Survey of Multicast Routing Algorithms and Protocols

Pragyansmita Paul and S V Raghavan[1]
(**pragyan@cs.iitm.ernet.in**, **svr@cs.iitm.ernet.in**)
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Chennai 600036 INDIA

Abstract

In the age of multimedia and high-speed networks, multicast is one of the mechanisms by which the power of the Internet can be further harnessed in an efficient manner. When more than one receiver is interested in receiving a transmission from a single or a set of senders, multicast is the most efficient and viable mechanism. In the protocol stack of the network, multicast is best implemented in the network layer in the form of a multicast routing protocol to select the best path for the transmission. The other layers of the protocol stack provide additional features for multicast.

This paper deals with how multicasting is implemented in the Internet (IPv4). With emphasis on the implementation of multicast at the network layer the implementation of additional features for multicast at the other layers of the protocol stack are presented. The network layer is concerned with routing of the data in an efficient manner with minimal duplication of data to the various receivers. The features of the routing protocols that have been proposed for best effort as well as QoS-based multicast are analyzed. Some of the issues and open problems related to multicast implementation and deployment are discussed along with an overview on how multicast service is deployed in some of the existing backbone networks.

1.0 Introduction

Data communication in the Internet can be performed by any of the following mechanisms: unicast, broadcast, anycast and multicast. Unicast is point-to-point communication. Broadcast is when data is forwarded to all the hosts in the network. Anycast is when data is to be transmitted to any one of the members selected to be part of a group. Multicast is when data is to be transferred to only a group of hosts on a network. In the age of multimedia and high-speed networks, multicast is one of the viable mechanisms by which the power of the Internet can be further harnessed in an efficient manner.

Steve Deering first suggested IP multicast in his PhD dissertation in 1988. The first usage of multicast on a wide scale was during an “audiocast” at the March 1992 IETF meeting in San Diego.

There have been a number of techniques proposed to implement multicast in the Internet and intranet. This paper will be surveying the various techniques that have been proposed, their pros and cons and the suitability of a particular technique for a given multicast scenario.

Multicast is most efficiently implemented and handled at the network layer. Multicast was initially implemented as IP-encapsulated tunnels forming the Multicast backBONE

[1] Author for further correspondence

(MBONE)[2]. Multicast data is routed over the network using either the IP-encapsulated tunnels or the multicast enabled routers[3]. This paper focuses on the routing algorithms and protocols that have been proposed for multicast communications.

Additional features for multicast can be implemented at the other layers of the protocol stack such as reliability in transport layer, intranet multicast in data-link layer and session information and log maintenance in application layer. These and a few other additional features that are presently being provided by the various layers of the protocol stack will be presented in this paper.

The rest of the paper is organized as follows. Section 2 deals with the features of multicast and the various results that have been found highlighting the efficiency of multicast over multiple unicast to each and every receiver that are part of the group. Section 3 presents the multicast routing algorithms and Section 4 presents the multicast routing protocols. Routing protocols maintain the state information and use the routing algorithm to select the most appropriate route. Section 5 deals with how features such as reliability and group management are added to multicast communication at layers other than the network layer. Section 6 discusses the issues and open problems in implementation and deployment of multicast in the existing Internet. Section 7 concludes with a discussion on the state of the art in the deployment of multicast.

2.0 Multicast

When the number of sender and receiver involved in a data communication is one-to- many or many-to-many or many-to-one, multicast is used as the means of data communication. The sender(s) and receivers are assumed to be part of a group. The features of a multicast group are described below:

- A host can be a member of any number of multicast groups.
- The membership to a multicast group is dynamic, the sender(s) and receivers can join or leave the group at any time. For scalability, the join and leave operation has to be simple without any side effects.
- To be a sender of a group, it is not necessary that the host is a member of the group.
- Each group is identified by a Class D (from 224.0.0.0 to 239.255.255.255 as shown in Table 1) address in IPv4 networks.
- Data communication is done using User Datagram Protocol (UDP). This is to avoid the overhead of reliability and flow control that is associated with Transmission Control Protocol (TCP).

The multicast groups can be classified either as permanent or transient groups. The transient groups remain in existence as long as there are members in the group. However permanent groups remain in existence even when the number of members in the group is zero. Apart from this, the multicast groups can be classified either as dense or sparse groups based on the distribution of the group members in the network.

With the advent of multicasting, many applications have emerged that can derive maximum benefit from multicasting of data. Some sample applications are videoconference, real-time

[2] MBONE is a virtual network developed to run on top of the physical Internet. IP-encapsulated tunnels connect the non-multicast-capable routers and the routers communicate using the DVMRP protocol (as described in Section 4).

[3] A generic router has four components: input ports, output ports, a switching fabric and a routing processor. The routing processor participates in routing protocols and creates a forwarding table that is used in packet forwarding.

multimedia applications and Distributed Interactive Simulation (DIS). The multicast applications can be divided into the following categories:

- Single-point to multi-point e.g. Audio-Video broadcasts, Database updates, Push applications
- Multi-point to multi-point e.g. Video-conferencing, Distance Learning, Multiplayer Games
- Multi-point to single-point e.g. Resource Discovery, Data Collection, Auctions

Address Range	Uses
224.0.0.0 to 224.0.0.255	Administrative functions and system level routing chores (always sent with TTL of 1)
224.0.1.0 to 238.255.255.255	Multicast end user applications within groups, intranets and Internet
239.0.0.0 to 239.255.255.255	Locally administered or site specific multicast applications

Table 1. Class D IPv4 addresses for multicasting

Multicast Communications

The data transfer associated with a multicast group needs to be handled differently by the intermediate nodes, namely the routers involved in the routing of the multicast packets from the sender(s) to the receivers. The need to handle multicast data differently coupled with the different types of applications using multicast and their varied requirements has led to the development of various routing algorithms and protocols. An ideal multicast routing algorithm will have the following features:

- The load on network should be minimal. This essentially involves avoiding loops and avoiding traffic concentration on a link or a sub-network.
- It should support reliable transmission.
- The routing algorithm should be able to select optimal routes, taking into consideration different cost functions, including available resource, bandwidth, number of links, node connectivity, price to be paid and end-to-end delay. It should further maintain optimality of the routes after any changes occur in the group or the network.
- It should minimize the amount of state that is stored in the routers, so that more groups can be supported in a network without any scalability issues.
- The data transmitted should reach only the members of the group.

Multicast routers[4] communicate among themselves using the standard routing protocols and deliver the multicast datagram from the sender(s) to the receivers. The host which wants to send data to a multicast group transmits the datagram using the local network multicast capability. The multicast router on receiving the datagram looks up its routing table and forwards it to the appropriate outgoing interface. When a host decides to join a particular multicast group, it sends the request to the local multicast router. The local multicast router makes an entry for this group (if it does not exist already) and propagates the information to other multicast routers to establish the multicast routes. Multicast routers use Internet Group Multicast Protocol (IGMP) [28] to gather member information for the multicast groups [2].

However, all routers in the Internet are not multicast capable. Solution is to use IP-encapsulation tunnels such hosts, as used in the MBONE. The multicast router at the source end of the tunnel encapsulates the datagram and forwards it. By encapsulation, it means that the router prepends another IP header with the destination address as the unicast address of the multicast router at the other end of the IP tunnel. Intermediate routers view it as a unicast

[4] A router that supports IGMP and one or more multicast routing protocol

datagram and forward it as per the unicast routing table. Destination router removes the outer encapsulated IP header and forwards the packet as a multicast datagram.

Once the data is delivered to the end-host by the network, the end-host must deliver a copy of the message to all the processes that belong to that group. Multiple processes on a given end-host can belong to the same multicast group.

Efficiency of Multicast

Multicast provides efficient communication and transmission, optimizes performance and enables truly distributed applications. Copies of message are made only when paths diverge at a router, that is, when the message is to be transferred to another route in the path to the receiver or when a receiver is attached to the router. The optimal multicast path is computed as a tree or a group of trees. The quality of the tree is determined by low delay, low cost and light traffic concentration.

The first effort at quantifying the cost advantage in using multicast was by Chuang and Sirbu [30]. It focuses on link cost such as bandwidth quantification and ignores node cost such as routing table memory, CPU usage. Where there is a direct relationship between the number of unicast packet hops and the number of receivers, the number of multicast packet hops remains nearly equal. It does increase with the increase in membership size, but at a slower rate than unicast.

The normalized multicast tree cost, that is, the ratio of the total length of the multicast distribution tree to the average length of the unicast routing paths was found to be a power-law.

$$L_m / L_u = N^k \quad \textit{Chuang Sirbu's Law}$$

where k is a factor ranging between 0 and 1. L_m is the total length of the multicast distribution tree, which is the sum total of the edge costs of all links. L_u is the average length of the unicast routing path while N is the number of routing nodes that have hosts subscribed to the multicast group rather than the number of subscribed hosts. This considers the resource consumption in the provider's network only. The cost of serving receivers on a shared broadcast capable subnet is assumed to be zero. k was found to be 0.8 for a range of generated networks – random[5], transit-stub[6] and hierarchical topological styles.

The above formulae was further analyzed in [39] for k-ary trees and general networks which are not k-ary. The asymptotic form of the Chuang Sirbu's law for k-ary trees is as stated below:

$$L_m \approx X \left(c - \frac{\ln (X / M)}{\ln k}\right), \text{ where } X = \frac{\ln (1 - (m/M))}{\ln (1 - (1/M))}$$

where m is number of randomly chosen distinct network locations, M is total number of possible receiver locations, k is the degree of the tree. This function is not similar in form to the Chuang Sirbu Scaling Law, but it produces results quite similar to it.

[5] The Waxman model, where nodes in the network are placed at random points in a two-dimensional grid. Links are added to the network by considering all possible pairs of nodes and then deciding whether a link should exist according to a probability function. The probability function is based on how far apart the two nodes are and how many links are expected to exist in the whole network.

[6] All the nodes in this model are routers. Graphs are repeatedly checked for connectivity, which ensures that resulting sub-graph is chosen at random from all possible graphs.

It is further generalized that the above asymptotic form of Chuang Sirbu's Law holds true for any network that exhibits the exponential increase in S(r). S(r) is a reachability function that denotes the number of distinct sites that are exactly r hops from the source. S(r) has an exponential increase if the number of hops reachable as we go from r to r+1 hops is proportional to the number of sites reachable in r hops.

The nature of S(r) in real networks was evaluated in [56]. The asymptotic form of the average number of joint hops in a shortest path multicast tree from a root to m arbitrary chosen group members is derived for the random graph $G_p(N)$ and the k-ary tree. It exhibits exponential behavior, in that the number of routers in the Internet that can be reached from a root grows exponentially in the number of hops with an effective degree of approximately 3.2 [51]. Intuitively, this means that multicast provides efficient transmission of data when the receivers are spread wide apart.

The pricing model for multicast was discussed in [30]. A flat rate pricing is suggested for dense-mode multicast and for applications such as webcasting. However a pricing model based on membership size is proposed for sparse mode multicast.

The bandwidth used by a multicast transmission is not directly attributable to a single receiver. Thus, there is a need for ways to distribute the cost among the various receivers. [15] suggests a One-Pass mechanism whereby the accounting control messages make a single pass from the source down the multicast tree to all receivers. Nodes allocate costs to members as the accounting message passes through them. The information used to make the allocation decisions comes from two sources – the multicast routing and the accounting message itself. [49] investigates the cost-sharing algorithms for multicast transmissions.

Properties of Multicast Tree

[55] is a study on how the number of links in a multicast tree changes as the number of multicast users in a group change. It is shown that the stability of a tree tends to a Poisson distribution for large N. It is further shown that Steiner trees are more instable than shortest path trees.

There were two assumptions made in the above studies. The first one is that packets are delivered along the shortest path tree. This assumption ignores shared trees. However, since most of the multicast deployed in the current Internet uses packet forwarding based on reverse shortest path, it is quite realistic. Second assumption is that the multicast group members are chosen uniformly out of the total number of nodes. If m and N are large, then deviations from the uniformity assumption are negligibly small as confirmed by the measurements of [52]. Thus the second assumption is also valid in the current Internet scenario.

A study has been conducted in [52] on the impact of the shape of the multicast tree towards multicast efficiency. It was found that that all multicast trees have similar characteristics in terms of key parameters such as depth, degree frequency and average degree. This will overcome the skepticism associated with the deployment of multicast, that the shape of the multicast tree changes over time with the arrival and departure of receivers and this will affect the performance of multicast communication.

The overall efficiency of a particular tree is determined by a few key properties, such as height, breadth and number of receivers. Faloutsos et. al. propose a number of power laws related to network properties. Medina et. al. suggest that the key properties of the network such as preferential connectivity, incremental growth, geographical distribution of nodes and locality of edge connections are responsible for the presence of the skewed distributions in networks. The skewed distributions and power laws also characterize properties of the multicast tree.

The power law relating outdegree 'd' and frequency 'f_d' is $f_d \propto d^O$ (O is negative). This means that majority of nodes in the Internet have few out-going links and that only a small number of nodes have high degree. Thus, the trees produced are more likely to be tall than wide.

A metric of the form $\delta = 1 - (L_m/L_u)$ is used as a metric to estimate the percentage gain in multicast efficiency over unicast. This metric focuses on the bandwidth efficiency of data transmission. It is independent of the group distribution and behavior. The efficiency alone cannot validate a generated tree; height and breadth of the tree also matter. Branching nearer to the receiver improves efficiency; branching close to the source reduces efficiency. Multicast efficiency also increases as the average path length increases [43].

3.0 Multicast Routing Algorithms

The data transmitted needs to be transferred from the sender(s) to the receivers. The sender(s) and receivers are mostly end-hosts. Intermediate nodes are the routers, which route/direct the data from the sender(s) to the receivers. A spanning tree has been considered one of the most efficient and viable mechanisms to perform the data transmission in such a scenario, since it minimizes duplication of packets in the network. Messages are duplicated only when the tree branches and this ensures data communication is loop-free. An efficient multicast routing algorithm will aim to build a Minimal Spanning Tree (MST).

The type of tree to be used depends on whether receivers are sparsely or densely distributed throughout the network; the number of receivers does not matter. The receivers might have a set of requirements like the cost or a given amount of delay that it can tolerate in the receipt of data. Different type of trees to handle such special cases has been proposed.

3.1 Source Tree

Source tree algorithms (also known as shortest path trees) build a separate tree for each source. Reverse Shortest Paths (RSP) connects each of the receivers to the source. RSP is constructed by using Reverse Path Forwarding (RPF) at the intermediate routers, as mentioned in Section 4.1.1.0. This is efficient for high data rate sources. It provides minimal delay at the expense of cost. It exhibits lesser traffic concentration. When source tree is used, a network with a large number of groups and with each group having a large number of sources, can stress the storage capability of the routers.

Source trees consume more bandwidth for each individual multicast group. However their demands are more evenly distributed than the center based trees, especially in networks with high outdegree. Thus a network can support more high bandwidth multicast groups, if source trees are used instead of center based trees [7].

3.2 Shared Tree

Shared tree algorithm builds a single tree to be used by all the sources. The data communication in the tree can be one way or bi-directional. This is efficient for low rate sources and is efficient in the amount of state information that needs to be maintained at each router. However, it exhibits higher traffic concentration.

Shared trees use a single location in the network called the core or the Rendezvous Point (RP) to which all packets from the sources are sent and from which packets are sent to all receivers. The paths from certain receivers to the source may be longer, which may cause additional delay. This will be a disadvantage for delay-sensitive and high bandwidth applications. The core is a potential bottleneck for data transmission. CBT and PIM-SM are examples of routing protocols making use of shared trees.

The selection of the node to act as the core is critical to the performance of the routing protocol. Finding the optimal center for a group is an NP-complete problem. Thus, the core has to be selected based on certain heuristics such as the following:

- Random Router (not necessarily member of the group)
- Random Member
- Topological Center of the entire network
- Topological Center of the multicast group (and not the network) (not necessarily member of the group)
- Topological center of the multicast group (necessarily member of the group)
- Random tree node (Only nodes belonging to the current multicast tree)
- Tree Center (Only nodes belonging to the center of the current multicast tree)

The center-based heuristics outperforms the others. However, tree-center heuristic significantly outperforms heuristics based on random selection [42]. For applications with localized participations, group-based core improves the performance over topological-based core. Choosing the topological center of the subgraph induced by the receivers or just choosing a receiver at random results in bandwidth and delay performance that is better than the performance of a topology center. However, group information based on a subgraph does not improve performance much if the receivers are widely distributed [12].

Two heuristics - Minimal Member Protocol and Hill-Climbing Protocol - have been proposed for the distributed algorithms to select the core of a multicast group [19]. In the former, as members join and leave, the center migrates. Center calculates its own weight according to some predefined function and multicasts its weight along with the list of group members/sources (if necessary) to all group members. In the latter, a path list holds the list of nodes in the “path” formed by traversing toward neighbors with better weights. The probing node queries its neighbor and so on.

Wall shows that a topologically centered tree gives a delay bound of twice that of source-specific trees. It improves to three-times if the root is moved to a group member. Wei and Estrin show that total bandwidth usage of center-specific trees lie somewhere between minimal Steiner tree and source-specific trees [19].

The choice of whether to use a dynamic center-location algorithm depends on the importance of minimizing tree cost versus the time and complexity required. Since multicast permits dynamic memberships, the optimal center keeps changing. The optimal center is unlikely to move very much for groups with a relatively large number of members at steady state, however, this does not hold true for small groups.

The following performance metrics are of interest in the case of core-based trees:

- Length of the path that a packet must follow to reach the core, also known as the reach cost.
- Number of messages transmitted from the core till the last member receives the packets.
- Length of the path that a join request must follow to reach the core.
- Number of member-to-core paths that use a link, a measure of the traffic concentration.

Shared tree offers more favorable scaling characteristics than all other multicast algorithms, in terms of network state maintenance, bandwidth efficiency and protocol overhead, by a factor of the number of active sources. Routers between a non-member sender and the delivery tree do not incur any cost pertaining to multicast as the packet is encapsulated and unicast to a core on the tree.

Doar and Leslie’s Naïve algorithm for constructing route for dynamic multicast groups computes the multicast routes by combining the shortest paths of the initial multicast group members. New members are joined to the nearest attachment point on the existing tree [7].

[18] proposes a one-to-many join mechanism to build a shared tree. Inter-domain joins are generated by egress nodes to a well-known multicast address that starts a one to many spanning join. Leaf routers request intra-domain joins to an egress node. If the node receiving the join request has no state for the group, it broadcasts the message out on its other interfaces via RPF. If it has state, the message is terminated and the on-tree node responds with a unicast join-request towards the egress node that initiated the join. This join-request installs temporary state along the path towards the initiator. The paths not chosen will eventually time-out.

3.3 Steiner Tree

A minimal spanning tree is a tree that spans all the group members and minimizes the total weight of the tree. Steiner tree minimizes the total cost of a shared tree. It minimizes cost at the expense of delay. Finding such a tree in a network is a NP-Complete problem. Since it is NP-Complete in nature, it is not possible to find an exact solution for the same. A number of approximate and heuristic solutions have been proposed for the same.

Kou, Markowsky and Berman (KMB) algorithm is an approximation of Steiner trees. A distributed version of KMB was proposed by Wall. The cost of a tree generated with the KMB algorithm averages 5% more than the cost of a Steiner tree. KMB trees have higher delay for larger groups than center trees. It has higher variations in delay than center trees. When the Steiner tree consists of only group members, the KMB tee is a Steiner Minimal tree. Since KMB needs the complete network topology, it is not practical for wide area networks. Both KMB and Wall’s algorithm assumes that the group is statically configured.

There are many networks in practice where the communication links are asymmetric and cannot be modeled by undirected edges. Such problems are modeled as directed Steiner tree problems [29]. Multimedia communication can tolerate only a limited delay in the data transfer from the sender to the receiver. Delay bounded Steiner trees is a solution for the same. A tree that has minimal cost under a given delay constraint is called a delay bounded Steiner tree.

3.4 Reduced Trees

Reduced trees are proposed in [31] as a solution for scalability of multicasting. The set of vertices in a tree can be partitioned into a set of members (of degree 1), relay nodes (of degree 2) and duplicating nodes (of degree at least 3). A reduced tree is a tree that is modified such that there are no relay nodes. This leads to around 80% reduction in the amount of state information that is maintained per group.

3.5 Incremental Distributed Asynchronous Algorithm for MST

A distributed algorithm proposed for updating a MST when a new node joins the group. Recomputing the MST when changes are made to the underlying network is unnecessarily expensive when the new MST coincides with the old one. This incremental algorithm makes use of the existing structure to avoid computing from scratch. The algorithm runs asynchronously and processors at each vertex of the network is required to know only information concerning its adjacent edges. Each message exchanged contains at most an edge weight and a few bits [4].

3.6 Bounded Shortest Multicast Algorithm (BSMA)

BSMA starts by computing a least-delay tree rooted at a given source and spanning all group members. It iteratively replaces super-edges in the tree with cheaper super-edges not in the tree, while not violating the delay constraint until the total cost cannot be further reduced. Super-edge of a tree is the longest simple path whose internal nodes are relay nodes and each relay node connects exactly two tree nodes.

3.7 Bauer Algorithm

It imposes constraints on the number of outgoing links that can be used for a group. The tree construction begins with an arbitrary starting point and an edge that is closest to the partial tree is added, one at a time. The heuristic is repeatedly applied to the network graph for different starting points. It defines and monitors a damage index to the multicast tree as members join and leave, and triggers tree rearrangement when the index exceeds a certain threshold.

3.8 Delay Variation Multicast Algorithm (DVMA)

Buffering at the source, at the switching nodes and at the receiver may be used as a tool to combat delay variation. Buffering at the source and switching element would require the source and switching element respectively, to maintain additional information about all destinations. Buffering at receiver is straightforward and cancels the effect of delay variation. However, providing bounds on delay variation while routing will result in a more efficient usage of buffering resources.

A tree that is bounded by both delay and delay variation is known as a delay variation-bounded multicast tree (DVBMT). Whenever the size of the multicast tree is greater than two, DVMBT is an NP-Complete problem. DVMA builds a DVMBT spanning tree. It assumes that the complete topology is available at each of the nodes. The algorithm starts with a spanning tree satisfying the delay constraint, which may not include some members. Next the algorithm searches through the candidate paths satisfying the delay and the delay variation constraint from a non-tree member node to any one of the tree nodes. On finding such a path, it adds the members to the existing tree. The spanning tree built by DVMA satisfies the delay constraint. Further it either satisfies the variation constraint or has the smallest value of variation among the trees considered by the algorithm [24].

3.9 ARIES / GREEDY / Edge Bounded Algorithm (EBA)

A Rearrangeable Inexpensive Edge-based online Steiner Algorithm (ARIES) is a heuristic for updating multicast trees dynamically in large point-to-point networks. GREEDY and EBA are some more heuristics that have been proposed for the same purpose.

ARIES monitors the accumulated damage to the multicast tree within local regions of the tree as nodes are added/deleted and triggers a rearrangement when the number of changes exceeds a certain threshold. It joins the new member to the existing tree by its shortest path to the tree. It uses a Geographic-Spread Dynamic Multicast Heuristic to decide the node to which the new member is joined. For each add request, it identifies the tree node closest to the new member and three nearby nodes in the existing tree. The heuristic then chooses the least cost configuration among four possible ways to connect the new member to the three identified nodes. If more than one cheapest alternative exists, GSDM chooses the one with greatest geographic spread. In addition, for each delete request, it deletes the nodes only if it is a leaf [25].

The GREEDY heuristic attempts to disturb the tree as little as possible. The new member is connected to the nearest tree node using the shortest path. A delete request leads to deletion of only leaf nodes and any non-member nodes in the path. EBA heuristic enforces bounds on the distance between nodes in the tree after each change and performs rearrangements when the distance exceeds a set bound.

4.0 Multicast Routing Protocols

The routing protocols are deployed at the intermediate nodes, namely the routers that make up the path from the sender(s) to the receivers. The routing protocols have two main responsibilities: to collect and maintain state information that can be used by the routing algorithms in selecting the best path to the receivers and to select the most appropriate path among the various paths available using a path selection procedure. Other than building the distribution tree, multicast routing protocols have the additional responsibility of group management. A multicast routing algorithm together with appropriate scheduling, forwarding and policing mechanisms can provide QoS guarantees for real-time multicast applications

The routing protocols are classified into dense and sparse mode protocols. PIM is one of the routing protocols that can operate in either of the modes. Sparse mode protocols offer better scalability and efficiency. However the core can be a single point of failure that can be overcome by having multiple RPs for a group or having failover/backup RPs. The essential difference between the two modes has been enunciated in Table 2.

Differing Characteristic	Dense-mode protocols	Sparse-mode protocols
Transmission mechanism	Broadcast and prune	Core / Rendezvous Point (RP) based (Centralized group management). Traffic restricted to the multicast group.
Distribution tree	Source Distribution tree (shortest path tree)	Source Distribution tree or Shared distribution tree or both
Existing Protocols	DVMRP, PIM-DM	CBT, PIM-SM
Group management	Maintains information of hosts that are (positive) or not part of the group (negative)	Maintains information of hosts that are part of the group
Routers where state is maintained	At all the routers, irrespective of whether it is on the multicast tree. State can be positive or negative.	Only in routers on the packet delivery tree
Storage overhead in terms of routing entries	A routing entry for each (source, group) pair – either positive or negative.	The shortest path entries, shared path entries and the negative cache entries for paths that are in the switching process.
Bandwidth overhead	Total number of unwanted data packets transmitted over all network links along with the periodic prune messages	Total number of PIM control messages.

Table 2: Dense Mode Protocols vs. Sparse Mode Protocols

While providing paid services over the Internet, there is need to guarantee the quality of service. The quality of service desired is specified as Quality of Service (QoS) constraints in terms of QoS metrics. The constraints can be classified as link constraints and tree constraints. The routing protocols have to build trees that satisfy the specified constraints or optimization criterion.

QoS support is usually associated with scalability problems since QoS requires connection-specific information. This holds true for point-to-point connections for which the routing tables do not maintain connection-specific state. However, multicast tables keep connection specific state in routers, namely the multicast group address that refers to a connection and the list of interfaces on which data is to be forwarded for the receivers. Adding QoS connection specific information is straightforward and increases the routing state only by a fraction. But providing QoS routing for multicast has the following disadvantages as well:

- Distributed continuous media applications have diverse QoS requirements. Multiple constraints make the multicast routing problem intractable.
- A multicast routing protocol has to handle a number of issues like state collection and updation, handling of dynamic topology and membership changes, tree maintenance and scalability. QoS further complicates the protocol design process. Further QoS Routing subject to two or more additive/multiplicative metrics is known to be NP-Complete.
- One has to consider how to collect/maintain QoS-related state at minimal cost, how to construct a QoS-satisfying route/tree and how to maintain QoS across routing domains.
- As the number of receivers, the number of different QoS requirements to be satisfied by the source increases. This leads to the problem of many heterogeneous QoS requirements.

We categorize the routing protocols based on whether the routing protocol finds the multicast tree based on best effort or a given QoS requirement. The routing protocols are described in the following subsections – Best Effort Multicasting (Summarized in Table 3) and QoS-based Multicasting (Summarized in Table 4).

4.1 Best Effort Multicast Routing Protocols

4.1.1 Distance Vector Multicast Routing Protocol (DVMRP)

DVMRP [1] is a distance vector style algorithm that builds source based multicast trees. When a DVMRP router receives a multicast packet, it sends the packet to all attached routers and waits for a response. Routers with no group members return a “prune” message, which eventually prevents further multicast messages for that group from reaching the router. The prune state is soft, that is, it will time-out within a set time interval. If after sending a prune and before the state can time-out, the host wants to join the group, it has to send a “graft” message upstream. DVMRP is inefficient when the number of receivers in the group is sparsely distributed.

DVMRP builds its own routing table instead of reusing the existing unicast routing table for RPF checking of incoming packets. A packet is assumed to have arrived on the RPF interface if a router receives it on an interface that it uses to send unicast packets to the source. If the packet arrives on the RPF interface, then router forwards it out the interfaces that are present in the outgoing interface list of a multicast routing table entry. If it does not arrive on RPF interface, it is silently discarded to avoid loop-backs. The advantage of RPF is that it does not require the router to know about spanning trees. This way, multicast adapts automatically and only is sent where it is wanted. RPF checking cannot be used to check the validity of a path in case of asymmetric paths[7].

[13] proposes a hierarchical distance-vector multicast routing protocol. This approach involves partitioning the MBone into non-overlapping regions, while using DVMRP as the inter-region routing protocol. Intra-region routing protocol may be accomplished by any of the multicast routing protocols.

4.1.2 Multicast Open Shortest Path First (MOSPF)

MOSPF is a link state routing protocol that builds the map of the network topology, including location of domains and tunnels. It selects the best path to the required receivers using Djikstra’s shortest path algorithm. It is meant to be in use within an Autonomous System (AS). When there are multiple sources or many groups, it is CPU intensive. It is best used when relatively few sources or groups are active at any given time. It does not work well in presence of unstable links, as it leads to frequent state update and the associated computations. MOSPF does not support tunneling. The path is calculated only “on-demand” and cached for later use. It constructs source based multicast trees. It can also be considered as a QoS routing algorithm that minimizes delay. It is one of the dense mode protocols that requires explicit join from the receivers [8].

4.1.3 Core Based Tree (CBT)

CBT [26, 27] builds a single bidirectional shared tree for the data transmission from the source(s) in the group to the receivers. When an intermediate node receives a packet meant for the group, it forwards it to the remaining members of the group that are downstream to the node. It does not need to forward it to the core. Core selection is one of the major issues in CBT and can be handled by the various heuristics proposed for core selection as in Section 3.2.0.

4.1.4 Protocol Independent Multicast (PIM)

PIM operates in two modes – Dense mode (PIM-DM) and Sparse Mode (PIM-SM). PIM-DM operates similar to DVMRP. Sparse mode protocols use explicit join messages to set up uni-directional shared distribution trees. Dense mode protocols use only source distribution trees and uses RPF checking to determine if a packet is to be forwarded.

In PIM-SM [35], a node is selected as the Rendezvous Point (RP) and all group communication takes places by sending the packets to it. It is not dependent on any particular unicast routing method. However, it uses existing unicast routing table for the routing decisions. Each of the sources in a PIM-SM multicast group send their packets to the RP. Since it builds unidirectional shared tree, only the RP can forward data to the members. Intermediate nodes should forward the data only to the RP. Any site interested in joining requests one of the RPs to set up a tunnel to the RP. All PIM-SM traffic is transported by unicast instead of multicast.

The PIM-SM router with highest IP address is Designated Router (DR) for the subnet and is responsible for sending Prune/Join messages to the RP. DR determines the RP for a group using a hash function. Information about RP is obtained by sending Bootstrap messages. The tree obtained is not necessarily optimal.

[7] Asymmetric Routing means that the unicast route from A to B may differ from the path from B to A.

PIM-SM allows switching of the receiver connectivity to the tree from Shared tree path to Source tree path. When a group has numerous highly active sources, the bandwidth of the shared links may not be able to accommodate all the traffic. The QoS requirements of a user may not be met along the Shared Tree, thus it has to switch to its shortest path to the RP. When a receiver switches from Shared tree mode to Source Tree mode, the number of packets dropped during the transition period is dependent on two factors: The delay difference between the Shared tree path and Source Tree path from the source to the receiver and the sending rate of the source [7].

4.1.5 Simple Multicast

Simple multicast and EXPRESS multicast are based on the Root Addressed Multicast Architecture (RAMA) architecture. RAMA architecture is applicable in cases where multicast applications have a single source or have a single primary source, which can be used as the core of the tree. The address of the root is appended to the multicast group address, which is unique over the Internet. This eliminates the need for coordinated multicast address allocation across the Internet. These are two protocols that take care of address allocation along with routing of data. Most routing protocols assume that unique address has been allocated to the group.

Simple multicast, a variation of CBT, considers the identity of a group to be a 8-byte combination (C, G) where C is the 4-byte “core node” IP address and G is the 4-byte regular multicast address. In case of multiple sources to the group, one of the source nodes is chosen as the “core” node and the shared tree is rooted at the core. A single bidirectional shared tree covers both intradomain and interdomain multicasting. It does not allow switching to source specific shortest paths. Receivers send join messages to the source. The exact method of core selection is not a part of the routing protocol. If the core selection is optimal, Simple Multicast is the most suitable for multiple sources applications.

4.1.6 EXPRESS multicast

Like Simple multicast, EXPRESS multicast is based on the RAMA architecture. It builds a source tree with the root located at the source. Receiver sends join messages to the source along the reverse path to the source. The group is identified by the 8-byte address (Source Address, Group Address). Since the source address uniquely identifies the group, the protocol can only be used for single source group communications unlike Simple multicast (also based on RAMA architecture) that can support multiple sources per group. It assumes sources learn about receivers via some mechanism outside EXPRESS; it does not support IGMP. EXPRESS multicast using IP multicast channel have been assumed to give the most scalable solution for single source applications.

4.1.7 Source Specific Multicast (SSM)

The network must maintain knowledge about which hosts in the network are actively sending multicast traffic. In SSM [60] the receiver informs the router to which it is sending the join request the list of source(s) of the group it is subscribing to. The receiver must subscribe or unsubscribe to (Source Address, Group Address) channels to receive or not receive traffic from specific sources. Receivers can receive traffic only from (Source Address, Group Address) channels that they are subscribed to. This is in contrast to IP multicast where receivers need not know the source(s) of the group to receive traffic from the group. The address range 232.0.0.0 through 232.255.255.255 has been reserved for SSM applications and protocols.

4.1.8 Centralized Multicast

Most routing protocols assume routers participate both in forwarding multicast packets and in control algorithms for routing, resource reservation and group management. Centralized Multicast [31] separates data and control flow and centralizes control in distinct control elements. The "control element gateway" is introduced for each domain to construct the portion of the multicast tree inside the domain. "Control element root controllers" are introduced for the Internet to construct the inter-domain portion of a multicast tree. The tree in the domain can be a bidirectional shared tree or a source-specific shortest path tee. The inter-domain multicast tree is bidirectional shared tree.

4.1.9 Border Gateway Multicast Protocol (BGMP)

Border-Gateway Multicast Protocol (BGMP) is implemented at the border routers of a domain. It constructs inter-domain bi-directional shared trees using a single root, while allowing any multicast routing protocol to be used within the domains. The root is located at the domain whose address range covers the group's address; which is typically the group initiator's domain. BGMP requires strict address allocation [34].

4.1.10 Multiprotocol Extensions to BGP (MBGP)

This protocol proposes using the BGP to setup and forward multicast routing state. This is to enable faster deployment of multicast, as BGP is widely in use in the current Internet.

4.1.11 Multicast Internet Protocol (MIP)

MIP constructs both group-shared and shortest-paths multicast trees. The operations can be sender or receiver initiated or both. It is independent of the underlying unicast routing algorithms used. Instead of using "soft state", MIP uses diffusing computations to update and disseminate multicast routing information. The protocol creates loop-free distribution trees independently of loops the unicast routing tables may have.

Every computation started by a router to create and maintain a multicast tree is propagated to other routers, as needed using a recursive query-response mechanism. It has the advantages of faster response time, as it does not depend on timers. Being event-driven, it does not incur any overhead traffic when the tree is stable [23].

4.2 QoS Based Multicast Routing Protocols

4.2.1 Kumar et. al. proposed solution

It generates two routing trees: a shortest path tree and a Steiner tree. It identifies a given number of destinations, say k, for whom the difference between the delay observed in the Steiner tree and the delay in the shortest path tree is largest. Such destinations are replaced from the Steiner tree to the shortest path tree. Average delay is the metric it tries to minimize between the various destinations.

4.2.2 Kompella et. al. distributed solution

This is an algorithm that builds the delay constrained least cost tree based on MST heuristic [6]. Link delays are assumed to be integers and delay constraint is bounded, so that the complete graph can be constructed in polynomial time. A delay-constrained spanning tree of the complete graph is constructed. It generates the routing tree starting from the source, the tree is incrementally expanded till the destination is reached, such that the selected edge (1) connects a node in the tree and a node outside the tree (2) does not violate the delay constraint (3) minimizes a selection function (cost / tradeoff between cost minimization and delay minimization). It proposes two heuristics to select the next non-tree node to be included in the tree – the cost-delay heuristic and the cost heuristic. Cost delay heuristic calculates the weight of a link as,

$$W(u, v) = \begin{cases} \dfrac{c(u, v)}{\Delta - (D(u) + d(u, v))} & \text{, if } D(u) + d(u, v) < \Delta \\ = \text{Infinity} & \text{, otherwise} \end{cases}$$

where c and d are cost and delay matrix respectively. D is the accumulated delay from source s to node u. Δ is the amount to which delay is constrained along the path. Cost heuristic selects the next node assuming that the weight of the node is the same as the cost of the path. There is a distributed version of the same algorithm.

Routing Protocol	**Source/ Shared tree**	**Dense/ Sparse**[8]	**Global/ Local Information**	**Intra/ Inter-domain**
DVMRP	Source	Dense	Not applicable	Intra
MOSPF	Source	Dense	Global	Intra
CBT	Shared (bidirectional)	Sparse	Local	Inter
PIM-SM	Shared (unidirectional) which can be switched to source	Sparse	Local	Inter
PIM-DM	Source	Dense	Not applicable	Intra
Simple	Shared (bidirectional)	Sparse (RAMA architecture)	Local	Inter
EXPRESS	Source	Sparse (RAMA architecture)	Local	Inter
SSM	Source	Sparse (receiver specifies the exact source in the group)	Local	Inter
Centralized	Shared (bidirectional) or source for intradomain. Shared bidirectional for interdomain.	Sparse	Local	Inter
BGMP	Shared (bidirectional)	Sparse	Local	Inter
MBGP	Source	Sparse	Local (full path)	Inter
MIP	Shared as well as source.	Dense as well as Sparse	Local	Inter

Table 3: Best Effort Multicast Routing Protocols

[8] Dense mode is broadcast and prune. Sparse mode is explicit join protocol.

4.2.3 MAMCRA

Multiple Adaptive Multiple Constraints Routing Algorithm (MAMCRA) [57] guarantees QoS to the multicast members in an efficient, but not always optimal manner. MAMCRA provides solutions for multiple constrained multicast. It can approximately solve Multiple Constrained Minimum Weight Multicast problem. It can be considered as a heuristic to solve multiple parameters Steiner tree.

MAMCRA solves the multiple constrained routing problem by the following mechanism: First the set of shortest paths from the source to all the multicast members is evaluated. Following this, the entire tree is optimized such that the length function is reduced, without violating the constraints. This is based on the fact that a multicast tree may not always guarantee the requested QoS constraints, while multiple unicast QoS paths will satisfy the constraints. A set of paths has to be maintained per tree. However, it is desired that a multicast tree should be constructed, either by fine-tuning MAMCRA or by renegotiating the constraints.

4.2.4 Policy Tree Multicast Routing

Policy Tree Multicast Routing (PTMR) builds multicast trees that even under asymmetric conditions readily comply with imposed policies and enable support for shortest path and QoS. A Policy tree characterizes PTMR architecture, which is the product of receiver-initiated source originating tree construction. Policy Route is a path from source to destination given by a sequence of domains that satisfies the policy requirements of the source and the involved domains and supports the requested service quality [33].

4.2.5 Constrained Bellman Ford Algorithm

The constrained Bellman-Ford Algorithm is used to connect one group member at a time to the source. It is based on Constrained Adaptive Ordering Heuristic, that states that after each run, the unconnected member with the delay constrained minimum-cost path to the source is chosen and added to the existing sub-tree.

4.2.6 Receiver Driven Layered Multicast (RLM)

Source-based rate-adaptation performs poorly in a heterogeneous environment that is typical of multicast, as there are as many target receiver rates as the receivers. Distributing one uniform representation of the signal results in low-capacity regions suffering congestion and high-capacity regions being underutilized. If the burden of rate-adaptation is moved from the source to the receiver, heterogeneity is accommodated.

Selective forwarding can be carried out using multiple IP-Multicast groups where each receiver specifies its subscription by joining a subset of the groups. The relationship among the information contained across the set of groups in a session can either be cumulative or independent. In cumulative, each layer provides refinement information to the previous layers. In independent, each layer is independent –source transmits multiple copies of the same signal simultaneously at different rates (different qualities), also known as Simulcast.

[17] extends the cumulative selective forwarding by adding a rate adaptation protocol. Each receiver runs a control loop: on congestion it drops a layer and on spare capacity it adds a layer.

4.2.7 Hop-by-Hop (HBH)

HBH [54] uses the unicast infrastructure to do packet forwarding but uses EXPRESS channel model (Source Address, Group Address) to identify a group. It constructs SPT instead of Reverse SPT, thereby providing best routes in asymmetric networks and being suitable for eventual implementation of QoS based routing. Asymmetric unicast routing affects most multicast routing protocols since majority of them construct Reverse SPT using RPF check. Thus, data packets from source to receiver follow the unicast route from receiver to source. If these paths have different characteristics, use of reverse SPT may be problematic for QoS deployment.

HBH uses a tree management algorithm that provides enhanced tree stability in the presence of group dynamics and reduces tree bandwidth consumption in asymmetric networks. It uses two tables, one Multicast Control Table (MCT) and one Multicast Forward Table (MFT). MFT stores the address of a next branching node instead of the address of a receiver (except for the branching router nearest to the receiver). This tree management scheme minimizes the impact of member departures on the tree structure.

4.2.8 Quality of Service sensitive Multicast Internet protoCol (QoSMIC)

QoSMIC [36] starts with a shared tree and individual receivers switch to source tree when necessary. It constructs trees based on the greedy heuristic that connects each user to the “closest” branch of the existing tree. The search for the “closest” branch can be done by a local search or a multicast tree search.

It uses dynamic routing information without relying on a link state exchange protocol to provide the routing information. Dynamic metrics help respond proactively to link congestion but this has scaling problems. Thus, dynamic metrics are not used in link state exchange protocol; instead it is used to evaluate and select from the alternate paths possible.

This protocol requires a Manager router of a group. The fundamental difference between a core router and a Manager is that the distribution tree is not rooted at the Manager. There can be multiple managers for efficient and scalable solutions with reduced set-up time. The managers can be changed during the lifetime of the group without any data loss. This protocol can be used at both the intra-domain and the inter-domain level. The protocol creates loop-free distribution trees independently of loops the unicast routing tables may have.

4.2.9 QoS-aware Multicast Routing Protocol (QMRP)

QMRP [48] is a routing protocol for non-additive metrics. Most of the QoS routing protocols depend on topology filtering[9] to handle non-additive metrics. High-level design of QMRP makes it operable on top of any unicast routing algorithm, both intra- and inter-domain. It is applicable only for groups with fixed membership. The termination mechanism detects the failure as well as the success of routing without the use of timeout. It always constructs loop-free multicast trees. It can be used to construct both sender-based trees and shared trees.

QMRP Protocol starts with a single path but, when necessary, it can expand the search by splitting at one or multiple points (multiple path routing) in a controlled manner according to perceived network conditions. QMRP finds a feasible path if one exists. Multiple path routing

[9] Topology filtering is when min/max QoS measures are treated by omitting all links (and possibly disconnected nodes) that do not satisfy the requested QoS measures. Essentially there are three types of metrics: additive (e.g., delay, delay jitter and cost), multiplicative (e.g., 1-loss probability=probability of successful transmission) and concave (also known as min/max)(e.g., bandwidth).

can lead to resource contention. To avoid this, two protocol parameters are used: Maximum Branching Level (MBL) and Maximum Branching Degree (MBD). MBL is number of nodes performing multipath routing. MBD is the maximum number of REQUEST messages allowed to be sent by a node in multipath stage. The probes proceed only along paths that do not violate QoS and optimization requirements.

4.2.10 Nearest Destination First (NDF)

At each iteration, the nearest unconnected destination to the partially constructed tree is found and added to the tree. The algorithm attempts to find the tree with the least average delay.

4.2.11 Destination-Driven Multicast (DDMC) Routing

The cost is reset to zero at the destination nodes, making each destination node behave like a new “source”. The reason for this is that any nodes reachable from a destination node incur only an incremental additional cost [22].

Routing Protocol	Source/ Shared Tree	Dense/ Sparse	QoS Metric	Global/ Local Information	Intra/ Inter-domain
Kumar et. al.	Shared as well as source	Sparse	Delay	Local	Inter
Kompella et. al.	Source	Sparse	Delay constrained minimum cost	Local	Inter
MAMCRA	Source	Sparse	Multiple additive metrics	Local	Inter
PTMR	Source	Sparse	Handles asymmetric link policy	Local	Inter
Constrained Bellman Ford	Source	Sparse	Delay constrained minimum cost	Local	Inter
RLM	Source	Sparse	Layered transmission of media	Local	Inter
HBH	Shortest Path Tree	Sparse	Handles asymmetric link policy	Local	Inter
QoSMIC	Shared (unidirectional) which can be switched to source	Sparse	Any metric for selection of route and any dynamic metric such as available bandwidth for selecting among alternate routes.	Local	Inter
QMRP	Source as well as shared	Sparse	Non-additive metric	Local	Inter
NDF	Source	Sparse	Average Delay	Local	Inter
DDMC	Source	Sparse	Cost	Local	Inter

Table 4: QoS based Multicast Routing Protocols

5.0 Multicasting at the other Layers

Multicast can be implemented at the different layers of the protocol stack such as the data link layer, network layer, transport layer and application layer. Multicasting in the network layer has been discussed in detail in Section [3] and [4]. Multicasting and additional features for the same as provided by the data link layer, transport layer and application layer are discussed in the following subsections.

5.1 Data Link Layer

The data link layer protocols such as Ethernet, FDDI, and token ring also provides support for multicast. Implementing multicast in the data link layer is sufficient if the multicast application is restricted to a Local Area Network (LAN). However, if the multicast traffic has to be extended beyond a single LAN, network layer multicasting has to be used. A mapping must be done from network layer multicasting to data link layer multicasting mechanisms for transfer of data across LANs using multicasting. The procedure involves three separate operations:

- Network multicast address resolution to LAN multicast address
- Copying and forwarding of messages
- Group membership registration

When a local router on a LAN receives a network-layer addressed such as IP Class D address multicast packet, it maps the multicast address to a data link layer multicast address such as Ethernet MAC address. Data link layer LAN protocols reserve portions of their address space for multicast. Address translation from the IP Class D address to an Ethernet MAC address is done by dropping the lower 23 bits of the IP multicast address in the low order 23 bits of the Ethernet multicast address. The multicast address range in the Ethernet is from 01:00:5E:00:00:00 to 01:00:5E:7F:FF:FF. Since the upper 5 bits of the multicast address are ignored in this mapping, the resulting address is not unique. 2^5 (≈ 32) different Class D IP addresses can map to each Ethernet address. Thus, the device driver or IP modules of the receiver must perform filtering based on the group address.

The switches[10] in the network can also learn about multicast membership in the network and use the information to transmit data to the appropriate ports of the switch. In Control Layer 2 multicasting, the switch detects the incoming multicast traffic and examines the destination MAC address to determine which port(s) should receive the traffic.

Cisco Group Management Protocol (CGMP) is a Cisco-developed protocol that allows Catalyst switches (Cisco products) to leverage IGMP information to make data-link layer forwarding decisions. With CGMP in place, IP multicast traffic is delivered to only those Catalyst switch ports that are interested in the traffic.

IGMP snooping requires the LAN switch to examine network layer information in the IGMP packet sent from host to the router. When the switch observes an IGMP request sent from host for a particular group, switch adds the hosts port number to the associated multicast table entry. Similarly when it observes the IGMP leave, it removes the host from the table entry. However it is very CPU intensive as every multicast packet has to be examined just to find an occasional IGMP packet.

[10] Routers build maps and give direction to packets towards destinations. Switching forwards packets between interfaces. Routers have the additional overhead of path determination.

5.2 Transport Layer

There have been proposals to implement multicast transport protocols, which implement reliability and synchronization for multicasting. One such proposal is Multicast Transport Protocol (MTP). MTP does not require acknowledgements of messages; it is a negative acknowledgement protocol. It provides the synchronization necessary for members to agree on the order of receipt of messages – the ordering and agreement protocol uses serialized tokens.

Reliable Adaptive Multicast Protocol (RAMP) is a transport level protocol. Some applications do not require in-order delivery; RAMP includes an option to allow out-of-order delivery of data. It has a NACK-based retransmission scheme. RAMP at the sender aggregates retransmission requests for a "retransmission hold time". If the number of receivers requesting retransmission is below a threshold, then the packet is unicast to each individual receiver, else it is multicast to the entire group. Another optimal approach would be to construct a new multicast group for the retransmission; however, it might turn out to be costly due to the group set-up time.

5.3 Application Layer

Though network layer multicast is known as the most efficient way to support multicast, application layer multicast handles features such as security, QoS and other operations better. [2] proposes a scheme called M-RTP for multicast RTP sessions. The idea behind this scheme is to set up the multicast RTP sessions over a set of unicast RTP sessions established between the various participants of the multicast session.

Any application interested in becoming a member of a group has to send an IGMP request to its local router. Application layer handles group management for the end hosts in multicast communication.

6.0 Some Issues and Open Problems

Multicast routing involves the issue of scalability. For unicast routing, router maintains information about individual networks. Each entry is short and contains a metric associated with each network that gives a measure of how to reach the network. For multicast routing, router maintains more information – information about individual networks and multicast groups. Each entry specifies the source sending to the group, multicast group address, interface on which data arrives, interfaces on which it is to be forwarded. It also has to maintain the state of each interface with respect to the presence of groups. Each multicast routing entry requires more memory than each unicast routing entry in the routing table.

The following factors influence the issue of scalability in multicast routing:

- The spread of participants
- The number of sessions that can be active at any given time
- The number of participants in a session
- The heterogeneity of the network bandwidth, which makes it difficult for the sender to decide on an optimal sending rate
- The type of data being sent as multicast data, which includes factors of service quality

Router table lookup speed has to increase. Lookup tables need to be designed for fast IP address lookup and fine-tuned for multicast applications. Appropriate data structure needs to be developed which will have minimal update and lookup time complexity, and which scales with the number of entries in the routing table. For faster deployment, ease of configuration in using multicast should be enhanced.

Multicast address allocation is another one of the unresolved issues. A multicast group initiator typically contacts an address allocation application such as sdr in MBONE and an address is randomly assigned from those that are not in use. Adequate address management mechanisms need to be deployed.

Multicast Address Set Claim (MASC) divides the address allocation into three phases – domain level (MASC), within a domain (Address Allocation Protocol AAP) and between hosts and network (Multicast Address Dynamic Client Allocation Protocol MADCAP can be used by hosts to request address from the Multicast Address Allocation Server MAAS). GLOP is another form of multicast address allocation. A "glop" (256 addresses) of addresses is assigned statically to each AS, with AS number encoded as part of address. The first octet is static, next two octets encode the AS number and final octet provides the range of addresses to be allocated. GLOP does not specify how addresses are allocated within the domain. Other issues include:

- Excessive processing of control information
- Poor response to network congestion
- Inability to handle high-priority traffic
- Suboptimal error recovery and retransmission procedures
- Authentication and Security of data transmitted by multicast
- Support for network management

7.0 Multicast: State of the Art (PIM/MBGP/MSDP)

Internet2 is a research effort aimed at the next generation internetworking. It supports multicast in its Abilene and vBNS backbone. Initially dense mode protocols were in use. Now they support PIM-SIM/MSDP/MBGP. The group information, namely the active sources, is shared among the various domains in the backbone using MSDP. MBGP is used for the sharing of the routing information across domains. PIM-SM is used to provide support for building multicast forwarding tree and provide join capabilities for local receivers. MSDP and MBGP are also used for monitoring the performance of the group.

By PIM-SM, traffic from all the sources for a particular group within a particular domain will reach the group's receivers, however any source outside the domain will remain disjoint. This is overcome by using MSDP. MSDP requires a node in each domain to inform the other domains the existence of active sources. Though this is not a scalable solution, the current Internet uses MSDP for distribution of information about the multicast groups. Full MBGP peering among the core vBNS Juniper m40 and Cisco routers (a multi-platform network) ensures shortest-path distribution trees across the vBNS backbone.

On the GEANT network (Since 1st December 2001, the GEANT network has taken over the previous pan-European research network TEN-155) backbone the multicasting is entirely native and sparse mode using PIM-SM. Multicast is deployed on the same physical infrastructure together with unicast data. All connections between all participants are done via PIM-SM/MSDP/MBGP, the same as in the Internet2 backbone networks Abilene and vBNS.

References

[1] D. Waitzman, C. Partridge, S.E. Deering, "*Distance Vector Multicast Routing Protocol*", Internet Request for Comment 1075, November 1988.

[2] Stephen. E. Deering, "*Host extensions for IP multicasting*", Internet Request for Comment 1112, August 1989.

[3] S. Armstrong, A. Freier, K. Marzullo, "*Multicast Transport Protocol*", Internet Request for Comment 1301, February 1992.

[4] Yung H. Tsin, “*Incremental distributed asynchronous algorithm for minimum spanning trees*”, in Computer Networks and ISDN Systems, Elsevier Science Publishers, North Holland, Vol. 26, 1993, pp. 227-232.

[5] R. Braudes, S. Zabele, "*Requirements for Multicast Protocols*", Internet Request for Comment 1458, May 1993.

[6] Vachaspathi P. Kompella, Joseph C. Pasquale, George C. Polyzos, “*Multicast Routing for Multimedia Communication*”, in IEEE/ACM Transaction on Networking, Vol. 1, No. 3, June 1993, pp. 286-292.

[7] Liming Wei, Deborah Estrin, “*A Comparison of Multicast Trees and Algorithms*”, Computer Science Department, University of Southern California, USA, Technical Report USCCS-93-560, September 1993.

[8] J. Moy, "*MOSPF: Analysis and Experience*", Internet Request for Comment 1585, March 1994.

[9] Stephen Deering, Deborah Estrin, Dino Farinacci, Van Jacobson, Ching-Gung Liu, Liming Wei, “*An Architecture for Wide Area Multicast Routing*”, in ACM SIGCOMM, Computer Communication Review, 1994, pp. 126-135.

[10] Liming Wei, Deborah Estrin, “*The Trade-offs of Multicast Trees and Algorithms*”, in Proceedings of the 1994 International Conference on Computer Communications and Networks, 1994, pp. 17-24.

[11] Hans Eriksson, “*MBONE: The Multicast Backbone*”, in Communications of the ACM, August 1994, Vol. 37, No.8, pp. 54-60.

[12] Kenneth L. Calvert, Ellen W. Zegura, Michael J. Donahoo, “*Core Selection Methods for Multicast Routing*”, in Proceedings of ICCCN, 1995, pp. 638-642.

[13] Ajit S. Thyagarajan, Stephen E. Deering, “*Hierarchical Distance-Vector Multicast Routing for the MBONE*”, in ACM SIGCOMM Computer Communication Review 1995, pp.60-66.

[14] Stephen E. Deering , “*Multicast Routing in Internetworks and Extended LANs*”, in ACM SIGCOMM Computer Communication Review 1995, pp. 89-100.

[15] Shai Herzog, Scott Shenker, Deborah Estrin, “*Sharing the “Cost” of Multicast Trees: An Axiomatic Analysis*”, in ACM SIGCOMM Computer Communication Review, August 1995, Vol. 25, pp. 315-327.

[16] Liming Wei, Deborah Estrin, “*Multicast Routing in Dense and Sparse Modes: Simulation Study of Tradeoffs and Dynamics*”, Computer Science Department, University of Southern California, USA, Technical Report 95-613, 1995.

[17] Steven McCanne, Van Jacobson, Martin Vetterli, “*Receiver Driven Layered Multicast*”, in ACM SIGCOMM Computer Communication Review 1996, pp.117-128.

[18] Ken Carlberg, Jon Crowcroft, “*Building Shared Trees Using a One-to-Many Joining Mechanism*”, in ACM SIGCOMM Computer Communication Review, January 1997, pp. 5-10.

[19] David G. Thaler, Chinya V. Ravishankar, “*Distributed Center-Location Algorithms*”, In IEEE JSAC, Vol. 15, No. 3, April 1997, pp. 291-303.

[20] Hussein F. Salama, Douglas S. Reeves, Yannis Viniotis, "*Evaluation of Multicast Routing Algorithms for Real-Time Communication on High-Speed Networks*", in IEEE JSAC, Vol 15, No. 3, April 1997, pp.332-345.

[21] Christophe Diot, Walid Dabbous, Jon Crowcroft, "*Multipoint Communication: A Survey of Protocols, Functions and Mechanisms*", in IEEE JSAC, Vol. 15, No. 3, April 1997, pp. 277-290.

[22] Anees Shaikh, Kang Shin, "*Destination-Driven Routing for Low-Cost Multicast*", in IEEE JSAC, Vol. 15, No. 3, April 1997. pp. 373-381.

[23] Mehrdad Parsa, J. J. Garcia-Luna-Aceves, "*A Protocol for Scalable Loop-Free Multicast Routing*", in IEEE JSAC, Vol. 15, No. 3, April 1997, pp. 316-331.

[24] George N. Rouskas, Ilia Baldine, "*Multicast Routing with End-to-End Delay and Delay Variation Constraints*", in IEEE JSAC, Vol. 15, No. 3, April 1997, pp. 346-356.

[25] Fred Bauer, Anujan Varma, "*ARIES: A Rearrangeable Inexpensive Edge-based on-line Steiner algorithm*", In IEEE JSAC, Vol. 15, No. 3, April 1997, pp. 382-397.

[26] A. Ballardie, "*Core Based Trees (CBT version 2) Multicast Routing*", Internet Request for Comment 2189, September 1997.

[27] A. Ballardie, "*Core Based Trees (CBT) Multicast Routing Architecture*", Internet Request for Comment 2201, September 1997.

[28] W. Fenner, "*Internet Group Management Protocol, Version 2*", Internet Request for Comment 2236, November 1997.

[29] Moses Charikar, Chandra Chekuri, To-yat Cheung, Zuo Dai, Ashish Goel, Sudipto Guha, Ming Li, "*Approximation Algorithms for Directed Steiner Problems*", in Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, 1998, pp. 192-200.

[30] John C. I. Chuang, Marvin A. Sirbu, "*Pricing Multicast Communication: A Cost-Based Approach*", presented at the Internet Society INET 1998 Conference, Geneva, Switzerland, July 21-24, 1998.

[31] S. Keshav, S. Paul, "*Centralized Multicast*", Computer Science Department, Cornell University, USA, Technical Report TR98-1688, 1998.

[32] Jena-Jacques Pansiot, Dominique Grad, "*On Routes and Multicast Trees in the Internet*", in ACM SIGCOMM Computer Communication Review, January 1998, pp. 41-50.

[33] Horst Hodel, "*Policy Tree Multicast Routing: An Extension to Sparse Mode Source Tree Delivery*", in ACM SIGCOMM, Computer Communication Review, April 1998, pp. 78-97.

[34] Satish Kumar, Pavlin Radoslavov, David Thaler, Cengiz Alaettinoglu, Deborah Estrin, Mark Handley, "*The MASC/BGMP Architecture for Inter-domain Multicast Routing*", in ACM SIGCOMM, April 1998, pp. 93 – 104.

[35] D. Estrin, D. Farinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, V. Jacobson, C. Liu, P. Sharma, L. Wei, "*Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification*", Internet Request for Comment 2362, June 1998.

[36] Michalis Faloutsos, Anindo Banerjea, Rajesh Pankaj, "*QoSMIC: Quality of Service sensitive Multicast Internet protoCol*", in ACM SIGCOMM, Vancouver, Canada, September 1998, pp. 144-153.

[37] K. Dubray, "*Terminology for IP Multicast Benchmarking*", Internet Request for Comment 2432, October 1998.

[38] Shigang Chen, Klara Nahrstedt, "*An Overview of Quality-of-Service Routing for the Next Generation High-Speed Networks: Problems and Solutions*", in IEEE Network, Special Issue on Transmission and Distribution of Digital Video, November/December 1998, pp. 64-79.

[39] G. Phillips, S. Shenker, H. Tangmunarunkit, "*Scaling of Multicast trees: Comments on the Chuang Sirbu Scaling law*", in ACM SIGCOMM 1999, September 1999, Harvard, MA, pp. 41-51.

[40] D. Thaler, "*Interoperability Rules for Multicast Routing Protocols*", Internet Request for Comment 2715, October 1999.

[41] P. Bagnall, R. Briscoe, A. Poppitt, "*Taxonomy of Communication Requirements for Large-scale Multicast Applications*", Internet Request for Comment 2729, December 1999.

[42] Eric Fleury, Yin Huand, Philip K. McKinley, "*On the Performance and Feasibility of Multicast Core Selection Heuristics*", in Networks, John Wiley, 2000, pp. 145-156.

[43] Robert C. Chalmers, Kevin C. Almeroth, "*Developing a Multicast Metric*", in Proceedings of IEEE Global Internet (GLOBECOM) 2000, Vol. 1, pp. 382-386.

[44] Kevin C. Almeroth, "*The Evolution of Multicast: From the MBone to Interdomain Multicast to Internet2 Deployment*", in IEEE Network, January/February 2000, pp. 10-2012.

[45] Bin Wang, Jennifer C. Hou, "*Multicast Routing and its QoS Extension: Problems, Algorithms and Protocols*", in IEEE Network, January/February 2000, pp. 22-36.

[46] Laxman H. Sahasrabuddhe, Biswanath Mukherjee, "*Multicast Routing Algorithms and Protocols: A Tutorial*", in IEEE Network, January/February 2000, pp.90-102.

[47] Christophe Diot, Bryan Lyles, Brian Neil Levine, Hassan Kassem, Doug Balensiefen, "*Deployment Issues for the IP Multicast Service and Architecture*", in IEEE Network, January/February 2000, pp. 78-88.

[48] Shigang Chen, Klara Nahrstedt, "*A QoS-Aware Multicast Routing Protocol*", in Proceedings of IEEE INFOCOM 2000, Tel-Aviv, Israel, March 26-30, 2000, pp. 1594-1603.

[49] Joan Feigenbaum, Christos H. Papadimitriou, Scott Shenker, "*Sharing the cost of Multicast Transmissions*", in Thirty-Second Annual ACM Symposium on Theory of Computing (STOC00), May 2000, pp. 218-227.

[50] J. Yu, "*Scalable Routing Design Principles*", Internet Request for Comment 2791, July 2000.

[51] P. Van Mieghem, G. Hooghiemstra and R. van der Hofstad, "*A Scaling Law for the Hopcount*", Delft University of Technology, Delft, The Netherlands, Technical Report 2000125.

[52] Robert C. Chalmers, Kevin C. Almeroth, "*Modeling the Branching Characteristics and Efficiency Gains in Global Multicast Trees*", in IEEE INFOCOM 2001, pp. 77-86.

[53] R. Cohen, G. Kaempfer, "*A Unicast-based Approach for Streaming Multicast*", in IEEE INFOCOM 2001, Anchorage, Alaska, April 2001, pp. 36-44.

[54] Luis Henrique M. K. Costa, Serge Fdida, Otto Carlos M. B. Duarte, "*Hop-by-Hop Multicast Routing Protocol*", in ACM SIGCOMM, August 2001, pp. 249-259.

[55] P. Van Mieghem, M. Janic, "*Stability of a Multicast Tree*", in IEEE INFOCOM 2002, pp. 1099-1108.

[56] Piet Van Mieghem, Gerard Hooghiemstra, Remco van der Hofstad, "*On the Efficiency of Multicast*", In IEEE/ACM Transactions on Networking, vol. 9, No. 6, pp. 719-732.

[57] Fernando Kuipers, Piet van Mieghem, "*MAMCRA: A Constrained-Based Multicast Routing Algorithm*", In Computer Communications, Vol. 25, No. 8, pp. 801-810.

[58] Homepage of vBNS backbone network. Located at http://www.vbns.net/

[59] Homepage of Dante network. Located at http://www.dante.net/geant/index.html

[60] Source Specific Multicast, available online at
http://www.cisco.com/univercd/cc/td/doc/product/software/ios121/121newft/121t/121t3/dtssm.pdf