuration Debug

The following tools are currently used to debug multicast problems [3]:

Mrinfo: shows the multicast tunnels and routes for a router/mrouted

Mtrace: traces the multicast paths between two hosts

RTPmon: displays receiver loss collected from RTCP messages

Mhealth: monitors tree topology and loss statistics

Multimon: monitors multicast traffic on a local area network

Mlisten: captures multicast group membership information

A presentation of these tools can be found in [3] and at URL: `http://www.cs.ucsb.edu/~almeroth/`.

8.3 Performance Measurement Tools

NIMI (National Internet Measurement Infrastructure)'s goal is to measure the global Internet. NIMI was designed to be scalable and dynamic. NIMI is scalable in that NIMI probes can be delegated to administration managers for configuration information and measurement coordination. It is dynamic in that the measurement tools are external to nimid as third party packages that can be added as needed.

More information can be found at URL: `http://www.ncne.nlanr.net/nimi/`

9 Multicast-enabled Applications and Libraries

9.1 Introduction

This section introduces already existing multicast-capable applications especially in the cooperative work area. It also covers some prototype/products offering advanced multicast services that enable the easy development of new applications. Many of the arising multicast applications are also multimedia applications that behave differently and have different requirements from unicast applications. Multicast is just one of the enabling technologies of multimedia applications [22].

9.2 Classification of Applications from a Data-Flow Point of View

Applications can be classified as follows:

- Does the receiving application need to receive all the data (e.g. file transfer) or is a subset of the data stream sufficient (e.g. multi-quality video streams)?
- Is the data stream known in advance (e.g. video on demand) or generated on the fly (e.g. an application sharing environment)?
- In the latter case is the data stream continuous (e.g. a CBR, Constant Bit Rate, video stream) or bursty (e.g. white board)?
- Are late arrivals possible (e.g. during a video-conference) or not (e.g. during a one-time file transfer)?

These features greatly influence the kind of transmission technique. This is especially true with multilayer schemes. For instance [39] is well suited to file transfer applications, where data is known in advance.

The case of CBR data flows is always studied through video-conference applications. Using a hierarchical data encoding provides an easy way to do packet scheduling, each transmission level corresponding to an increased image quality.

Many collaborative work applications like a white board (e.g. `wb`, `mb`, section 9.4 and 9.5) or an application sharing environment (e.g. doing X11 message multiplexing like `XTV` [2]) will generate bursty data streams on the fly. The amount of data exchanged can be rather low during a certain span of time and then, because of a participant action (e.g. he launches a new X Windows application under `XTV`'s control), the tool generates on the fly large amounts of data. DSG (Destination Set Grouping) is well suited to this situation.

9.3 RTP/RTCP for End-to-End Feedbacks

RTP (real-time transport protocol) [61] provides end-to-end network transport functions suitable for applications transmitting real-time data, such as audio, video or simulation data, over multicast or unicast network services. These services include payload type identification, sequence numbering, time stamping and delivery monitoring. Yet RTP does not guarantee quality-of-service ! This latter function, if required, must be provided by some external means (e.g. using diff-serv or int-serv architectures).

RTP comes along with a control protocol (RTCP) to allow monitoring of the data delivery in a scalable manner, and to provide minimal control and identification functionality. For instance, RTCP enable a source to be aware of the losses experienced by (a subset of) the receivers. This information can then be used to adapt the flow to the current network conditions.

Most of the current MBONE tools (see below) are using RTP/RTCP.

9.4 The MBONE Tools

The following list reports the tools commonly referred to as "MBONE tools".

LBNL Audio Conferencing Tool, VAT: The LBNL audio tool, vat, is a real-time, multi-party, multimedia application for audio conferencing over the Internet. Based on RTP.

`http://www-nrg.ee.lbl.gov/vat/`

Robust Audio Tool, RAT: Another open-source audio conferencing and streaming application. Includes many real-time error recovery mechanisms.

`http://www-mice.cs.ucl.ac.uk/multimedia/software/rat/`

Most flavors of UNIX and Windows supported.

Video Conference, VIC: The LBNL very popular video conferencing tool.

`ftp://ftp.ee.lbl.gov/conferencing/vic/`

An improved version is available at URL:

`http://www-mice.cs.ucl.ac.uk/multimedia/software/vic/`

Most flavors of UNIX and Windows supported.

INRIA Video-conferencing System, ivs: IVS is a software system to transmit audio and video data over the Internet using standard workstation. It includes PCM and ADPCM audio codecs, as well as a H.261 codec. Development stopped. Its successor is called Rendez-Vous.

`http://www-sop.inria.fr/rodeo/ivs.html`

WhiteBoard, WB: The most popular whiteboard tool. Available (in binary format only) at URL:

`http://www-nrg.ee.lbl.gov/wb/`

Network Text Editor, NTE: A whiteboard working only in text mode. Most flavors of UNIX and Windows supported.

Session Directory, SDR: A popular tool for announcing and being informed of MBONE sessions. Enables the automatic launching of MBONE tools with proper parameter configuration.

The latest release of most of the above tools is available at URL:

`http://www-mice.cs.ucl.ac.uk/multimedia/software/`

9.5 The Mash Project

The Mash research project seeks to define an comprehensive architecture for multimedia communication and collaboration over the Internet using IP multicast. It can be seen as an experimental follow-up to some of the above MBONE tools. It includes the following applications and building blocks:

Scalable, Reliable Multicast Framework (libsrm): libsrm is a reliable multicast toolkit that can be customized by applications with different reliability semantics.

MediaBoard: MediaBoard is a white-board tool that employs the Scalable Reliable Multicast (SRM) protocol machinery for data delivery.

SCUBA: Scalable, ConsesUs-based Bandwidth Allocation: SCUBA is a mechanism for real-time multimedia bandwidth sharing that exploits receiver interest. The scheme uses a distributed algorithm to establish consensus among all participants in the conference on the sharing of the session bandwidth using a voting mechanism.

Media Archival and Playback: The archive portion of the MASH project is divided into two areas: exploring formats for new data types, and developing a collaboration archive architecture.

Coordination and Control Architecture: Coordination is centered around the "Coordination Bus" that glues together sub-components of the MASH system to create new applications. A developer can mix and match components from the MASH toolkit by composing them across the Coordination Bus to build an arbitrary application.

The tools resulting from this continuing work, even if recent and rather elaborated, are not widely used compared to the MBONE ones. Mash should more be seen as a framework for research in the collaborative work and multicast transmission environment.

9.6 The Various Multicast Libraries (Free and Commercial)

Here is a list of multicast libraries and multicast file transfer tools. Some are research prototypes, others are commercial products :

MCL (MultiCast Library): Generic multicast library, very easily used and well suited to layered multicast transmissions schemes. Work under progress.

Free, source code, in C, available for various architectures.

`http://www-rp.lip6.fr/~roca/mcl/`

MDPv2 (Multicast Dissemination Protocol): Protocol framework and software toolkit for reliable multicasting of data objects.

Free, source code, in C, available for various architectures.

`http://manimac.itd.nrl.navy.mil/MDP/`

OmniCast: OmniCast (Starburst Software) is a one-to-many content distribution software for guaranteed, reliable, multicast distribution.

Commercial Product.

`http://www.starburstcom.com/index.html`

PGM for FreeBSD: L. Rizzo's PGM implementation of PGM This is a PGM Host implementation for FreeBSD. Its use requires that multicast routers support PGM features (in other words use CISCO routers everywhere `;-)`

Free, source code, in C.

`http://www.iet.unipi.it/~luigi/pgm.html.`

WhiteBarn's PGM: The WhiteBarn PGM implementation (WhiteBarn, Inc.)

Source code free for noncommercial use only.

`http://www.whitebarn.com`

RMDP (Reliable Multicast Transport Protocol): RMDP is a library providing a reliable layered multicast service for service for bulk-data transfers (e.g. a file). Based on FEC and a layered congestion control scheme.

Free, source code, in C and Java versions.

`http://www.cs.ucl.ac.uk/external/L.Vicisano/rmdp/`

RMF (Reliable Multicast Framework): RMF is a (more or less) generic framework to ease the implementation of various reliable multicast styles of protocols.

Free, source code, in C++ and Java versions.

`http://www.tascnets.com/mist/RMF/`

RMTP-II (Reliable Multicast Transport Protocol): Reliable multicast protocol. Product distributed by Talarian Corporation.

Commercial Product.

`http://www.talarian.com/rmtp-ii/`

References

[1] IEEE Project 802. *IEEE P802.1p: Supplement to MAC Bridges: Traffic Class Expediting and Dynamic Multicast Filtering.* IEEE, Incorp. in IEEE Standard 802.1D, Part 3: Media Access Control (MAC) Bridges: Revision, 1998.

[2] H. Abdel-Wahab and Mark Feit. Xtv: A framework for sharing x window clients in remote synchronous collaboration. In *IEEE TriComm '91: Communications for Distributed Applications and Systems*, April 1991.

[3] K. Almeroth. *Managing IP Multicast Traffic: A First Look at the Issues, Tools, and Challenges*, February 1999. IP Multicast Initiative White Paper.

[4] K. Almeroth. The evolution of multicast: from the mbone to inter-domain multicast to internet2 deployment. *IEEE Network, Special Issue on Multicasting*, January 2000.

[5] D. Applegate, R. Bixby, V. Chvatal, and W. Cook. On the solution of traveling salesman problems. *Documenta Mathematica - extra volume ICM - III*, pages 645–656, 1998.

[6] Ballardie, R. Perlman, C. Lee, and J. Crowcroft. Simple scalable internet multicast. technical report, University College London (UCL), April 1999.

[7] T. Bates, R. Chandra, D. Katz, and Y. Rekhter. *Multiprotocol Extensions for BGP-4*, February 1998. RFC 2283.

[8] S. Bhattacharyya, D. Towsley, and J. Kurose. The loss path multiplicity problem in multicast. In *IEEE INFOCOM'99*, March 1999.

[9] Supratik Bhattacharyya, Christophe Diot, Leonard Giuliano, Rob Rockell, John Meylor, Dave Meyer, and Greg Shepherd. *A Framework for Source-Specific IP Multicast Deployment*, July 2000. Work in Progress <draft-bhattach-pim-ssm-00.txt>.

[10] S. Biswas, R. Izmailov, and B. Rajagopalan. *A QoS-Aware Routing Framework for PIM-SM Based IP-Multicast*, June 1999. Internet-draft: draft-biswas-pim-sm-qos-00.txt.

[11] Bommaiah, A. McAuley, R. Talpade, and M. Liu. *AMRoute: Adhoc multicast routing protocol*, August 1998. work in progress; <draft-manet-amroute-00.txt>.

[12] Brad Cain, Steve Deering, Isidor Kouvelas, and Ajit Thyagarajan. *Internet Group Management Protocol, Version 3*, June 2000. Work in Progress: <draft-ietf-idmr-igmp-v3-04.txt>.

[13] D. Cavendish, A. Fei, M. Gerla, and R. Rom. On the construction of low cost multicast trees with bandwidth reservation. In *High Performance Computing and Networking*, 1998.

[14] D. Cavendish, A. Fei, M. Gerla, and R. Rom. On the maintenance of low cost multicast trees with bandwidth reservation. In *IEEE Globecom'98*, 1998.

[15] A. Clerget. A tag-based udp multicast flow control protocol. Technical Report 3728 3728, INRIA, July 1999.

[16] Anca Dracinschi and Serge Fdida. Congestion avoidance mechanism for unicast and multicast. In *European Conference on Universal Multiservice Networks, ECUMN'2000*, October 2000.

[17] Anca Dracinschi and Serge Fdida. Efficient congestion avoidance mechanism. In *The 25th Annual IEEE Conference on Local Computer Networks (LCN'00)*, November 2000.

[18] D. Estrin, D. Farinacci, A. Helmy, V. Jacobson, and L. Wei. *Protocol Independent Multicast Version 2, Dense Mode Specification*, May 1997. Work in Progress: <draft-ietf-idmr-pim-dm-spec-05.txt>.

[19] Tony Speakman et al. *PGM Reliable Transport Protocol Specification*, April 2000. Work in Progress, <draft-speakman-pgm-spec-04.txt>.

[20] D. Farinacci, Y. Rekhter, D. Meyer, P. Lothberg, H. Kilmer, and J. Hall. *Multicast Source Discovery Protocol (MSDP)*, July 2000. Internet-draft: draft-ietf-msdp-spec-06.txt.

[21] S. Fdida. *Multimedia Transport Protocol and Multicast Communication.* Kluwer Academic Publishers, W. Effelsberg, O. Spaniol, A. Danthine, and D. Ferrari (eds.), book chapter 1, 1996.

[22] S. Fdida, O. Fourmaux, and R. Onvural. Enabling multimedia networks. *Electronic Journal on Networks and Distributed Processing (RERIR/EJNDP)*, pages 28–32, March 1997.

[23] Bill Fenner, Mark Handley, Hugh Holbrook, and Isidor Kouvelas. *Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)*, July 2000. Work in Progress, <draft-ietf-pim-sm-v2-new-00.txt>.

[24] S. Floyd and V. Jacobson. Random early detection gateways for congestion avoidance. *IEEE/ACM Transactions on Networking*, 1(4):397–413, August 1993.

[25] S. Floyd, V. Jacobson, S. McCanne, C. Liu, and L. Zhang. A reliable multicast framework for light-weight sessions and application level framing. In *IEEE SIGCOMM'95*, 1995.

[26] O. Fourmaux and S. Fdida. Multicast for RSVP switching. *Telecommunication System Journal*, (11):85–104, 1999.

[27] P. Francis. Yoid: extending the multicast internet architecture. Unrefered Report; http://www.aciri.org/yoid/, September 1999.

[28] M. Handley, B. Whetten, R. Kermode, S. Floyd, L. Vicisano, and M. Luby. *The Reliable Multicast Design Space for Bulk Data Transfer*, March 2000. Work in Progress, <draft-ietf-rmt-design-space-01.txt>.

[29] H. Holbrook and Brad Cain. *Source-Specific Multicast for IP*, March 2000. Work in Progress <draft-holbrook-ssm-00.txt>.

[30] H. Holbrook and D.R. Cheriton. Ip multicast channels: Express support for large-scale single-source applications. In *ACM SIGCOMM'99*, September 1999.

[31] J.C. Hou, H-Y. Tyan, B. Wang, and Y-M. Chen. *QoS Extension to CBT*, February 1999. Internet-draft: draft-hou-cbt-qos-00.txt.

[32] C. Huitema. The case for packet level fec. In *Protocols for High Speed Networks (PfHSN'96)*, October 1996.

[33] V.P. Kompella, J.C. Pasquale, and G.C. Polyzos. Multicast routing for multimedia communication. *IEEE/ACM Transactions on Networking*, 1(3):286–292, June 1993.

[34] S-J. Lee, W. Su, J. Hsu, M. Gerla, and R. Bagrodia. A performance comparison study of ad'hoc wireless multicast protocols. In *IEEE INFOCOM'00*, March 2000.

[35] X. Li and M. Ammar. Bandwidth control for replicated-stream multicast video distribution. In *HPDC-5*, August 1996.

[36] C.R. Lin and K-M. Wang. Mobile multicast support in ip networks. In *IEEE INFOCOM'00*, March 2000.

[37] J. C. Lin and S. Paul. RMTP: A reliable multicast transport protocol. In *IEEE Infocom'96*, 1996.

[38] M. Liu, R. Talpade, and A. McAuley. Amroute: Adhoc multicast routing protocol. Technical Report TR 99-1, CSHCN, 1999.

[39] M. Luby, J. Gemmell, L. Vicisano, L. Rizzo, J. Crowcroft, and B. Lueckenhoff. *Asynchronous Layered Coding (ALC): a scalable reliable multicast protocol*, March 2000. Work in Progress: <draft-ietf-rmt-pi-alc-00.txt>.

[40] M. Luby, J. Gemmell, L. Vicisano, L. Rizzo, J. Crowcroft, and B. Lueckenhoff. *Reliable multicast transport building block: Forward Error Correction codes*, March 2000. Work in Progress: <draft-ietf-rmt-bb-fec-00.txt>.

[41] M. Luby, L. Vicisano, and T. Speakman. *Heterogeneous multicast congestion control based on router packet filtering*, June 1999. Work in Progress, presented at RMRG meeting, Pisa.

[42] D. De Lucia and K. Obraczka. *Congestion Control Mechanism for Reliable Multicast*, September 1997. presentation during Reliable Multicast (RM) meeting.

[43] A. Mauthe, D. Hutchinson, G. Coulson, and S. Namuye. From requirements to services : Group communication support for distributed multimedia systems. In *Second International Workshop, IWACA'94*, 1994.

[44] A.J. McAuley. Reliable broadband communications using a burst erasure correcting code. In *ACM SIGCOMM'90*, September 1990.

[45] S. McCanne, V. Jacobson, and M. Vetterli. Receiver-driven layered multicast. In *ACM SIGCOMM'96*, October 1996.

[46] T. Montgomery. A loss tolerant rate controller for reliable multicast. Technical Report IVV-97-011, NASA, August 1997.

[47] J. Moy. *Multicast Extensions to OSPF*, March 1994. RFC 1584.

[48] J. Nonnenmacher, E. Biersack, and D. Towsley. Parity-based loss recovery for reliable multicast transmissions. In *ACM SIGCOMM'97*, September 1997. also in IEEE Transactions on Networking, 1998.

[49] T. Ott, J. Kemperman, and M. Mathis. *The stationary behavior of ideal TCP congestion avoidance.* in preprint. http://popeye.snu.ac.kr/ schoi/TcpPerf.html.

[50] S. Paul. *Multicasting on the Internet and its Applications.* Kluwer Academic Publishers, 1998.

[51] B. Rajagopalan and R. Nair. Multicast routing with resource reservation. *Journal of High Speed Networks*, 7(2), July 1998.

[52] J. F. Rezende, A. Mauthe, S. Fdida, and D. Hutchison. Fully reliable multicast in heterogeneous environments. In *Protocols for High Speed Netowrks - PfHSN'96*, 1996.

[53] L. Rizzo and L. Vicisano. Effective erasure codes for reliable computer communication protocols. *ACM Computer Communication Review*, 27(2), April 1997.

[54] V. Roca. Packet scheduling for heterogeneous multicast transmissions. In *Protocols for High Speed Networks (PfHSN'99)*, August 1999.

[55] V. Roca. On-demand layer addition (odl): Making multi-layer multicast transmissions cheaper. technical report 0239, INRIA, February 2000.

[56] S.P. Romano, C. Deleuze, J.F. Rezende, and S. Fdida. Integrated QoS architecture for IP switching. In *Proceedings of the 3rd European Conference on Multimedia Applications, Services and Techniques*, pages 312–326, May 1998.

[57] G.N. Rouskas and I. Baldine. Multicast routing with end-to-end delay and delay variation constraints. *IEEE Journal on Selected Areas in Communications*, 15(1):1–9, April 1997.

[58] L. Sahasrabuddhe and B. Mukherjee. Multicast routing algorithms and protocols: a tutorial. *IEEE Network*, pages 90–102, January 2000.

[59] H.F. Salama, D.S. Reeves, and Y. Viniotis. Evaluation of multicast routing algorithms for real-time communication on high-speed networks. *IEEE Journal on Selected Areas in Communications*, 15(3):332–345, April 1997.

[60] T. Sano, N. Yamanouchi, T. Shiroshita, and O. Takahashi. Flow and congestion control for bulk reliable multicast. In *IEEE INFOCOM'98*, February 1998.

[61] Schulzrinne, Casner, Frederick, and Jacobson. *RTP: A Transport Protocol for Real-Time Applications*, July 2000. Work in Progress, <draft-ietf-avt-rtp-new-08.txt>.

[62] D. Thaler, Deborah Estrin, and D. Meyer. *Border Gateway Multicast Protocol (BGMP): Protocol Specification*, March 2000. Internet-draft: draft-ietf-bgmp-spec-01.txt.

[63] D. Thaler and M. Handley. On the aggregatability of multicast forwarding state. In *IEEE INFOCOM'00*, March 2000.

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

[65] J. Tian and G. Neufeld. Forwarding state reduction for sparse mode multicast communications. In *IEEE INFOCOM'98*, February 1998.

[66] T. Turletti, S.F. Parisis, and J. Bolot. Experiments with a layered transmission scheme over the internet. In *IEEE INFOCOM'98*, February 1998.

[67] L. Vicisano, L. Rizzo, and J. Crowcroft. Tcp-like congestion control for layered multicast data transfer. In *IEEE INFOCOM'98*, February 1998.

[68] Lorenzo Vicisano, Mark Handley, and Jon Crowcroft. *B-MART: Bulk-data (non-real-time) Multiparty Adaptive Reliable Transfer Protocol*, 1997. Technical Report available at: $http://research.ivv.nasa.gov/RMP/links.html$.

[69] D. Waitzman, C. Partridge, and S. Deering. *Distance Vector Multicast Routing Protocol*, November 1988. RFC 1075.

[70] B. Wang and J. Hou. Multicast routing and its qos extension: problems, algorithms, and protocols. *IEEE Network*, pages 22–36, January 2000.

[71] B. Wang and J.C. Hou. QoS-based multicast routing for distributing layered video to heterogeneous receivers in rate-based networks. In *IEEE INFOCOM'2000*, 2000.

[72] H. Wang and M Schwartz. Achieving bounded fairness for multicast and tcp traffic in the internet. In *ACM SIGCOMM'98*, September 1998.

[73] B. Whetten, T. Montgomery, and S. Kaplan. *A High Performance Totally Ordered Multicast Protocol.* Theory and Practice in Distributed Systems - K. P. Birman, F. Mattern, A. Schiper (Eds.) Springer Verlag LNCS 938, 1995.

[74] B. Whetten, L. Vicisano, R. Kermode, M. Handley, S. Floyd, and M.Luby. *Reliable Multicast Transport Building Blocks for One-to-Many Bulk-Data Transfer*, March 2000. Work in Progress, <draft-ietf-rmt-buildingblocks-02.txt>.

[75] Paul P. White. RSVP and integrated services in the internet: A tutorial. *IEEE Communications Magazine*, 35(5):100–106, May 1997.

[76] D. Zappala. Alternate path routing for multicast. In *IEEE INFOCOM'2000*, March 2000.