A COMPARATIVE ANALYSIS

OF INTER-DOMAIN MULTICAST ROUTING PROTOCOLS

by

ALEJANDRO ESTEBAN AVELLA

Electronics Engineer, Universidad Simon Bolivar, Caracas, Venezuela, 1994

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Master of Science

Interdisciplinary Telecommunications Program

1998

THE UNIVERSITY OF COLORADO LIBRARY

The University Library or any other Library which borrows this thesis for the use of its patrons is expected to secure the signature and home address of each user

Author of thesis, **type** in your name and permanent address here.

Alejandro Avella
La Castellana, Av. El Pedregal, Qta. San Jose
Caracas, 1060, Venezuela
South America

Theses which have been approved for master and doctoral degrees and deposited in the University of Colorado Library are open for inspection. They are to be used only with due regard for the rights of the authors. Bibliographical references may be noted and short passages may be copied. Extensive copying or publication by someone other than the author of the thesis requires the consent of the author and the Dean of the Graduate School of the University of Colorado. In every case proper credit must be given to both the author and to the University in all written or published work.

This thesis has been used by the following persons whose signatures indicate their acceptance of the above restrictions.

Date **Signature** **Address**

A COMPARATIVE ANALYSIS

OF INTER-DOMAIN MULTICAST ROUTING PROTOCOLS

by

ALEJANDRO ESTEBAN AVELLA

Electronics Engineer, Universidad Simon Bolivar, Caracas, Venezuela, 1994

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Master of Science

Interdisciplinary Telecommunications Program

1998

This thesis entitled:

A comparative analysis of inter-domain multicast routing protocols

written by Alejandro Esteban Avella

has been approved for

Interdisciplinary Telecommunications Program

by:

Stan Bush

Jon R. Sauer

Jon Sauer

Gerald Mitchell

Date: 12/2/98

The final copy of this thesis has been examined by the signators, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline

Avella, Alejandro Esteban (M.S., Telecommunications)
A comparative analysis of inter-domain multicast routing protocols
Thesis directed by Professor Stanley E. Bush

It has been suggested that further work is needed in the area of multicast routing and that an evaluation of current multicast routing protocols is encouraged [Deering-98, Deering-98c]. This thesis presents a comparison of the two most important proposals for Inter-Domain Multicast Routing (IDMR). Advantages and disadvantages of each protocol are pointed out and goals that a new proposal should meet are stated.

Researchers have proposed many protocols for IDMR. The IETF has advanced two protocols to the experimental standard status: Core Based Trees (CBT) [Ballardie-97] and Protocol Independent Multicast - Sparse Mode (PIM-SM) [Estrin-98]. The IETF hopes that the market will decide which is the best solution, however it does not provide criteria on how to compare the protocols. This thesis summarizes the key features of CBT and PIM-SM, and makes a comparison based on extensive criteria. Other recent proposals are introduced but only these two are compared.

The criteria used to compare the protocols are composed of five parts: protocol status, basic characteristics, technical criteria, operational criteria and overall assessment. More importance is given to the operational criteria, since this thesis simulates the type of analysis a network manager would do to compare the protocols.

This thesis argues that the Protocol Independent Multicast - Sparse Mode (PIM-SM) proposal is the best solution currently available. This protocol should be chosen as the inter-domain multicast routing protocol to be deployed in the enterprise if a solution is needed immediately. However, it is also shown that neither PIM-SM nor CBT meet all the requirements of a good inter-domain multicast routing protocol. Specifically, the inability of scale (flooding and no aggregation) and lack of support of policies and heterogeneity are the main drawbacks of the protocols. This author argues that none of these protocols will be the final solution adopted in the Internet. The Border Gateway Multicast Protocol [Kumar-98, Thaler-98] has started to move in the right direction and should be the protocol to watch in the near future.

As background information, this thesis reviews internetworking basics, unicast routing, multicast basics, multicast operations, multicast market and multicast routing.

Keywords: PIM-SM, CBT, multicast, routing, IDMR, DVMRP, MOSPF, PIM-DM, MBONE, RIP, OSPF, BGP, IGMP.

DEDICATION

To my family

ACKNOWLEDGEMENTS

I would like to thank to Stan Bush, Jon Sauer and Gerald Mitchell for their guidance as thesis advisors. Thanks go also to Jon Crowcroft and Evi Nemeth for their insightful reviews.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 - Introduction

Most of the communications on the Internet are based on the one-to-one paradigm i.e. a sender establishes a session with only one receiver. Examples of this model include the web, ftp and telnet. Many new applications do not fit this one-to-one model. Multicast technology allows one-to-many communication and enables the use of multimedia applications, such as multiparty teleconferencing over the Internet. Multicast routing efficiently supports one-to-many or many-to-many communication and it saves bandwidth by sending a single packet of a message from a source to multiple receivers. Intermediate routers in the network replicate packets only when needed.

Researchers have proposed several multicast routing protocols, but there is no agreement on a single standard. In the past decade many proposals have appeared on how to implement multicast capabilities in the Internet [Deering-88, Waitzman-88, Deering-91, Ballardie-93, Ballardie-95, Moy2-94, Ballardie-97, Estrin-98]. This thesis explores these proposals and compares the two current proposals for inter-domain multicast routing from the IETF.

1.1 - Purpose

The purpose of this thesis is to present a comparative analysis of the existing approaches for inter-domain multicast routing, in order to provide conclusions under what conditions which alternatives would be best.

1.2 - Thesis statement

This thesis argues that the Protocol Independent Multicast - Sparse Mode (PIM-SM) proposal is the best solution currently available. This protocol should be chosen as the inter-domain multicast routing protocol to be deployed in the enterprise if a solution is needed immediately. However, it is also shown that neither PIM-SM nor CBT meet all the requirements of a good inter-domain multicast routing protocol. Specifically, the inability of scale (flooding and no aggregation) and lack of support of policies and heterogeneity are the main drawbacks of the protocols. This author argues that none of these protocols will be the final solution adopted in the Internet. The Border Gateway Multicast Protocol [Kumar-98, Thaler-98] has started to move in the right direction and should be the protocol to watch in the near future.

1.3 - Importance of this topic

This thesis is an invaluable resource for network managers trying to deploy an inter-domain multicast routing protocol. It demonstrates that the best alternative available today is PIM-SM and it demonstrates why other alternatives should be discarded. Since many of the alternatives have been proposed relatively recently, there is a lot of confusion for network managers on what the alternatives are, what the advantages and disadvantages of each of the alternatives are and which alternatives they should choose. This thesis clarifies all the above issues.

Many new applications are emerging that are one-to-many in nature. The driving force of new applications will force network managers to deploy IP multicast in their networks. Intra-domain routing protocols such as the Distance Vector Multicast Routing Protocol (DVMRP) and the Multicast Open Short Path First (MOSPF) have been used for years and the choices between the options have been documented. However, to the best of this author's knowledge, no previous work has been done analyzing the available options for inter-domain routing from the point of view of a network manager. Some of previous comparisons are described in

Chapter 7. This thesis then becomes a very helpful guide to compare the alternatives for inter-domain multicast routing protocols.

Emerging applications such as the transmission of real-time multimedia training, collaborative applications, videoconference over the Internet, etc.) are quickly driving the deployment of IP multicast on the enterprise. These applications are being developed fast and they are requiring more and more resources from the network. IP multicast is the underlying technology that enhances the IP protocol to support one-to-many communications and it is crucial to understand it and deploy soon in the enterprise so that these new applications can be used and give a competitive advantage over other companies.

Without the use of IP Multicast, applications are forced to use replicated unicast packets,[1] which is a waste of bandwidth and cannot scale to billions of recipients. Applications save bandwidth using IP Multicast, since they do not have to replicate packets for every recipient. Only a stream is created and the network is in charge to replicate the packets only where needed.

Some day millions of multicast sessions will be running at any given time. At this point the Internet will be clogged with traffic and then the deployment of multicast will be desperately needed. The native support of IP multicast on all the routers of the Internet will be absolutely necessary. It is then when the scalability[2] of the

[1] As July of 1998, Real Progressive Networks (http://www.real.com) is using replicated unicast for their Real Audio software product.

[2] Scalability means how well the protocol behaves as the number of nodes in the network increases. In the case of multicast routing, the two parameters that affect the scalability of multicast routing are the number of sources and groups

multicast routing protocols will be really tested. This day may be sooner than we think and that is why it is important to have a scalable multicast routing protocol.

If the Multicast Backbone (MBONE) [Eriksson-94] becomes a production network and gets commercialized and not a researcher's toy anymore, then many corporations will be forced to deploy a scalable multicast routing protocol. Then an understanding of the options available for inter-domain multicast routing will be essential for network managers.

Imagine an Internet where you can have millions of channels to watch and where you can select what you want to see or hear from a web page. This is a dream that will be only available if the problem of scalability of multicast routing is solved.

1.4 - Scope

This thesis presents a comparative analysis of proposed solutions for inter-domain multicast routing. The thesis supports the argument that PIM-SM is the best current available choice for inter-domain multicast routing for most enterprises.

In order to provide context to the problem, unicast routing is covered. As a background, this thesis provides a description of the basics of multicast. Intra-domain multicast approaches such as DVMRP, MOSPF and Protocol Independent Multicast - Dense Mode (PIM-DM) are covered in the background chapters but are not included in the comparison, since the comparison is focused on inter-domain multicast routing protocols. The main approaches for inter-domain multicast routing

present in the network at any point in time. For a more in depth definition of scalability see section 8.6.1.

are described before the comparison is presented. Several proposals are discarded even before the comparison is done (See Chapter 6).

This thesis deals primarily with computer science and networking topics. It covers some of the business aspects of IP Multicast in the introduction and throughout the chapters.

This thesis will not deal with the problem of reliable transport multicasting, which is a transport level (Layer 4) problem. Neither will it deal with the problem of deploying multicast over different data link technologies such as IP Multicast over ATM or IP Multicast over Frame Relay, etc. In the same way it will not deal with the QoS and real-time issues in Multicast. Finally, it will not touch the flow control problems in Multicast applications. These are all topics that could be used as related research topics since there are several resources on the topic and many problems that still need an answer. For more information on these other topics review the references from Ammar's Multicast Tutorial [Ammar-97].

Knowledge of TCP/IP, IP addressing, related protocols such as ICMP, ARP, DNS, etc. and the basics of networking are assumed. However, Chapter 2 covers the minimum that needs to be understood to understand multicast routing.

1.5 - Methodology

The basic approach is to compare proposed solutions for inter-domain multicast routing. Based on a literature search of the alternatives, this would consist of: developing a consistent criterion for comparing the alternatives; a list of the alternatives; a consistent comparison of the alternatives, and, a conclusion about under what conditions which alternatives would be best.

Simulation using the OPNET network tool was explored. However, the simulation turned to be more difficult than expected. The work accomplished with OPNET is presented in Appendix A.

1.5.1 - Literature search

The first step performed was to collect all relevant information on the topic. Two surveys of multicast technology were used to start the research:

- INRIA's survey done in April 1997 [Diot-97] and,
- Mostafa Ammar's Tutorial on multicast done in September 14 of 1997 [Ammar-97],

Another key source of information was the Inter-Domain Multicast Routing (IDMR) working group of the IETF. This group is currently doing research on a multicast routing protocol that provides scalable multicast routing across the Internet.

The Multicast Backbone Deployment (MBONED) group, which is in charge of the deployment of IP Multicast technologies in the Internet, was used as source of information too.

Email communication with researchers that are currently working on this area was conducted in order to obtain and validate the proposal of this thesis. Some of the researchers that were consulted include: Mostafa Ahmad (Georgia Tech), Deborah Estrin (USC), Anthony Ballardie (University College of London), Jon Crowcroft (University College of London), Liming Wei (Cisco), Dino Farinacci (Cisco) and David Thaler (Microsoft).

1.5.2 - Evaluation criteria

The first step consisted of a thorough review of related work on similar comparisons on multicast routing protocols done in the past.

In 1995, Ballardie published his Ph.D. thesis and he compared four routing protocols: DVMRP, MOSPF, CBT and PIM [Ballardie-95]. The evaluation criteria he used included the three factors that most influence the scalability of a multicast algorithm: a) Group state information; b) bandwidth consumption/link utilization and; c) processing costs.

Estrin and Wei did a comparative analysis using computer simulations between Short Path Trees and Core Based Trees [Wei-95]. The comparison criteria were based on three factors: End-to-End delay, Cost (bandwidth consumption and tree state information) and traffic concentration. They concluded that both approaches have good properties depending on the application.

People from Harris Corporation and Naval Research Laboratory have been working the last few years creating models for multicast routing protocols using OPNET[3]. They have published their work in two journals [BillHartz-96] [BillHartz-97]. They also presented their work in the IDMR meeting during the IETF Conference in April 95 and their slides are available on-line.[4] The comparison criteria that they used include end-to-end delay, network resource usage, join time, the size of the multicast routing tables and the impact of the timers introduced by the protocols.

In March 98, Deering and Perlman [Deering-98] suggested that more research was needed on the problem of scalability for multicast routing. As a way to evaluate different multicast routing protocols they suggested that a matrix could be created and each routing protocol would be rated against the following criteria:

[3] For more info on the OPNET Modeler visit http://www.mil3.com

[4] Their April 95 IETF presentation is available at:

ftp://cs.ucl.ac.uk/darpa/IDMR/IETF-APR95/CBTvsPIM-cain.ps number of sources, number of receivers, number of groups, amount of data, burstiness, duration, topological distribution, etc. [Deering-98]

This thesis uses a combination of the criteria used in previous work plus criteria more relevant to a network manager trying to deploy these protocols. Chapter 8 describes the criteria to be used to compare the protocols. Chapter 7 describes in more detail previous comparisons.

1.5.3 - List of alternatives

The alternatives considered for inter-domain multicast routing are:

- Core Base Trees (CBT) [Ballardie-95]
- Protocol Independent Multicast Sparse Mode (PIM-SM) [Estrin-98]

Chapter 6 explains why these are the chosen alternatives for the comparison and also points out the alternatives that were discarded from the comparison and explains the reasons for this decision.

1.6 - Written outline

This thesis is divided in two main parts: Background and Comparatively Analysis.

Part I is the background and it has seven chapters. These chapters are descriptive and give the reader enough background to understand multicast routing protocols. Chapter 2 describes the basics of internetworking. Chapter 3 covers unicast routing. Chapter 4 is an introduction to the basics of IP Multicast. It also presents a roadmap for deploying IP multicast from the operations point of view and a brief survey of the products and services offered for related multicast technologies. Chapter 5 covers the main proposals for intra-domain multicast routing.

Part II is the comparatively analysis of inter-domain multicast routing protocols. It contains five chapters. Chapter 7 summarizes related work on comparisons of multicast routing protocols. Chapter 6 is a survey of inter-domain multicast routing proposals and explains the reasons for choosing PIM-SM and CBT as the two protocols to analyze in this thesis. Chapter 8 sets the criteria to be used for the comparison. Chapters 9 and 10 contain the analysis of CBT and PIM-SM based on the criteria presented in chapter 8. Chapter 11 is the summary and conclusions from this thesis. Chapter 12 has recommendations on future work on IP multicast.

The appendix includes information on the simulation performed using OPNET, useful web sites and a glossary.

Chapter 2 - How the Internet Works

The problem of inter-domain multicast routing, which is the main focus in this thesis requires a good understanding of the basic components of the Internet and how it works. This chapter covers the basic technologies of the Internet. It could be skipped if you feel that you have good grounding in internetworking.

This chapter describes the main components of the Internet and how they interact with each other. It describes all that is needed to make the World Wide Web (WWW) work. This chapter provides a high level introduction, more in-depth descriptions can be found at [Tanenbaum-96] [Peterson-96] or other computer communications books.

2.1 - The Cloud

Have you ever wondered how you get back a web page on your computer when you type a URL on your browser? From your point of view it is a point-to-point connection between your computer and a web server that contains the page you are interested in.

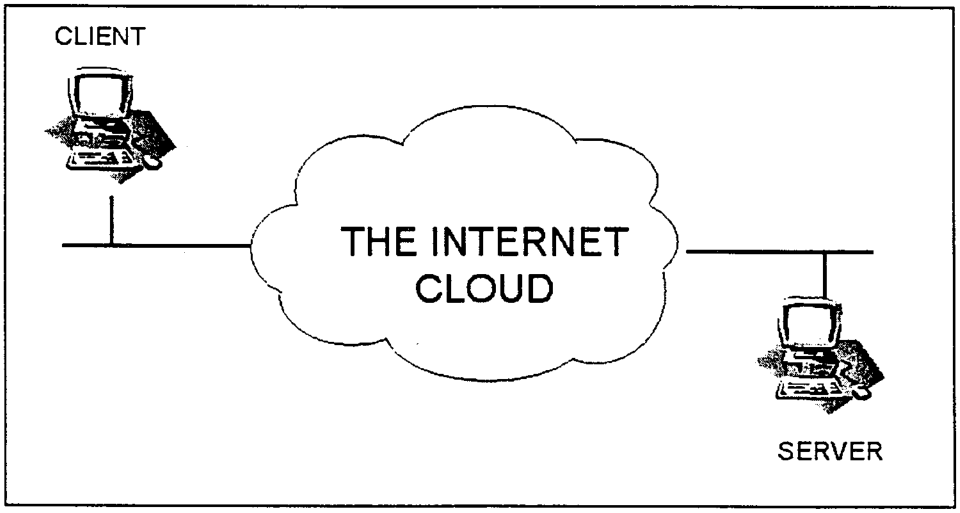


Figure 2. 1 - The Cloud

From the user's point of view it is almost magic. You just type something and all of the sudden you get a page with pointers to virtually millions of other places. The connection is like a cloud, or a black box. Users don't really know how the connection is done and even more they don't care how they are connected, they just want to get to the information.

A user connects through a client application to the cloud. The cloud routes the user's packet to the server, in which another application is running in an infinite loop (a daemon program) that communicates with the client application (See Figure 2. 1).

2.2 - Connecting computers (LAN and routers)

Assume you are a network engineer. You have been assigned the task of configuring the "**cloud**" of a building. The building has 300 computers and you have been assigned the task to connect all these computers. It is clear, you need a **network**, in order to interconnect all these computers. But, which network?

After doing some investigation you find out that one of the cheapest solutions is to connect all the computers by using **10 BaseT Shared Ethernet**. Also, you find out that you will need six Ethernet segments because of distance limitations. So, you start doing your work and you get a logical connection of the six Ethernet segments as shown in Figure 2. 1. The **client** is connected to one of these Ethernet segments and the **Server** is connected to another of these Ethernet segments in the basement of the building.

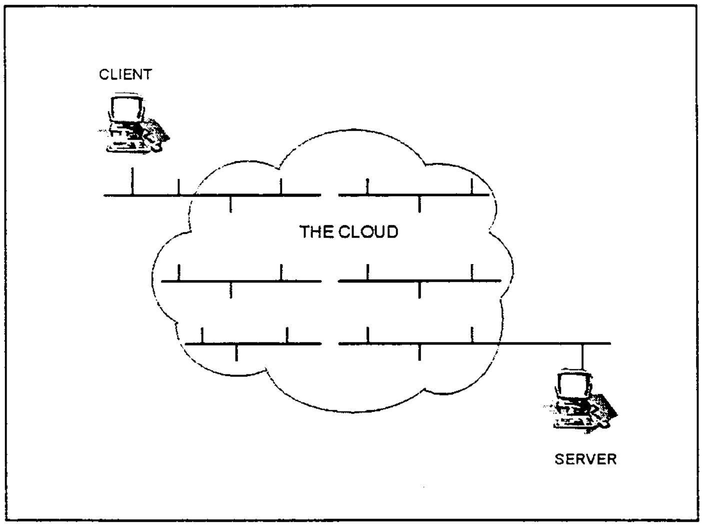


Figure 2. 2 - Six Ethernet Segments

2.3 - Interconnecting networks (routers and bridges)

The interconnection of these six Ethernet segments could be accomplished using a variety of networking equipment. The traditional options for internetworking include: repeaters, bridges, routers and gateways.

Repeaters regenerate electrical signals between segments. They don't understand the contents of the signals. Repeaters work at the physical layer of the OSI model. They are good to extend the distance of a network. You get a single network in terms IP address, noise and broadcast carrying across the network. ~~Bridges~~ are generally cheaper than other interconnection options.

Repeaters

A **bridge** interconnects two networks and filters layer 2 frames based on the MAC address providing isolation between segments. Bridges can't see the payload inside of a data-link frame and they create separate IP subnetworks. They filter noise and Ethernet collisions. They require minimal configuration. They connect similar data-link technologies (e.g., Ethernet-to-Ethernet, Token Ring to Token Ring, etc.).

Bridges do not scale because of two reasons: broadcasts and the spanning tree algorithm. They do not handle heterogeneity since they need the same MAC level address

A **router** interconnects several networks, using different data-link technologies. It works at the network Layer of the OSI model (Layer 3). They create separate networks using different IP addresses. They require considerable effort to configure.

A **gateway** acts as a translator between networks using incompatible protocols, such as TCP/IP and SNA or SNA and X.25. A gateway does everything a router does plus has the ability to convert upper layer protocols. For example, one of the most common types of gateways is a device between a local area network and a mainframe.

Of the four options to interconnect the six Ethernet segments, routers are a good choice that provides flexibility and allows better managing of the network. So, in order to obtain full connectivity around the building you decide that you will use a mesh with three routers, each of which will connect two Ethernet segments. The connections between the routers are going to be **serial point-to-point links**.

A **router** it is just another computer that has no hard drive and has two or more network interface cards ("**NIC cards**") (generally called **interfaces**). In our example, we will need three routers each equipped with two Ethernet interfaces and two serial interfaces. So, now we can connect the six segments using the configuration shown in Figure 2. 3.

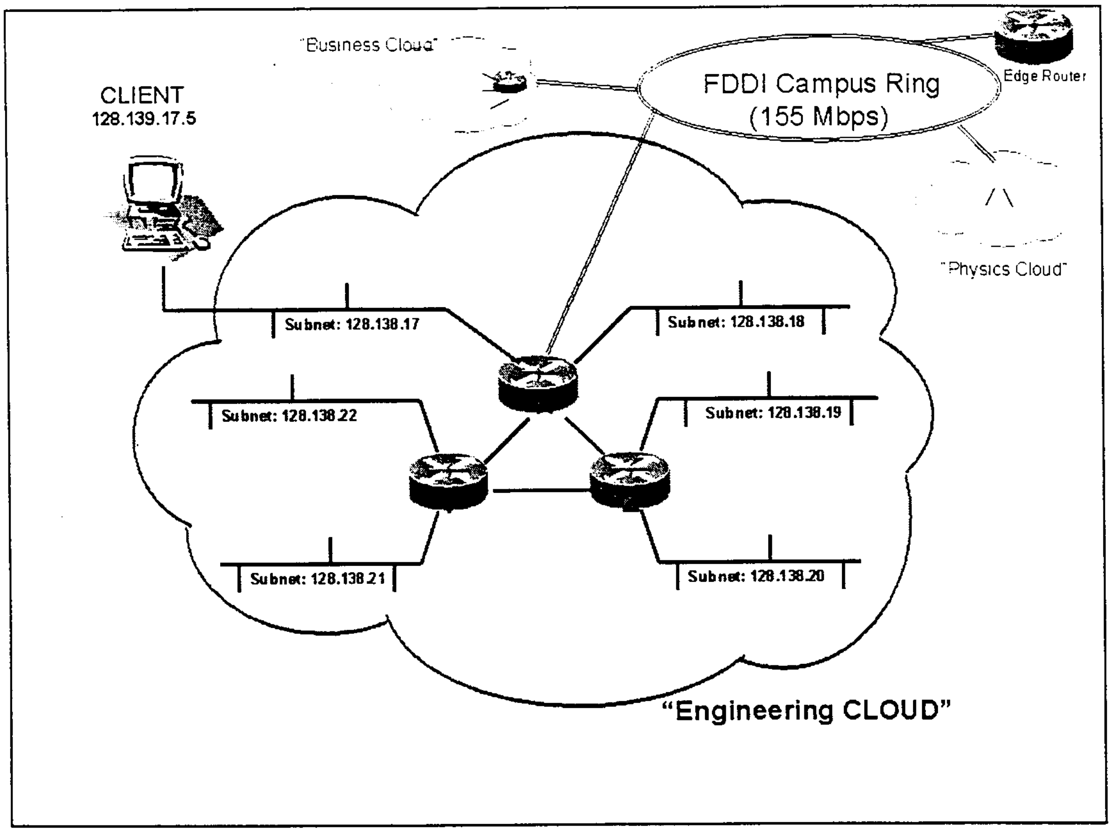


Figure 2. 3 - Router Mesh Connecting the Six Ethernet Segments

2.4 - Giving names to the devices (IP Addressing)

Ok, now you have a network. You have to assign unique names to each single **NIC card** that is on the network. For a host, there is generally going to be only one NIC card but for a router there is going to be two or more interfaces that you will have to name.

Each single interface has already been named by the manufacturer with a unique Medium Access Control address **(MAC address)**. But you need to give the interface another name. In the Internet world this name is a unique Internet Protocol address **(IP addresses)**.

In other words, every single interface in your network will have two identities (a MAC address and an IP address). Just like human beings that have a real name given by their parents and a social security number assigned by the government.

Your job is then to assign IP addresses to every single NIC card on the network. But, which address?. The answer is that you have to assign a network address that is in a block that has been pre-assigned by the **Internic**[1] to your network.

For example, here at the **University of Colorado at Boulder** we have been assigned a class B address (128.138). This gives us a block of 16,384 IP addresses to distribute throughout the university. This is clearly many more than we need for the 300 hundred computers in the business building. However, if we are not careful we can waste a lot of IP addresses if they are not properly assigned.

2.5 - Efficiently use of the address space assigned (sub-netting)

Before you start assigning addresses. Take into account that every single Ethernet segment has to be assigned the same network ID prefix on the IP address, so that the routers can work properly. An analogy to this name convention is that all the addresses of houses on the same street have to have the name of the street in it, so that the postman can find the houses properly.

Initially you were given the IP address block 128.138 and you can play with all the addresses that start with this **network ID**. With the address space you have been assigned you can play in a number of ways, for example:

a) One Big network with 16,384 hosts attached to it. This is not feasible using Ethernet because the maximum number of hosts per segment is 1,024.

b) Two networks each of 8,192 hosts. This is not very practical either.

[1] Internic is the organization that assigns IP addresses and DNS names (http://www.internic.net)

c) After several trials you decide a good solution may be to have 128 sub-networks each having a maximum of 128 hosts per segment.

The ability to take an IP address block and divide it into smaller segments of IP addresses is called **sub-netting**. You can divide the IP address block by using **a sub-net mask**.

The sub-net mask that does the trick of dividing your IP block into 128 networks each having a maximum of 128 hosts is 255.255.255.0. Now in all of your IP addresses the third byte on all your IP addresses is going to identify the sub-network that a particular host belongs (this is like "the street" on which this host is located from our previous analogy). The fourth byte is going to identify the host in that particular sub-network (this is like the house number on the street analogy). Figure 2. 4 illustrates the concept of subnetting.

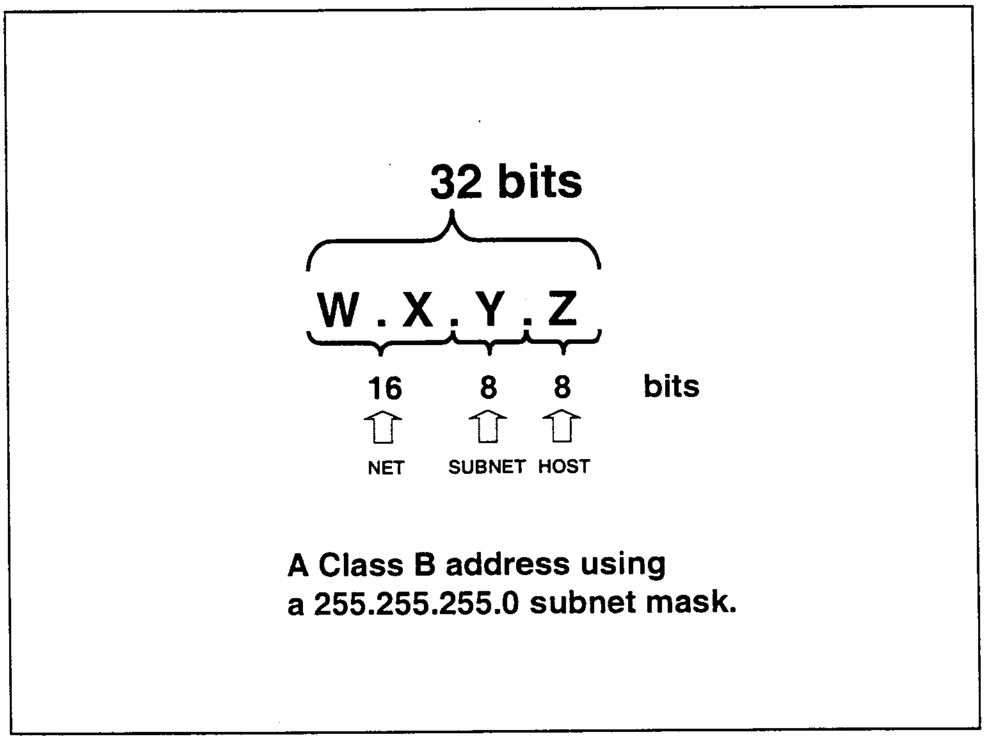


Figure 2. 4 - Sub-netting of a Class B Address Using the Mask 255.255.255.0

Let's name each of the six sub-networks. It is just as easy as assigning a number to the third byte of the IP address. At the end of this process we will have six sub-networks.

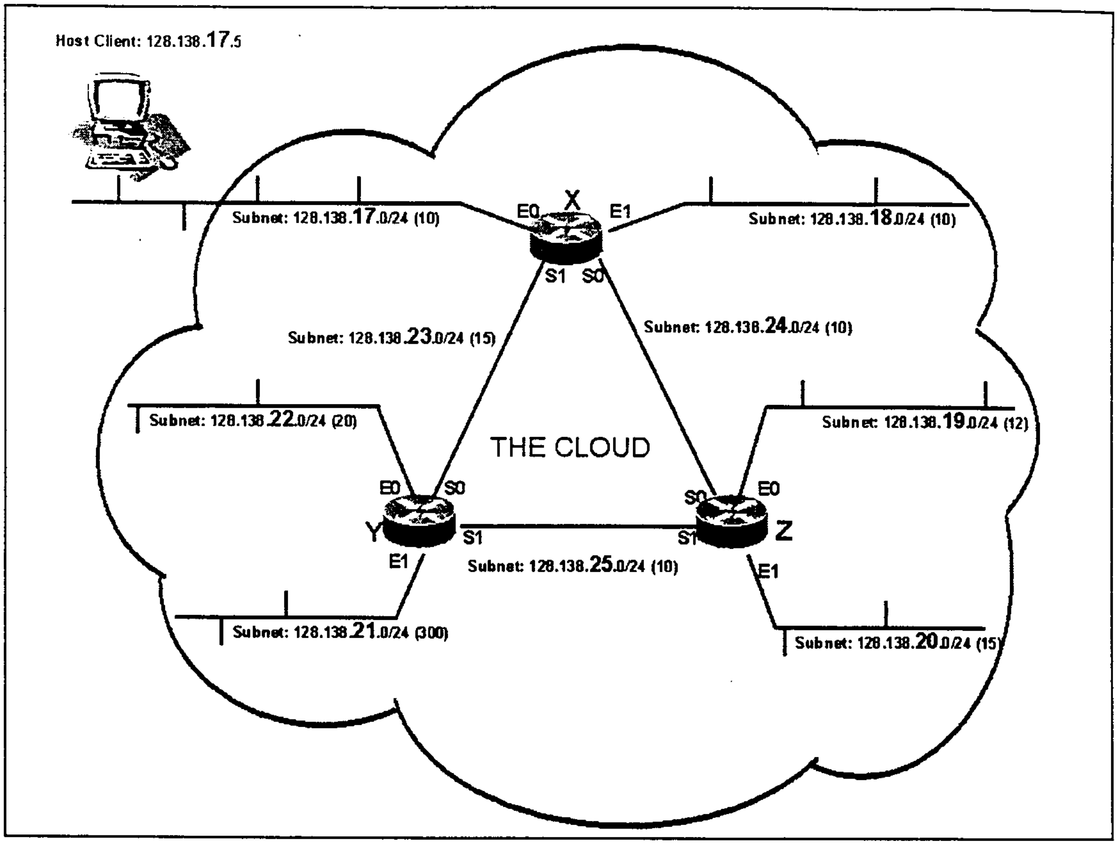

Host Client: 128.138.17.5
Subnet: 128.138.17.0/24 (10)
E0
X
E1
Subnet: 128.138.18.0/24 (10)
S1
S0
Subnet: 128.138.23.0/24 (15)
Subnet: 128.138.24.0/24 (10)
Subnet: 128.138.22.0/24 (20)
THE CLOUD
Subnet: 128.138.19.0/24 (12)
E0
S0
Y
S1
E1
S0
E0
S1
Z
E1
Subnet: 128.138.25.0/24 (10)
Subnet: 128.138.21.0/24 (300)
Subnet: 128.138.20.0/24 (15)

Figure 2. 5 - Sub-network Numbering.

Now we need to assign the IP addresses to every single interface on the network. Table 2. 1 shows the IP address assignments for the client, the server and their respective connections to the routers. Note that the Network ID and the subnet ID for all the interfaces that share a segment are the same. This is a key attribute that makes routing possible.

Interface	**IP address**
Client	128.138.*17*.5
Router X, Ethernet Interface Zero (E0)	128.138.*17*.1
Server	128.138.*20*.7
Router Z, Ethernet Interface One (E1)	128.138.*20*.1

Table 2. 1 - IP Address assignments for some of the interfaces of the network shown in Figure 2. 5.

2.6 - Sending IP Packets

The network is already built, but how do the packets get to their destination?

The client computer asks how do I get to www.yahoo.com? This, in Internet terms, is a "**Domain Name System (DNS)** Request". The DNS server comes back with the network ID address for the domain yahoo.com. If the DNS server on campus domain does not have an entry for this domain, then Colorado's DNS server sends a "DNS request message" to one of the nine root servers available on the Internet. Eventually, the DNS system will come back with the IP address for the domain yahoo.com.

Imagine sitting on the client machine trying to access a web page. The client wants to send an application message to the server so that the web page can be seen. In the case of a web surfer, this message would be a "HTTP GET message".

My computer needs to find the path to the server and send this packet in the right direction. My computer needs to find out whether the next hop towards the destination is another host in the same network - e.g., on the same Ethernet segment-- or a router that will further forward the packet.

More specifically, the question is: Is the 128.138.20.7 (server's IP address) in my own Ethernet segment? How does my computer know? It just does a comparison of its own network ID (including sub-net ID) with the network ID (including the subnet ID) of the destination IP address. It does the comparison using the sub-net mask assigned to my computer and a logical "AND" operation.

Local IP address:	128.138.17.5	Dest. IP:	128.138.20.7
Subnet mask	255.255.255.0	Subnet mask:	255.255.255.0
NetworkID+SubnetID:	128.138.17	Net ID + sub netId:	128.138.20

Since the two network IDs are different this means that the server is in another sub-network, which means the computer does not know how to reach this computer and it needs to send the packet to the router so that the packet may be forwarded in the right direction. The computer needs to know who is its default router to send this packet to the server. The default router or also called the **default gateway**, needs to be configured by the network administrator on each computer in the network.

If the two network IDs were the same then the server is in the same Ethernet segment that the computer. The computer will send the packet directly to the server. It does this by finding the MAC address of the server using the **Address Resolution Protocol (ARP).**

2.7 - Finding the path to the destination (Routing)

Now, we have a packet that has arrived at router X in Figure 2. 5. The router has a very simple forwarding task. It examines the **destination IP Address** of the packet and checks to see if that IP address is on the same network of one of its interfaces. For this particular example it will do the following comparisons:

The router will ask itself: Is 128.138.20.7 (server's IP address) in one of my direct attached networks (two Ethernet segments and two serial lines)?

For this particular example the answer is No. If the destination IP address were in the same network, then the router would use ARP to find the MAC address of the destination host and create the Ethernet frame with that MAC address as the destination address of the Ethernet Frame.

From the point of view of router X, It only has 4 options to send this packet:

Send packet through interface E0, if destination host is in this subnet.

Send packet through interface E1, if destination host is in this subnet.

Send packet through interface S0, to router Z.

Send packet through interface S1, to router Y.

If the destination IP address is not in any of the attached interfaces, then the router consults its **routing table**, to see which is the next router that is closest to the destination. In our particular example, the question the router will ask to itself is: Do I send the packet to router Y or do I send the packet to router Z?

But, how is the routing table created? A network administrator could eventually create the routing table manually for every router on its network (**static routing**). But this might be cumbersome if your network is huge because it requires a lot of manual configuration. So, a better approach is needed.

Routing Protocols create routing tables dynamically (dynamic routing), so that this is not an administration nightmare for the network administrator. There are many different types of routing protocols. The next chapter will cover the three most used routing protocols in use in today's Internet, which create routing tables "on the fly".

A routing protocol learns about the **topology** of the network dynamically and it stores all the knowledge about the network topology in the router's routing table.

2.8 - Connecting to other networks on campus

The engineering building connected alone has some value, but what makes the network really valuable is the ability to access information that is available on other networks around the world. In order to achieve interconnection between networks, a network manager will probably use a router. See Figure 2. 6.

Router Z in the engineering building needs to be connected to the campus backbone. The campus backbone is running on a FDDI ring at 155 Mbps. This means that we will have to configure Router Z to have a FDDI interface, and connect this interface to the ring.

On the campus ring are connected several other routers that are the point of connection of each campus building with other buildings and also with the outside world.

2.9 - Connecting the campus to exterior world

The most common type of connections to the exterior world available from ISPs or telephone companies are 56 Kbps lines, T1 lines (1.544 Mbps) or Frame Relay (up to 2 Mbps). A router cannot be connected directly to such external leased lines. A Channel Service Unit/Data Service Unit (**CSU/DSU**) is needed. This unit is used to convert the RS-232 signal coming from the serial port of the router into the correct framing to go over the leased line. In Figure 2. 6, Router W is connected to the outside world through a CSU/DSU. The router that connects to the exterior world is called an **edge router**.

A Channel Service Unit (CSU) is the first device that the external leased line encounters on the customer premises. The main function of the CSU is to protect the carrier's network from events that happen on the customer's network. Among other functions a CSU provides electrical termination, line conditioning, loopback tests and equalization for the telephone line.

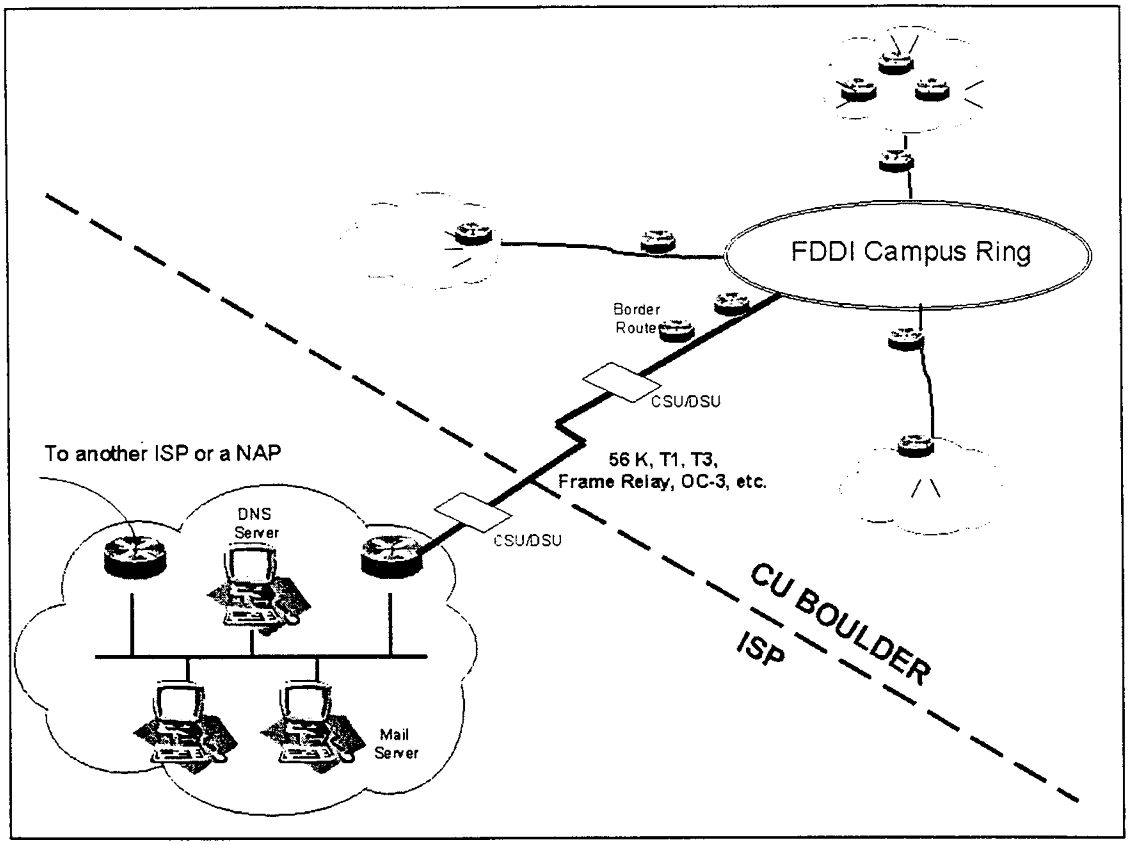


Figure 2. 6 - Connecting the Campus to an External ISP

A Data Service Unit (DSU) sits between a CSU and customer's equipment such as routers, multiplexers or terminal servers. The main function is to translate the customer's equipment signaling to the carrier's signaling. For example, a serial port of a router running RS-232 would be adapted to the framing and rates of a T1 line. Note that this translation requires buffers since RS-232 is asynchronous and a T1 is synchronous[2]

[2] Asynchronous means that the transmission is based on stop and start bits and the line is idle if nothing is being transmitted. Synchronous means that the transmission is based on clock synchronization and the line is continuously used whether there is information to send or not.

An ISP will charge a monthly recurring fee for the connection to its router, and thus to the Internet. On top of this charge a customer has to contract a line from its premises to the ISP's Point of Presence (POP). Typically, the connections available from the local Phone Company are either T1 lines or Frame Relay connections.

The router that connects to the exterior world is often called a **border router**.

2.10 - Interconnecting ISPs

Today's Internet is a collection of Internet Service Providers (ISPs) connected by Network Access Points (NAPs) (See Figure 2. 7). ISPs are also connected to each other in arbitrary ways by bilateral peering agreements. An ISP typically has a number of high-speed TCP/IP routers in a number of cities. All these routers are connected via leased high-speed data lines from long distance exchange carriers. The connection of these routers forms a national backbone. Typically, backbones are formed from 45 Mbps links (DS-3) or faster leased lines.

The Network Access Point (NAP) is an exchange point for Internet traffic. Internet Service Providers (ISPs) connect their networks to the NAP for the purpose of exchanging traffic with other ISPs. With the NSFNet, both directions of traffic used the same link. In today's network each direction may follow a different route and this complicates bandwidth reservation for quality of service purposes and also multicast routing.

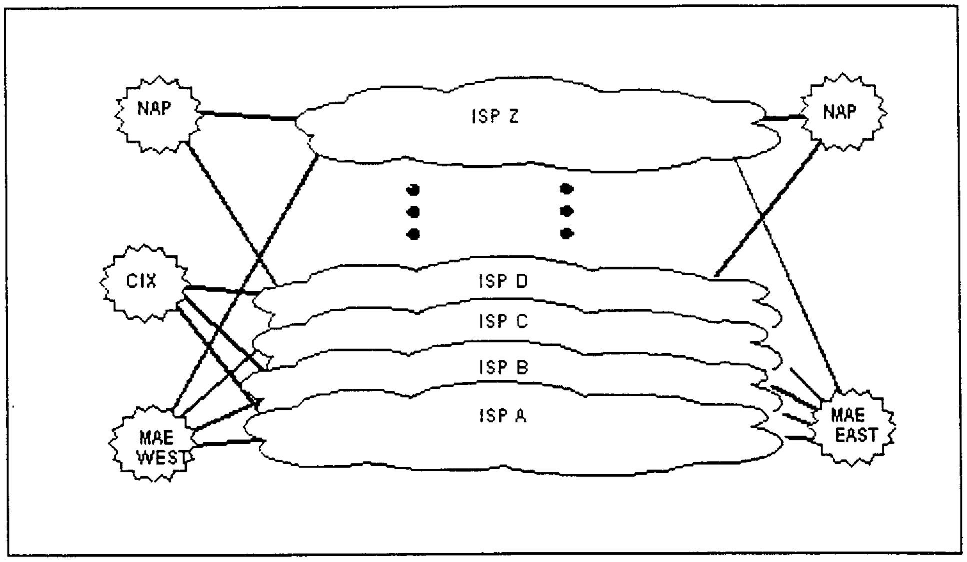


Figure 2. 7 - Interconnection of ISPs

The heart of today's Internet is based on four major official Network Access Points (NAPs) in the US:

- **MAE-West:** It is located in San Francisco, California and it is operated by PacBell.
- **Chicago NAP:** It is located in Chicago, Illinois and it is operated by BellCore and Ameritech.
- **New York NAP:** It is located in Pennsauken, New Jersey and it is operated by SprintLink
- **MAE-East:** It is located in Washington, DC and it is operated by MCI-WorldCom.

In each NAP there is a **routing server** that maintains a database of information regarding the issues of interconnection. In these NAPs, private backbone operators interconnect with each other. They can also establish bilateral relationships. This is the concept of **peering**. When a provider peers with another provider they agree to exchange traffic.

Peering agreements are bilateral agreements between individual ISPs. A peering agreement establishes a relationship between ISPs to allow each other's traffic to transit their backbones. Each ISP must negotiate independently and needs to contact every ISP they wish to peer with individually.

Besides the four official NAPs, plenty of ISPs peer privately, bypassing congested public peering points to prevent their Internet traffic from bogging down. These intermediate interconnections allow them to have better routes between two points and avoid bottlenecks. Peering reduces the traffic on a provider's network since packets can be routed to the other provider's network. This concept is also known as "**hot potato routing**".

2.11 - A practical example

In order to illustrate today's structure of the Internet consider a connection to Yahoo's web site (See Figure 2. 8). Suppose a browser at the University of Colorado makes a request to Yahoo's web site. Packets from this session need to transverse networks owned by three different transit ISPs plus the internal networks of the University and Yahoo's company. From the traceroute readings it seems that Qwest and MCI have a peering arrangement to exchange traffic at Denver, CO and MCI and GlobalCenter have a peering arrangement to exchange traffic at San Jose California. None of the packets seems to be exchanged at public Network Access Points (NAPs).

```
>traceroute www.yahoo.com

traceroute to www7.yahoo.com (204.71.200.72) 30 hops max, 38 byte packets
 1  128.138.129.1 (128.138.129.1)  1 ms  1 ms  1 ms
 2  cuatm-gw.Colorado.EDU (128.138.138.10)  1 ms  1 ms  1 ms

 3  ncaratm.inet.qwest.net (204.131.62.14)  2 ms  1 ms  1 ms

 4  border2-hssi1-0.Denver.mci.net (204.70.29.5)  7 ms  23 ms  8 ms
 5  core1-fddi-1.Denver.mci.net (204.70.3.113)  4 ms  4 ms  5 ms
 6  core2.Sacramento.mci.net (204.70.4.49)  28 ms  77 ms  61 ms
 7  border8-fddi-0.Sacramento.mci.net (204.70.164.67)  38 ms  27 ms  26 ms
 8  globalcenter.Sacramento.mci.net (204.70.123.6)  36 ms 34 ms   35 ms

 9  fe1-0.cr1.SNV.globalcenter.net (206.251.5.12)  35 ms 34 ms  34 ms

10  www7.yahoo.com (204.71.200.72)  35 ms (ttl=240!)  39 ms (ttl=240!)  35 ms (ttl=240!)
```

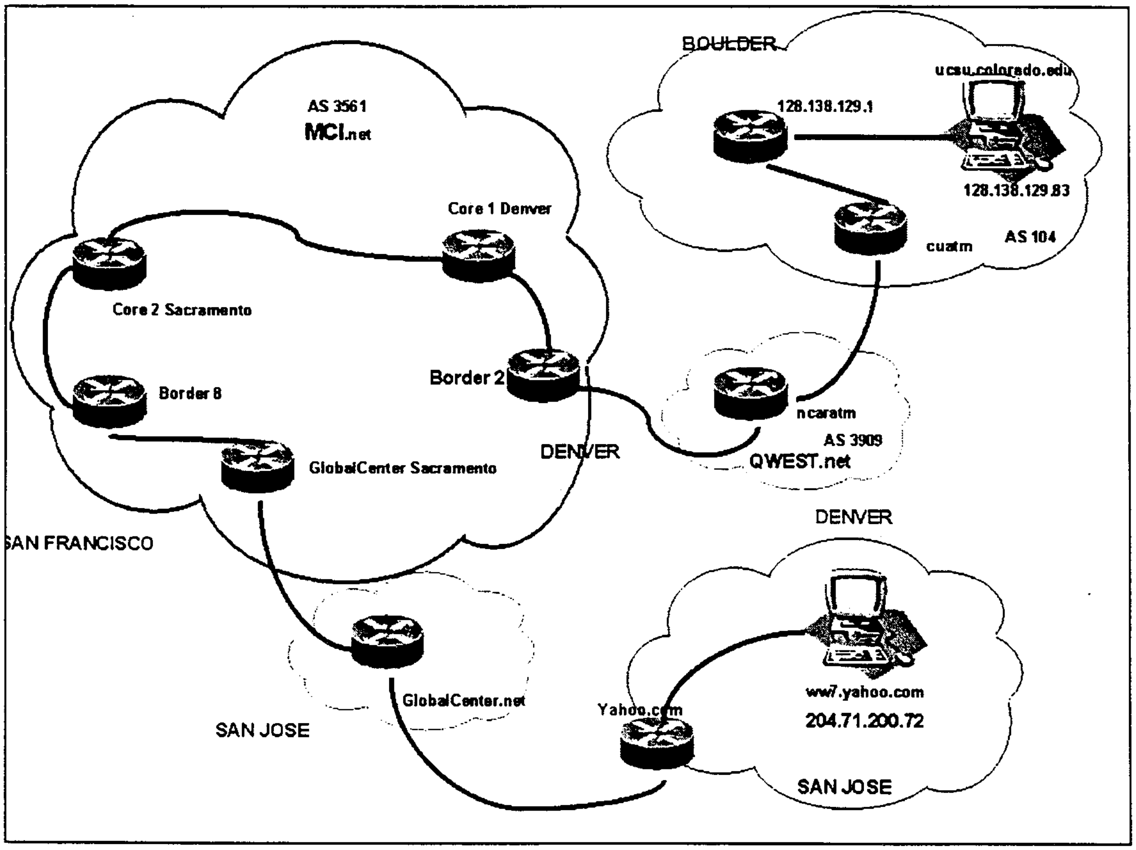


Figure 2. 8 - ISPs Transversed to Get to Yahoo's Web Site

2.12 - Chapter Summary

This chapter has presented a brief overview of what are the main components of today's Internet and how it is interconnected. A clear understanding of the Internet topology is basic to understanding routing protocols, which is the main subject of this thesis. If more background is needed the reader is directed to [Tanenbaum-96] or some other computer communications book.

The topology of how internal networks are connected to the Internet is important in understanding the issues in inter-domain multicast routing. Usually organizations are connected to the Internet through a leased line such as T1 line. This leased line will typically come from an Internet Service Provider (ISP), which also needs connection to other networks on the Internet. ISPs interconnect between them in Network Access Points NAPs. NAPs are very congested in today's Internet.

The next chapter covers the main intra-domain protocols used for unicast routing: RIP, OSPF and BGPv4.

Chapter 3 - Unicast Routing

Several multicast routing proposals are based on extensions of existing unicast routing protocols. For example, the Distance Vector Multicast Routing Protocol (DVMRP) [Waitzman-88] is based on the Routing Information Protocol (RIP).

A clear of understanding of unicast routing is essential before the extensions of these protocols for multicast are introduced in the following chapters. This chapter introduces the three major protocols using in TCP/IP networks for unicast routing (RIP, OSPF and BGP). It also describes how a router works. Figure 3. 1 shows the location of these 3 protocols in the TCP/IP stack.

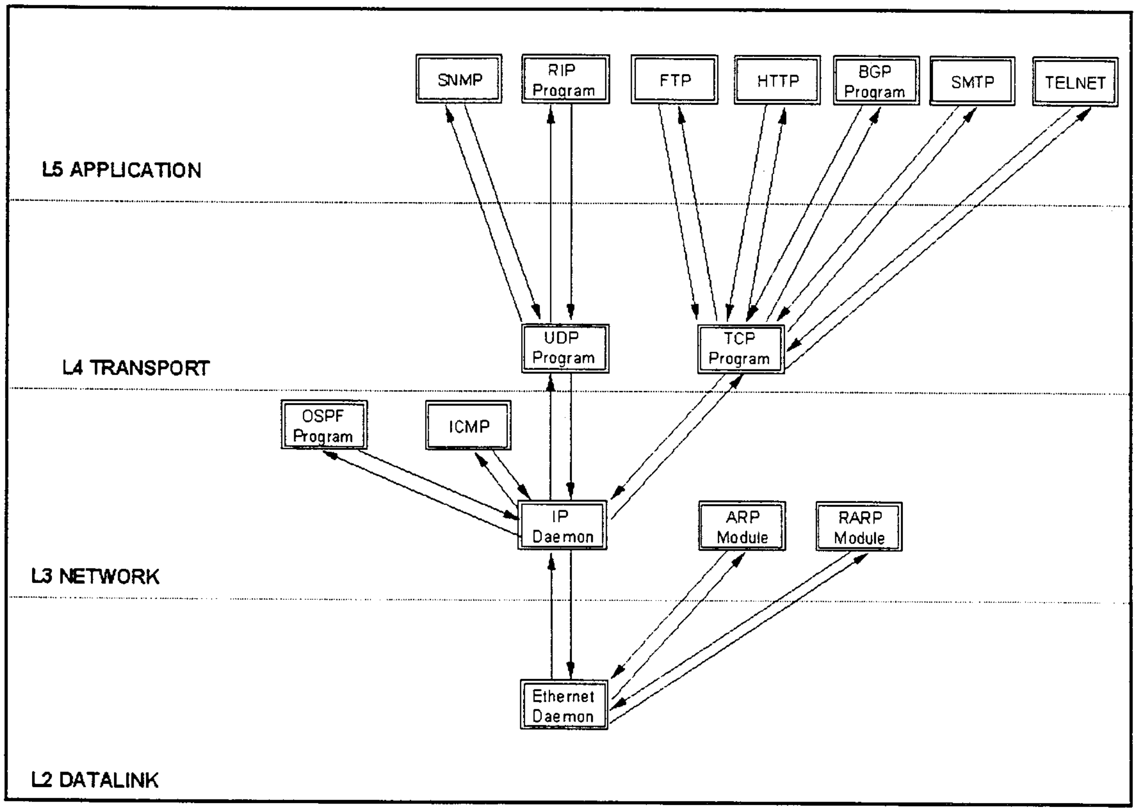


Figure 3. 1 - Routing Protocols and the TCP/IP stack

3.1 - Intra-Domain vs. Inter-Domain Routing

The Internet is a collection of **Autonomous Systems (ASs)**[1]. Autonomous systems run intra-domain routing protocols such as RIP and OSPF within their boundaries. ASs interconnect with each other using an inter-domain routing protocol. The most common inter-domain routing protocol is the BGP-4.[2] (See Figure 3. 2)

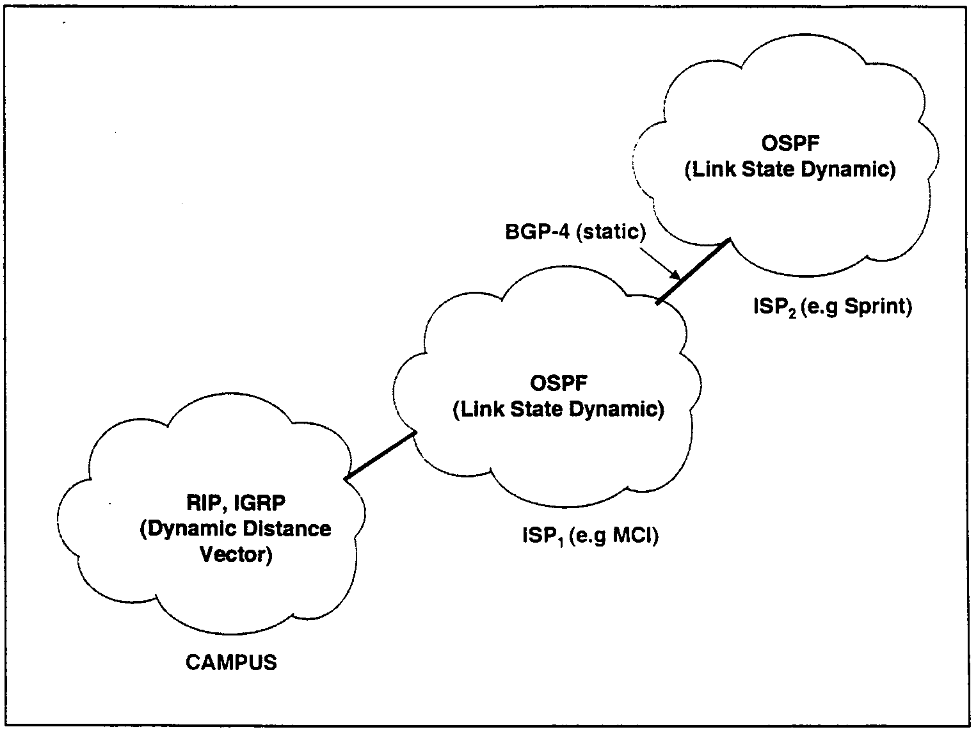


Figure 3. 2 - Unicast Routing Protocols used in the Internet

[1] An autonomous system is an internetwork that is under the control of a single authority e.g. a university, a corporation, etc.

[2] Intra-domain protocols are also known as Interior Gateway Protocols (IGP). Inter-domain routing protocols are also known as Exterior Gateway Protocols (EGP).

Each AS is assigned a unique 16 bit identifier (the AS number). For example, the University of Colorado at Boulder is AS 104.[3] AS numbers range from 1 to 65,535. The ASN is a number that uniquely identifies a organization. Some of the most common ASN are the ones of big ISPs (See Table 3. 1). There were around 700 registered ASs in the Internet at the beginning of 1994. As October of 1998, there are 11,554 Autonomous Systems.[4]

ASN	ISP
3561	MCI
1239	Sprint
701	UUNET
174	PSI
1673	ANS
1	BBN
4200	AGIS
4969	Net Access

Table 3. 1 - Common ASN numbers

An internal router usually runs one intra-domain routing protocol, while an external router needs to speak an internal routing protocol plus an external routing protocol. The purpose of these routing protocols is to adapt the routes to the changing conditions of the network. Before defining the protocols that a router runs let's describe how a router works.

[3] This information can be obtained with the `whois` command.

```
>whois -h whois.arin.net "ASN 104"
```

A web interface is available at: http://www.arin.net/whois/arinwhois.html

[4] As October 1, 1998 there were 11,554 registered ASN. This number was obtained by typing this command:

```
>whois -h arin.net "as 11554"
```

3.2 - What does a router do?

A router is a computer with more than one Network Interface Card (NIC), generally called interfaces. These interfaces connect to networks that support various physical and data link technologies such as Ethernet, Token Ring, FDDI, BRI, ATM and serial point-to-point links. These networks are referred to as its "**directly-connected networks**". Each of the networks has an IP address and a network mask associated with it.

A router enables communication between networks that are not directly connected. The main job of a router is to **forward packets** coming from an input interface to the appropriate output interface(s). [5] This forwarding decision is based on the IP destination address of the incoming packet and the information in a routing table stored on the router.

The routing table contains a list of destination networks and how to reach them. The destination network could be directly connected or not. If the destination is directly connected then it's just a matter of creating a level 2 frame and sending the frame to the destination through the interface specified on the routing table. A router that does not have a direct physical connection to the destination checks its routing

[5] In the case of unicast routing the packet will be forwarded only in one interface. In multicast routing the packet could be forwarded in more than one interface. For packet switching (e.g. the TCP/IP) the decision of which interface a packet should go out on is made per packet, since each packet contains source and destination addresses and moves independently through the network. While for virtual circuit networks (e.g. X.25) the routing decision is done with the first packet and all subsequent packets follow that route.

table and forwards the packet to the "next hop router". This forwarding mechanism enables an IP packet to find its destination through the network.

The routing table is created as the result of the communications between routers. Routers communicate routing update messages or link state advertisements (LSAs) (discussed later in this chapter). The process by which routing tables are built is known as **routing**.

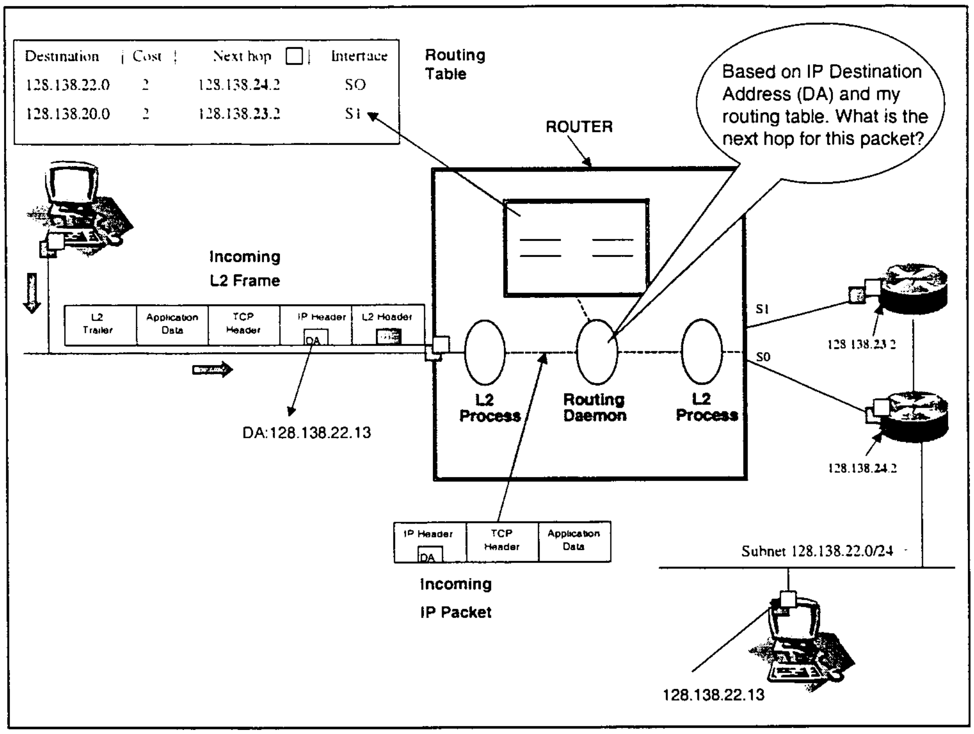
The forwarding action is performed by a program that runs continuously in routers that is called **routing daemon**. The Unix BSD implementation of a routing daemon is called **GateD** [6]. This program implements various routing protocols such as OSPF, RIP, HELLO, BGP, ISIS, etc. GateD is designed to handle dynamic routing with a routing table built from information exchanged by routing protocols.

Figure 3. 3 describes how a router works. The routing daemon makes a routing decision based on information stored in a **routing table**. This table tells the router how to forward packets. Routers create routing tables using **routing protocols**, which is the main topic of this chapter.

When a router receives a layer 2 frame, it strips down the header and the trailer of the data-link layer frame and gives packets to the routing daemon. This process examines the packet's IP destination address and it determines that it either knows or does not know how to forward the packet. If it knows how to forward the packet, it changes the destination physical address (L2 addresses) to that of the next

[6] For more information on Gated visit: http://www.freebsd.org/cgi/ports.cgi?query=gated. Source Code for GateD is available at: ftp://ftp.gated.merit.edu/net-research/gated/gated-3-5-10.tar.gz. A description of GateD is available at: ftp://ftp.gated.merit.edu/net-research/gated/gated.ps.gz

hop and transmits the frame. If the router does not know how to forward the packet then it typically drops the packet.



Destination	Cost	Next hop □	Intertace
128.138.22.0	2	128.138.24.2	SO
128.138.20.0	2	128.138.23.2	S1

Figure 3. 3 - A router routing a packet

If the next hop is the destination host, then the packet has arrived at its destination and we are done. If the next hop is another router then the process is repeated and the packet is forwarded somewhere else. This procedure of forwarding packets only knowing what is the next hop is known as the **hop-by-hop** nature of IP routing.

Note that in all this forwarding process the destination IP address remains untouched as it transverses networks, while the MAC address is changed as it goes from one network to the other. (See Figure 3. 4)

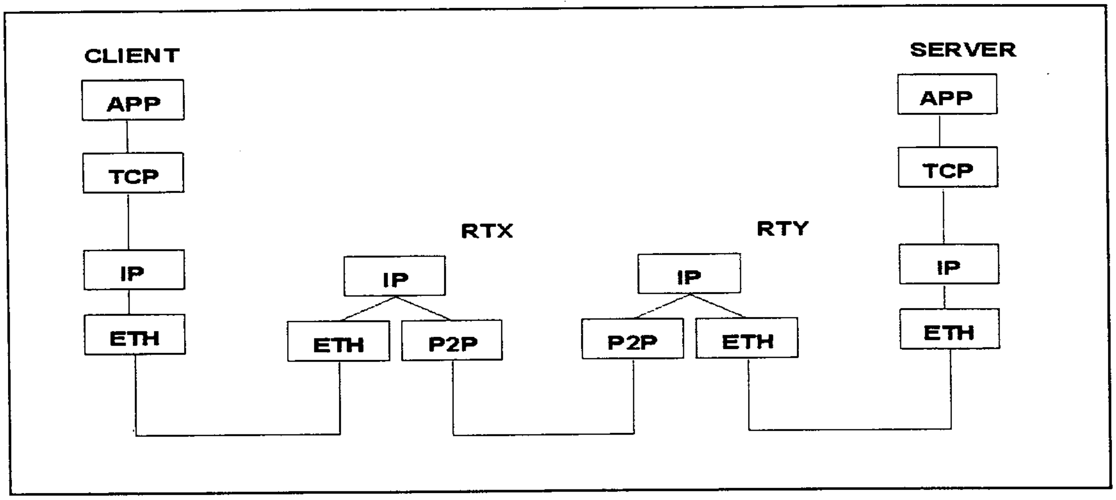


Figure 3. 4 - A packet traversing networks.

Routing tables can be created statically or dynamically. If a network administrator establishes the route packets should follow manually then it is called **static routing.** If the routing table is updated by routing daemons based on changes in topology and traffic then it is called **dynamic routing**.

Static routing is fast and sometimes is the best routing solution. However, it does not adapt to topology changes such as node or link failures, addition of new nodes or links or congestion in links. Also, as the size of the networks grows, the task of maintaining this routing table becomes unmanageable.

Dynamic Routing allows creating routing tables on the fly. Routing tables are updated as a result of changes in network topology. The following section covers a basic introduction to the two main intra-domain dynamic routing protocols: RIP and OSPF.

3.3 - Routing Information Protocol (RIP)

3.3.1 - Overview

RIP is intended for use as an intra-domain unicast routing protocol in small networks. The protocol is limited to networks whose longest path involves 15 hops. RIP uses the hop count as a metric. This means that RIP does not consider bandwidth or congestion when choosing a route. RIP is based on the fact that it is possible to calculate optimal routes for the entire system by using information directly from neighbor routers.

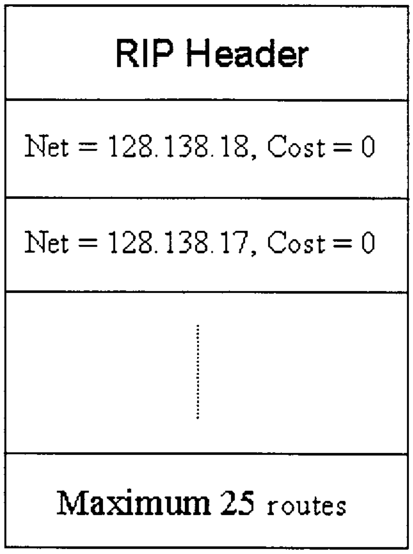
The Routing Information Protocol (RIP) [Hedrick-88] has been one of the major dynamic routing protocols employed in the Internet, due in large part to the fact that it is included in Berkeley UNIX in a program called `routed` (pronounced "route dee"). RIP automatically creates and maintains a routing table.

3.3.2 - RIP Advertisements

Routers running the RIP protocol communicate with each other using **RIP advertisements** (called **response command** in [Hedrick-88]). These update messages describe the routing table as it currently exists in that router. A router sends these updates at *regular time intervals*, usually every 30 seconds; a router also sends an update message whenever an update from another router causes it to change its routing table (these are called **triggered updates**). These advertisements are the mechanism with which routers learn about network topology. Figure 3. 5 illustrates the format of a RIP advertisement.

The IP source address of a RIP update is the IP address of the interface of the router that is sending the update in this network. The source address is then taken by the neighbor router as the "next hop" column of the routing table for a particular entry.

RIP runs on top of UDP and it listens only to the UDP port 520. The maximum datagram size for RIPv1 is 512 octets. That allows for RIP advertisements to carry up to 25 destination networks in the payload of the IP packet. Multiple RIP advertisements are used if nodes have larger routing tables.



RIP Header

Net = 128.138.18, Cost = 0

Net = 128.138.17, Cost = 0

Maximum 25 routes

Figure 3. 5 - RIP Advertisement (RIP response command)

3.3.3 - Algorithm

The main idea behind RIP is a per router exchange with other directly-connected routers of global routing tables. These updates can be periodical or triggered by changes.

The following is a brief description of how RIP works:

1) When a router is turned on, it sends a **request command** throughout its interfaces. This command asks neighbor routers for information in order to populate

its routing table[7]. Normally requests are sent as broadcasts, from UDP source port of 520.

2) Each router keeps a table with an entry for every possible network destination in the domain. The table contains the next router interface and the metric for that route. The table also contains timers and flags per each entry. A router times out a route when it hasn't heard from the router from this route for a certain amount of time. The default for this timer is 180 seconds (6 times the advertisement interval).

3) Each router advertises periodically the contents of its entire routing table to its neighbor routers. RIP updates contains destination network addresses and the metric for each destination. RIP advertisements are also sent when there is a change in the routing table. These advertisements are sent as broadcast (e.g. Destination address 255.255.255.255).[8]

4) When a router receives a RIP advertisement it compares all the routes advertised by its neighbor node with its routing table. For each advertised route it performs the following algorithm:

If the entry is not in router table then it adds the entry. It sets the destination to the address advertised. It adds one to the metric advertised to account for the link just transversed. The next hop column is populated with the IP address of the router

[7] Request messages are also used by diagnostics programs to get all or parts of the routing table.

[8] The fact that RIP update messages are broadcast forces every host in the network to process these broadcasts. RIPv2 uses multicast so that only RIPv2 routers listen to RIP advertisements.

interface that sent the RIP update. It initializes the timeout timer. It sets the route change flag and the router sends a triggered update.

If the entry is in router table and the entry came from the neighbor router, then the cost is compared to the new entry with the cost of the entry in the routing table.

If cost of new entry is less than cost of routing table entry then take this entry and replace the old entry with the new one (We have found a shorter path to this particular destination). Also, start the timeout timer and change the route change flag so that a trigger update will be sent to neighbor routers. Otherwise, discard the advertised entry.

5) The algorithm converges once every node has a complete routing table. That is, the routing table has entries for every possible network in the domain. We can say that each router has discovered the network topology dynamically.

Every single router runs this algorithm on the network. After several advertisements every single router will have created a routing table. This routing table has an entry for every single network in the autonomous system that says what is the next router towards a particular destination network and how many hops away this network is located. Also, the router's physical interface to use, timers and flags are kept per each entry. The routing table is updated according to RIP updates received from directly connected routers. Figure 3. 6 illustrates a typical RIP routing table.

Destination	**Next Router Interface**	**Number of Hops**	**Physical Interface**	**Timers**	**Flags**
128.138.17	Directly Connected	2	E1	t1,t2,t3	U
128.138.18	Directly Connected	4	E0	t1,t2,t3	-
128.138.21	128.138.13.1	2	S0	t1,t2,t3	U

Figure 3. 6 - Typical RIP Routing Table

RIP maintains only the best route to a destination. When new updates provide a better route, then the information in the update replaces the old route information.

3.3.4 - Stability Features

RIP specifies several techniques to deal with network topology changes such as link failures, router additions, etc. These include hop limit count, hold-downs, split horizon and split horizon with poison reverse.

The **hop limit count** establishes a maximum of 16 for the cost of a particular route. Once the hop count reaches 16 (Infinity), then the route is eliminated from the routing table. This technique eliminates routing loops created by link or router failures.

Hold-downs tell routers to not change a route that has recently been updated from a RIP advertisement coming from another router. The idea is to prevent the propagation of incorrect information due to the fact that it takes time to propagate network changes throughout the network and routing updates may contain contradictory information.

The idea of **split horizon** is that when a router sends a routing update to its neighbors, it does not send those routes it learned from each neighbor back to that neighbor. The idea is that it does not make sense to claim reachability for a destination network to the neighbor from whom the route was learned. The split horizon rule helps to prevent two node routing loops.

Split Horizon with poison reverse sends the routes back to its neighbors with a cost of infinity so that that route will not be chosen. Poison Reverse updates are intended to defeat larger routing loops. In general, it is better to use poison reverse since it breaks routing loops quicker.

3.3.5 - Improvements for RIPv1

It is undesirable for the update messages to become synchronized, since it can lead to unnecessary collisions on broadcast networks (e.g. Ethernet). Sally Floyd and Van Jacobson proposed a solution for this problem [Floyd-93].

For non-broadcast interfaces such as Frame Relay or X.25, Gerry Meyer proposed a solution based on acknowledged transmissions of updates and the use of only triggered updates [Meyer-94].

Support of multiple metrics is a desirable feature for RIP since the hop count does not distinguish between interfaces of different speed or congestion.

Researchers have proposed the use of "source-tracing" to avoid routing loops that are created in RIP [Cheng-89] [Rajagopalan-89]. However, their proposals require RIP to change in a way that makes it not backward compatible with older versions of RIP. For this reason, their proposals have not been standardized.

Increasing infinity to a larger number than 16 has been proposed several times. Again, this cannot be accomplished since it is not backward compatible with older implementations of RIP.

3.3.6 - RIPv1 is obsolete (RIPv2 and RIPng)

RIP Version 1 [Hedrick-88] has been declared a historic protocol [Halpern-96]. The main reason to declare it obsolete is that it does not support CIDR (see section 3.9). That is the RIP routing table does not contain a prefix mask. The prefix length is inferred from the address. The lack of subnet masks is a serious problem for routers since they need a subnet mask in order to forward a packet.

Another problem that RIP has is that is limited to small domains (diameter of 16 hops or less). The reason for this limitation is that sometimes RIP will have to count to infinity (16) in order to purge bad routes. This delays the convergence of routing. So, if you increase the value for infinity then the convergence time is going to be even poorer.

RIPv2 [Malkin-94] overcomes the limitations of RIPv1. RIPng [Malkin-97] is an extension of RIPv2 to support IPv6. The following paragraphs briefly describe each of them.

RIPv2

RIPv2 was defined for the first time in January of 1993 [Malkin-93], and the draft was re-issued in November of 1994 [Malkin-94].

RIPv2 includes a mechanism for authenticating the sender of the routing information and it expands the information carried in RIP advertisements. RIPv2 uses multicast to reduce the overhead imposed by RIPv1 broadcasts. The main contribution of RIPv2 is that it now can be used in environments using VLSM and CIDR.

The entries in RIP update messages used to contain a destination network and a metric. These route entries now carry additional information: a route tag field, a subnet mask and a next hop address.

The **route tag** field provides a mechanism of separating RIP routes for this domain from "external" RIP routes, which may have been imported from an EGP or another IGP. For example, routes imported from BGP-V4 will have the Autonomous System number as the route tag, indicating that this route was learned from an external autonomous system.

The **subnet mask** is now included on RIP update messages. This permits the support of CIDR.

The **next hop** allows for optimization in an environment that uses multiple routing protocols. This field is included to populate the next hop field of the routing table with an address other than the source of the originator of the RIP update message. If this value is 0.0.0.0, then RIPv1 works as RIPv2, that is the next hop field is the originator's IP address of the RIP advertisements received. By specifying the Next Hop Field it is possible to eliminate some hops and achieve a better path.

Multicasting to only RIPv2 routers instead of broadcasting to all hosts and routers in a subnetwork reduces the load on hosts or routers not running the RIP protocol. Multicasting to the group of RIPv2 routers allows exchange of information that RIPv1 routers would not understand. This is useful since a RIPv1 router may misinterpret a route because it does not have its corresponding subnet mask.

RIPv2 is **still limited to small networks** with diameters less than 15. The counting-to-infinity method is still used to break routing loops.

RIPng

RIPng supports IPv6-based networks. It is limited to use in small networks because of the hop limit count of 16 as RIPng and RIPv2. RIPng is defined in [Malkin-97].

Instead of passing a subnet mask as in RIPv2, RIPng passes the **IPv6 prefix length** (the length in bits of the network number). The next hop field, not present in RIPv1 update messages which was added in RIPv2, was eliminated for RIPng. However, the functionality was preserved and it is now done with a **new next hop mechanism. Authentication** has been removed from the RIP update messages since IPv6 has built in security.

The size of the payload of a routing update message is no longer arbitrarily limited to 512 bytes. It is now limited by the MTU of the transversed network minus the IPv6 (40 bytes), UDP (8 bytes) and RIPng (4 bytes) headers. The number of entries that might be transmitted is then this number divided by the size of route entry (20 bytes = 16 bytes of IPv6 address + 4 bytes for route tag, prefix length and metric). For example, for an Ethernet the number of routes that can be transmitted in one RIP advertisement is about 72 routes.[9] This number is a lot more than the maximum of 25 routes permitted in RIPv1 and RIPv2. This is a performance enhancement that allows having less overhead in the network.

The routing table is composed of the IPv6 prefix destination network, the prefix length, the metric (sum of the cost to get to this destination), IPv6 address of the next router along the path (i.e. the next hop), the route change flag and various timers associated with the route.

RIP advertisements are sent either as a broadcast in the network or as a multicast to the entire RIPng routers group. RIP advertisements have a 4 byte header plus a number of routing entries. These routing entries are composed of three parts: the IPv6 address of the destination, the prefix length, the metric and a route tag.

The next hop field is not a separate field of a route entry of the update message as it was in RIPv2. Normally, the next hop field is the IP address of the router that advertised this routing update message.

The next hop is specified as a special entry of the update. It is identified by the value of 0xFF in the metric field of a route entry. Subsequent route entries use the next hop value advertised as the value to use to populate its next hop field for

[9] (1500-40-8-4)/20 ~ 72 route entries.

this destination network. By using the next hop field in RIP advertisements, packets will follow better routes when different routing protocols are used in the domain.

RIPng is also an application program that runs on top of UDP. In this case it uses port number 521.

3.3.7 - RIP Summary

RIP is simple to implement by software developers and it is also easy to configure. This is the main reason for its success. It is useful in small network environments. The routing loops formed in RIP are somewhat solved with the use of split horizon or triggered updates, but they still do happen and last for minutes causing problems in the network.

In RIP, each router broadcasts (destination address network, metric) pairs to that router's neighbors. RIP uses the hop count from a source to a destination as the metric. A hop is a subnetwork linked to routers. RIP selects as the best route the one that has the smallest metric (i.e., the least number of hops). It does not take into consideration congestion, bandwidth or delay.

RIP version 1 did not have a provision for passing around a network mask. The mask needed to be inferred based on whether the address is class A, B or C. RIPv1 supports fixed length subnet masks. The mask is always the same for a given network number. RIPv2 added the ability to support VLSM and CIDR.

3.3.8 - More information on RIP

The reader is directed to the following references for a more in depth coverage of RIP: [Huitema-93] [Hedrick-88] [Malkin-94] [Malkin-97].

3.4 - RIPv1 Shortcomings

RIP, although popular, has many shortcomings. A few of the most notable are:

- Lack of Support of CIDR
- Only works in small network topologies
- Limited metric use
- Slow Convergence Time
- Lack of security

3.4.1 - Lack of Support of CIDR

A major drawback of RIP is that it does not support Variable Length Subnet Masks (VLSM) which is the basis of the Classless Inter-Domain Routing (CIDR)[10]. RIP does not propagate subnet mask information on its routing updates. A router will apply its locally defined subnet mask to the IP destination address of a packet that needs to be forwarded, and it will get the wrong network ID to look up into the routing table. As a result, the packet will be incorrectly forwarded.

In other words, RIP only supports the classfull approach for IP address making the protocol a bad choice, because it does not efficiently use the address space available.

RIP domains are considered flat networks. This means that all routers are at the same level, which implies that is not possible to aggregate routes. In other words, each router needs to be aware to every single router on the network by having state on its routing tables. This approach clearly does not scale for large networks and even less for the Internet.

[10] CIDR is covered on section 1.9

RIP version 2 [Malkin-94] has added support for VLSM and CIDR, but it still carries all the other deficiencies described in this section.

3.4.2 - Only works in small network topologies

RIP uses the number of hops as the metric to choose between routes. It does not allow for routes beyond 15 routers. In large internets, such as university and corporate environments, many routers are required to connect multiple network segments together. In these types of networks it is often a problem for network administrators to guarantee that the 15-metric barrier will not be exceeded. This problem is also known as the **count-to-infinity** limitation.

3.4.3 - Limited metric use

RIP does not factor the speed of a link or "cost" into route computation, since RIP advertisements do not transmit the speed of a particular interface. This lack of information often results in RIP making sub-optimal routing decisions due to lack of information. For example, a low-speed link could be chosen as the preferred route over a high-speed link, since the routing decision is only based on the number of hops or links between routers.

Since it uses a fixed metric to compare alternate routes it does not take into consideration real-time parameters such as measured delay, reliability or load.

3.4.4 - Slow convergence time

Slow Convergence time means that there is a delay in responding to network changes. Convergence refers to the point in time at which the entire network becomes updated to the fact that a particular route has appeared or disappeared.

RIP works on the basis of periodic updates and hold-down timers. If a route is not received in a certain amount of time, the route goes into hold-down state and gets aged out of the routing table. The hold-down and aging process translates into minutes in convergence time before the whole network detects that a route has disappeared.

RIP's method of table updating contributes to delays in responding to network changes. A RIP router forwards its entire routing table to its neighboring router(s). The receiving router must process that table by comparing each entry with entries it currently has. In large networks this task is CPU and memory intensive. This computation is done on every routing table exchange every 30 seconds, whether the tables have changed or not. The receiving router cannot know if a change in the network has occurred until it has completed the table comparison.

The slow convergence time in RIP creates routing loops. A routing loop occurs when a packet travels in circles between two routers. This happens when there is a misconfigured router, or while the routing tables have not been stabilized after changes in topology of the network.

3.4.5 - Lack of security

RIP does not have any security features defined that make the protocol a bad choice when internal sub-networks need to be defined. If a domain has internal areas, then routing updates need to be authenticated in order to prevent mis-configuration from routing updates coming from other areas.

Also, the lack of security presents a possible point of attack from hackers that could advertise wrong routing information.

3.5 - Open Shortest Path First - (OSPF)

3.5.1 - Overview

In order to overcome the shortcomings in RIP, the Internet Engineering Task Force (IETF) proposed a new routing protocol called Open Shortest Path First (OSPF) in 1989 [Postel-92]. Since then version 2 has been documented in four different documents: [Moy-91] [Moy-94] [Moy-97] [Moy-98]. The last of the RFCs is the more current one. It is useful to observe the evolution of the protocol. This can be done by reading the appendixes of these documents. The latest specification of the OSPF Management Information Base (MIB) is documented in [Baker-95].

OSPF is the intra-domain routing protocol recommended by the IETF [Gross-92]. The idea behind this recommendation is to have a common intra-domain routing protocol in order to facilitate interoperability between different vendors.

OSPF is based on the "distributed map" concept: all nodes have a copy of the network map, which is regularly updated. From this map the best routes to other destinations are calculated. The whole idea of the OSPF routing protocol is to maintain a synchronized copy of the topological database in all of the routers of the network.

OSPF is based in a distributed database, a flooding procedure, a definition of adjacency and special records for external routes.

OSPF is a dynamic routing protocol for TCP/IP networks. OSPF runs directly over IP (Protocol 89). It is designed to be run inside a single Autonomous System. Each OSPF router maintains a database describing the Autonomous System's topology. From the **topological database**, each router calculates a routing table that has the shortest distance to every destination. The resulting tree gives the entire route to any destination network or host. However, only the next hop to the destination is stored in the routing table and it is used in the forwarding process. This calculation is done using the Shortest-Path First algorithm (**Dijkstra Algorithm**).

Routers communicate with each other using Link State Advertisements (LSAs). Each LSA describes a small piece of the OSPF routing domain (See section 3.5.7). The collection of LSAs received from other routers in the domain form the topological database. This database is identical in a single area. The **reliable flooding algorithm** is used in order to keep the topological database in an area synchronized (identical). OSPF multicast protocol packets reduce traffic in a domain.

The name link-state comes from the fact that router advertisements tell other routers, among other things, the state of its links (whether a link is up or down). The collection of all the link-state-advertisements coming from other routers forms the topological database from which the routing table is calculated.

OSPF features include the following:

- Fast convergence time after topology changes have occurred.
- Support for CIDR.
- Security Support through packet authentication.
- Equal-cost multipath that allows load balancing.
- Hierarchy provided with areas supports big ASs.
- External routes are imported into the OSPF routing tables.

3.5.2 - Features

OSPF recalculates routes quickly in the face of topological changes. Its convergence time is fast after topological changes occur, and minimum OSPF traffic is exchanged while calculating new routes, since only changes are transmitted among routers.

Traffic is also reduced since periodic advertisements of the protocol occur every 30 minutes.

OSPF packets are sent to multicast groups, which better utilizes network resources.

It has been proven that the link-state algorithm self-stabilizes quickly [Gross-92]. OSPF does not generate much traffic and it responds quickly to topology changes or link failures. However, the amount of information stored in a router (all the LSA coming from every other node on the OSPF domain) can be quite large. This limitation makes OSPF have scaling problems. OSPF can be used for large corporation networks but it can not be used for the global Internet.

OSPF supports precise metrics and, if needed, multiple metrics. OSPF also has the ability to set different routing metrics. Network administrators can configure the cost to express a function of bandwidth, delay, dollar cost, reliability or other factors.

OSPF supports host-specific routes and subnet routes as well as network-specific routes.

OSPF has global knowledge and a map of the world at each node.

OSPF supports equal-cost multipath, which allows load balancing of traffic. That is, if multiple equal-cost routes to a destination exist, they are all discovered and used. This means that a router potentially has several available next hops to a given destination, which offers the ability to load balance traffic between several paths. Load balancing is the ability to assign the same cost to routes to the same place, which will cause traffic to be distributed evenly over those routes.

Support for CIDR and VLSM. OSPF transmits **subnet masks** in the LSAs. This feature permits variable length subnet masks (VLSM) in a domain and it also supports CIDR. A subnet mask indicates the portion of the IP address that identifies the attached network.

Security Support. All OSPF routing protocol packets are authenticated providing a level of security and the ability to have a hierarchy in an AS. Each OSPF packet is authenticated. The OSPF packet header (which is common to all OSPF packets) contains a 16-bit Authentication type field, and 64 bits of Authentication data. Each area in the OSPF domain must run a single type of authentication. Currently, there are two types of authentication: clear password and cryptographic.

Type of Service (TOS) based routing. OSPF can calculate a separate set of routes for each IP Type of Service.

Stub area support. To support routers having insufficient memory, areas can be configured as stub. This allows routers in that area to not receive the huge amount of external routes on the Internet. As October of 1998, that number was around 55,000 external routes. [11]

Incremental updates. When an external route changes only its entry in the routing table is recalculated. This saves a considerable amount of CPU memory.

3.5.3 - Issues

The size of an OSPF link state database can be quite large, especially in the presence of thousands of external LSAs (around 55,000 in today's Internet). This imposes requirements in the amount of memory required for a router. For example,

[11] The "CIDR Report" by Tony Bates counts the number of routes that a router in today's Internet needs to be aware of. This number is around 55,000 routes as October of 1998. For more info visit: http://www.employees.org/~tbates/cidr-report.html

about 3.5 Mbytes of memory would be required to store 55,000 LSAs each of 64 bytes in length.

A router having problems with the size of routing tables could do the configuration in a stub area, where the only information needed is how to get to the AS Border Router.

OSPF has a very limited ability to express policy

3.5.4 - Graph Representation of network

The Topological Database maintained by each node of the OSPF domain is expressed as a bi-directional graph. The graph consists of vertices and edges:

1. Vertices represent

    a) router

    b) network of two types:

    - Transit: if it can carry "external" data. That is, data that neither originates nor terminates on a host attached to this network.
    - Stub: if it is not a transit network

2. Edges represent

    a) Direct point-to-point links between two routers

    b) Connection of a router to a network

3.5.5 - OSPF Terminology

Routers in an OSPF domain are classified in four types: internal routers, area border routers, backbone routers, and AS boundary routers (ASBR). Internal routers are within an area. Area Border Routers connect two or more areas. Backbone routers are on the backbone. AS Boundary Routers talk to routers in other ASs.

In an OSPF domain routes could be of four types: intra-area, inter-area, external type 1 and external type 2.

External routes learned from an inter-domain routing protocol such as BGPv4, are also flooded in an OSPF domain. These external routes appear in the routing tables of each router in the domain. For every external destination, the router advertising the shortest path is discovered, and the next hop to the advertising router becomes the next hop to the external destination.

Virtual links serve to connect separate components of the backbone. The two endpoints of a virtual link are area border routers (See RT10 and RT11 in Figure 3. 7). A network administrator manually configures virtual links. The configuration is just the address of the other end area border router and the area that the two routers have in common (called the **transit area**). By defining Virtual Links OSPF removes topological restrictions on area layout in an Autonomous System.

OSPF introduces the concept of **neighbor routers**. Two routers are considered to be neighbors, if each of them has an interface to a common network. Neighbors are discovered dynamically by the Hello protocol. Two neighbor routers could become **adjacent** for the purpose of exchanging routing information. Not every pair of neighboring routers becomes adjacent. Adjacencies control the distribution of routing protocol packets. Routing protocol packets are sent and received only on adjacencies. Router connected by point-to-point networks and virtual links always become adjacent.

A **transit area** is an area supporting one or more virtual links.

Routes describing hosts attached directly to routers are referred to as **host routes**. The mask for host routes is 255.255.255.255, which indicates the presence of a single node.

A **cost** is associated with the output side of each interface (See Figure 3. 7). This metric is manually configured by a network administrator and it could represent a combination of network characteristics such as throughput, delay, monetary expense or reliability. By default, each interface is assigned a numeric integer cost based on bandwidth. The cost of a path is the sum of the costs of intermediate links.

Each multi-access network that has two or more routers connected to it, needs to have a **designated router (DR)**. This router will generate network link advertisements for the network as a whole. This feature reduces the amount of traffic that needs to be exchanged among OSPF routers and also the size of the topological link state database. The Hello protocol elects a designated router in a network. Each multi-access network also elects a **backup designated router (BDR)**, in case the DR fails.

3.5.6 - OSPF Packets

There are 5 OSPF **messages types**: keepalives ("hello message") database description and request, send and acknowledgements of LSA messages. They share a common protocol header. (See Table 3. 2)

Type	Packet Name	Function
1	Hello	Discover/maintain neighbors. Establish adjacencies.
2	Database Description	Summarize topological database contents
3	Link State Request	Obtain topological database from a neighbor
4	Link State Update	Topological Database advertisement. There are 5 types of advertisements.
5	Link State Ack	Acknowledgment of the reception of a LSA.

Table 3. 2 - OSPF Packet Types

Hello packets are sent periodically by each router on all OSPF configured interfaces. A router sends Hello messages to tell neighbor routers that it is still alive and reachable. If a neighbor replies, the link between them is said to be "up".

Otherwise, the link is said to be "down". Generally, they are sent each 10 seconds in a point-to-point link and every 30 seconds on NBMA interface. Routers send these packets to multicast address 224.0.0.5 (allSPFrouters). The function of Hello packets is to establish and maintain neighbor relationships, elect the Designated Router (DR), establish adjacencies and describe optional capabilities in OSPF.

Database Description (DD) packets are exchanged between routers to initialize their topological database. When adjacency between a DR and an internal router is initialized. A DD packet describes the content of the link-state database. Multiple packets are used to describe the database. In the exchange, one router is designated as MASTER and the other as SLAVE. The router designated as MASTER is the one that sends Database Description Packets. The processing of a DD packet depends on neighbor state.

A **Link State Request packet** specifies a list of link state packets that a router wishes to receive. These packets are sent if parts of the topological database are missing or if parts of the database are out of date. This usually happens after exchanging Database Description packets.

A **Link State Update packet** carries a collection of LSAs from one router to its adjacent router. Several LSAs are included in a single Link State Update packet. These packets are multicast on physical networks that support multicast, if the physical network does not support multicast then the packets are sent unicast to all adjacent routers.

A **Link State Acknowledgement packet** is used to acknowledge each LSA received by the router. Multiple LSAs can be acknowledged in a single packet. Link state Acks are sent to multicast addresses. If the state of the router is "DR" or "BDR" then acknowledgments are sent to the group address 224.0.0.5 (All-OSPF-Routers).

If the state of the router is "DR other" then it is sent to the group address 224.0.0.6 (All-DesignatedRouters-Routers)

3.5.7 - Types of LSAs

Table 3. 3 is a summary of the 5 types of Link State Advertisements.

Advertisement Name	Contents	From	To
Router links (Router LSA)	State and cost of router's interfaces to an area	All routers	The particular area that the router belongs to
Network links (Network LSA)	Routers connected to a network	Network's Designated Router (DR)	The particular area that the router belongs to
Network Summary links (IP Network Summary LSA)	Route to networks in other areas inside of the AS.	Area border routers (ABR)	The particular area that the router belongs to
AS boundary routers (ASBR) Summary links (ASBR Summary LSA)	Routes to AS boundary routers.	Area border routers (ABR)	The particular area that the router belongs to
AS external links (External LSA)	Route to a destination in another AS	AS boundary routers (ASBR)	The whole AS

Table 3. 3 - Types of Links State Update packets.

Router links and network links advertisements describe how an area's router and networks are interconnected. Summary Link advertisements provide a way of condensing an area's routing information. AS external advertisements provide a way of advertising external routes throughout the Autonomous System.

Each router originates router links advertisements. Designated routers for a network originate network link advertisements. Area border routers originate a single summary link advertisement (a single IP network or AS Boundary Router). ASBR originate a single AS external link advertisement for each external network that they know about it.

All advertisement types, except AS external links advertisements are flooded throughout a single area only. AS external link advertisements are flooded throughout the entire Autonomous System, with the exception of stub areas.

All link state advertisements (with the exception of network links advertisements (Type 2)) specify metrics. In router link advertisements (Type 1), the metrics indicate the costs of the described interfaces. In summary, links (Types 3 & 4) and AS external link (Type 5) advertisements, the metric indicates the cost of the described path.

On broadcast networks, the Link State Update packets are multicast. On non-broadcast, multi-access networks, separate Link State Update packets must be sent, as unicasts, to each adjacent neighbor.

Router link advertisements advertise the cost of each interface that is advertised.

OSPF defines two types of external LSAs. E1 considers the total cost up to the external destination. In this case OSPF metrics are comparable to external metrics. E2 considers only the cost of the outgoing interface to the external destination. In this case, OSPF metric and BGP metrics are different.

External and summary LSAs are 36 bytes long. Router LSAs and Network LSAs are generally three times as large as an AS external link (around 108 bytes).

3.5.8 - OSPF Hierarchy (Areas connected by a backbone)

Areas

OSPF allows sets of networks to be grouped together into an area (see Figure 3. 7). The topology of an area is invisible from the outside of the area. This allows the creation of a hierarchy inside an AS that reduces the size of routing tables and the size of LSAs exchanged between routers.

When no OSPF areas are configured, each router in the AS has an identical topological database. From the topological database, each router calculates its routing table with shortest distances to all destinations from this router (the shortest path tree rooted at each router). When areas are configured, each area runs a separate copy of the basic link-state algorithm. This means that each area has its own topological database. A router has a separate topological database for each area that it is connected to.

?

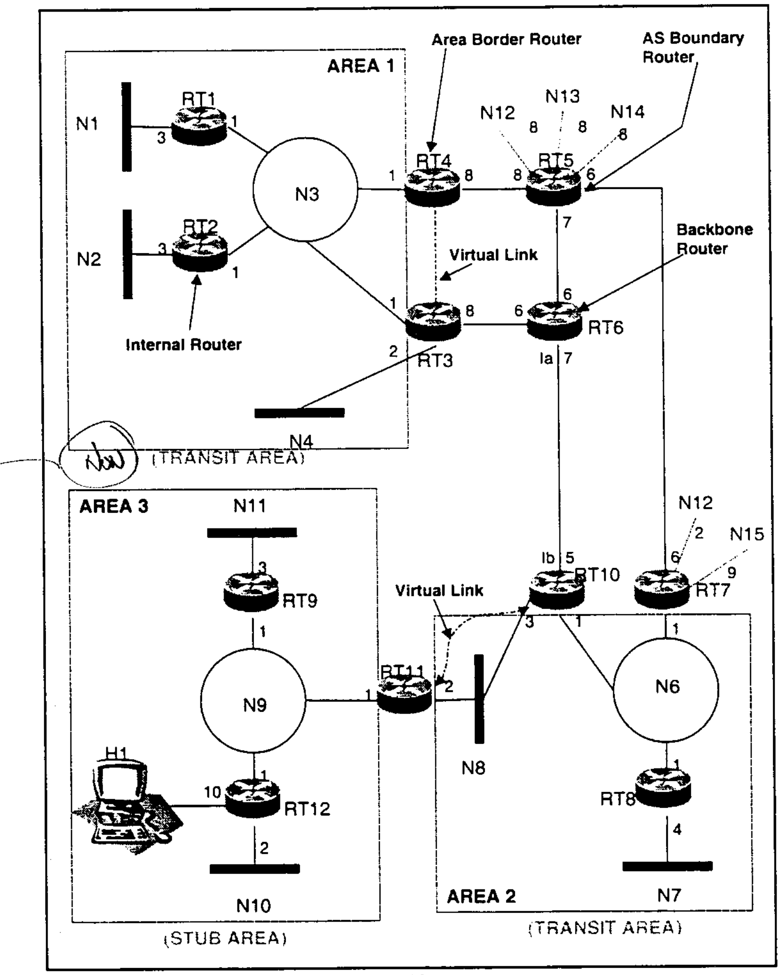


Figure 3. 7 - OSPF Components[12]

[12] This figure is a replica of the network used as an example in [Moy-94].

Areas are connected by an internal backbone in the AS. The backbone consists of those networks not contained in any area, its attached routers, and those routers that belong to multiple areas. The backbone feature forces an AS to have a star configuration (areas connected to a central backbone).

OSPF allows configuring certain areas as "**stub areas**". An area can be configured as a stub area because all external traffic must travel through a single border router. AS external advertisements are not flooded into stub areas. This reduces the topological database size, and therefore the memory requirements, for a stub area's internal routers.

Every AS has a **backbone area**, called area 0. Any router that is connected by two or more areas is part of the backbone. The backbone enables the exchange of routing information between area border routers. All areas must be connected to the backbone. Every area border router hears the area summaries from all other area border routers. After this, each area border router calculates the distance to all the networks outside of its areas by examining the collected advertisements, and adding in the backbone distance to each advertising router.

3.5.9 - Types of Networks supported

OSPF supports three kinds of connections or networks:

a) Point-to-point networks (56k lines, T1, T3, etc.),

b) Broadcast multi-access networks (mostly LANs such as Ethernet, FDDI, etc.)

c) Non-Broadcast Multiple Access (NBMA) subnetwork technologies (most packet-switched WANs such as ATM, Frame Relay, SMDS, and X.25 private and public networks).

A **multiple-access network** is one that can have multiple routers on it, each of which can directly communicate with all the others. A multi-access network is represented in the directed graph by a node for the network itself plus a node for each router.

For NBMA networks some configuration information is necessary for the correct operation of the Hello Protocol. Point-to-point networks do not need to be assigned IP addresses in OSPF.

3.5.10 - How it works

When a router boots, it sends **Hello Messages** on all its point-to-point links and multicasts them on broadcast multi-access links. On NBMA interfaces, it needs some configuration information of routers on the other side of the network. From the responses each router **learns who its neighbors are**.

Then a Designated Router and Backup Designated Router is elected on *each broadcast multi-access and NBMA networks. This election is not done in point-to-point links.* Once the DR and BDR are elected then each router **brings up adjacencies** with the DR and BDR. In this way communication is only needed between routers and the DR, which is less than a communication with every single router in the network. Only **adjacent routers** exchange routing information. The BDR takes on the role of the DR in case of failure of the DR.

Once two routers become adjacent they describe their topological database to each other. The description of the entries that each router has is done with the **Database Description DD Packets**. Each DD packet gives the sequence numbers of all the LSA stored currently by the sender. By comparing its own values with those of the sender, the receiver can determine who has the most recent values.

The adjacent router with the least current information, sends **Link State Request packets** asking for the missing or updated LSAs. Each router performs this "database comparison" with its adjacent routers. At some point, every router in the area will have an identical copy of the topological database and we say that the **databases are synchronized**. This process of interchange of LSAs among routers is referred as **reliable flooding**.

During normal operation, each router floods the other routers in the area **Link State Update packets** containing LSAs to each of its adjacent routers every 30 minutes. These LSAs populate the topological database. The Link State Update packets are acknowledged to make them reliable. Each of these packets have a sequence number, so a router can see whether an incoming Link State Update packet is older or newer that what it currently has in its topological database. Link State Update Packets are also sent when there is a change in the topology of the network, e.g., a point-to-point link went down.

OSPF broadcasts local knowledge of its interfaces to all of the other OSPF-routers. The flooding procedure is reliable ensuring that all routers in an area have the same topological database.

Whenever a LSA arrives at a router, a router uses the information to update its topological database, by marking links "up" or "down". If there are changes in the topological database, then each router runs the Shortest Path First algorithm (Dijkstra's algorithm) and each of them creates a routing table that finds the shortest path from the router to all of the destinations advertised among routers of the OSPF domain.

A router that connects to two areas needs the databases for both areas and must run the shortest path algorithm for each one separately.

1) R1 and R2 connected by Ethernet

2) R1 and R2 come up and do DD exchange

3) They exchange their Router LSA (Via LSA request and LS UPDATE)

3.5.11 - The Topological database

The collection of link state advertisements forms the topological database. Each router in an OSPF domain maintains the autonomous system's **topological database**. The topological database is populated with link state advertisements (LSAs) (router links, network links, summary links and AS external link advertisements) coming from other routers in the OSPF domain. This database describes a directed graph of the autonomous system. The vertices of the graph consist of routers and networks. The edges could be physical point-to-point connections between routers or a router's interface to a network.

A router has a separate topological database for each area that it belongs. All routers belonging to the same area have identical topological databases for the area.

Installing a new link state advertisement in the topological database may cause the routing table to be recalculated.

3.5.12 - The Routing Table

There is a single routing table in each router. The OSPF routing table contains all the information necessary to forward an IP data packet toward its destination. Each routing table entry describes a set of paths to a particular destination. When forwarding an IP packet, the entry providing the best path to a destination is selected. This matching entry provides the next hop towards the packet destination.

The main fields of the routing table are now described. Some other fields are defined in the RFC.

Destination Type: Whether the destination is a network, an area border router, or an AS boundary router.

Destination: It is the IP address of the destination.

Address Mask: Allows to identify the network in the destination field.

Area: Indicates the area from which this entry was learned.

Cost: Indicates the degree of preference for this route.

Next Hop: Indicates the interface to be used to forward this packet and also on multi-access networks it indicates the IP address of the next router toward the IP packet's destination.

3.5.13 - Designated Router (DR) Election

The Designated Router originates the network LSA. All routers synchronize their databases with the DR. The synchronization is done by sending and receiving LSAs to/from the Designated Router during the flooding process. A Backup Designated Router is also elected in case the DR fails. The Designated Router is elected based on the router with the higher router priority configured.

3.5.14 - Database Synchronization

Bringing Up Adjacencies

When a router's interface becomes operational it begins sending a hello packet (OSPF packet type 1).

OSPF routers communicate with adjacent routers. Each separate adjacency is described by a neighbor data structure and **neighbor finite state machine (FSM).** This FSM describes the state of a conversation with a neighboring router. This state machine has 8 states: down, attempt, Init, 2Way, Exstart, Exchange, Loading and Full. The states describe the process of bringing up an adjacency, i.e., to establish a relationship with a neighbor router to exchange the contents of the topological database.

This neighboring router may become an adjacent router for this router.

Hello packets are sent out each functioning router interface. They are used to discover and maintain neighbor relationships, assuring 2-way communication between neighbors. They are also called keepalive messages. On multi-access networks, Hello Packets are also used to elect the Designated Router and Backup Designated Router, and in that way determine what adjacencies should be created. The Hello packet contains a list of all the routers from which Hello Packets have been seen recently. It also indicates which is the current designated and backup router for a network.

Reliable Flooding

The transmission of routing updates is done with the use of a technique known as **reliable flooding**. Individual components of the link state databases (the LSAs) are refreshed infrequently (every 30 minutes), at least in the absence of topological changes.

Link State Update packets (OSPF Packet Type 4) contain several distinct advertisements, and floods each advertisement one hop further from its point of origination. Each Link State Update packet must be acknowledged separately, through a Link State Acknowledgement packet.

OSPF uses a **reliable** router-to-router protocol to quickly and efficiently propagate routes. It does not use the reliability coming from TCP, since it runs directly over IP. OSPF propagates only the changes to its tables. The messages between OSPF routers contain either a change or a simple "no change" statement. The receiving router instantly knows if there has been a change to the status quo.

These LSAs are sent using reliable flooding. This means that the advertisements are sent to all other nodes, so that every node receives a copy of the LSA. The sequence number of the LSA is used to determine which LSA to store in case that a node receives duplicate LSA from the same node. The TTL field is used to age entries on the routing tables that are old.

Routing update messages contain information about how to reach certain destinations. By analyzing routing updates a router can build a detailed picture of the network topology. **Link state advertisements (LSAs)** inform other routers of the state of the sender's link. Where state means whether the interface is up or down, its IP address, etc.

Once a node has a copy of the LSA from every other node, then the node has a complete map for the topology of the network, and from this map OSPF calculates a routing table with the best route to each destination.

3.5.15 - Routing Table Calculation (Dijkstra)

The link state algorithm relies on two mechanisms: reliable dissemination of link-state information and the calculation of routes from the sum of accumulated link knowledge. This section describes the algorithm used to calculate the routing table. Dijkstra algorithm constructs a short-path tree, routed at the router making the calculation. The following is a description of Dijkstra algorithm:

Defining the following variables for a router running Dijkstra algorithm:

- N is the set of nodes in the graph
- L(i,j) is the cost of the link between i and j
- S is the node executing the algorithm
- W is node being analyzed
- M is the set of nodes incorporated so far in the algorithm

- C(n) is an array of costs from this node to every other node in the network

We can calculate the shortest path tree from this router as follows:

```
M={s}
For each n in N-{s}
    C(n) = l(s,n)
While (N ≠ M)
 M = M∪{w} such that C(w) is the min. for all w in (N-M)
    For each n in (N-M)
        C(n) = Min( C(n), C(w)+l(w,n) )
```

3.5.16 - Flushing Advertisements

There is a procedure for flushing old or unreachable advertisements. Entries are deleted from the topological database once they reach an "age" of one hour (MaxAge constant). This means that if an advertisement is not received in two consecutive periodical advertisements (every 30 minutes) then the advertisement is deleted from the topological database.

3.5.17 - Importing external routes

External routes are imported into OSPF domains with "External LSAs". Each external route is imported into the Autonomous System in a separate "External LSA". The reason for importing routes one at a time, is that this approach reduces the amount of flooding traffic (since external routes change often, and then only changes are flooded throughout the domain).

The number of external routes typically exceeds the number of internal routes by far. OSPF routing tables are then composed mainly of routes to networks outside of the AS.

Routes can be imported using two different types of metric: Type 1 and Type 2 metric.

It is possible to specify a forwarding address in an External LSA. This means that data traffic to the advertised destination should be forwarded to the forwarding address, instead of the AS boundary router. *This is useful because it eliminates extra-hops in order to get to an Autonomous System Border Router.*

Some external routes can be tagged with the Autonomous System that the route belongs. This feature is useful to better manage external routes in the AS.

3.5.18 - Latest OSPF features

The main changes since Moy-94 are the support of Point-to-MultiPoint interface and Cryptographic Authentication.

The Point-to-MultiPoint interface was added as an alternative to OSPF's NBMA interface when running OSPF over non-broadcast subnets. Unlike the NBMA interface, Point-to-MultiPoint does not require full mesh connectivity over the non-broadcast network.

Cryptographic authentication was added to OSPF, which adds more security to routing packets against hackers. A message digest algorithm is used to provide authentication, currently the algorithm used is MD5 [Rivest-92].

3.5.19 - More information on OSPF

For more information see [Moy-98c].

3.6 - How OSPF overcomes RIP limitations

This section explains how OSPF tries to fix RIP's limitations.

3.6.1 - Using a better metric

OSPF does not use the hop count as the routing metric to assign to links. Instead it uses the speed of an interface as the metric. OSPF requires the network administrator to assign a cost to each circuit between routers. This cost is based on the speed of the given interface; the higher the link speed, the lower the interface or route cost. Aggregate route cost is cumulative. As each router's circuit costs are added into the equation, the total route cost is reflected in the OSPF routing table.

3.6.2 - Making it work for big network topologies

The fact that the hop count is not used also implies that the 15 maximum number of hops that route can have is also eliminated. No limits on the number of hops a route can take. This implies that OSPF can be used for big networks such as the ones found on large enterprises.

3.6.3 - Improving response on network changes

OSPF advertises the state of its links to all the routers on the network only when changes happen. RIP has triggered updates but RIP also advertises its complete routing table to its neighbor routers periodically (e.g. every 30 sec.).

Since router advertisements are sent every time a change occurs and not periodically, then the convergence time of the OSPF protocol is greatly reduced.

3.6.4 - Authentication of routing messages

RIP advertisements do not have a password field, which makes the protocol vulnerable to a network halt when wrong routing updates are transmitted from a router or from a malicious user. The lack of a password field also prevents network administrators from defining internal areas in a domain.

OSPF packets contain 64 bits of authentication information in the header. Administrators can set these bits and restrict unauthorized routers from communicating with an OSPF router. Route authentication has two distinct advantages: 1) It can prevent accidental misconfiguration of the network, and 2) multiple, independent OSPF networks can coexist in the same AS by using different authentication keys.

3.6.5 - OSPF supports VLSM and CIDR

The subnet mask is exchanged on the LSA that is flooded on the OSPF domain. As a result networks using VLSM are now recognizable and then CIDR is supported. This feature makes OSPF a better choice for big networks.

3.6.6 - Additional Hierarchy

OSPF divides a domain into different areas, which creates additional hierarchy that can be used for big Intranets. With this hierarchy, routing updates can be aggregated. As a consequence, the number of entries on routing updates is greatly reduced and also the number of Link State Update packets is reduced. This characteristic allows OSPF to scale to bigger networks, since routers only need to know how to reach an area.

3.7 - RIP vs. OSPF Comparison

RIP is limited and simple, while OSPF is powerful and complex.

RIP exchanges number of hops to a destination, while OSPF maintains a "map" of the network that is updated quickly after any change in the topology.

Immediately after the transmission of new information through the flooding protocol and the local computation of the Dijkstra algorithm in each router, all routes in the network are sane - no intermediate loops, no counting to infinity. Since there are no routing loops in OSPF then this property alone make OSPF a much superior protocol.

The main difference between RIP and OSPF can be summarized as follows. In RIP, each node tells only to its directly connected neighbors its entire routing table. In OSPF each node tells all the routers in the network only the state of its directly connected links.

OSPF does not interchange routing tables among routers. It interchanges LSAs from which it calculates the routing table for the router. LSA are small and rarely need to be sent (every 30 minutes or with network topology changes), so they consume very little available bandwidth. RIP on the other hand exchanges requires routers to exchange entire routing tables every 30 seconds. This approach does not scale to bigger networks, as the Internet grows, routing updates grow larger and consume significantly more bandwidth.

OSPF has become increasingly important in recent years because it uses less network capacity than RIP updates. Additionally, it converges on a stable configuration after changes occur more quickly than RIP. OSPF is the protocol recommended by the IETF [Gross-92].

OSPF does have its own shortcomings, however. The most notable is the fact that OSPF requires a great deal of processing power to keep the link state information up to date. But in today's routers, fast CPUs and ample memory are prevalent. The beauty of the protocol is that it is robust and keeps network traffic to a minimum, unlike RIP, which is constantly broadcasting tables, thus burdening the network with router-to-router traffic.

The Dijkstra calculation is of order (n * log(n)), where n is the number of router in a single area of the routing domain. Note that the complexity of Dijkstra is greatly reduced by dividing an autonomous system in to areas. The Bellman-Ford algorithm, which is the basis of RIP, converges in O(N*M) where N is the number of nodes, and M is the number of links in the graph.

RIP requires more bandwidth than OSPF, because of the need of periodic updates every 30 seconds. OSPF only transmit changes when there is a change on the network topology and it floods content of the link state databases every 30 minutes.

OSPF requires more memory than RIP. OSPF keeps tracks of all the external routes that are currently available on the Internet (using External LSAs), while RIP does not store external routes in its routing table. External LSAs are 36 bytes long. These are stored with some supporting data, which increases the size of these LSAs to about 64 bytes. As November of 1998, an OSPF router would have to keep about 55,000 external routes[13] this would require around 4 MB of memory in the router.

OSPF is a multi-path routing protocol. That is, for any given destination there could be several routes to the same destination. This feature allows OSPF to load balance the traffic among several links. RIP is a single path routing protocol that does not allow having several paths to the same destination and there is no way to achieve load balancing on links of equal cost to a destination.

Table 3. 4 summarizes the differences between RIP and OSPF.

[13] See The CIDR Report at: http://www.employees.org/~tbates/cidr-report.html

	RIP	OSPF
Where they get their information	From adjacent routers (Locally)	From all routers (globally)
When they change the routes	RIP advertises routes every ΔT sec through messages known as "RIP advertisements" or with network changes.	Periodical advertisements and also when a router's state changes.
Type of algorithm	Distance Vector	Link State
Name of the Algorithm	Bellman-Ford	Dijkstra
Convergence Time	Slow	Fast
Complexity	Simple	Complicated
Type of advertisements	Each router tells *directly* connected routers (neighbors) what networks it knows how to reach (world).	Each router tells *all* routers in the network (world) what it knows about the network status of its interfaces (neighbors).
Single Path or Multipath	Single Path	Multipath
Hierarchy	Flat domain	Allows intra-domain hierarchy with the use of areas.
To be used at	Intra-Domain	Intra-Domain
Memory and CPU requirements	Low	High

Table 3. 4 - RIP vs. OSPF

3.8 - BGP- 4

3.8.1 - Motivation

RIP and OSPF are protocols designed to be used inside autonomous systems; they are not designed to route between autonomous systems, since they do not provide mechanisms to segregate enterprises into different administrations that are technically and politically different.

A routing table that grows linearly with the number of hosts clearly will not work in the global Internet. Hierarchical addressing and route aggregation are used to limit the number of entries in routing tables and routing advertisements.

Even though OSPF provides better routing scalability, which enables it to be used in bigger and more complex topologies, OSPF should be restricted to interior routing. The amount of computations needed for a router will become too large for a router to handle when used for the global Internet.

The Exterior Gateway Protocol (EGP) [Mills-84] was the inter-domain routing protocol used in the early days of the Internet. EGP had several limitations that created the need for a better protocol. The main limitation of EGP was that it constrained the topology of the Internet to have a tree like structure as the one in the early days of the NSFNET. This restriction motivated the design of a new protocol capable of support a non-hierarchical connection of ISPs. EGP was also inefficient dealing with routing loops and was not able to configure routing policies.

In order to use resources of another AS, one needs explicit authorization before using that AS's resources for relaying packets. A routing protocol was needed that understands the complex relationships of the world. BGP was able to understand these policies and that is one of the reasons it has been widely accepted.

3.8.2 - History

In 1989, The Border Gateway Protocol (BGP) version 1 [Lougheed-89] was proposed to overcome EGP limitations. Since then several versions of BGP have appeared. Version 2 was introduced in 1990 and it is described in [Lougheed-90]. Version 3 was defined in [Lougheed-91] in 1991. The latest version of BGP is version 4 and it was defined in March of 1995 in [Rekhter-95].

EGP and BGP versions 1, 2 and 3 are considered obsolete. BGP-v4 is today's de-facto standard used for inter-domain routing. Most EGP routers have been upgraded to BGP.

BGPv4 was the first version of BGP that handled aggregation (CIDR) and supernetting with the use of VLSM. BGPv4's support of CIDR plus better filtering and policy setting capabilities have accelerated the rapid deployment of this protocol.

BGP was initially deployed in 1993, since then it has been widely spread to the point that is currently considered the de-facto standard for inter-domain routing.

3.8.3 - Overview

BGP is an inter-domain routing protocol and it is usually run between ISPs. It does not require a specific intra-domain routing protocol to be run internally by each AS. It does not impose restrictions on the underlying Internet topology as EGP did. The key features of the protocol are the notion of path attributes and aggregation of network layer reachability information (NLRI).

The essence of BGP is to advertise the subnetworks of an AS to the Internet. BGP tells routers outside of an AS (upstream providers or "peers") about what routes (IP networks in this AS) they "know how to get to" inside its AS.

Figure 3. 8 illustrates a border router advertising routes to the Internet. The border router at the University of Colorado is telling the world that network "128.138" is in its campus. Once the advertisement is made, if a packet is destined to "128.138" at UUNET network, it would be routed through Sprint network. Note that the traffic flow is in the opposite direction than the BGP advertisements.

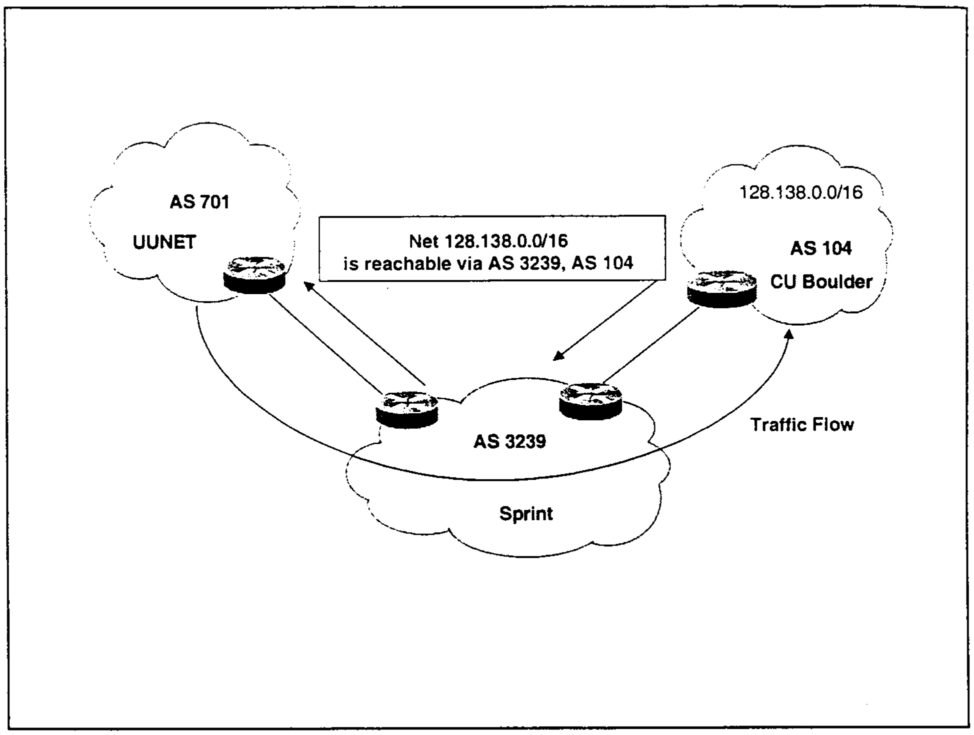


Figure 3. 8 - BGP basics.

BGP allows a system administrator to set up policies that are particular to specific ASs. BGP routing decisions reflect political control. For example, a BGP route update says: "This AS can be reached through this path", but it is not necessarily the most optimal path.

BGP is a policy-based protocol. That is, the routing decisions are based on policy factors such as who you signed your contract with, who is your default route, etc. BGP finds a permissible path among organizations. Since it is based on policy, the algorithm to find routes between nodes is less complex.

BGP is more concerned with reachability than optimality. BGP is only concerned whether a router is reachable through a particular route, and it is not concerned whether this is the shorter path to a destination.

A BGP speaker advertises reachability information for all the networks within an AS, and in the case of a transit AS, the speaker also advertises the networks that can be reached through this AS. BGP advertisements are of the form: "You can reach this network at this AS or through this AS".

BGP does not create routing loops. If a path contains the AS of the router running BGP, then the route it is discarded. This eliminates any routing loop.

BGP is a **path vector protocol**. The term path vector refers to the fact that BGP routing updates carry a sequence of AS numbers that need to be transversed to reach a particular destination. In other words, each router tells its neighbor the exact path to reach a destination network.

BGP is a reliable protocol since routers communicate with each other using TCP sessions. Two routers speaking BGP are called neighbors or peers. BGP runs on top of TCP in order to provide reliability to the BGP routing messages. BGP uses port 179. TCP provides an ordered reliable transport. This ensures that reliability is taken care of by TCP and does not need to be implemented in BGP itself.

BGP does not require that the entire ASs in the Internet use the same metrics. Every AS is free to setup its own policies in order to control traffic.

3.8.4 - How BGP works

In order to run BGP, there must be a priori agreement between the ASs for accepting to relay transit traffic. Each Autonomous System has to have at least one border router that needs to be configured to run BGP and to recognize a BGP peer in the other AS.

A border router is an IP router that is charged with the task of forwarding packets between ASs. This router is usually called a BGP speaker, because it runs BGP.

In the case of a stub AS where there is only a "BGP speaker", it maybe not worthwhile to run BGP (static routes to the ISP may be enough). For multihomed AS, there is more than one "BGP speaker", and then getting table with external destinations is helpful to route outgoing traffic.

From the point of view of a BGP router, the world consists of other BGP routers and the lines connecting them. A BGP speaker connects to a single BGP speaker in another AS. This relationship needs to be manually configured by network administrators. Two BGP neighbors communicating between ASs must be in the same physical network. These relationships between ISPs are called "peering". The name comes from the fact that their BGP routers open a TCP connection between them and the routers become BGP peers. When BGP neighbors belong to different ASs, it is said that they are running external BGP (eBGP).

If a domain has more than one BGP speaker, then these routers need to establish BGP neighbor relationships with every single router inside of the AS. In other words, BGP routers need to be connected logically in a mesh topology. This type of BGP is called internal BGP (iBGP). These routers do not need to share the same layer 2 subnetwork (e.g. the same Ethernet segment) as in the case of external BGP (eBGP). In order to alleviate the requirement of a logical full mesh inside an AS, a network manager can configure "Confederations" and "Route Reflectors" which decrease the number of internal peerings required in a single AS.

eBGP is used to exchange routes between different Autonomous Systems, while iBGP is used to exchange routes between the same Autonomous System.

The protocol operates in terms of messages, which are sent over TCP connections. The first thing that happens is that a TCP connection is opened between the two BGP routers. These two routers are called neighbors or peers. A BGP router exchanges routes (IP networks that the router knows how to reach) with another BGP router. The fact that BGP uses TCP simplifies the protocol since reliability is done one layer above.

After this TCP connection is established, each router sends an OPEN message to negotiate the peering relationship's parameter. If the OPEN message is accepted by the BGP neighbor, it answers with a KEEPALIVE message.

During this "initialization process", routers negotiate the value of the Hold Time parameter, which is the number of seconds for the Hold Timer. The hold timer is run by each BGP router and it indicates the number of seconds that may elapse between the receipt of successive Keepalive or Update messages. If the hold timer expires the connection to the BGP neighbor is dropped. The Hold Time parameter is negotiated between the peers, by choosing whichever router has a lower Hold Time configured.

Also, during the "initialization process", BGP routers negotiate the version of BGP they will run, inform the BGP neighbor about the AS that this router belongs to and sends its "name", which is the BGP identifier.

After the connection has been established neighbor routers are ready to interchange information. Initially the entire BGP database is exchanged between the two routers. That is the two BGP databases need to be synchronized. This synchronization is done through UPDATE messages

The **routes** exchanged by BGP routers are an association of a destination network numbers with the attributes describing the path to that destination network.

BGP neighbors communicate between them using UPDATE messages (see Figure 3. 9). The UPDATE message tells a neighbor which destination networks it can reach through this AS. More specifically, the UPDATE message sent a BGP neighbor contains among other things the following information:

- AS_Path: The AS that this router belongs to.

- Next_Hop: The IP address of the interface where the router sends this UPDATE message.

- NLRI: A list of all the sub-networks in this AS.

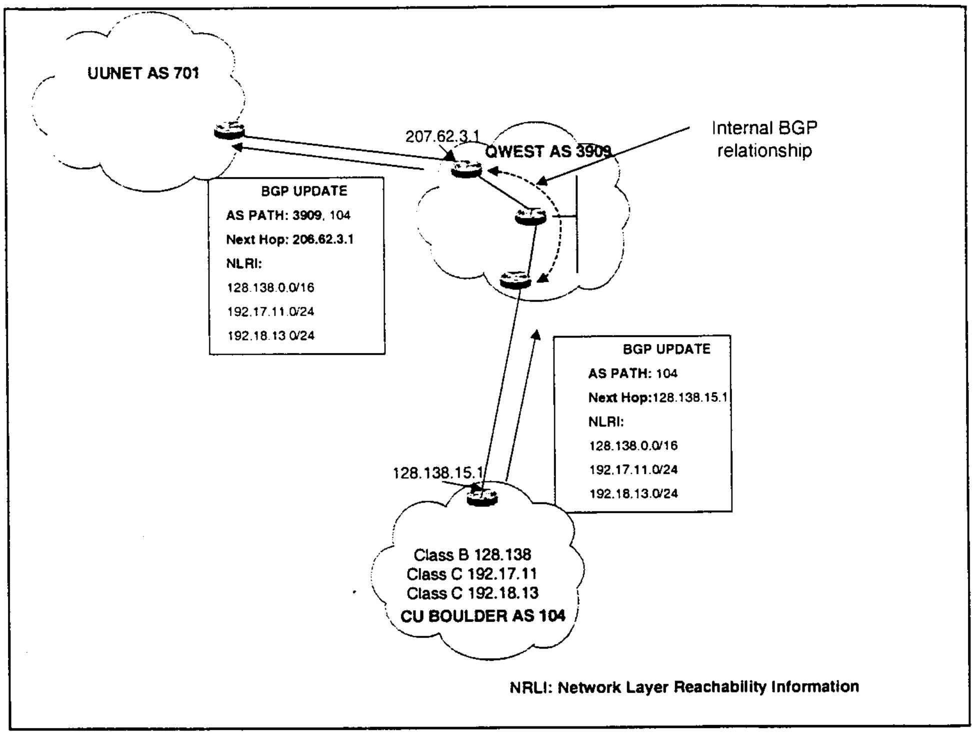


Figure 3. 9 - BGP Update message transmission between BGP peers.

This UPDATE message tells the BGP neighbor that the subnetworks listed in the NLRI are reachable via this BGP router and that the only autonomous system transversed is this AS.

The way the AS_Path attribute gets populated is as follows: Every time a router is advertised by a BGP router, it is stamped with the AS number of the router doing the advertisement. As a route moves from Autonomous System to Autonomous System, it builds the AS_Path attribute for this route.

Before a BGP router exchanges information with an external AS, BGP ensures that networks within the AS are reachable. This is done by a combination of internal BGP peering among routers within the AS and by redistributing BGP routing information to an IGP that runs within the AS such as RIP or OSPF.

Once routers have a complete copy of the BGP database, BGP constructs a graph of autonomous systems based on information exchanged between BGP neighbors. This graph is a map of the relationships of different ASs. The directed graph is referred to as a tree. The nodes on the tree are identified with AS numbers that are assigned by the Internic [14].

Each router maintains a database of the subnetworks that it can reach and the preferred route for reaching that subnetwork.

After that, only changes are transmitted among peers. This is called "Incremental updates". This is better than using periodical routing updates from the point of view of CPU overhead and bandwidth consumption. The incremental update is broadcast to all other routers implementing BGP. Only the best path is advertised to peers, although alternates paths are also kept in the BGP database.

Once a neighbor receives an UPDATE message and it changes its routing database, it sends another broadcast message to all of its neighbors informing of the new route.

If a router receives an UPDATE that includes the router's AS number in the AS_Path field, then the UPDATE is dropped. In this way, routing loops and the count-to-infinity problem are avoided. When the UPDATE is sent to an internal neighbor the router does not add the common AS number to the AS_Path field, so that this message is considered valid by the neighbor.

If routes go down or there is a change in topology, then BGP needs to tell its neighbors. This is done with UPDATE messages and the WITHDRAWN information

[14] For more information on the Internic visit: http://www.internic.net

elements. The WITHDRAWN routes contain a list of the destination networks that have gone down.

If there are no changes to transmit between BGP neighbors routers send KEEPALIVE packets to each other to assure that the connection is still alive and that the router is up. This is called the neighbor reachability procedure.

3.8.5 - Packet Types

OPEN

A BGP router sends an OPEN message after the TCP connection has been established, in order to negotiate parameters of the communication.

KEEPALIVES

A KEEPALIVE message is sent as an acknowledgement to the BGP neighbor once an OPEN message is received. They are also sent after a connection between peers has been established to determine whether peers are reachable.

They are sent periodically, usually at a rate of one third of the hold time. This allows missing two consecutive keepalive messages and still considering the connection between peers alive

UPDATES

An UPDATE message tells a neighbor about the routes it knows how to reach. This allows BGP to construct a map of the topology of the interconnection of ASs.

The Network Layer Reachability Information (NLRI) is one of the information elements that is carried in a BGP UPDATE message. This element contains information on the networks that are reachable through this AS. The NLRI consists of al list of pairs <length, prefix>, where length is the number of masking bits that a particular IP address has (prefix).

The Path attribute provides BGP with the capability of detecting routing loops and the flexibility to enforce policies.

The two most important attributes are the **AS_Path** attribute and the **Next_Hop attribute**. The AS_Path attribute is a sequence of AS numbers a route has transverse before reaching the BGP router. The Next_Hop attribute tells the router what is the next router in the path to a destination AS.

Other attributes defined in BGP are used to setup policies. The remaining attributes are: Origin, Multi Exit Disc, Local Preference, Atomic aggregate, Aggregator.

NOTIFICATION

Notification messages are error messages that are sent whenever there is a problem in the communication between the two neighbors.

Table 3. 5 summarizes the messages used in BGP.

BGP Message	**Function**
OPEN	Opens a neighbor relationship with another router.
KEEPALIVE	Used to (1) acknowledge an Open message and (2) periodically confirm that the neighbor is alive.
UPDATE	Used to (1) transmit information about a single route and (2) list multiple routes to be withdrawn.
NOTIFICATION	Notifies a neighbor about errors.

Table 3. 5 - BGP Messages

3.8.6 - BGP Attributes

The BGP attributes are a set of parameters that can be modified to affect the BGP decision process about which path is best. This allows total control from the network administrator's perspective. A BGP router uses these attributes to control which routing information it accepts, prefers or distributes to other neighbors. The two most important attributes are the AS-Path and the Next-Hop attribute. They specify how to reach a network.

AS-Path: This attribute lists the ASs that a datagram must transverse if it follows this route. For example, a BGP speaker may advertise: "In order to get to 128.138.0.0/16 you have to transverse AS 3909 and AS 104. The network 128.138.0.0/16 is one of the networks advertised in the NLRI field (See Figure 3. 9).

Next Hop: The IP address of the border router that should be used as the next hop to the destination networks listed in the NLRI field. This IP address will be used to forward datagrams for a given destination network.

Other attributes such as Local preference, MED, Origin and Communities allow network managers to specify policies.

Local Preference: It is a weight used internally in an AS. This attribute is used in multihomed ASs in order to specify which AS should be used in order to reach another AS. The higher local preference is preferred. This attribute has significance to routers in other ASs.

Multi-Exit Discriminator (MED): It is a hint to external neighbors about the preferred path into an AS when there are multiple entry points into the AS. It is a cost for internal routes in an AS. It is used among ASs. It is used when a border router in another AS has multiple entry points into another AS, in order to select the best entry point. The lower MED valued is preferred. This metric used to be called the Inter-AS metric in previous versions of BGP.

Origin: This attribute describes how the route to this destination network was learned. There are 3 options:

- IGP: The route was learned from local routes. In this case the AS_Path attribute will be a single element and it is going to be local AS number.

- EGP: The route was learned from a remote AS. In this case the AS_Path attribute will be a list of AS, and this BGP router will append the local AS to the list.

- Incomplete: The route was learned by some other means. For example, by redistributing OSPF routes into the BGP router.

Community: A community is a group of destinations that share some common property. This attribute is useful in applying policies. This attribute was added in 1996 and it is defined in [Chandra-96].

Atomic_Aggregate: Used to implement CIDR. Present if this route was not the most specific one known by the advertiser.

Aggregator: Used to implement CIDR. Indicates who formed the route if the route is an aggregate of smaller routes.

Other BGP attributes used in more advanced configurations include: Originator ID, Cluster List, Destination Preference (DPA), Advertiser and rcid_path. The Originator ID and the Cluster List are used for BGP Router Reflectors and they are defined in [Chen-96]. The rcid_path is defined in [Haskin-95].

3.8.7 - BGP Decision Process

A BGP router may have received information on how to reach a network from several BGP neighbors. In case of more than one route to reach a network, BGP must decide which is the preferred route. It is here where a network manager can configure his preferences and policies so that BGP behaves the way he wants.

Once the best path has been chosen this is the path that is propagated to other BGP peers. Other routes are kept in the BGP database, in case that the best routes become unavailable.

The decision of which route is best is based on the path attributes described in the previous section. For example, a route is better than another is the number of AS that need to be transversed (AS_Path attribute) is smaller.

The decision process is a module of the BGP program that returns a "distance" for a given path. The "distance" is determined by the previously entered configuration from a network manager. Any route violating a policy constraint automatically gets a score of infinity. The router then adopts the route with the shortest distance. The output of the decision process is the set of routes that will be advertised to all peers.

When there are two routes to the same network but with different specificity (different prefix length), then the most specific route always wins. For example, if there is a route 128.138.0.0/16 and another 128.138.0.0/20, then the second route is selected by BGP.

The following algorithm selects among routes based on the BGP attributes and some other parameters. BGP selects only one path as the best path in its routing table and it then propagates the path to its neighbors. Other routes are also kept in case the preferred route becomes unavailable.

The BGP Path selection process selects the best path to a given AS. Once the best path is selected this route is put into the BGP routing table and it is then propagates to BGP neighbors. The selection process consist of the following rules [Halabi-97]:

1) Do not consider iBGP path if not synchronized

2) Do not consider path if next hop is inaccessible.

3) Highest weight (local to router)

4) Highest local preference (global within AS)

5) Prefer the route originated by this BGP router

6) Shortest AS path

7) Lowest origin code (IGP<EGP<incomplete)

8) Prefer external path over internal path (MED)

9) Prefer the path through the closest IGP neighbor

10) Prefer the path with the lowest router id

3.8.8 - BGP Finite State Machine (FSM)

This section provides more detail in how a neighbor relationship is established. It describes the states from the beginning of establishing a TCP connection to the point where UPDATE messages can be transmitted between a pair of BGP neighbors. Figure 3. 10 illustrates the states of the BGP Finite State Machine. There are six states, a brief description follows:

Idle: BGP is waiting for a start event, which is caused by manual configuration from a network manager. After the Start event, BGP tries to establish a connection with its neighbor. BGP then transitions to a Connect State.

Connect: BGP is waiting for the TCP connection to be established. If the TCP connection fails, then BGP transitions to the Active state. If the TCP connection is successful then it sends an OPEN message to its peer and BGP transitions to OpenSent state.

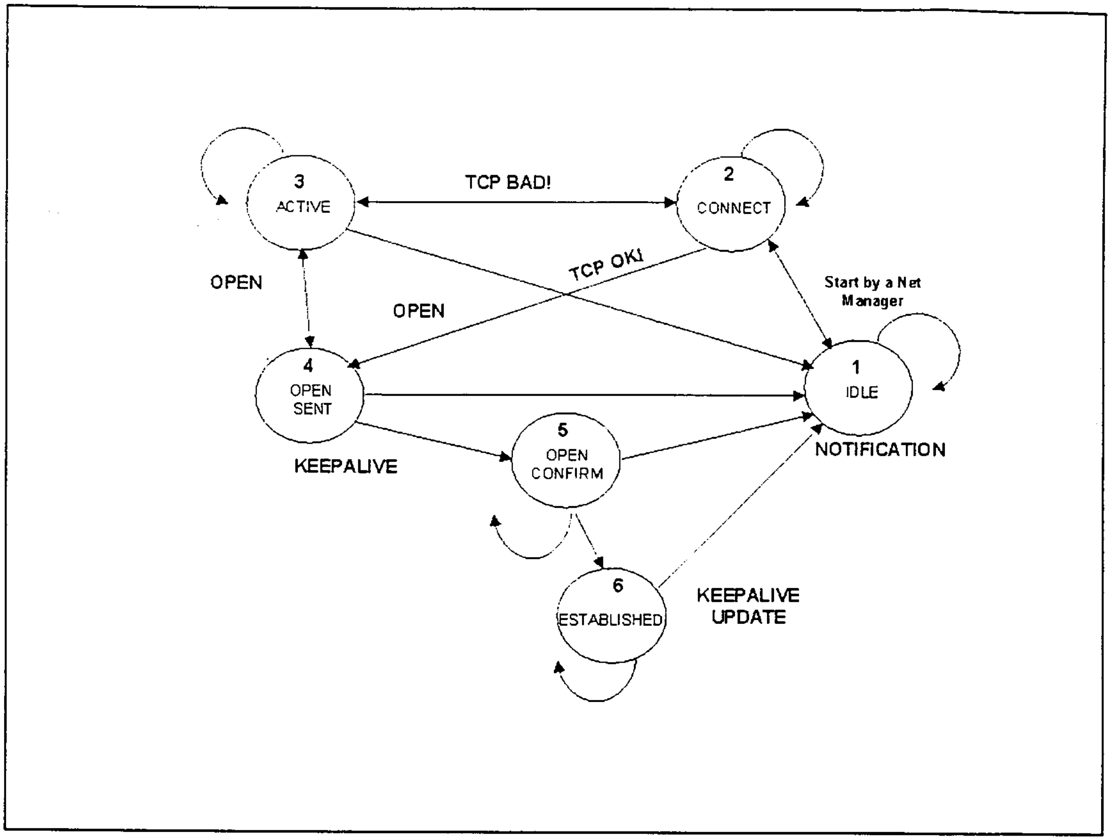


Figure 3. 10 - BGP Finite State Machine [Halabi-97]

Active: BGP keeps trying to open a TCP connection. So, in reality the session is inactive.[15] If the connection is achieved then BGP transitions to OpenSent. If "connect retry timer" expires then BGP transitions to an Idle state.

OpenSent: BGP is waiting for an OPEN message from its peer. Once this OPEN message is received, it checks parameters such as the BGP version and the Hold Time parameter, and sends a KEEPALIVE message as an acknowledgement.

[15] This is a bad name for this state, since network managers may get confused while reading the debug information in routers. If the state stays in Active, it actually means that the TCP connection has not been achieved.

It also realizes if the neighbor is an internal peer (same AS) or an external peer (different AS) BGP then transitions to OpenConfirm

OpenConfirm: BGP waits for a KEEPALIVE or NOTIFICATION message. If a KEEPALIVE is received, the state will go to Established. If a NOTIFICATION error is received BGP will transition to Idle.

Established: This is the final stage of the neighbor negotiation. At this stage, BGP starts exchanging UPDATE packets.

3.8.9 - Policies

BGP allows network managers to configure "policies" in routers. A policy allows configuring the flow of packets based on political, security, economic or other considerations. For example, a network manager may configure the following policies:

a) I don't want my traffic go through this ISP because it is too expensive.

b) Do not sent confidential traffic outside of my network.

c) Do not allow transit traffic on my network.

d) Minimize the number of ASs transversed in order to get to a destination.

Policies are manually configured in each BGP router. They are not part of the BGP specification itself. Some of the common methods used by network managers to control the sending and receiving of updates are: Prefix filtering, AS_Path filtering, Route_map filtering and Community filtering.

3.8.10 - Synchronization with an IGP

A BGP router that has learned of a path toward a given network should update the AS routing tables, either by inserting an external link in the OSPF database or by adding an entry in the RIP routing table.

The rule says: Do not advertise a prefix until a matching router has been learned from an IGP. This ensures consistency of information through the AS and avoids black holes within the AS.

A **black hole** occurs when a BGP router advertises a network that it does not know how to reach. If a network manager misconfigures a BGP router (e.g. mistyped an IP address), then this router may start advertising a network that does not belong to that AS. If this advertising is more specific (the length of the mask is larger), then the real advertisements for that network will be discarded. As a consequence, packets destined to that network will be misrouted to the misconfigured BGP router.

3.8.11 - BGP Summary

The main function of BGP is to exchange network reachability information with other ASs. This information is used to create a graph of AS connectivity without routing loops and with which the policies from a particular AS could be configured.

3.8.12 - More information on BGP

For more specific information please see [Halabi-97] [Huitema-93].

3.9 - CIDR

With the exponential growth of the Internet in the early 90's, routers could no longer exchange network information, since routers had little memory. There were too many addresses to exchange and it was taking too long to do table lookups on routers.

Classless Inter-domain Routing (CIDR) was introduced to control the growth of the global IP routing tables beyond the ability of current equipment and to better utilize the IPv4 address space. Specifically it proposes a solution to the exhaustion of the class B network address space. It was proposed as patch technology before the introduction of IPv6, which is the long-term solution for the problem, proposed by the IETF. CIDR was proposed by the IETF in 1993 in [Fuller-93].

CIDR allows aggregating routes, making it possible to reduce the amount of information transmitted between and stored in each router as the network grows. It provides aggregation using a single entry in a routing table to tell how to reach a group of different networks. It also provides mechanisms to efficiently use the address space available on IPv4 addresses.

Aggregation is the process of combining the characteristics of several different routes in such a way that a single route could be advertised. Aggregation reduces the amount of information that BGP routers must store and exchange with other BGP routers.

CIDR is a short hand notation for routers that replaces thousands of addresses with a simple reference to another backbone provider that services those addresses.

In general, it is possible to aggregate routes if IP addresses are assigned taking into consideration which network provider gives service to an organization before assigning the address to that organization. One way to achieve this goal is to give blocks of addresses to regional network providers and let them distribute the addresses. Note that for this scheme to work, all of the networks that share a prefix should be getting Internet service from the same provider so that only with the prefix it is possible to route packets in the right direction.

The growth of routing tables is due to the fact that many organizations are connecting to the Internet. The classfull approach of handling Internet addresses (Class A, B and C) does not offer a solution to deal with such a growth. It creates bigger routing tables and it forces the transmission of bigger routing tables.

The address assignment inefficiency is caused because the IP addresses are assigned in blocks of 256 addresses (Class C), 65,535 addresses (Class B) or 16,777,215 addresses (Class A). This approach is not very flexible. For example, a mid-sized organization in need of 535 addresses would be assigned a Class B address only using 535 addresses and wasting 60,000 addresses.

The usual network masks used on the Internet are the straight class A mask (255.0.0.0), the straight class B mask (255.255.0.0) and the straight class C mask (255.255.255.0). These masks contain 8, 16 and 24 bits respectively. With the use of **Variable Length Subnet Masks (VLSM)** routers can aggregate networks that share the first X bits of the IP address, and then considerably reduce the size of routing tables.

The concept of CIDR is a move from the traditional IP classes (A, B, C) toward the concept of IP prefixes. The **IP prefix** is an IP network address with an indication of the number of bits (left to right) that constitute the network ID of an IP address. For example, the IP prefix 128.138.0.0/16 indicates the IP network address 128.138.0.0 using a 16-bit mask. This is equivalent to the IP network 128.138.0.0 with a subnet mask of (255.255.0.0).

CIDR creates new problems. For CIDR to work an AS should get its IP address block from its provider. If the AS wants to change from one provider to another, then the AS network manager is forced to change all its internal IP addresses, which is in most cases a huge job, so network managers are forced to stay with their current provider. A way to work around this problem is to use Network Address Translation (NAT), which maps private and global addresses, allowing it to have internal addresses that are different from the addresses advertised outside the domain.

BGP-4 is an inter-domain routing protocol that can deal with the fact that network IDs can be of any length. In other words BGP-4 supports CIDR. Next section highlights the major characteristics of BGP-4.

3.9.1 - More information on CIDR

For more information on CIDR see [Fuller-93], [Huitema-93] and [Halabi-97].

3.10 - Chapter summary

This chapter has introduced the basics of how a router works and the protocols used for unicast routing ("one-to-one routing"). A very clear understanding of unicast routing is essential, since many of the proposals for multicast routing ("one-to-many routing") are based on extensions to the protocols used for unicast routing.

This chapter has presented the two most important intra-domain protocols used on the Internet (RIP and OSPF). It also provided an introduction to the most used inter-domain routing protocol (BGP-4).

The most important idea to obtain from this chapter is that intra-domain protocols are used inside of organizations and they find optimal paths between nodes, while inter-domain routing protocols are used between autonomous systems and they find permissible paths among AS. The complexity of Inter-domain routing is in the order of AS while the complexity of intra-domain routing is in the order of networks inside of an AS.

Also, it is important to note that none of these protocols support multicast ("one-to-many routing"). The extensions of these protocols to support multicast are presented in chapter 5. However, before going into that topic it is necessary to have an introduction to multicast. The next chapter is an introduction to the basics of IP multicast.

Chapter 4 - Introduction to Multicast

4.1 - What is IP Multicast

IP multicast is an extension to the IP protocol to support communications between one source and many recipients[1]. IP multicast extensions enable one stream of data to be received by several recipients without packet duplication. With IP multicast only one copy of the data is sent from the sender to all the receivers. This characteristic translates in bandwidth saving and ability to scale to groups of thousands of receivers.

Multicast is supported by many data link (layer 2) technologies. For example, there is support for multicast on Ethernet. If the scope of an application is limited to a single LAN, then it is feasible to use the support of multicast that is provided by the data link layer.

However, many useful applications span several local area networks and possibly several data link technologies such as Ethernet, Token Ring, FDDI, ATM, Frame Relay, SMDS or other networking technologies. For this reason, it is best to implement multicast at the network layer.

4.2 - Motivation

The Internet as we know it today uses point-to-point TCP connections between a sender and a receiver. The information gets replicated when a server transmits the same information to many recipients (e.g. a seminar, a

[1] Some people refer to multicast as multipoint communications.

videoconference, etc.). This limits the number of recipients that can join the group severely and wastes network resources.

Figure 4. 1 shows how bandwidth is wasted in many places using unicast to transmit the same information several times. The server has to send the information repeated as many times as clients require the feed. Intermediate links have to carry repeated information too. And finally, LAN bandwidth is also wasted with repeated information. With multicast the information is only transmitted once from the server, and replicated only where needed. LAN segments only see one copy of the server feed, even though there are several clients in the same segment.

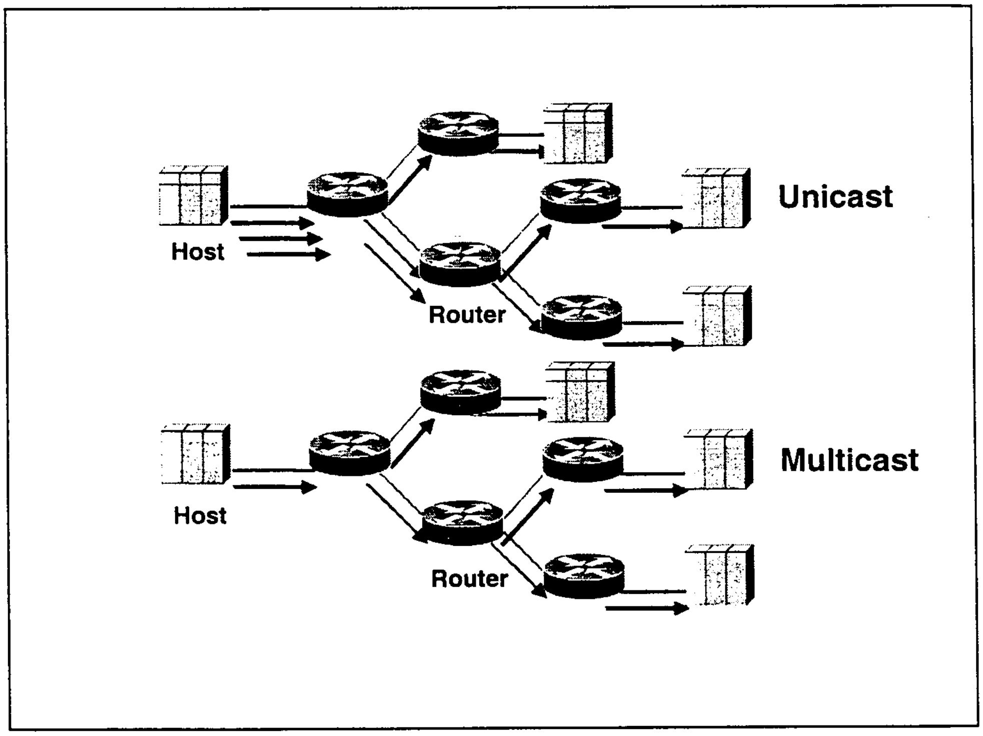


Figure 4. 1 - Unicast vs. Multicast Transmission (Source: [Cisco10-98])

4.3 - Problems with traditional methods

The problem of sending information from one source to many destinations, but not all of the nodes in an internetwork can be solved by using multiple unicast sessions or by sending broadcast packets. These two approaches are not bandwidth efficient. In the first case, a sender sends one copy of the message per recipient. In the second approach, the message is sent to all the nodes in the network, which means they have to process it and thus are wasting their resources.

Broadcast and unicast solutions have design flaws that do not permit them to scale to large groups. Some of the problems are:

- Bandwidth is wasted with duplicated traffic.
- Replicated unicast imposes loads in hosts and routers.
- Broadcasts cause unnecessary processing for host's CPU.

There are two examples of applications that solve the problem of one-to-many communications without using IP multicast: Real Player[2] and PointCast[3].

Real Audio and Real Video are plug-in programs that are added to a web browser that make possible to send audio and video streaming over the Internet. These applications send a separate flow of IP packets for every receiver wanting to receive the feed. As the number of receivers increase it becomes more and more difficult for the server to send that many packets and also the bandwidth is quickly consumed.

PointCast is a push technology that allows a broadcaster to send information to a web browser, without any action taken from the receiver, just like broadcast media such as TV and radio work. PointCast also uses replication of unicast IP packets for every single receiver. PointCast has severe scaling problems and harms the performance of the network. The use of these applications in a network

[2] For more information go to: http://www.real.com

[3] For more information go to: http://www.pointcast.com

produces a serious inefficiency in the use of the bandwidth available and the network is likely to collapse soon, as more people start to use these new applications.

4.4 - Summary of standardization process

Steve Deering is considered the inventor of IP Multicast. He defined the host extensions for IP Multicast and defined a group membership protocol for LANs called Internet Group Membership Protocol (IGMP). He described his work in a series of RFCs (see **Error! Reference source not found.**). RFC-1112 is the most current document and it defines multicast extensions for hosts.

RFC	Year	Highlights	Status
966	1985	Deering and David Cheriton first proposed a multicast extension to the Internet Protocol. [Deering-85]	Obsolete
988	1986	Deering refined his previous proposal and defined version 0 of IGMP. Host extensions for IP multicasting. [Deering-86]	Obsolete
1054	1988	Another refinement to his proposal. [Deering2-88]	Obsolete
1112	1989	Deering defined the final version of the host's extensions for IP multicasting and defined version 1 of IGMP. As of this writing, this is the definitive document on the subject [Deering-89].	Standard

Table 4. 1 - Host Extensions for IP Multicast RFCs

IGMP has also been improved through the years. IGMP version 1 is defined in the appendix of RFC-1112 [Deering-89]. Since 1989 two more versions have been introduced (see Table 4. 2)

Spec.	Year	Highlights	Status
RFC-1112	1989	Appendix 1 of RFC-1112 describes IGMP version 1. It defines only two messages: Query and Report. Scoping with TTL [Deering-89].	Standard
RFC-2236	1997	IGMP version 2 defines Leave Group Message and Group specific query message. Introduces a querier election mechanism. [Fenner-97]	Standard
Internet Draft	1997	IGMP version 3 enables hosts to listen only to a specified subset of the hosts sending to the group [Cain-97].	Work in progress

Table 4. 2 - IGMP RFCs

Steve Deering finished his Ph.D. dissertation at Stanford University in 1991 [Deering-91]. His thesis compared six new algorithms for multicast routing. He proposed a service model for multicast that is widely used today and it also proposed a scheme for inter-domain routing that served as a base for later proposals such as CBT and PIM-SM.

He presented some of his ideas in several conference papers [Deering-88] [Deering-90]. In these papers he proposed extensions to two common routing algorithms - distance vector routing and link state routing. Also, he suggested modifications to the single-spanning-tree algorithm for multicasting in large extended LANs. He also presented his ideas related to inter-domain multicast routing.

Intra-Domain Proposals for multicast routing started to appear around the same time. An extension of RIP to support multicast was proposed in 1988 [Waitzman-88]. Other researchers proposed an extension to OSPF in 1994 [Moy2-94]. PIM-DM is a multicast routing protocol that is independent of the unicast routing protocol used. The latest specification for PIM-DM is available in [Deering-98b]. See Table 4. 3 for a summary of intra-domain multicast routing protocols. Chapter 5 describes these protocols in greater detail.

RFC	Year	Highlights	Status
1075	1988	It uses Reverse Path Multicasting (RPM) to create a distribution tree. This document has been declared historic [Coltun-98] and the current implementation is defined in [Pusateri-98].	Obsolete
1584	1994	Defines multicast extensions for OSPF.	Standards Track
-	1998	Defines Protocol Independent Multicast - Dense Mode	Work in progress

Table 4. 3 - Intra-Domain Multicast Routing Protocols

Based on Deering's work many other proposals started to appear for Inter-Domain Routing protocols. In particular Anthony Ballardie's proposal in the 1993 SIGCOMM conference has gained considerable attention [Ballardie-93]. His proposal was called Core Based Trees (CBT) and was further developed in his Ph.D. thesis from the University College London [Ballardie-95]. The CBT protocol was formally specified in [Ballardie-97].

Ballardie's proposal generated more work and in 1994 a new approach based on CBT was created. This new approach was called Protocol Independent Multicast - Sparse Mode and it was first covered in [Deering-94], further refined in [Deering-96] and formally specified in [Estrin-98]. Table 4. 4 is a summary of the two main proposals for Inter-Domain Multicast Routing from the IETF. Chapter 6 summarizes other proposals for inter-domain multicast routing.

RFC	Year	Highlights	Status
2189	1997	Defines CBT protocol specification [Ballardie-97]	Experimental
2362	1998	Defines PIM-SM [Estrin-98]	Experimental

Table 4. 4 - Inter-Domain Multicast Routing Protocols

Work in multicast routing is a hot area of research at this time. Two groups are currently working in the IETF defining IP Multicast:

- Inter-Domain Multicast Routing (IDMR), in the Routing Area, which is in charge of defining the routing protocols that make multicast a reality.
- Mbone Deployment (MboneD), in the Operations Area, which is in charge of the technical and engineering details of deploying IP multicast to the enterprise.

Other groups are also putting forth effort in standardizing IP Multicast. Among these groups, the two most important ones are the IRTF and the IEEE:

- The Internet Research Task Force (IRTF), which is chartered by the IETF to do long term research about Internet issues, it is currently putting a lot of effort on the issue of **reliable multicast**[4].
- The IEEE is involved in the standardization of IP Multicast over switched LANs (e.g. switched Ethernet), and has created the 802.1p specification.

4.5 - Critique to RFC-1112

The IP Multicast Model proposed by Deering in RFC-1112 [Deering-89] does not require a sender to be part of a group. This may cause undesirable situations such as a malicious user sending unwanted content to a group. The initial idea of Deering was to simplify as much as possible the host model, however it would be better to limit somehow the ability of a host to send to a group.

A second problem of the IP Multicast model is that a host may join any group without prior consent from the sender. This is in contrary to a possible business model for Internet broadcasting, in which users are charged for the content they receive on pay per view basis.

A review of this model should be done since routing protocols are based on the IP multicast model. If there are design flaws in this model then the routing protocols and applications that use this model inherit its problems.

4.6 - IP multicast basics

When a source node wants to send a message to some subset of the other nodes, but not all of them, the type of communication is called **multicast**. The other

[4] For more information visit the "Reliable Multicast Research Group" web site at http://www.east.isi.edu/rm/

two types of communications can be seen as specific cases of multicast. When the destination group is only one node, then we have a **unicast** type of communication. When the destination group is all the existing nodes in the internetwork, then we have a **broadcast** type of communication.

Applications running in hosts that want to join a multicast group issue commands that specify a Class D address that identifies a multicast feed. The Internet Group Management Protocol (**IGMP**) is the protocol used by routers to communicate with hosts in a LAN to learn about the groups that these hosts want to subscribe to.

A tree is a unique path including routers and subnetworks from a source to a number of receivers. Some routing protocols create trees that have the root in the source (**source-based trees)**. Some other routing protocols create trees that are rooted in an intermediate router in the network that is called a rendezvous point. These type of trees are called **shared-based trees**.

A **group** is a set of hosts that are listening to a source. The hosts could be in different LANs. When a host wants to listen to a source then it **joins** a group, and when a host loses interest in the transmission it **leaves** the group. This means that **membership** to groups is dynamic, that is, a host could join or leave a group at any point of the transmission. Just like, we switch TV channels or radio stations. This dynamic property of the group members means that the **distribution tree** has to grow or shrink dynamically. When you shrink a tree, the action is called **pruning** the tree.

Figure 4. 2 shows a source based tree for the server on LAN F. There are members of the group only on LANs A, D and E (leaves). The tree is pruned for LANs B and C. Since LAN E is multi-homed, then a router must be elected to forward the feed for that group. The other router is pruned. At the end of the process, routers have state in their routing tables that let them know how to forward multicast packets in the right directions.

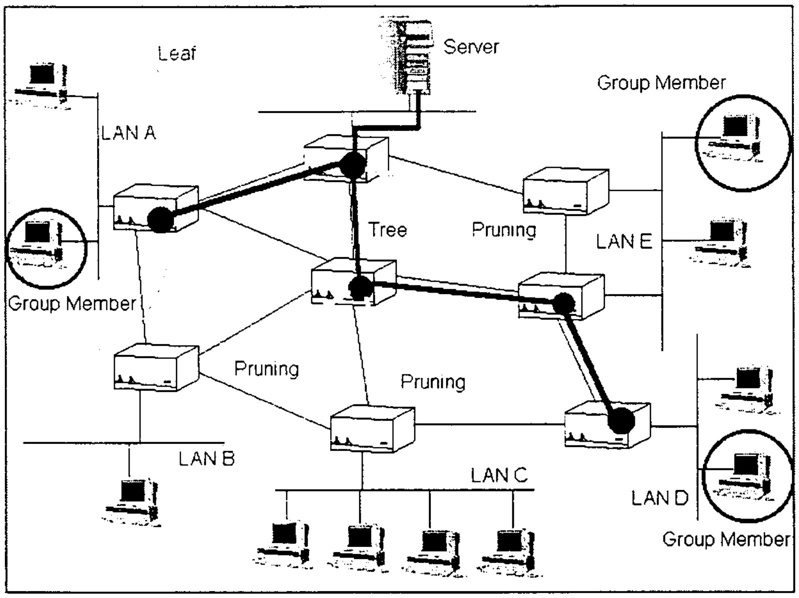


Figure 4. 2 - IP Multicast basic jargon (tree, group, leafs, prune)

4.7 - Benefits

By using IP multicast instead of a regular replicated unicast for one-to-many and many-to-many communications the following benefits are achieved:

- The main benefit of IP multicast is that it saves bandwidth, which indeed reduces networking costs and increases response times of network applications.
- It enables the fast delivery of multimedia information. IP multicast enables new applications such as videoconferencing over the Internet and video streaming which are bandwidth intensive. Without IP multicast the support for these applications has serious scaling problems.
- By using IP multicast the server has less load, since it only sends the feed once and the network redistributes the content to only the receivers that are part of the

group. This means that a cheaper server can be used. In the same way, the load imposed on intermediate routers along the path is also greatly reduced.

- IP Multicast enables Internet broadcasting to a large group of participants.

4.8 - MBONE

Researchers have been using a "testbed network" that exploits the benefits of IP Multicast for about six years. They have constructed this network as a cooperative and volunteer effort. Researchers have called this overlay network the Multicast Backbone (MBONE)[5]. The idea is to be able to test multicast capabilities before multicast is widely available in hosts and routers around the world.

The MBONE is composed of islands that can directly support IP multicast, such as multicast LANs like Ethernet, linked by "IP tunnels". These tunnels are created with the encapsulation of multicast packets in unicast packets between workstations having operating system support for IP Multicast[6] and running the "mrouted" multicast routing daemon. The encapsulation is added on entry to a tunnel and stripped off on exit from a tunnel. In this way, routers along the way that do not understand multicast have no problem, since they just are routing unicast packets.

At the present time, only portions of the Internet have multicast-capable routers. The size of the MBONE, compared to the Internet as a whole, is relatively small. As February of 1995, the Internet was home to 48,500 networks. At the same time the MBONE only comprised about 1,700 networks (roughly 3.5 percent of the

[5] For more information on the MBONE read [Eriksson-94] and visit http://www.mbone.com

[6] Windows 95, Windows NT 4.0 and major flavors of UNIX support the host extensions of multicast defined in the RFC-1112 [Deering-88].

Internet) [Savetz-96]. Casner produced a map of the MBONE in May of 1994 [Casner-94]. See Figure 4. 3. At that time there were 20 countries and only 901 routers running "mrouted".

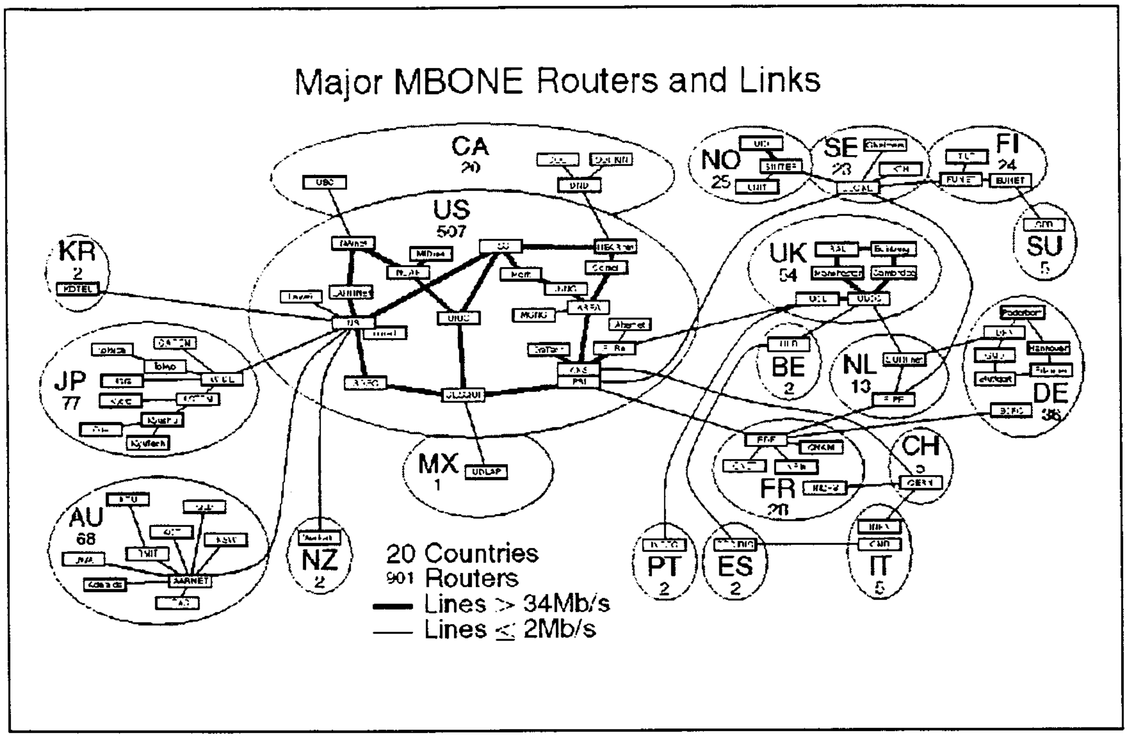


Figure 4. 3 - MBONE Map as May of 1994 (Source: [Casner-94])

The first videoconference over the Internet was held on February 18 of 1992 as a meeting of the IETF Audio/Video Transport (AVT) group [Casner-92]. Since then many other events have been transmitted on the Internet, such as NASA events, Rolling Stone Concerts, etc. For a summary of MBONE events see [Stein-98]

For more information on operational details of the MBONE see [Casner-93].

4.9 - IP Addressing for IP Multicast

Each feed that is transmitted on the MBONE is identified by a Class D address. It is a very similar concept to the one used in radio stations, in which each station is assigned a frequency band. The difference is that Class D addresses are not pre-assigned; they are dynamically assigned.

There are 268 million class D addresses from the range of 224.0.0.0 to 239.255.255.255 of the IP address space. The high-order 4 bits of a Class D address are set to 1110, and the remaining 28 bits are set to a specific multicast **group ID** (see Figure 4. 4).

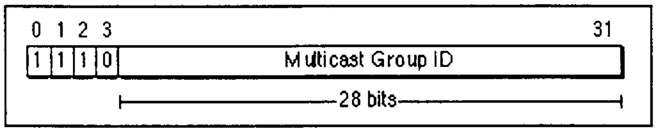


Figure 4. 4 - Class D IP Address format (Image Source: Semeria 98)

The block of multicast addresses ranging from 224.0.0.1 to 224.0.0.255 is reserved for the use of routing protocols and other low-level topology discovery or maintenance protocols. Table 4. 5 shows some of the pre-assigned multicast addresses. The remaining groups ranging from 224.0.1.0 to 239.255.255.255 are assigned to various multicast applications dynamically or remain unassigned. From this range, 239.0.0.0 to 239.255.255.255 are to be reserved for site-local "administratively scoped" applications, not Internet-wide applications. Figure 4. 5 shows the range of 268 million addresses and their distribution.

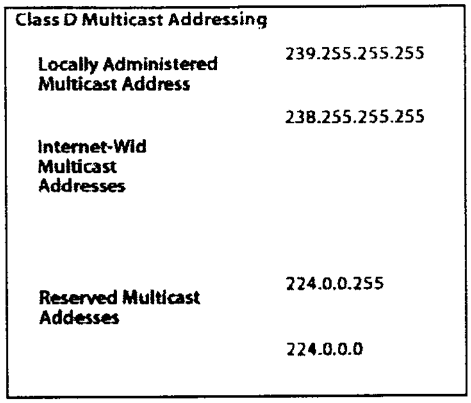

Class D Multicast Addressing
Locally Administered Multicast Address
239.255.255.255
238.255.255.255
Internet-Wid Multicast Addresses
224.0.0.255
Reserved Multicast Addesses
224.0.0.0

Figure 4. 5 - Distribution of Class D Addresses

Multicast Address	Use
224.0.0.1	All multicast systems in the LAN (hosts and routers)
224.0.0.2	All multicast routers
224.0.0.3	All DVMRP routers
224.0.0.5	All OSPF routers

Table 4. 5 - Pre-assigned multicast addresses [Reynolds-94]

There are 16 million reserved addresses and 16 million addresses to be used inside of an Autonomous System. The remaining 236 million can be assigned dynamically Internet wide. This number is extended using scoping as discussed in the next sections. The next section introduces how these addresses are dynamically allocated in the Internet.

4.10 - Multicast Address Allocation

When a sender has a feed to send, it has to dynamically get a multicast address from the pool of available addresses. An early attempt to solve this problem was to make the application randomly select an address until an unused one was found. But this requires that the probability of two applications choosing the same address to be very small.

The behavior is that when allocating from an address space of size N, if the number of addresses allocated is less than the square root of N then the probability of collision is negligible, but when there are more than square root of N addresses allocated, then the probability of collision is 1 [Jacobson2-94]. See Figure 4. 6. Since the IPv4 multicast address space is ≈ 2^{28}, about 2^{14} (around 16,000 addresses) can be dynamically allocated before the probability of collision is 1.

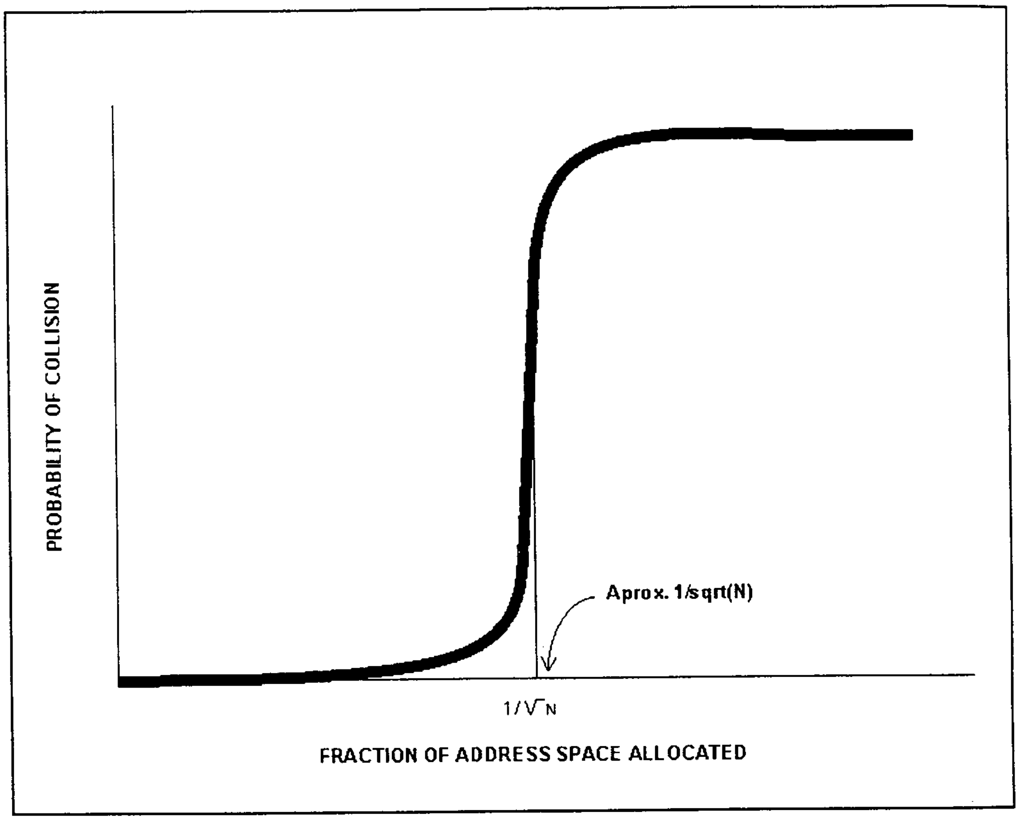


Figure 4. 6 - Probability of collision in an address space of N addresses (Source: [Jacobson-94b])

4.11 - Joining a multicast group

Hosts wanting to receive a multicast feed need to tell their local router that they are interested in listening to that "channel". They tell their local routers about the group membership interest using IGMP.

The Internet Group Management Protocol (IGMP) runs between hosts and their immediate neighbor multicast routers only. In order to communicate group membership to other networks inside of the AS, routers use one of the multicast routing protocols described in chapter 5. For communication of group membership between AS, routers use one of the inter-domain routing protocols described in chapter 6.

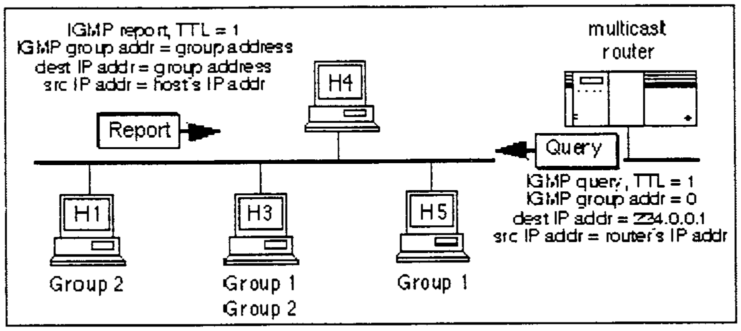


Figure 4. 7 - Router issues an IGMP Query to determine group membership. Host responds with IGMP Report

Multicast routers that run IGMP use **IGMP host-query** messages to keep track of the hosts that belong to multicast groups. These messages are sent to the all-systems group address 224.0.0.1. The hosts then send **IGMP report** messages listing the multicast groups they would like to join. When the router receives a packet addressed to a multicast group, it forwards the packet on those interfaces that have hosts that belong to that group (See Figure 4. 7)

IGMP **version 1** is most common [Deering-89]. In IGMP 1 implementations, hosts cannot explicitly exit a multicast group, but instead simply stop reporting interest in a given group. The router then halts retransmission of the traffic for that group upon expiration of a time-out mechanism.

IGMP **version 2** is defined in [Fenner-97]. Among its improvements, IGMP 2 lets a host inform a multicast-aware router when it no longer wants to receive traffic for a given multicast group. This decreases the leave latency and saves bandwidth in the local LAN.

IGMP **version 3** is a preliminary draft specification published in [Cain-97]. The main improvement that will be introduced is that hosts can listen only to a specified subset of the hosts sending to the group.

4.12 - Receiving Multicast datagrams

Once a router forwards a datagram to a LAN, then hosts in that LAN need to be prepared to listen for that group-ID address. This requires a mapping between the Group ID multicast address and the data link address. For each data link layer protocol a different mapping is needed. These mappings have been specified recently for almost all data link technologies such as Ethernet, Token Ring, FDDI, ATM, Frame Relay, etc.

An application running in a host can program its NIC card to listen to that multicast address.

In Ethernet, the class D addresses are mapped into MAC layer addresses by taking out the last 23 bits of the MAC address and substituting that with the last 23 bits from the IP address. At the data link layer (level 2), the MAC addresses are marked with one bit to be multicast addresses. Figure 4. 8 illustrates how the multicast group address 224.10.8.5 (E0-0A-08-05) is mapped into an Ethernet (IEEE-802) multicast address (01-00-5E-0A-08-05). The first 3 bytes are reserved and they are always (01-00-5E) [Reynolds-94].

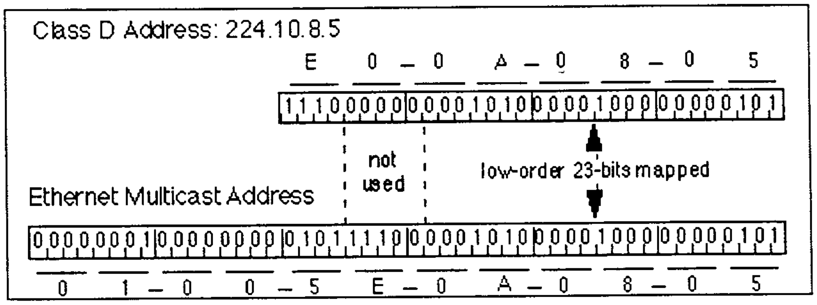


Figure 4. 8 - Mapping of an IP address (L3) to an Ethernet Address (L2) (Image Source: [Semeria-98])

4.13 - Multicast Operations

This section provides a high level overview of what needs to be upgraded in the current network in order to provide IP multicast natively.

4.13.1 - Roadmap of the deployment

Many things have to be upgraded in your network before you can fully utilize all the benefits derived from IP Multicast. For example, router, switches, TCP/IP stack, NIC cards, desktop operating systems and application software needs to support IP multicast. If you have recent equipment (say two years old) it is likely that it already supports multicast and it may be very easy to do the upgrade. However, if you have older hardware or software it may be hard to accomplish the task of deploying IP Multicast in your enterprise. This section will give you some hints on where to start and what to do in order to start the process.

Probably the best way to start is to deploy a pilot project in a single LAN segment of the Intranet, just to get familiar with the new technology and to make sure everything will work fine in your environment. After that, you can continue to support IP multicast for a WAN connection and test it with the LAN test previously deployed. This will give you a good understanding of the technologies involved and also will give you on-site training to your network staff, so that when the applications force you to deploy IP multicast you will be already prepared.

Some of the key points that have to be reviewed to upgrade a network to support multicast are:

1. First, each sending and receiving host's operating system and TCP/IP stack must support multicast, including IGMP. Second, each host's network adapter driver must implement multicast. Third, routers and switches must be multicast-capable

or at least not multicast destructive. Fourth, applications must be multicast-enabled.

2. Check the equipment you have. Many newer routers already support IP multicast, although reconfiguration or upgrades may be required.
3. Pick the right multicast routing protocol for your internetwork. Different protocols apply, depending on the underlying unicast routing protocol and density of the multicast population. This thesis offers a good starting point on how to choose the protocol and proposes a criteria to compare the alternatives.
4. Evaluate your LAN switches or consider buying one. Multicast filtering switches can substantially reduce LAN traffic.
5. Keep track of new developments done by the IETF and other standards organizations. Standards are being developed for enhanced services, like reliable and real-time delivery of IP multicast traffic.
6. Check what forward-thinking ISPs are doing. Some of them are already testing multicast service over their backbones.

4.13.2 - Problems deploying IP multicast

Through last year, IP multicast technology made a big improvement in market penetration. However, the bandwidth-conserving technology is not yet widely deployed in the Internet or within corporate networks [Nerney-97]. Some of the reasons of why IP multicast has not been widely deployed yet are explored in this section. IP multicast has not been deployed in commercial networks as quickly as expected for several reasons:

- User's lack of interest in killer applications that require IP multicast. At this point in time, there is no real pressure from users to have IP multicast deployed.

- Insufficient support in routers and other network equipment. Many network managers have hesitated to multicast-enable their networks because upgrading requires new, expensive hardware and is a delicate and time-consuming process. Multicast requires all sending and receiving hosts as well as the network infrastructure to be multicast-enabled or not multicast destructive.
- Support for IP Multicast is not well tested in a number of platforms and environments.
- Developers, knowing that most network devices do not support IP Multicast, have not created applications that support it.
- The majority of the applications in the current Internet (e.g. www, email, ftp, telnet, etc.) are point-to-point based and don't get any immediate benefit from IP multicast.
- Firewalls block UDP for security reasons, and those block IP Multicast traffic too.
- Some applications need a reliable multicast transport protocol. At this point, there is no single solution for this problem. A survey of the alternatives for reliable multicast protocols can be found in [Obraczka-98, Tascnets-98].
- For ISPs the major problem is that there is no Inter-Domain Routing Protocol available at this moment. DVMRP, PIM-DM and MOSPF do not scale to the Internet. PIM-SM and CBT are supposed to solve this problem, but have many open issues as seen in later chapters of this thesis. A multicast version of BGP has been recently proposed [Thaler-98].
- It might be hard to get a tunnel to connect to the MBONE. This is because the MBONE is a cooperative volunteer effort from researchers around the world and it depends on how lucky you are in order to setup a tunnel to a multicast router that is connected to the MBONE.

- For ISPs, there is a business problem since IP Multicast reduces the amount of traffic that transverse their networks. This is bad for them since it means fewer dollars when charging by usage. However, it also increases the number of receivers that want to listen to a multicast feed. So, in reality, the deployment of multicast could mean more traffic for the ISP. A traffic study should be accomplished to evaluate the impact of IP Multicast and its effect on the business model of ISPs.
- In the transition period, solutions like Multicast tunnels are being used. These solutions take time to configure and may have to be torn down once there is native support for IP Multicast in the Internet.

As a summary, the deployment of IP Multicast could be time consuming and it is influenced by many factors that are not only technical issues. The deployment of IP Multicast will be accelerated once an application appears that requires the use of IP Multicast to work in a proper way. Next, section explores some of these applications.

4.14 - Some Existing Applications

Of all the applications, Webcasting (Broadcast of voice and video over the Internet) seems like a good candidate to be the killer application that will drive the deployment of IP Multicast. The applications for IP multicast reside wherever there is a need to distribute information in a one-to-many relationship. Some of the examples include:

- Webcasting: Voice and Video "Broadcasts" over the web
- Multimedia Conferencing.
- Distance Learning
- Multimedia Kiosks

- Pushing web content
- Software distribution
- One to many email
- News distribution
- Stock quotes distribution
- Video Streaming
- Database replication
- GroupWare
- Large File transfers
- Client Replicated Databases

The World Wide Web served as the killer application that induced the wide deployment of the TCP/IP protocol suite in the Internet. So far, there is no application that has forced the wide deployment of IP multicast in the Internet. There have been some applications such as Real Audio and PointCast that have the range of being killer applications, but they have been deployed using IP unicast. The performance is good for end-users using unicast as long as the traffic levels are kept low enough. As more traffic is added, IP multicast will have to be deployed.

In the author's opinion, the biggest killer application that will force the deployment of IP multicast is when broadcast stations, such as CNN, start to send live programming through their web site. At that point the delays experienced by users of a clogged network will force broadcasters to invest in IP Multicast technologies, so that the transmissions can be done smoothly.

4.14.1 - Free multicast tools

There is a number freely available multicast tools for multimedia conferencing in the MBONE. The most common applications are "vat" and "vic" developed by the

Network Research Group at the Lawrence Berkeley National Laboratory in collaboration with the University of California, Berkeley. These programs enable a user to participate in audio and videoconferences over the Internet[7].

4.14.2 - Commercial products

Some of the companies that are developing products for IP Multicast include: 3Com, Ascend, Cisco, Cabletron, GlobalCast, Hewlett-Packard, Hughes, Intel, Media4, Microsoft, Microspace, Newbridge, StarBurst, UUNET, @Home, and many others.

The market for IP Multicast enabled applications and networking equipment is quickly growing and it is likely to continue to grow more in the next few years. More and more companies have realized the benefits of IP multicast in their networks. For an up-to-date snapshot of the IP Multicast market visit the "IP Multicast Initiative" web site[8].

4.15 - Chapter Summary

This chapter has presented a summary of the basic technologies that are the basis of the IP Multicast model. Each individual multicast group is identified by a particular **class D IP address**. Each host can join or leave a multicast group through the use of the **Internet Group Management Protocol (IGMP).** Routers keep track of these groups dynamically and build multicast distribution trees that create paths

[7] For other free tools visit:

http://www.merit.edu/net-research/mbone/index/titles.html and

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

[8] The URL is http://www.ipmulticast.com

from each sender to all receivers. When a router receives a multicast packet for a group, it forwards that packet on all the interfaces that appear on the outgoing list for that group in the multicast routing table.

Operational challenges were explored in this chapter. Many things need to be upgraded to support multicast, and it is likely that many problems will appear while trying to deploy IP Multicast in the enterprise.

Market applications were also summarized in this chapter. This author expects explosion of commercial products in the next few years, especially after a killer application for multicast appears. In this author's opinion the killer application for multicast could be TV-like Internet Broadcasting.

Chapter 5 - Intra-Domain Multicast Routing Protocols

IP Multicast needs to deal with three basic problems to make the communication feasible: addressing, dynamic registration and multicast routing. This thesis deals primarily with the problem of multicast routing. This chapter covers the three protocols that the IETF has standardized: DVMRP, MOSPF and PIM-DM. These protocols are based on well-known multicast algorithms. Next section covers the main algorithms known to build multicast distribution trees.

5.1 - Multicast Algorithms

Many multicast routing protocols use algorithms that are based on the Reverse Path Forwarding algorithm proposed by Dalal and Metcalfe in 1978 [Dalal-78].

Routers have to run algorithms that will allow them to create a tree that interconnects a source with several destinations. IGMP is not concerned with the delivery of multicast packets between neighboring routers or across an internetwork. IGMP only deals with group membership on a Local Area Network. To provide an Internet-wide delivery service, it is necessary to define multicast routing protocols. A multicast routing protocol is responsible for the construction of multicast packet delivery trees and performing multicast packet forwarding. This section explores a number of different algorithms that may potentially be employed by multicast routing protocols:

- Flooding
- Spanning Trees
- Reverse Path Broadcasting (RPB)
- Truncated Reverse Path Broadcasting (TRPB)

- Reverse Path Multicasting (RPM)
- Steiner Trees
- Core-Based Trees

Reverse Path Forwarding (RPF) is the name of the algorithm proposed by Deering in his thesis [Deering-91]. A RPF algorithm allows multicast frames to reach all sub-networks without traveling back to the source. The name "reverse path forwarding" refers to the fact that forwarding is based on the router's knowledge of the shortest path back to the source. Since then, many new variations have been proposed to the algorithm. The most current and efficient one is **Reverse Path Multicasting (RPM)**. Even some other approaches have been proposed such as Steiner Trees and Core Base Trees.

Steiner Trees based solutions have been proposed by researchers. The one that is considered the most optimal is described in [Kou-81]. Other proposals are described in [Takahashi-80] [Rayward-86] [Winter-87] [Cimet-87] [Noronha-94]. In this thesis the Steiner tree proposals will not be analyzed since it has been proved that Steiner Trees are not practical [Hwang-92] [Doar-93] [Bauer-95] [Diot-97]. Core Based Trees have gained more acceptance and they are used by many current routing protocols (e.g. PIM-SM and CBT). Core Based Trees were first proposed by Wall in his Ph.D. dissertation [Wall-80].

Table 5. 1 summarizes the advantages and disadvantages of all these algorithms. The development of such algorithms is an area of extensive research today.

Algorithm	Advantages	Disadvantages
Flooding	• Simple.	• Does not scale. • Inefficient use of router memory.
Spanning Trees	• No loops. • A lot of experience by developers.	• Can centralize traffic on a small number of links. • May not provide the most efficient path.
Reverse Path Broadcasting (RPB)	• Simple. • Reasonably efficient. • Does not require the router to know the entire spanning tree. • No need for a mechanism to stop the forwarding process.	• It does not take into account group membership.
Truncated Reverse Path Broadcasting (TRPB)	• Eliminates unnecessary traffic on leaf sub-networks.	• It does not take into account group membership.
Reverse Path Multicasting (RPM)	• Spanning tree only spans to routers and sub-networks with group members along the shortest path.	• It still does not scale. • Requires periodic flooding. • State required for each (source, group) pair.
Steiner Trees	• Create optimal trees.	• Not practical.
Core-Based Trees	• State only for each group. • Uses explicit join, which saves network bandwidth.	• Traffic concentration and bottlenecks at core routers. • May create sub-optimal routes.

Table 5. 1 - Advantages and disadvantages of multicast algorithms

5.2 - Multicast Routing Protocols

Multicast Routing creates a delivery tree from a source to all the receivers that want to be included in a group. Multicast routing protocols use the group membership information gathered by the Internet Group Management Protocol (IGMP), and unicast routing tables to tell upstream routers about their desire to be included in a group. There are three protocols that have been proposed by the IETF to be used internally in a domain for multicast routing:

- Distance Vector Multicast Routing Protocol (DVMRP) was first formally defined in RFC 1075 [Waitzman-88]. DVMRP has scaling problems because of the necessity to flood frequently. The problem has been somewhat solved with the implementation of pruning in newer versions of DVMRP [Pusateri-98].

- Multicast extensions to OSPF (MOSPF) is defined RFC 1584 [Moy-94]. It does not work well in environments that have many active sources or environments that have unstable links [Maufer-98].
- Protocol Independent Multicast-Dense Mode (PIM-SM) is an Internet draft from the IDMR group [Estrin-96]. This is an improved version of DVMRP that does not require any particular unicast routing protocol. This protocol is intended for internal networks that are densely populated. It does not solve the problem of inter-domain multicast routing.

5.3 - DVMRP

DVMRP is a distance vector multicast routing protocol that builds per source-group multicast distribution trees using the Reverse Path Multicasting algorithm ("broadcast and prune") [Deering-90]. DVMRP was developed after RIPv1, however the infinity in DVMRP is 32 hops and it supports CIDR. Also, it makes special use of the Poison-Reverse advertisements to indicate "child-parent" relationships. DVMRP is implemented in a multicast routing daemon called "mrouted" [Deering]. The most current version is 3.8[1].

5.3.1 - Motivation

In 1988, a standard existed to support multicasting over IP networks [Deering2-88]. However, no routing protocols to support internetwork multicast were

[1] Mrouted can be obtained for free in this web site:

http://www.merit.edu/net-research/mbone/index/titles.html#mrouted available. DVMRP proposed the first solution to the problem. DVMRP was developed to experiment with the algorithms proposed in [Deering-88].

5.3.2 - History

DVMRP is almost ubiquitously available on the MBONE, mainly because it was the first multicast routing protocol invented. It was first defined in the RFC 1075 in November of 1988 [Waitzman-88]. Originally, the entire MBONE ran only DVMRP. Nowadays, DVMRP still remains as the core multicast routing protocol of the MBONE, however many other multicast routing protocols are also running.

The original specification was derived from RIP and used the **truncated reverse-path broadcasting (TRBP)** algorithm to construct the distribution tree. The current version of DVMRP implements the **Reverse Path Multicast (RPM)** algorithm [Deering-90].

DVMRP version 1 [Waitzman-88] has been declared obsolete [Coltun-98]. The most current version of DVMRP is version 3 and it is currently an Internet draft of the Inter-Domain Multicast Routing (IDMR) working group of the IETF [Pusateri-98].

5.3.3 - DVMRP Terminology

DVMRP introduces the concept of a **virtual interface**. A virtual interface is either a physical interface (e.g. Ethernet, FDDI, etc.) or a tunnel. A **tunnel** is the encapsulation of an IP multicast packet into a unicast packet in order to transverse non-multicast-capable routers. Tunnels typically use either IP-in-IP encapsulation [Perkins-96] or Generic Routing Encapsulation (GRE) [Hanks1-94, Hanks2-94]. IP-in-IP encapsulation is assigned IP protocol 4. Each tunnel needs to be configured by a network manager by specifying the destination router, a cost and a threshold. Figure 5. 1 shows a tunnel from an autonomous system to the MBONE.

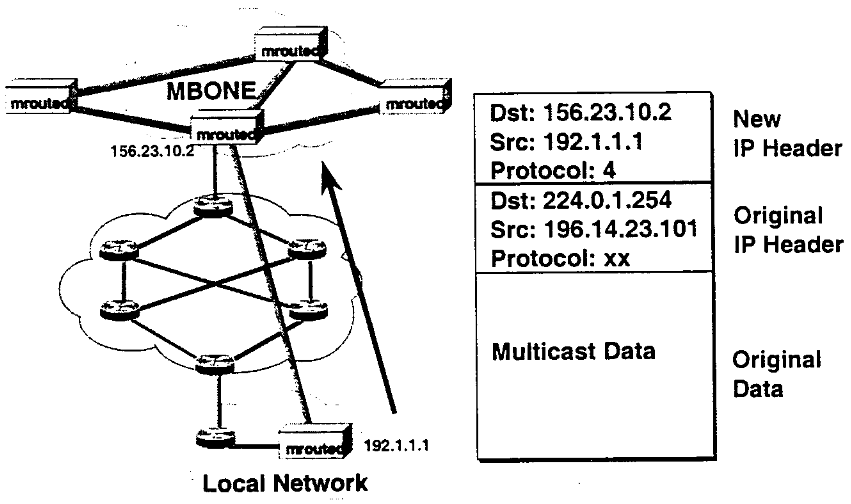


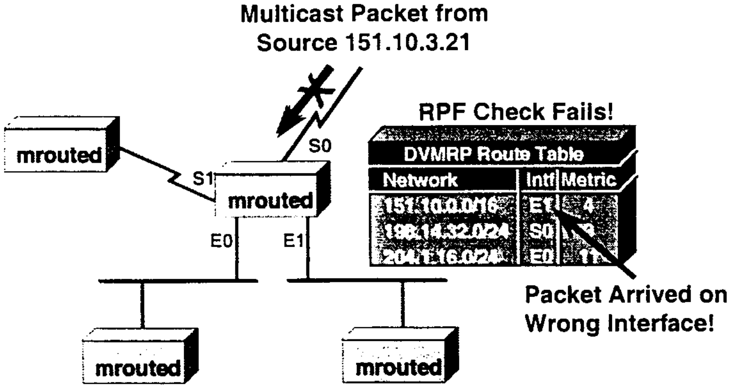
Figure 5. 1 - DVMRP Tunnel (Image Source: [Cisco10-98])

The threshold of a tunnel allows specifying scope to a multicast packet. If a multicast router receives a packet with a TTL that is lower than its threshold then it discards the packet. This mechanism allows restriction of multicast feeds to certain domains and also permits re-use of multicast address. Another option to specify scopes for a group is to use administratively scoping [Meyer-98].

DVMRP performs a test called the reverse path forwarding check (**"RPF check"**) to every multicast packet that arrives to the router. The rule says, "A router forwards a multicast packet if received on the interface used to send unicast packet to the source". The interface with the shortest distance to the source is called the **RPF interface**. If the "RPF check" fails, the packet is dropped. If the "RPF check" succeeds, the packet is forwarded to each interface in the outgoing interface list.

Figure 5. 2 illustrates a packet from source 151.10.3.21 that is dropped because it arrived in the wrong interface.

Incoming Packets Must Pass RPF Check



DVMRP Route Table

Network	Intf	Metric
151.10.0.0/16	E1	4
198.14.32.0/24	S0	[illegible]
204.1.16.0/24	E0	11

Figure 5. 2 - "RPF Check" (Image Source: [Cisco10-98])

5.3.4 - Message Types

All DVMRP messages are encoded as IGMP Type 0x13, with the code field in the IGMP header indicating the DVMRP packet type. There are separate packets for neighbor discovery (Probes), distribution of multicast source information (Route Reports), tree maintenance (Prunes, Grafts, and Graft Acks) and debugging messages (Ask Neighbors 2 and Neighbors 2). Table 5. 2 has a summary of the messages used in DVMRP. Figure 5. 3 shows how the DVMRP messages are encapsulated inside of IGMP packets.

Code	Packet	Description
1	Probe	For DVMRP neighbor discovery
2	Route Report	For multicast routing exchange
7	Prune	For pruning multicast delivery trees
8	Graft	For graft multicast delivery trees
9	Graft Ack	For acknowledgement of Graft messages
5	Ask Neighbor 2	Request for list of DVMRP neighbors
6	Neighbor 2	Response to above request

Table 5. 2 - DVMRP Message Types

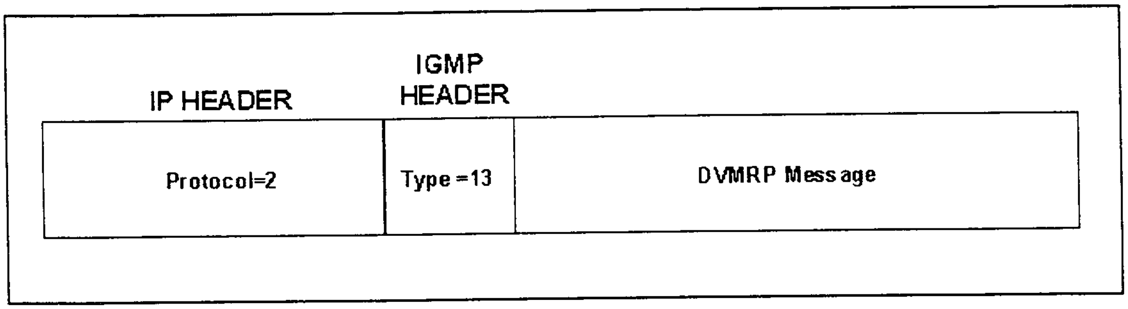


Figure 5. 3 - DVMRP Packet Encapsulation

5.3.5 - How it works

DVMRP routers dynamically discover DVMRP neighbors using the "**Probe**" messages. Probes are multicast to the "All-DVMRP-Routers" (224.0.0.4) group address. At the end of this process there is a two-way neighbor adjacency between a DVMRP routers and all of its neighbors.

Initially, when a DVMRP router receives a multicast packet, it performs the "RPF check". If the "RPF check" succeeds then it **floods** the packet to all of its downstream interfaces for this group (i.e. all interfaces except the RPF interface). If a router realizes that it does not have group members in its downstream interfaces[2], then it sends a "**Prune**" message towards the source. This temporarily stops the transmission of multicast packets to this branch of the tree. Eventually, the prune

[2] A router knows group members on an interface through the IGMP protocol.

state in the upstream router times out and multicast packets start to flow again downstream. This process is repeated over and over again and it is the reason DVMRP is called a "broadcast and prune" protocol.

If a new group member appears in a subnetwork then a **"Graft"** message is sent upstream towards the source to extend the tree to this new group member. This message invalidates the previous "Prune" message and starts the flow of multicast packets for this branch again. The "Graft" message is re-transmitted until a **"Graft-Ack"** message is received.

Each DVMRP router periodically broadcasts **"Route Reports"** to its neighbors, which contain a list of sources and the distance from these sources to the router. With this information, a DVMRP router can calculate the previous hop on each multicast source's path (The previous hop on a source's tree is the DVMRP router that is advertising the shortest distance from the source). A DVMRP router also notices if it is a child or a parent router for a particular group[3]. If it is a child router, then it sends a **"Poison Reverse"** to its upstream router indicating that it is expecting to receive traffic from the parent router for these sources. "Poison Reverse" is an entry on a "Route Report" message with a cost of infinity or more (>32).

From the "Route Report" messages a routing table is built. This table contains "source prefixes" and "from-gateways" instead of "destination prefixes" and "next-hop router" fields that unicast routing protocols build. The source prefix indicates a network that has a source sending to a group. The "From Gateway" field

[3] If a router has a better metric for a source's network then it is the parent for this group. Otherwise, the router is a child for this source.

indicates the previous-hop router leading back toward a particular source prefix. See Figure 5. 4.

Source Prefix	From Gateway	Metric
128.138.0.0/16	128.138.1.15	5
172.1.15.0/24	172.1.15.12	3
198.1.1.0/24	198.1.1.13	7

Figure 5. 4 - Sample DVMRP Routing Table

The routing table does not contain group membership information. Based on this routing table, the database of known groups from IGMP and received prune messages, DVMRP builds a multicast forwarding table. This table tells a router how to forward a multicast packet. It contains the source network and the group plus the Incoming interface and the outgoing interface list. See Figure 5. 5.

Source Prefix	Group	Incoming Interface	Out going Interface List
128.138.0.0/16	224.1.1.1	1	2,3,4p
172.1.15.0/24	224.1.1.1	2	3,5
198.1.1.0/24	224.1.1.1	1Pr	2p 3p
145.2.1.0/24	235.1.1.1	1	3,4

Figure 5. 5 - Sample DVMRP forwarding table[4]

In subnetworks with several routers on it, a designated router (DR) is elected to avoid duplicate packets coming from the same source. The selection of the DR is based on the best route metric back to the source network (with the lowest IP address used as tie-breaker).

[4] A "Pr" in the Incoming Interface field indicates that a prune message has been sent to the upstream router.

5.3.6 - Evaluation

DVMRP has several **design flaws** that prevents it to scale to the entire Internet:

- It is a distance vector routing protocol, which means that it is slow to adapt to topology changes (i.e. slow to converge), and it is limited in its network diameter to only 32 hops[5].
- It must maintain source specific information when not on-tree. This has a serious scaling consequence: Keeping state in off-tree routers is a waste of valuable router memory. If there are millions of sources, then the router might consume all the memory for sources that their LANs are not interested in. In other words, source-specific state does not scale well as the number of sources increases.
- Multicast traffic is periodically broadcast across the entire internetwork. This is done to update group membership of leafs that have been pruned. Broadcast is does not scale for the global Internet.
- Another problem is that DVMRP does not have a hierarchy in place, and logically the MBONE is still one (extremely) large, flat routing domain. This creates big routing tables, since every router on the MBONE must be aware of all the other multicast routers present on the MBONE.

5.3.7 - Summary

DVMRP is a "flood and prune" protocol that creates source-based trees. It has a built-in unicast routing protocol that allows a router to locate source networks and to establish child-parent relationships between routers.

[5] RIP is limited to 16 hop diameter networks. The value of infinite was increased in DVMRP to allow bigger networks.

DVMRP is not suitable for use on a wide scale, whether a large Intranet or the Internet. Ideally, it should be used in small networks. However, DVMRP has spread to almost all MBONE sites because at the time the MBONE started (1992) there were no other multicast routing protocols to choose from. DVMRP is used in the MBONE today as an inter-domain multicast routing protocol even though it has serious scaling problems.

5.3.8 - More information on DVMRP

For more information on DVMRP read the documents referenced in this section. DVMRP is also well introduced in [Cisco10-98, Semeria-96, and Maufer-98].

5.4 - MOSPF

MOSPF relies on IGMP as the protocol hosts use to join multicast groups. A receiving host sends an IGMP message to a local MOSPF designated router that keeps a database of directly attached members in each group. Once an MOSPF designated router learns of a new group member on its attached subnets, it sends out a special **"group membership Link State Advertisement (LSA)"** message that is propagated to all other MOSPF routers within the OSPF area. When a MOSPF router receives one of these multicast LSAs, it adds the group membership information to its link-state database (this database is the basic OSPF link state database with the MOSPF multicast extensions). By propagating the locations of multicast group members, MOSPF routers create a detailed map of the multicast topology (See Figure 5. 6).

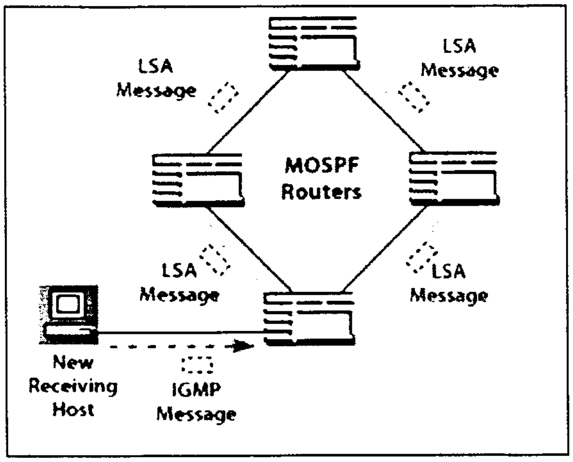


Figure 5. 6 - MOSPF LSA propagation (Image Source: [Bay-96])

5.4.1 - History

The protocol is based on the work of Steve Deering in [Deering-91]. MOSPF specification is defined in the RFC 1584 [Moy2-94]. The main motivation behind the design was to create a version of the link state routing protocol that could handle multicast. This protocol has only been implemented Proteon and 3Com routers. An analysis of the protocol is available in [Moy4-94].

5.4.2 - MOSPF Terminology

MOSPF is based in same concept of **areas**, which defines a two level hierarchy in a domain (See section 3.5).

In MOSPF, a **wildcard multicast receiver** is a router that receives all multicast datagrams regardless of destination. This new type of router is introduced because routers in a non-backbone area do not know about group membership in other areas.

5.4.3 - Message Types

MOSPF uses the same message types used in OSPF [Moy-94]. It also defines a new type of link state advertisement that propagates group membership in an OSPF area. The new message is called "**group-LSA**" and it describes the location of group members in the OSPF domain.

5.4.4 - How it works

MOSPF creates a source-based tree. This tree is created by augmenting the OSPF link state database with the "group-LSAs" and defining additional calculations to be performed in the topological database. The topological database besides having a complete map of the domain now has information on group membership for each network segment in the domain.

When a host joins a group it sends an IGMP Host Membership to its designated router. This router in turn floods a "group-LSA" to all other OSPF routers in the domain. The "group-LSA" received by MOSPF routers is stored in the topological database for later use.

The multicast routing calculation is performed "on-demand" fashion. This means that the source-based tree is only calculated when the first multicast packet appears to the router. Later multicast packets for the same (source, group) pair use the same calculation performed for the first packet.

The algorithm used to calculate the source-based tree is a modification of the Dijkstra algorithm used in OSPF. The difference is that now the shortest path is calculated from one source to many destinations and that the calculation must be done for every (source, group) pair in the internetwork. If type of Service (TOS) routing is beign used, then the tree calculation is performed for each type of service too.

5.4.5 - Evaluation

The flooding of group-LSA is the main disadvantage of MOSPF. This impedes the use of this protocol in the global Internet.

MOSPF requires OSPF as the unicast routing protocol running in an AS. If an autonomous system is running OSPF, then MOSPF is the logical choice as the protocol to be used as the multicast routing protocol. This is the wisest decision since MOSPF takes advantage of the good properties of link-state routing (See section 3.6).

If connection to the MBONE is required then interoperability issues with DVMRP must be evaluated. This should not be a problem since the rules of interoperation have been defined [Thaler-98c].

MOSPF calculates a tree for each (source, group) using the Dijkstra algorithm. This may generate too many calculations for some low-end routers with little processing capacity. Trials should be done to evaluate the impact of deploying MOSPF in those routers.

5.4.6 - Summary

MOSPF is an enhancement of OSPF v2 [Moy-94], enabling the routing of IP multicast packets. MOSPF requires a unicast routing table created by OSPF. MOSPF uses explicit joins and creates a source-based tree. This protocol is meant to be used internal to a single autonomous system.

5.4.7 - More information on MOSPF

For more information in OSPF see [Moy3-94, Moy3-94, Maufer-98, Moy-98c].

5.5 - PIM-DM

PIM-DM is a multicast routing protocol that does not require having information from a particular unicast routing protocol, and that is where it gets its name. It can use routing tables from RIP, OSPF or any other unicast routing protocol. PIM-DM is very similar to DVMRP. The main difference is that DVMRP has a built-in unicast routing protocol, while PIM-DM uses whatever unicast routing protocol is being used in the domain. PIM-DM is likely to be used in Intranets where source(s) and receiver(s) are close together such as in campus. Due to periodic re-flooding PIM-DM is more applicable when bandwidth is plentiful.

Figure 5. 7 shows the basic operation of PIM-DM and how it creates a multicast distribution tree. The tree is a source-based tree and it is constructed by using the **Reverse Path Multicast (RPM)** algorithm ("flood and prune" paradigm). Initially, when a source wants to send a feed to the network it sends the feed to its directly attached multicast router. Then, this router re-transmits the presence of this new feed to other routers attached to the network. If a multicast router receives this message, and it does not have hosts attached downstream, then it sends back a **prune** message.

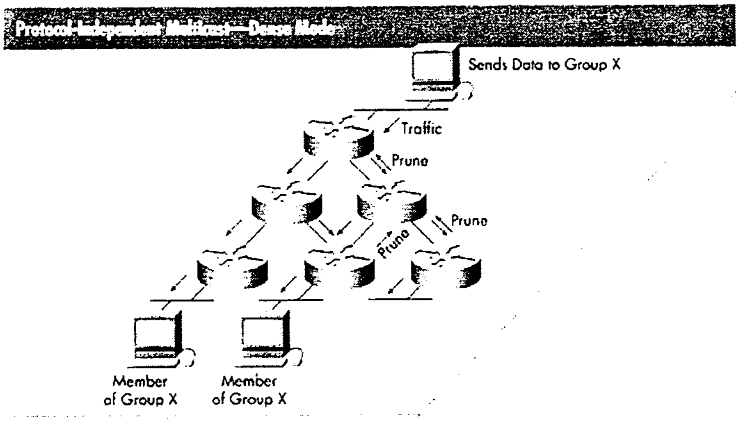


Figure 5. 7 - PIM-DM Operation (Image Source: Bay 96)

5.5.1 - History

The foundation of this protocol is largely built on Deering's early work on IP multicast routing [Deering-91]. The architecture for PIM-DM initially appeared in [Deering-94]. It was proposed in conjunction with PIM-SM, but it was later specified as a separate protocol. The most current version of the protocol specification is an Internet draft of the IETF [Deering-98b].

5.5.2 - PIM-DM Terminology

A PIM-DM neighbor router is a router that is in the same network (e.g. point-to-point link, Ethernet, etc.) and it is also running PIM-DM.

PIM-DM dense mode performs the "**RPF check**" before forwarding multicast packet. The rule states that a packet is only forwarded if the packet arrived on the interface that this router would use to send a unicast packet to the source. The interface that passes this check is called the **RPF interface** (see section 5.3.3).

5.5.3 - Message Types

PIM-DM v1 packets used to be inside of IGMP packets (using IGMP Type of x014). The Code field in the IGMP header determined the type of PIM-DM message. PIM-DM v2 packets have been assigned the protocol number 103. PIM-DM v2 runs directly on top of the IP protocol.

In PIM-DM, there are message types to detect neighboring routers (Hellos), to perform maintenance of source-based trees (Joins, Prunes, Grafts and Graft Acks) and to elect the designated forwarder for a LAN (Asserts). Table 5. 3 has a summary of all the messages used in PIM-DM.

Message	**Description**
Hello	Maintain neighbor adjacencies.
Join	Join the tree.
Prune	Leave the tree.
Graft	Re-join the tree.
Graft Ack	Acknowledgement of a Graft message.
Assert	Determine designated forwarder for a multi-access LAN.

Table 5. 3 - PIM-DM Message Types

5.5.4 - How it works

When a source has information to send to a group, it just creates a packet for multicast group and sends it to its LAN. The designated router for a LAN, then floods the data to all of its downstream interfaces and it creates (source, group) state. Routers in the domain further forward the packet downstream. They perform the "RPF Check" and forward the packet in all of its other interfaces except the "RPF interface" (See section 5.3.3). At the end of this flooding stage, all of the routers in the domain have created (S,G) state.

If a packet is received on an interface that is not the "RPF interface", then a "Prune" message is sent to the upstream neighbor that sent this multicast packet. A

"Prune" message is also sent upstream if there are no group members on a LAN. A "Prune" maybe overridden by a subsequent "Join" message.

In the case of multi-access networks with more than one router in the LAN, duplicate packets from the same (source, group) may occur. In those cases, a designated forwarder for the LAN is elected by using "Assert" messages. The router that is closer to the source wins and becomes the designated forwarder. In case of a tie, the router with the higher IP address becomes the designated forwarder for the LAN. This process is called the "removal of equal-cost paths".

If a new group member appears on a LAN that previously sent a prune, a "Graft" message is sent upstream in order to re-join the tree. The upstream router sends a "Graft Ack" message to indicate that this branch has re-joined the tree.

5.5.5 - Evaluation

The fact that PIM-DM is a broadcast and prune protocol limits the use of this protocol to be used internally in a domain. Flooding does not scale to the entire Internet. This is a bad characteristic of this protocol.

It is good that PIM-DM is independent of the unicast routing protocol that is running in a domain. This makes this protocol a better choice for domains with several unicast routing protocols running at the same time (e.g. a domain running RIP and OSPF concurrently).

The protocol independence in PIM-DM creates more bandwidth overhead than DVMRP, because some branches will get packets they would not get if there were more information on the topology of the domain [Deering-98b]. This factor is of little importance if bandwidth is plentiful inside of the domain. The flexibility gained using PIM-DM maybe a good trade-off for some domains.

5.5.6 - Summary

PIM-DM is a flood and prune multicast routing protocol that creates a source rooted tree for each (source, group) pair. The protocol is independent of a particular unicast routing protocol and it is likely to be used internally in a domain where receivers are densely populated in a domain. This protocol can be thought of as DVMRP without the "Routing Update" messages.

5.5.7 - More information on PIM-DM

Besides the references mentioned in this section, PIM-DM is also well covered in [Maufer-98, Cisco10-98].

5.6 - Vendor Support

Few router vendors support intra-domain multicast routing protocols. Table 5. 4 is a summary of the three major protocols and which vendors currently support them.

Protocol	**Vendor**
DVMRP	Proteon, Inc., Cisco, 3Com, Bay Networks
MOSPF	Proteon, Inc. , 3Com
PIM-DM	Cisco, Bay Networks

Table 5. 4 - Vendor support for Intra-domain routing protocols

5.7 - Chapter Summary

DVMRP uses implicit join and creates a source-based tree. It can be used for small networks, but it does not scale well for the general Internet. It has a built-in unicast routing protocol that is based on the distance vector algorithm.

MOSPF requires a routing table created by OSPF and it creates a source-based distribution tree. It is based on the link state algorithm.

PIM-DM is a protocol based on Reverse Path Multicasting algorithm. PIM-DM does not require having information from a particular unicast routing protocol, and that is where it gets its name. It can use routing tables from any unicast routing protocol. PIM-DM is very similar to DVMRP, and is likely to be used in Intranets where you have small number of users.

None of the protocols presented in this chapter scale to the entire Internet. The main reason for this is that they use **flooding** in order to create the multicast distribution tree. DVMRP and PIM-DM periodically broadcast multicast packets to links that do not lead to group members. MOSPF floods group membership to links that do not lead to group members through the group-LSA flooding mechanism. The issue of scalability has motivated further work in the area of inter-domain multicast routing. Next chapter is survey of proposals for inter-domain multicast routing protocols.

Chapter 6 - Alternatives for Inter-Domain Multicast Routing Protocols

This chapter presents some of the alternatives that researchers have proposed for interdomain multicast routing. The idea is to present a summary of the key features of each proposal. The last section of the chapter discusses the alternatives chosen for comparison in this thesis and the reasons for the decision. DVMRP, PIM-DM and MOSPF are not included in this list because they are considered proposals for intra-domain multicast routing protocols.

Table 6. 1 summarizes all the major proposals for inter-domain multicast routing and the first year that they were proposed.

Year	Protocol	References
1993	Core Based Trees (CBT)	[Ballardie-93] [Ballardie-95] [Ballardie-97] [Ballardie-98]
1994	Protocol Independent Multicast - Sparse Mode (PIM-SM)	[Deering-94] [Deering-96] [Estrin-98]
1995	Hierarchical Protocol Independent Multicast (HPIM)	[Handley-95]
1995	Hierarchical Distance-Vector Multicast (HDVMRP)	[Thyagarajan-95]
1996	Conference Steiner Multicast (CSM)	[Aggarwal-96] [Aggarwal-98]
1997	Yet Another Multicast (YAM)	[Carlberg-97]
1997	Multicast Internet Protocol (MIP)	[Parsa-97]
1997	Ordered Core Based Trees (OCBT)	[Shields-97]
1997	Alternate Path Routing an Pinning	[Zappala-97]
1998	Hierarchical Multicast Routing (HIP)	[Shields-98]
1998	Policy Tree Multicast Routing (PTMR)	[Hodel-98]
1998	Domainserver Hierarchy	[Komandur-98]
1998	Multicast Source Discovery Protocol (MSDP)	[Farinacci-98]
1998	Adhoc Multicast Routing Protocol (AMRoute)	[Talpade-98]
1998	Border Gateway Multicast Protocol (BGMP)	[Kumar-98]

		[Thaler-98]
1998	QoS Multicast Internet Protocol (QOSMIC)	[Faloutsos-98] [Banerjea-98]
1998	Single-Source Multicast (EXPRESS)	[Holbrook-98]
1998	Centralized Multicast	[Keshav-98]
1998	Simple Multicast	[Perlman-98]
1998	Static Multicast	[Sola-98] [Ohta-98] [Sola-98b] [Sola-98c]

Table 6. 1 - Proposals for Inter-Domain Multicast Routing

The following sections provide a summary of the key features of each of these protocols.

6.1 - CBT

Anthony Ballardie initially proposed the Core Based Trees (CBT) protocol in a SIGCOMM conference in 1993 [Ballardie-93]. Then, he further documented his work in his Ph.D. thesis [Ballardie-95]. As the result of the IDMR working group, CBT was proposed as an experimental RFC in RFC-2189 [Ballardie-97] and 2201 [Ballardie2-97]. Version 3 is also been worked out by the IETF in [Ballardie-98] and [Ballardie-98b].

CBT uses explicitly joined-shared trees. The core is the root of the tree. CBT does not depend in any specific unicast routing protocol. In CBT the same tree is shared for all the senders to a group. This characteristic considerably reduces the state that is kept in routers. CBT runs directly over IP using IP protocol 7.

6.2 - PIM-SM

PIM-SM was initially proposed in a SIGCOMM conference in 1994 [Deering-94]. Then, the proposal was further refined in [Deering-96]. The IDMR group at the

IETF continued to work in the protocol and recently finished a protocol specification for PIM-SM in June of 1998 in its RFC-2362 [Estrin-98].

PIM-SM was designed to overcome CBT limitations. PIM-SM maintains the traditional IP multicast service model of receiver-initiated membership. It uses explicit joins that propagate hop-by-hop from members' directly connected routers toward the rendezvous point of distribution tree. It builds a shared multicast distribution tree centered at a **Rendezvous Point** (RP), and then builds source-specific trees for those sources whose data traffic exceed an specified threshold. It is not dependent on a specific unicast routing protocol; and the protocol adapts to underlying network conditions and group dynamics. PIM-SM is optimized for environments where there are many multi-point data streams and each data stream goes to a relatively small number of the LANs in the internetwork. It runs directly over IP, using IP protocol 103.

6.3 - HPIM

Handley et al. proposed a modification to PIM-SM in [Handley-95]. They proposed a new protocol called Hierarchical PIM (HPIM). HPIM builds upon the work of PIM, but differs from PIM in several ways. The clearest difference is that HPIM does not require advertisement of rendezvous points (RP) to the senders and receivers of a group.

HPIM uses a hierarchy of RPs for a group. A receiver would send joins to the lowest level RP, which in turn would join a RP at the next level, and so on. The number of levels in the hierarchy depends on the scope of the multicast group. The trees built by HPIM are bi-directional.

The name of the proposal came from the fact that they proposed a statically configured hierarchy. The next step of their proposal would be to dynamically configure the hierarchy. The basic idea that they proposed is that each RP has a certain level.

One of the problems of HPIM is that it uses a hash function to pick the next RP on the hierarchy. This strategy could eventually create a very sub-optimal distribution tree.

6.4 - HDVMRP

Thyagarajan and Deering proposed a two-level hierarchical routing model for the Mbone, employing DVMRP as an inter-region routing [Thyagarajan-95]. Their proposal is called Hierarchical Distance Vector Multicast Routing Protocol (HDVMRP). It claims that it can interconnect any of the existing multicast routing protocols. In HDVMRP, routers flood data packets to the boundary routers of each autonomous system (AS) and this boundary routers respond with prunes if there are no members of that group in the AS.

Like DVMRP, HDVMRP has the problem that packets are broadcast to the entire Internet, even to autonomous systems that are not interested in the group transmission. In addition, each router running HDVMRP needs to keep a router entry for each sender sending to a group. That is the routing table will be in the order of O(SxG). Finally, HDVMRP requires encapsulating data packets for them to transit a domain, which is another undesirable characteristic since this adds more overhead to the protocol.

The approach proposed in H-DVMRP has been abandoned due to technical difficulties related to its scaling properties. However, the specification of H-DVMRP has been used as a model for the new proposals of multicast routing protocols.

6.5 - CSM

Aggarwal et al. proposed a multicast routing protocol called: Conference Steiner Multicast (CSM), that is suited for domains that wish to support mobile hosts but it requires OSPF as the unicast routing protocol [Aggarwal-96] [Aggarwal-98]. CSM is targeted towards (sparse) multicast conferencing and online discussion groups. CSM is based on the use of a shared, heuristic Steiner minimal tree for interconnecting group members. The main characteristic of CSM is that it proposes a mechanism to support mobile hosts that it is not supported by other multicast protocols.

6.6 - YAM

Carlberg and Crowcroft proposed a protocol called Yet Another Multicast (YAM) routing protocol [Carlberg-97]. Their protocol operates independently of any unicast routing protocol. Trees are built in an on-demand basis through the use of one-to-many joining mechanism. This protocol creates shared trees considering multiple routes. It supports multiple paths and it uses static information.

The YAM protocol implements the greedy heuristic. The greeding routing scheme has been shown to outperform the commonly used shortest path routing in various analytical [Takahashi-80] and experimental [Waxman-88] [Doar-93] [Waxman-93] studies. On average, the studies suggest a 10-30% advantage of the greedy approach in the efficiency (cost) of the distribution tree.

The fact that YAM floods control messages makes it a protocol that does not scales for large networks.

6.7 - MIP

Parsa et al. presented a multicast routing protocol called Multicast Internet Protocol (MIP), which offers a mechanism to construct both shared and shortest path multicast trees [Parsa-97]. MIP can be sender-initiated or received initiated which makes a feasible choice for a broader number of applications with different group dynamics and sizes. The main characteristic of their proposal is that MIP maintains a tree with no loops.

The MIP protocol introduces novelties on administrative and correctness issues, and guarantees loop-free shared trees and source-based trees.

6.8 - OCBT

Shields and Garcia-Luna-Aceves presented a multicast routing protocol that improves some of the deficiencies found in CBT [Shields-97]. They called their approach Ordered Core Based Tree (OCBT).

They showed that CBT could produce routing loops during periods of routing instability. OCBT eliminates this problem and reduces the latency of tree repair following a link or core failure. OCBT allows flexible placement of the cores, which improves the scalability of the protocol. They also proposed a hierarchy of cores (very similar to HPIM idea).

A problem with OCBT is that it is not independent of the multicast routing protocol that is running internally in the domain.

6.9 - Alternate Path Routing and Pinning

Zappala et al. [Zappala-97] suggested a multicast protocol that provides an alternate router to avoid a bottleneck link. They introduced extensions to interdomain multicast routing to scalably compute and install alternate paths and non-opportunistic, or pinned, routes. They also presented a multicast setup protocol for installing alternate paths, which they claim prevents loops. They called this setup protocol "MORF".

6.10 - HIP

Shields and Garcia-Luna-Aceves introduced a new protocol called the HIP protocol that uses OCBT as the inter-domain routing protocol in a hierarchy that can include any multicast routing protocol at the lowest level [Shields-98].

The HIP protocol introduces the idea of "virtual routers", so that an entire domain appears as a single router on a higher-level shared tree. HIP then routes between domains using the OCBT protocol.

6.11 - PTMR

Hodel proposes an extension to protocols like PIM-SM, called Policy Tree Multicast Routing (PTMR) [Hodel-98]. His proposal creates multicast distribution trees that work in asymmetric conditions, support policies, enable shortest path and QoS criteria.

6.12 - Domainserver Hierarchy

Komandur et al. [Komandur-98] recently proposed the Domainserver Hierarchy. This multicast routing protocol is a scalable multipoint-to-multipoint multicast over ATM networks using the Public Network-to-Network Interface (PNNI) hierarchical routing protocol.

6.13 - MSDP

Farinacci et al. proposed a mechanism to connect multiple PIM-SM domains together [Farinacci-98].

6.14 - AMRoute

Talpade et al. have recently proposed the Adhoc Multicast Routing Protocol (AMRoute) [Talpade-98]. This protocol allows for robust IP Multicast in mobile adhoc networks by exploiting user-multicast trees and dynamic cores.

6.15 - BGMP

The Border Gateway Multicast Protocol (BGMP) was presented in the last SIGCOMM Conference held in Vancouver in September of 1998 [Kumar-98] [Thaler-98]. They described an architecture for inter-domain multicast routing that consists of two complementary protocols: MASC and BGMP. The Multicast Address-Set Claim (MASC) protocol dynamically allocates to domains multicast address ranges from which groups initiated in the domain get their multicast address. The Border-Gateway Multicast Protocol (BGMP) constructs inter-domain bi-directional shared trees.

BGMP is a proposal for inter-domain multicast routing based on BGP-type routing, which focuses on connecting heterogeneous multicast routing domains and which allows ASs to control multicast transit traffic. BGMP uses shared trees in a slightly different way than other multicast routing protocols (e.g. PIM-SM and CBT). Each domain "owns" an address space and is the root of the distribution tree with such an address.

BGMP does not offer support for Quality of Service (QoS) and it uses reverse shortest paths routing.

6.16 - QoSMIC

Faloutsos et al. proposed Quality of Service Sensitive Multicast Internet protoCol (QoSMIC) [Faloutsos-98]. The primary characteristic of QoSMIC is that supports QoS-sensitive routing. It also supports multiple paths and dynamic routing information.

6.17 - Single-Source Multicast (EXPRESS)

Holbrook and Cheriton are proposing a protocol called Single-Source Multicast (a.k.a. EXPRESS) [Holbrook-98]. They propose to allocate a portion of the class D address space for EXPRESS. Each (S,G) pair in this range defines a unique single-source group that they call a channel. Their work is not published yet.

6.18 - Centralized Multicast

Keshav and Paul have recently proposed an approach for multicast routing called Centralized Multicast [Keshav-98]. They proposed to centralize control flow in distinct control elements.

6.19 - Simple Multicast

Radia Perlman et al. recently proposed a simple approach for inter-domain multicast routing. This proposal is very similar to CBT and PIM-SM, but it differs in several aspects:

- No need for routers to map cores with multicast addresses.
- The core is a member of the group
- Eliminates the complication of switching to a per source tree introduced by PIM-SM.
- The shared tree can be bi-directional.

6.20 - Static Multicast

Sola et al. proposed an inter-domain multicast routing protocol that scales for the whole Internet [Sola-98] [Ohta-98] [Sola-98b] [Sola-98c]. Their proposal is called

"Static Multicast" and it is based in PIM-SM or CBT, and DNS. The use of DNS gives a hierarchical mechanism to find RPs or Cores. They also proposed a scalable method for multicast address allocation based on DNS.

6.21 - Alternatives chosen

Among all the alternatives for inter-domain multicast routing presented by researchers, only two have achieved the RFC experimental level. They are CBT and PIM-SM. These two alternatives are the maturest technologies currently available. Because of their maturity, these are the two protocols that are going to be compared in depth in this thesis. Other alternatives are too new and are still in the development stage. It is not worthwhile to compare them at this point in time.

6.22 - Chapter summary

This section presented an abstract of the proposals found for inter-domain multicast routing. It also, presents the reasoning for choosing CBT and PIM-SM as the two protocols to analyze in this thesis.

Chapter 7 - Directly Related Work

There has been some previous work comparing CBT vs. PIM-SM. This chapter is a summary of the most important comparisons done to date in multicast routing. These comparisons are very technical and do not take into consideration factors the point of view of network managers. For example, none of the comparison take into account the ability to do billing, the interoperability of this new protocol with the existing infrastructure, vendor support, etc.

7.1 - Wei's comparison (1994)

Wei and Estrin [Wei-94] investigated the trade-offs between five different types of trees (see Table 7. 1). The two most important types that they compared are: Optimal Delay Core Based Trees (CBT) and Shortest Path Trees (SPT). Working protocols use these types of trees. For example, DVMRP uses Shortest Path Trees while the PIM-SM and CBT protocols use Optimal Delay Core Based Trees.

They compared the several types of trees using the following criteria:

- Low delay: It is desirable that the delay of the multicast tree to be minimal.
- Low cost: The cost of total bandwidth consumption should be low. He did not consider the cost of tree state information (i.e. routing table size).
- Traffic concentration: This is the amount of traffic coming from different sources that share a link.

Tree Type	Description
SPT	Shortest Path Tree are rooted at the source and provide minimal delay at the expense of cost (i.e. larger state is required).
KMB	Kou, Markowsky, Berman algorithm [Kou-81] that approximates Steiner Minimal Trees.
Optimal-Cost CBT	A Core Base Tree in which the center is placed with the minimum cost criteria.
MSPT	Member-Sender-rooted Shortest Path Tree. This is a tree in which the center is a member or a sender of the group.
Optimal-Delay CBT	A Core Based Tree in which the center has minimum maximum-delay or average delay among all group members and senders.

Table 7. 1 - Tree types compared in [Wei-94]

Wei enumerated the parameters that can affect the performance of a distribution tree, namely:

1. Reasonableness of a graph, which is the proportion of short and long links.
2. The node degree, which is the average number of nodes connected to each node.
3. Group Size: the number of nodes that form a group.
4. Number of active senders to a group.
5. Distribution of senders and receivers.
6. Graph Size, which is the number of nodes in the graph.

Wei concluded that source based trees and shared based have desirable characteristics depending on type of application that is running in the network. For delay sensitive applications it is best to use source based trees. If the application is not time sensitive then it is better to use shared based trees since they consume less memory in routers.

Based on the above conclusion, PIM-SM designers decided to include both types of trees in the protocol [Deering-94]. PIM-SM starts with a shared-based tree that requires less state per router and allows switching to a source based tree if the source is delay sensitive and requires high bandwidth.

They also concluded that one of the disadvantages of a center-specific tree is that over large groups, certain links may become bottlenecks and in the case of a large number concurrent senders, traffic concentration may occur.

7.1.1 - Critique of Wei's comparison

Wei's work has several problems. The following is an analysis of his work:

1. Wei did not take into consideration the cost of tree state information. This means that they did not take into consideration one of the most important aspects of any routing protocol of today's Internet. The size of the routing tables is one of the aspects that most influences the applicability of a protocol for the global Internet
2. They assigned the delay of a link to be the distance between two end nodes. This means that they do not take into account the transmission rate of a link or the propagation delay. This factor invalidates some of his conclusions regarding the delay of each tree type.
3. Some of his simulations use the topology of the old ARPANET, which is not the current structure of the Internet. Today's Internet is based on peering agreements and exchange of traffic among ISPs at private peering points and NAPs. His work should be extended to take into account the current topology of the Internet as described in chapter 2 (also see [Halabi-97]).

7.2 - "Georgia Tech" comparison (1994)

Calvert, Madhavan and Zegura developed a framework to do systematic comparisons for interdomain multicast routing schemes [Calvert-94]. They applied this framework to two practical schemes: DVMRP and CBT. They concluded that CBT had the potential to make more efficient use of resources, with modest performance penalty.

The framework features a simulation environment for measuring routing algorithm performance, realistic random graph models on which to compare the algorithms, and a visualization tool for networks and multicast routes.

The criteria used were bandwidth and delay. The parameters that they varied to run the simulations were: number of sources transmitting to a group, the group size and the average node degree (Number of nodes connected to each node) and the network size. All the simulations were done over a graph of 400 nodes and then they analyzed the effect of having a bigger network by increasing the network size to 900 nodes.

They concluded that DVMRP is generally superior to CBT in terms of maximum delay, while CBT can be superior in terms of bandwidth requirements. They concluded that even with a poor selection of the core in CBT it performs reasonably well.

7.3 - "Naval Postgraduate School" comparison (1994)

Shulka, Boyer and Klinker proposed a mechanism to locate the center of core-based trees with reservation of resources to guarantee QoS [Shulka-94]. As a complement of their work, they presented simulations in various topologies and showed that, with their center location mechanism, core-base trees yield lower tree cost than source-based trees for many concurrent senders with only a modest increase in path length.

They performed simulations that studied the effect on CBT and SPT as the number of concurrent senders increased. The topologies varied from 10 to 100 nodes distributed over 1 to 10 clusters. For example, one of the simulations was done over a topology with 3 clusters (each representing an autonomous system) each having 30 nodes evenly distributed. The average node degree was kept in the interval [3,5].

Their simulation took into account the presence of asymmetric links and implemented a simple center-location mechanism. The simulation required a priori knowledge of participants' locations for the purpose of the core placement mechanism.

7.4 - Ballardie's comparison (1995)

Chapter 4 of Ballardie's thesis has a comparison of the scalability of different proposals for multicast routing [Ballardie-95]. Ballardie compared 4 multicast routing protocols, namely: DVMRP, MOSPF, PIM-SM and CBT. He used very clear criteria that had 3 items that he considered the most important to make a comparison: group state information, bandwidth consumption and processing costs.

Ballardie concluded that Core Based Trees (CBT) offer the most favorable scaling characteristics for the typical case where there are some senders within or outside a group of receivers. Also, he concluded that CBT may not be suitable for all multicast applications, but it will be satisfactory for many.

7.4.1 - Critique to Ballardie's comparison

Since he was the proponent of CBT his analysis is biased towards his proposal and it makes CBT to appear as the best solution.

When he compared PIM-SM he included PIM-DM as part of the architecture. In this author's opinion, this makes his comparison hard to understand and it loses focus. PIM-DM is an intra-domain routing protocol while PIM-SM is an inter-domain routing protocol.

7.5 - Petitt's comparison (1996)

Chapter 3 of Pettit's thesis contains a qualitative comparison of multicast routing protocols [Petitt-96]. The criteria he considered to make the comparison was: state, control traffic overhead, data distribution overhead, scalability, join latency, complexity of implementation, delay link reutilization and convergence time.

He compared seven protocols: DVMRP, HDVMRP, MOSPF, CBT, PIM-DM, PIM-SM and HPIM. He wrote an overview and an evaluation for each protocol based on the criteria mentioned before.

7.6 - Harris Corporation's comparison (1997)

Harris Corporation performed a simulation of the two protocols that have been proposed for inter-domain multicast routing: PIM-SM and CBT. They examined the performance of these protocols in the Distributed Interactive Simulation (DIS) environment. Examples of DIS environments are: tank battle simulation and exchanging experimental data and weather maps. Their work has appeared in several magazines [Billhartz-96] [Billhartz-97] and it was also presented in two of the IDMR working group meetings in 1995 [Carlberg-95] [Billhartz-95]. For the purposes of this summary their latest work is considered [Billhartz-97].

They varied four parameters in their simulations: groups per host, group distribution type, join/leave dynamics and traffic generation rate. They based their comparison in 7 metrics: End-to-end delay, network resource usage (# of hops), overhead traffic percentage, join time, traffic concentration, routing table size and implementation difficulty.

They used three networks as the basis of their simulations: AAI Network (a real network of 11 sites and links with bandwidths in the range from 10 Mbps to 600 Mbps.), Mesh Network (a mesh of 19 sites with bandwidths in the range from 10

Mbps to 600 Mbps) and a Stressed Mesh Network (a mesh of 19 sites with links with a bandwidth of 3 Mbps).

They concluded that CBT is the best multicast routing protocol for environments with a large number of groups, each with many senders. Table 7. 2 summarizes their findings:

Criteria	**CBT**	**PIM-SM**
End-to-End Delay	Low	Low
Network Resource Usage	Moderate	Moderate
Overhead Traffic Percentage	Proportional to number of joins per second	Proportional to number of joins per second
Join Time	Low	Low
Traffic Concentration	High	Very High (Shared Based Tree) Low (Source Based Tree)
Routing Table Size	Linear with number of groups	Proportional to the product of number of groups and mean number of senders per group.
Implementation Difficulty	Low to moderate	Complex

Table 7. 2 - Harris' comparison of multicast routing protocols. Source: [Billhartz-97]

Their work was presented in the IDMR meeting of the IETF in 1995. The minutes summarizing their work is presented [Fenner-95]:

" - SPTs incur 5-20% less delay than CBTs.

- PIM SM (shared tree) consistently shows much longer delays (> 50 %) than CBT.

- PIM SM (with S,G state to RP) is identical in delay to CBT.

- Control packet overhead:

=> PIM SM with SPTs incurs about double the overhead of CBTs

=> PIM SM shared tree mode incurs about the same overhead as CBT.

- PIM and CBT join times were shown to be low and about equal.

- PIM SM with SPTs can involve up to a 50% increase in join latency.

- Regarding the question: Is PIM a superset of CBT?

The conclusion to this question is that PIM is a superset of CBT, but there appear to be distinct advantages to using CBT.

- Conclusions:

=> the end-to-end delay was on average 10% lower with PIM SPTs than the CBT delay.

=> in shared-tree mode, there is no advantage to using PIM over CBT.

=> in terms of b/w utilization (overhead), both protocols were shown to be about equal."

7.7 - Helmy's simulation (1997)

Helmy, a graduate student at the University of Southern California, proposed a method for analyzing the robustness of multicast routing protocols in a systematic fashion [Helmy-97]. He called his method Systematic Testing of Robustness by Examination of Selected Scenarios (STRESS). He only tested his method with PIM-SM. He was able to identify several protocol design errors in PIM-SM, and suggest solutions to these errors.

7.8 - Chapter Summary

This chapter presented a summary of other comparisons on multicast routing that researchers have done in the past. The approach and conclusions of their work were reviewed.

In this author's opinion, the best work accomplished so far in the area of comparison of Inter-Domain Routing Protocols was accomplished by Harris

Corporation [Billhartz-97]. They compared the two protocols analyzed in this thesis and concluded that the best protocol to be used for the Distributed Interactive Simulation (DIS) was CBT.

Most of the work summarized in this chapter was done using simulations. The real test of the protocols is when there are real implementations running in real routers. It is only then that the protocol could be said to have the desired characteristics. Such testing would need to wait until the protocols are supported by commercial vendors and deployed by network managers.

Next chapter presents the criteria used to compare CBT and PIM-SM. This criteria is based on the work summarized in this chapter.

Chapter 8 - Criteria For the Comparison

This chapter describes the criteria used to compare PIM-SM vs. CBT. The criteria are divided in five main parts: Protocol status, basic characteristics, technical criteria, operational criteria and overall assessment.

The protocol status is a summary of how mature the technology is. It includes pointers to the protocol specification, year of the initial spec, testing of features of the protocol, etc. The basic characteristics section of the protocol summarize key features of the protocol. The technical criteria captures the way of thinking used by protocol designer. The operational criteria represent the view of a network manager trying to deploy these protocols. The final piece of the criteria is an overall assessment on the advantages and disadvantages of each of the protocols analyzed.

Some of the benchmarks defined in this chapter are based on the benchmarks defined in [Dubray-98, Tascnets-98, Billhartz-97, Wei-94]. The following sections describe in detail the criteria to be used for the comparison.

8.1 - Protocol Status

8.1.1 - Specification

This is a document that formally describes a protocol. It also describes the status of this document. This information tells whether the protocol is proprietary or standard. It includes the year of the specification.

8.1.2 - Status

This is the status of the RFC (e.g. experimental, standard, etc) or the general consensus on the validity of the proposal, if it is a paper.

8.1.3 - Availability

This describes the current release of the protocol and who can get access to it.

8.1.4 - Supported Platforms

A list of support hardware/software architectures.

8.1.5 - Management Information Base (MIB)

This is a pointer to a specification for a MIB if such a document exits. MIBs are defined so that a protocol could be managed remotely.

8.1.6 - Implementations

This is the number of implementations known at the moment. This includes a list of individuals and organizations that are actively involved in defining and implementing the specified protocol.

8.1.7 - Features tested

This section describes the features of the protocol that has been tested.

8.1.8 - Operational experience

This is the level of operational expertise with the protocol. This is an estimate of the number of network engineers who have deployed the protocol.

8.1.9 - Router Vendor Support

This section summarizes major commercial companies supporting the protocol

8.2 -Basic Characteristics

This section highlights the key features of a protocol and provides an easy way to compare the different characteristics of the protocols.

8.2.1 - Tree Type

Some routing protocols create source based trees; some others create shared based trees, while there are some other types that create both types of trees.

8.2.2 - Uni/Bi-directional

Multicast routing protocols can create state in one direction or in both directions. In unidirectional trees traffic can only flow from sources to receivers while in bi-directional trees, traffic can flow in both directions.

8.2.3 - Loop-free-ness

Multicast routing protocols, which depend on a unicast routing protocol, can suffer from the same transient loops as the unicast protocols do. Ideally the protocol should guarantee to be loop free for the multicast distribution tree.

8.2.4 - RPF-Check

Some protocols perform a Reverse Path Forwarding (RPF) check on the received multicast packets. This mechanism checks whether the packet is received on the interface that a unicast packet would use to reach the source.

8.2.5 - Hard State vs. Soft State

Some protocols create state in routers that is refreshed with some acknowledgement mechanism (Hard State), while in other protocols, state disappear if it is not received it before a time period (Soft State).

8.2.6 - Protocol Independence

It is desirable for the protocol to be independent from unicast routing protocols and intra-domain routing protocols. This gives the protocol more flexibility and support for heterogeneous domains.

8.2.7 - RFC-1112 Compliant

The protocol needs to comply with the specifications of the IP Multicast Service model specified in RFC-1112 [Deering-89]. Specially, with the requirement that senders need not be members of a group to send data[1] and senders can send data to a group without a previous setup.

8.3 - Parameters

It is desirable to know what happens with the protocol, as the conditions in the network vary. For example, it would be interesting to know how the routing tables grow as the number of groups present in the Internet grow. Imagine these parameters as the x-axis in a two-dimensional graphic. These parameters are used as a reference for the technical criteria described in the next section.

8.3.1 - Number of sources

How does this protocol behave as the number of sources sending to a group increases? For example, what happens with the size of the routing table and the control traffic overhead?

[1] Not requiring senders to be members of a group accommodates a big range of applications. For example, many small sensors reporting data to a set of servers without the overhead of receiving each other's traffic.

8.3.2 - Number of receivers

How does this protocol behave as the number of receivers increases? Ideally the protocol should support big groups in the order of hundreds of thousands of receivers.

8.3.3 - Numbers of groups

How does this protocol behave as the number of groups present in the Internet increases?

8.3.4 - Amount of data

This is the amount of multicast data packets that a sender can send to a group without breaking the protocol.

8.3.5 - Burstiness

The effect of bursty sources in the routing protocol.

8.3.6 - Duration

This is an assessment on the impact of the duration of a group session on the routing protocol behavior.

8.3.7 - Topological Distribution

This is an assessment on how the geographical distribution of the recipients and senders affects the behavior of the protocol.

8.3.8 - Node Degree

Node Degree is the average number of connections of each node in the network. For example, a network with a node degree of 3 means that each node in the network has in average three links to other nodes. It is interesting to know how the protocol behaves as the network becomes more interconnected.

8.4 - Technical Criteria

The technical criteria to compare routing protocols are based on what are the requirements of the applications. Batsell summarized very well application requirements for 3 types of applications that use multicast: Distributed Interactive Simulation (DIS), Distance Learning and group VideoConference. Table 8. 1 is a summary of the application requirements presented in [Batsell-95].

Requirement	Application		
	DIS	**Distance Learning**	**Video-Conference**
Senders	Many	Small number	Small Number
Receivers	Are also senders	One or few	Are also senders
Number of groups per application	Many	One or few	One or few, but many groups can be present over network
Data transmission	Real-time	Real-time	Real-time
End-to-end delay	Small	Not Critical	Moderate
Set-up	Real-time	Non-real-time	Non-real-time
Join/leave dynamics	Participants rapidly join/leave	Receivers rapidly join/leave	Participants rapidly join/leave
Scalability Requirements	Large networks, many groups, many senders and receivers per group	Large networks, many receivers per group	Large networks
Multicast tree	Can rapidly move over the physical topology.	Rooted at source and includes current receivers	Includes participants, can slowly move over the physical topology.

Table 8. 1 - Application Requirements for multicast [Batsell-95]

Based upon the requirements of the applications, this comparison of the PIM-SM and CBT multicast routing protocols considers the parameters described in the following sub-sections. Some of the benchmarks are of interest to end-users running multicast applications such as: End-to-end delay and setup time, while other metrics are of most interest to network managers such as: traffic concentration, router memory and link bandwidth overhead.

8.4.1 - Link bandwidth overhead

This is the amount of control overhead packets introduced by the protocol. This also takes into account packet replication created by the protocol.

8.4.2 - CPU utilization

The amount of CPU cycles used by the protocol. The algorithm must try to reduce the CPU utilization as much as possible

8.4.3 - Router memory

The amount of routing state (router entry for groups and senders) created by the protocol should be minimized, because the amount of memory available in a router required to store routing tables is limited. Memory is getting cheaper everyday, so this criterion it is losing its previous importance.

8.4.4 - End-to-end delay

This is the average amount of time to deliver a packet from a source to a number of destinations.

8.4.5 - Join time

This is the time that elapses between a host sending an IGMP Report and the reception of the multicast feed.

8.4.6 - Leave time

This is the amount of time that elapses between the transmission of an IGMP Leave message to the cease of a multicast feed.

8.4.7 - Convergence Time

The amount of time that is required for all routers in a domain to recalculate their routing tables after topology changes should be small.

8.4.8 - Traffic characteristics

It is desirable that the routing protocol distributes the traffic evenly across the network. In general, it is a bad idea to concentrate the traffic in few links since it limits the scalability of the protocol.

8.4.9 - Address Allocation

The protocol should present a good solution to the problem of assigning class D addresses dynamically. That is, the probability of address collision, as well as the delay in obtaining an address to assign to a group, should be small.

8.4.10 - Address Aggregation

It is desirable that the protocol provides some form of address aggregation so that the routing tables could be reduced. The idea is to something similar as CIDR but for multicast. This problem is hard since receivers could be located anywhere in the Internet.

8.5 - Operational criteria

This section captures the factors a network manager would use to compare the protocols.

8.5.1 - Ease of configuration

This section is a subjective analysis on how easy is to configure the routing protocol.

8.5.2 - Ease of management

This section is a subjective mark on how easy to observe the protocol behavior and its impact to the network.

8.5.3 - Robustness

The protocol should perform well in the event of hardware failures, high load conditions and incorrect implementations. For example, the protocol should recover even from the failure of a core router in the case of shared trees. In other words the protocol should be reliable.

8.5.4 - Price

It includes some of the routers that implement the protocol and how much they cost.

8.5.5 - Interoperability

This section points out the issues of the deployment of this protocol in relation to other protocols that are already running on the network.

8.5.6 - Installation time

This is an estimation of how much time could take to deploy the protocol for an intermediate network engineer.

8.5.7 - Billing capability

This section discusses how this protocol solves the billing of multicast packets.

8.5.8 - Impact on existing network infrastructure

This is an overall assessment on how the deployment of this protocol may affect a production network.

8.5.9 - Multipath routing

It is desirable that several routes be provided from each source to each receiver. This enables policy routing.

8.5.10 - Quality of Service (QoS) support

The protocol should support arbitrary QoS requirements from users. To achieve this, the protocol has to consider multiple paths, and handle link asymmetry, e.g. for satellite connections. Multiple paths may be necessary for policy reasons. Also, the protocol should accommodate diverse application types with minimal user input.

8.5.11 - Mobility support

With the increase in mobile devices, it would be desirable for a routing protocol to support mobile environments.

8.5.12 - Heterogeneity

Ideally an inter-domain multicast routing protocol should support heterogeneous domains. That is, a domain might use any multicast routing protocol internally. Usually protocols support heterogeneity by defining a hierarchy that allows domains to run different multicast routing protocols. This is an important characteristic of an IDMR protocol since it is difficult, if not impossible, that all domains run the same protocol. A perfect example of a protocol that supports heterogeneity is the IP protocol, which can run of top of any data-link technology invented or to be invented.

8.5.13 - Policy support

The protocol should allow different autonomous systems to impose different sets of policies. A good protocol should provide mechanism to limit the traffic that a domain is willing to carry.

8.5.14 - Security

This is a description of the authentication mechanisms of the protocol. What are the mechanisms for authenticating routing messages and what other forms of protection are defined in the protocol.

8.5.15 - Complexity

This is a qualitative assessment of how complicated it is to develop and manage this protocol. In general, it is better to have simpler protocols since fewer problems arise and they are easier to debug and troubleshoot.

8.5.16 - Third-party dependency

The protocol should minimize the third-party dependency. For example, shared trees require traffic to go through a core. This may be a bad thing if the core is located in a domain that has slow links.

8.6 - Overall Assessment

This section is a summary of the good and bad properties of the protocol.

8.6.1 - Scalability

This section describes how the bandwidth, CPU and memory are used as the routing environment grows. This specifies how large the set of receivers can be. In particular, this metric takes into consideration three main factors: number of sources, number of receivers and number of groups. For example, it is desirable that multicast routing protocols support up to 100,000 receivers.

8.6.2 - Suitability

This section defines for which environments the protocol is well suited.

8.6.3 - Advantages

This section summarizes the advantages of this protocol over other alternatives.

8.6.4 - Disadvantages

This section summarizes the disadvantages of this protocol over other alternatives.

8.7 - Chapter Summary

This chapter presented a summary on how the protocols will be studied. Of all these criteria, there are factors that are more important than others. For example, for some applications it may be more important to reduce protocol overhead than maintaining optimal distribution trees. Low bandwidth consumption is one the primary concerns for inter-domain multicast routing protocol. Therefore, multicast traffic aggregation and low control overhead traffic are critical factors.

Other important criterion is the support of heterogeneous environments and policies. These policies could be asymmetrical and even source specific. Meyer described some of the concerns of why multicast routing protocols have not been widely deployed [Meyer-97].

The next two chapters present the analysis of alternatives considered. The analysis follows the criteria presented in this chapter.

Chapter 9 - CBT Protocol Analysis

This chapter presents an analysis of Core Base Trees (CBT), based on the criteria presented in chapter 8. The analysis starts with a summary of the of the CBT protocol.

9.1 - Summary of CBT

9.1.1 - Motivation

In 1992, the IETF broadcast the first conference over the MBONE [Casner-92]. Since that meeting all the IETF meetings are transmitted real time over the Internet. Anyone interested in the meeting can join and listen, but support for IP Multicast is needed. That is, a mechanism is needed to route packets from a source to many destinations. The multicast routing protocol used in 1992 (and it is still used) was DVMRP [Waitzman-88]. However, this protocol is known to not scale to the whole Internet because of its flooding mechanism. CBT was the first proposal of a multicast routing protocol with the intention to scale to the entire Internet.

9.1.2 - CBT Terminology

Core Based Trees create a shared tree that has a node called the "Core", which is the root of the tree. Routers in the direction of the core are called "upstream routers", while routers in the direction of hosts are called "downstream routers".

9.1.3 - History

The CBT protocol was formally introduced in 1993 in a SIGCOMM Conference [Ballardie-93]. However, talk about this protocol started in 1992 at the Inter-Domain Multicast Routing Protocol (IDMR) working group of the IETF. Early drafts of CBT can be found following the email threads at [IDMR-email]. Ballardie's Ph.D. thesis is another very good reference for CBT [Ballardie-95].

9.1.4 - Overview

CBT uses explicitly joined-shared trees. The core is the root of the tree. CBT does not depend in any specific unicast routing protocol. In CBT the same tree is shared for all the senders to a group. This characteristic considerably reduces the state that is kept in routers. CBT runs directly over IP using IP protocol 7.

9.1.5 - How CBT works

Each LAN has a designated router. The DR keeps track of group memberships and alert upstream routers about the desire of joining a group by hosts in its directly connected links. If there is more than one router in a LAN, then one router must be elected as the designated router. The election is performed using HELLO messages.

When a host joins a group (through an IGMP Report message), the DR on the LAN forwards this request to the group's core. The message sent is a JOIN_REQUEST. The DR knows the core's IP unicast address either by manual configuration or dynamically through the "bootstrap" mechanism, which dynamically discovers which core is serving which groups.

As the JOIN_REQUEST is forwarded towards the core, each intermediate router creates "transient join state". This "transient joint state" is the multicast group and the JOIN_REQUEST's incoming and outgoing interfaces.

Once the core receives this request, it sends a JOIN_ACK back to the DR confirming the join. The JOIN_ACK message is forwarded according to the "transient join state" previously created in the router. After that, hosts start receiving traffic for that group immediately. The JOIN_ACK can also be sent by the first router that is member of the shared tree.

As the JOIN_ACK returns to the DR, state is created in intermediate routers, specifying which outgoing interfaces should be used to forward a packet destined to a particular multicast group. This new group is either a known group or an unknown for the CBT router. If the group is known by the router then an entry exists on the router for this group, and a new interface is added to the list of outgoing interfaces. If this is an unknown group for the router then a new entry is added to the CBT routing table. This entry simply consist of the group and the interface on which the JOIN_REQUEST was received (this is the interface that leads toward a group member). Figure 9. 1 illustrates the process of the creation of a shared tree in CBT.

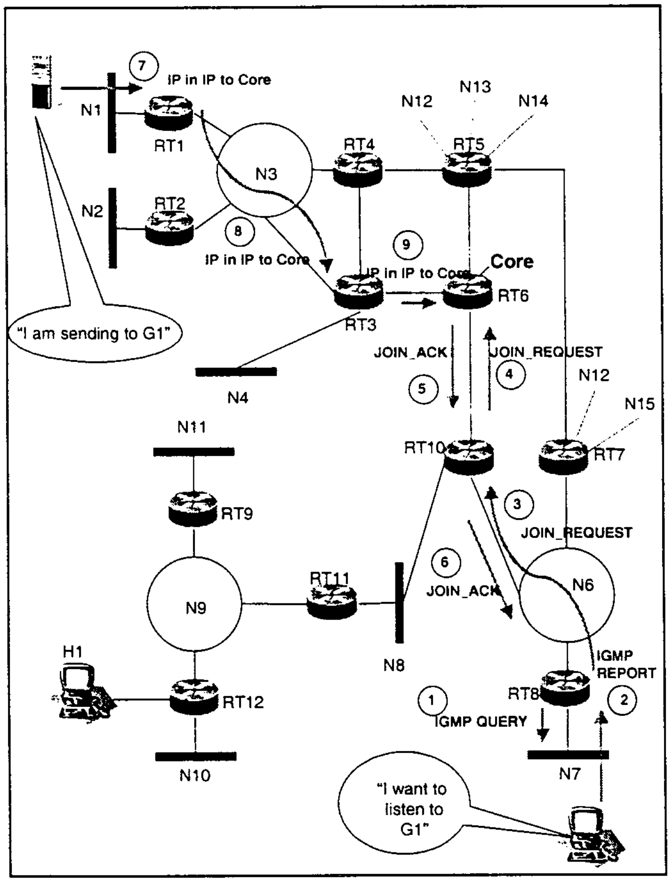


Figure 9. 1 - CBT Functional Overview.

When a host leaves a group it sends an IGMP Leave message to its DR. If there are no more members of a group in a LAN, then DR sends a QUIT_NOTIFICATION upstream to the core.

Just to make sure that everything is OK, each child router sends a "keepalive" message to its parent router. In CBT these "keepalive" messages, are called ECHO_REQUEST and ECHO_REPLY. A child router periodically unicasts an ECHO_REQUEST message to its parent router, which is then required to respond with a unicast ECHO_REPLY message.

If a router or a link goes down, then all the routers that are downstream of the router that failed are removed from the tree. This is done by sending a FLUSH_TREE message. After receiving a FLUSH_TREE message, each router is responsible for attempting to reattach itself to the group's core, in order to rebuild the shared delivery tree.

CBT Routing Table

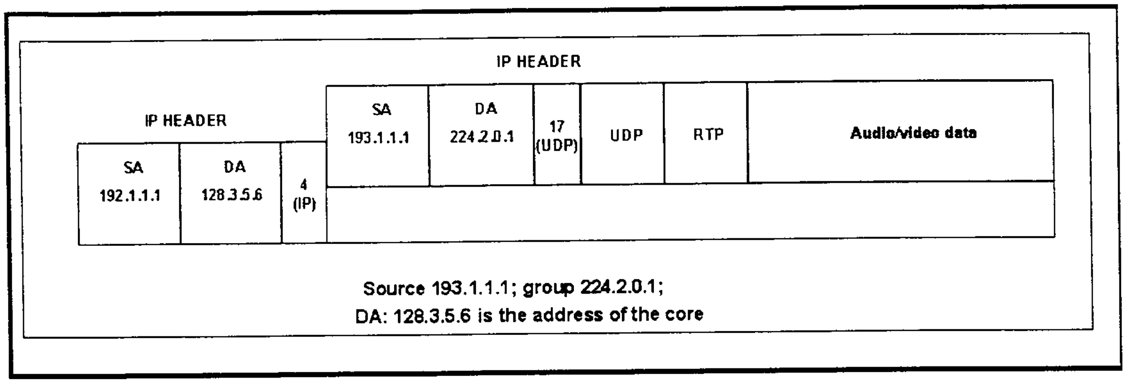
CBT constructs a routing table that has the following format:

```
(group, {outgoing interface list})
```

When a packet arrives with a Class D address in the destination address of IP header (group id), then the multicast routing table is searched using the group id obtained from the destination address of the IP packet. The router entry that matches the group-id identifies which outgoing interfaces should be used to forward the packet in the direction of group members.

Non-member sending

When a host that is not member of a group wants to send information to a group, it unicasts a packet to the core, using the core's IP unicast address. Note that members of a group send to the group by sending an IP packet using the group address in the destination field of the IP header. This technique is called "IP-in-IP" Tunnel. It consists of an IP multicast packet encapsulated by an IP unicast header. See Figure 9. 2.



Source 193.1.1.1; group 224.2.0.1;
DA: 128.3.5.6 is the address of the core

Figure 9. 2 - IP-in-IP encapsulation is used to send to the core of the tree.

9.1.6 - Message Types

There are seven messages in CBT. The HELLO message is used to elect the DR in a LAN; the remaining messages are used to maintain the shared tree. Table 9. 1is a summary of all CBT messages.

Packet Name	Description
HELLO	Elect a designated router
JOIN_REQUEST	Desire to join a group
JOIN_ACK	Confirmation of the join sent by the core
QUIT_NOTIFICATION	No more members of a group in this LAN
ECHO_REQUEST	Periodic keepalive message to maintain the tree
ECHO_REPLY	Response to an ECHO_REQUEST
FLUSH_TREE	Destroy tree state present in the router

Table 9. 1 - CBT Message Types

9.1.7 - CBT Summary of key features

The main key feature of CBT is that it creates a shared tree for all the sources sending to a group, reducing the amount of state that is needed to be kept in routers to O(G). CBT's state is bi-directional. This feature is unique among all existing multicast routing protocols. Data may flow in either direction along a branch.

Each group has exactly one core associated with it, but each core could serve as a core for several groups. CBT is an explicit join protocol, this means that a host will not receive traffic for this group unless they specifically asked for it.

9.1.8 - More information on CBT

Besides all the references provided in this chapter, CBT is also well explained in the following references [Moy-98c] [Maufer-98].

9.2 - Protocol Status

9.2.1 - Specification

As of this writing, the most current documents describing CBT are two RFC documents written in September of 1997, which describe version 2 of CBT [Ballardie-97] [Ballardie2-97]. These are the references that are going to be used in this thesis. There is also Internet-Drafts describing CBT version 3 [Ballardie-98] [Ballardie-98b].

9.2.2 - Status

As of November of 1998, CBT is an experimental standard, according to the IETF classification of protocols.

9.2.3 - Availability

CBT Version 2 source code is freely available at this ftp site:

ftp://ftp.labs.bt.com/Internet-Research/cbt-2.0.tar.gz

9.2.4 - Supported Platforms

At this time there is only one known supported platform: FreeBSD 2.2.[67]

9.2.5 - Management Information Base (MIB)

There is an Internet Draft specifying a MIB for CBT and it is available at [Ballardie-97c]

9.2.6 - Implementations

There is only one known implementation (see section 9.2.3). There are no commercial routers that implement CBT as of November of 1998.

9.2.7 - Features tested

No major testing has been performed with CBT.

9.2.8 - Operational experience

None.

9.2.9 - Router Vendor Support

CBT is not implemented in any commercial router.

9.3 - Basic Characteristics

9.3.1 - Tree Type

CBT creates a shared-based tree. There is no option to switch to a source-based tree as it is provided in PIM-SM. Shared create less state in routers, but the paths created are sub-optimal and that increases the end-to-end delay that packets experience from a source to all receivers in the network.

9.3.2 - Uni/Bi-directional Shared Trees

CBT creates bi-directional trees. This is a good property since it allows any member of a group to send data to the group, without the need of creating state per each source.

9.3.3 - Loop-free-ness

CBT v1 did not guarantee a loop free tree. Shields demonstrated that if a core fails loops could be formed [Shields-97]. This prompted the release of version 2 of CBT [Ballardie-97]. So, in theory, the current version does not create loops, however operational experience is needed to prove the absence of loops.

CBT depends on information coming from a unicast routing protocol, which may have problems itself (e.g. the slow convergence time of RIP). This fact may affect the operation of CBT.

9.3.4 - RPF-Check

CBT does not perform the RPF check[1]. This is a good characteristic of the protocol since it decreases the overhead imposed in the CPU of the router. PIM-SM does perform "RPF check", which requires to perform a "RPF lookup" which may be a costly task for some routers.

9.3.5 - Hard State vs. Soft State

CBT uses "hard states". Messages are acknowledged and repeated after a time-out. This increases the control traffic overhead in the protocol. The "soft state" mechanism used in PIM-SM is a better approach.

[1] "RPF Check" is explained in section 5.3.3

9.3.6 - Protocol Independence

CBT is not dependent on a particular unicast routing protocol. This is an improvement over other protocols that are tied to a specific unicast routing protocol (e.g. MOSPF requires OSPF). However, it requires CBT to be running in all the autonomous systems. It does not allow having different multicast routing protocols running in each domain. This is a bad characteristic since the Internet is a collection of AS, and every domain should be able to choose the protocol that meets their needs and be able to interoperate with other domains.

9.3.7 - RFC-1112 Compliant

CBT uses IGMP to discover group membership in LANs and it is compliant with the multicast model proposed in RFC-1112. The fact that it complies with the IP Multicast Model is good since it permits this protocol to inter-operate with other protocols. However, RFC-1112 has some problems itself that are then inherited by CBT (See RFC-1112 critique in section 4.5).

9.4 - Technical Criteria

9.4.1 - Link bandwidth overhead

CBT needs to advertise the set of routers that are candidates to be core. These advertisements are done to the entire Internet, which clearly does not scale. This is one of the big disadvantages of CBT.

No bandwidth is wasted due to pruning, since CBT assumes that receivers do not want to listen to a feed unless they explicitly join the group. This is an advancement over "Flood and Prune" protocols such as DVMRP, which waste bandwidth periodically to discover group members on pruned branches. The CBT architecture does not send multicast packets to parts of the network not interested in them.

CBT needs to be built with control traffic. This traffic is short-lived and sporadic. Flood and Prune protocols (e.g. DVMRP) build the tree as multicast data packets start to flow, which requires more state in the routers but it does not require "tree building" traffic.

In order to maintain the tree, CBT routers exchange keepalive messages that are 56 bytes long. This traffic may consume considerable bandwidth of a link if the time between keepalives is not chosen properly by a network manager.

The overhead introduced by the header in CBT is less than 1% of the bandwidth for most typical links [Ballardie-95].

9.4.2 - CPU utilization

The CPU of a router is used for two things: control traffic to build and maintain the trees and data forwarding of multicast packets.

The control traffic is mainly keepalives, since the construction of the tree is sporadic and it only happens at the beginning when receivers and senders are joining the group. The burden of sending these keepalives is dependent on the frequency of the keepalives and the number of group trees traversing a router. Ballardie proposed that in order to bound the processing burden, as the number of groups increases, the frequency of the "keepalives" should be decreased [Ballardie-95].

The CPU utilization associated with data packet forwarding is dependent on the number of group trees traversing a router. With CBT, there is no need to make a RPF lookup as it is with other multicast routing protocols such as DVMRP and PIM dense mode.

9.4.3 - Router memory

Routers in the unicast path between the non-member sender and the shared delivery tree do not need to maintain information about the multicast tree. That is, "off-Tree" routers do not to maintain group state.

"On-Tree" routers must maintain an outgoing interface list per group. Unlike source-based tree approaches, forwarding is not dependent on the packet's IP source address, and therefore CBT routers do not need to maintain state per each sender to the group. Since there is a need to keep an entry in the multicast routing table for each group, then CBT scales O(G).[2]

This means that less memory is needed in CBT than in PIM-SM. PIM-SM uses a combination of shared and source-based trees which in the worst case scenario create routing tables that are in the order of the product of sources times groups (O(SxG)). The property to create less state in routers is one of the main advantages of CBT.

9.4.4 - End-to-end Delay

The end-to-end delay for multicast packets from sources to receivers is not optimal because the tree that CBT builds is a shared tree. Packets must go first to the core of the tree before reaching receivers. Sometimes the path could be very bad, e.g. if the core router is another continent.

CBT does not discuss the placement of a core, it proposes to either use statically configured cores or to have the Designated Router for a host act as the core. A core that is not in an optimal place could create a tree with high end-to-end delay. For example, imagine a core in Australia with senders in the USA. Packets

[2] O(G) means in the order of multicast groups present in the Internet.

must travel to Australia in order to reach recipients that are in USA. This may not be an acceptable delay for applications that are time sensitive such as video-conferencing. The fact that the paths are not optimal and that CBT creates trees with high end-to-end delay is one of the main disadvantages of CBT.

9.4.5 - Join time

When a host wants to join a group it sends an IGMP Report to its local router. The join is forwarded to the core with a JOIN_REQUEST message. The core acknowledges the new branch of the tree sending a JOIN_ACK back to the LAN with a group member.

Depending on where the core is located, the join delay could be considerably long. Besides the propagation delay, the join time is increased by intermediate router queues.

The join time parameter is of little importance compared to other parameters. The join time only affects the performance seen by an end-user - i.e. users might experience a little bit of delay before the join a group.

9.4.6 - Leave time

There is no previous work that measures how much time a host has to wait to stop receiving multicast traffic.

9.4.7 - Convergence time

There is no previous work that focus on measuring how much time elapses between when a change occurs in the network and the time that all the routers know about the change. Further research is needed in this area. It might be possible that the protocol behaves inadequately in the presence of link or router failures.

9.4.8 - Traffic characteristics

CBT concentrates traffic around the core. This limits the scalability of the protocol, because there will be some links around the core highly congested. This is a disadvantage of CBT.

9.4.9 - Address allocation

There is no definition of an address allocation scheme as part of CBT. It is assumed that the address for a multicast group is obtained by some other means, e.g. the Session Directory Protocol (SDP). Routing and addressing are very related. CBT should propose a solution to the problem of address allocation as part of its architecture.

9.4.10 - Address aggregation

CBT does not provide any solution on address aggregation. It is not easy to provide a common aggregate for receivers that are located in many different networks. Aggregation is considered an area of active research at this time for CBT designers. Without aggregation routing tables become too large. This is a problem that also appears in PIM-SM that requires further research.

9.5 - Operational criteria

9.5.1 - Ease of configuration

Since there are no commercial routers that support CBT. It is not possible to tell how easy or difficult it is to configure CBT. The shareware version requires more work to configure and it is not as well documented as a commercial product. Also, if problems arise almost only the author can help you, whereas with a commercial product there is typically customer support. This is a disadvantage of CBT that might have a considerable effect on the decision made by a network manager.

9.5.2 - Ease of management

It is hard at this point in time to describe how easy it is to manage CBT. The protocol is not commercially available in routers, as a consequence very little could be said on how easy it is to manage the protocol.

9.5.3 - Robustness

CBT integrates failure recovery as a part of the protocol. This complicates the protocol, but provides answers to the problem of what happens when the core fails. The core as a single point of failure is a clear disadvantage of CBT, even though there is a mechanism to recover from core failure.

9.5.4 - Price

There are no commercial routers that implement CBT as of November 1998. At this moment the CBT daemon is free and it is available on-line (See section 9.2.2). The fact that the protocol could be installed for free is a good reason that may incline a manager to deploy CBT in his network. This may help a network manager to get familiar with multicast routing. However, if a network manager is completely new to multicast, it is a better idea to deploy DVMRP (see Section 5.3), since there is considerable operational experience in the Internet community with this protocol.

9.5.5 - Interoperability

Since the protocol has not been widely deployed, it is not possible to say at this time how well it works with other multicast protocols. The protocol should at least be able to inter-operate with DVMRP. Thaler wrote a document that specifies interoperability rules for multicast routing protocols [Thaler-98c].

9.5.6 - Installation time

CBT could take some time to install since the instructions come in the "README" file and there is no customer support for the protocol in case of questions on configuration. It seems that there is more documentation and help available for PIM-SM than for CBT.

9.5.7 - Billing capability

CBT does not offer a solution on how to do billing. This is one of the main weaknesses of the CBT protocol. ISPs are not willing to deploy a protocol that will not fit their business model.

9.5.8 - Impact on existing network infrastructure

The impact of deploying CBT might be big. For example, routers may need to be upgraded with new memory in order to support CBT. Also, the increase in network traffic caused by the deployment of multicast could cause the need to upgrade the links to support higher data rates. Note that even though multicast is supposed to save bandwidth, it also enables new applications that wouldn't be possible without multicast. As a result, there might be a net increase in the bandwidth used in the network. In summary, the deployment of any multicast routing protocol is something that requires careful thought.

9.5.9 - Multipath routing

CBT does not support multipath routing. The protocol is complicated enough in order to support a tree with a path for each host part of the group. The support of more paths for the same group would increase the size of the routing tables and the control overhead traffic.

Multipath routing might be necessary for some bandwidth intensive applications, so that traffic could be routed over several equal cost paths and in turn increasing the effective data rate of the multicast transmission.

9.5.10 - Quality of Service (QoS) support

There is no mention of QoS support in the CBT standard. This is bad since more and more applications are requiring a certain level of service and QoS support would become a must for a routing protocol in the near future.

9.5.11 - Mobility support

There is no support for mobile hosts in CBT. With the increase in sales in mobile computers, routing protocols would also need to support routes to mobile hosts. This is a disadvantage of CBT.

9.5.12 - Heterogeneity

There is no support for heterogeneity in CBT, since all CBT routers run in a flat domain. Every single domain must run CBT in order for the protocol to work. This is a disadvantage of this protocol, because ideally the protocol should allow the network manager of a domain to deploy whatever multicast routing protocol he considers convenient.

9.5.13 - Policy support

There is no support for policy in CBT. This is a big disadvantage of the protocol. Network managers need to be able to control the multicast traffic in the network. Just this criterion alone might be enough to discard CBT as a possible option. For example, ISPs need to have the ability to define policies in order to control multicast traffic.

9.5.14 - Security

A sender can send to a group without authorization. This may create problems in the future when multicast traffic is charged to end-users. This is a problem that it is inherited from the IP Multicast Model proposed by Deering in RFC-1112 (see section 4.5)

Encryption is needed to protect the privacy of the content in the transmission to a group. There may be conferences that are private that are just for specific individuals. CBT currently does not support encryption in the protocol.

9.5.15 - Complexity

This protocol has seven different message types. It only creates shared trees, which makes it simpler. The fact that the tree is based on a core makes it easy to troubleshoot but less reliable.

9.5.16 - Third-party dependency

Since the protocol is based on the concept of core routers, it creates a third-party dependency. This is something most ISPs don't like since they would rather have control over everything. For example, if the core fails they rely on the core router that belongs to another ISPs, which is not under their control but it affects their clients. This is a big disadvantage of any protocol based on shared trees. PIM-SM also suffers from this problem.

9.6 - Overall Assessment

9.6.1 - Scalability

There is no proof that the protocol scales to a large number of receivers, groups, sources or large network topologies. Since there is practically no operational experience with the protocol, it is hard to predict its scalability.

Simulation studies typically run simulations for relatively small network topologies because of the time it takes to run a simulation. For example, the maximum number of hosts in a topology generated by Wei were 500 nodes [Wei-94]. Harris study's used even smaller topologies (19 nodes) [BillHartz-97]. Other parameters such as the number of receivers, groups and sources are also kept in small numbers for the same reason of the time that it takes to run a simulation.

Ideally, a multicast routing protocol should be able to support groups with millions of receivers. This might be a more important requirement once multicast is deployed natively in the entire Internet.

9.6.2 - Suitability of the protocol

CBT is suited for intra-domain multicast routing, even though their designers hope to be used in the global Internet. The main reason for its inability to scale to the entire Internet is the flooding of the core set.

Harris's study concluded that CBT is better suited for applications with many senders, such as Distributed Interactive Simulation (DIS) [Billhartz-97]. The reason for this conclusion is that CBT requires less state in routers, since it uses shared trees.

9.6.3 - Advantages

The main advantages of CBT are:

- **Less state information:** All recipients of a group are reached over a single tree (versus multiple sender-rooted trees as with PIM-SM running source trees). This implies that there is less state information stored in routers. The state is bound by the number of groups present in the internetwork. In mathematical terms, state scales to O(G).
- **Better bandwidth utilization:** CBT does not send traffic to networks that do not have members. DVMRP uses implicit joins in which flooding to the entire Internet is required. A router with no members in its attached interfaces has to send prunes back to the source if it does not want the transmission. This process is repeated periodically and it is a waste of bandwidth. CBT is a big improvement over "flood and prune" protocols, such as PIM-DM and DVMRP. Even though this protocol is better than others, it still has the problem of the flooding of the core-set (see disadvantages below).
- **Scalability:** CBT supports bigger network topologies than DVMRP. However, the flooding of the RP-Core Set is a characteristic that impedes the deployment of CBT as a global solution for the Internet.
- **Unicast independence:** CBT does not depend on a specific unicast routing protocol. This is a good property of CBT. However it is still tied to the use of CBT as the multicast routing protocol, which makes CBT a less desirable protocol.
- **It is free:** CBT is a freeware program that can be downloaded from the Internet (see availability section (9.2.3) in this chapter). This is a good reason to start with this protocol and not with PIM-SM that is now a commercial product.

- **Simple:** The protocol is a lot simpler than PIM-SM because it does not define a mechanism to switch to a source-based tree.

9.6.4 - Disadvantages

The main disadvantages of CBT are:

- **Flooding of the core set:** The main disadvantage of CBT is that this protocol requires flooding the candidates to be a core to the whole network. This strategy does not scale in the Internet.
- **No support for policies:** A manager cannot specify policies using this protocol. This is a big disadvantage because there is no way for a network manager to control multicast traffic.
- **Immaturity:** CBT is not commercially available from any router vendor. This is a disadvantage, since ISPs need to deploy protocols that have been tested and that have few problems. It is also desirable to count on a company that offers customer support, so that they can fix problems quickly. Since there is no commercial support, it is harder to configure this protocol.
- **Sub-optimal paths:** There is a possibility that the path between some sources and some receivers may be sub-optimal. This is because the shared tree requires packets to go through the core, which could create the need for a packet to travel extra hops. These sub-optimal paths increase the end-to-end delay that a packet will experience and might be a disadvantage for some applications.
- **Traffic Concentration:** CBT concentrates traffic around the core. This is because the traffic of all the sources of a given group will transverse the same set of links. Multiple cores are used for different groups in order to alleviate this problem.

- **No heterogeneity:** Ideally an inter-domain multicast routing protocol should be able to support different protocols in each AS. CBT requires that every single domain runs the same routing protocol.
- **No security:** Receivers can join any group that is announced on the Internet. The sender does not know who is listening to its transmission. Security is provided by the application layer through encryption.
- **No aggregation:** There is a need to create a mechanism to aggregate class D address, so that state in routers and control traffic is decreased. This is a problem also in PIM-SM. At this point, this is an area of research.
- **Core Failure:** The protocol is too dependent on the core router. If the core fails the whole distribution tree goes down. CBT has a mechanism to recover from core failures, however this complicates the protocol. Moreover, little testing on real scenarios has been performed.
- **No support for billing:** There is no answer in CBT to the problem of billing between ISPs.
- **Third party dependency:** It is bad that the protocol relies on a core router that could belong to another autonomous system. This gives less flexibility to ISPs.
- **No support for "advance features":** There is no support for multipath routing, Quality of Service or mobility in CBT.
- **Core location:** If the core is in a place that is not optimal the multicast distribution tree can create delays that are not acceptable for some applications.

9.7 - Chapter Summary

This chapter presented a summary and analysis of CBT. It first summarized the key features of CBT and how it works. Then, an analysis was presented based on five sets of criteria: protocol status, basic characteristics, technical criteria, operational criteria and overall assessment.

The main characteristic of CBT is that it creates shared trees that are rooted at a core router. Even though CBT is an improvement over previous multicast routing protocols, it cannot be used as a scalable solution for the global Internet because of the flooding of the core set problem. Besides it does not support policies, which is one of the most important characteristics of any inter-domain routing protocol.

Chapter 10 - PIM-SM Protocol Analysis

This chapter presents a summary and an analysis for the Protocol Independent Multicast-Sparse Mode protocol. The analysis is based on the criteria presented in chapter 8.

10.1 - Summary PIM-SM

10.1.1 - Motivation

DVMRP [Waitzman-88] builds a source-based tree that minimize delay but creates more state in routers, while CBT [Ballardie-98] builds shared trees that reduce state in routers but increase the delay since paths are not optimal. The main motivation behind PIM-SM designers was to create a compromise between these two proposals. In essence what they did was to combine the two known ways of creating distribution trees (source and shared trees).

10.1.2 - PIM-SM Terminology

PIM-SM defines a **rendezvous point (RP)**, which is the core of the shared tree. It is just another name for the core of the tree.

There is also the concept of upstream routers in the direction of the core and downstream routers in the direction of group members.

PIM-SM introduces the concept of candidate RPs. These routers are other possible RPs in case of a failure of the RP. It also defines the **bootstrap router (BSR)** which is a router that keeps track of available RPs (called the RP set) and originates bootstrap messages.

10.1.3 - History

The protocol was first formally introduced in the 1994 SIGCOMM Conference [Deering-94]. The proposal was refined in 1996 in [Deering-96]. The IETF's IDMR working group standardized the protocol in RFC-2117 [Estrin-97]. The latest specification available is the RFC-2362 [Estrin-98].

10.1.4 - Overview

PIM-SM maintains the traditional IP multicast service model of receiver-initiated membership. It uses explicit joins that propagate hop-by-hop from members' directly connected routers toward the rendezvous point of distribution tree. It builds a shared multicast distribution tree centered at a Rendezvous Point (RP), and then builds source-specific trees for those sources whose data traffic exceed an specified threshold. It is not dependent on a specific unicast routing protocol; and the protocol adapts to underlying network conditions and group dynamics. PIM-SM is optimized for environments where there are many multi-point data streams and each data stream goes to a relatively small number of the LANs in the internetwork. It runs directly over IP, using IP protocol 103.

Senders register with the RP, which further forwards the traffic down to all registered members. There is only one RP chosen for a particular group. The RP could be statically configured or dynamically learned.

Forwarding is based on the Reverse Path Forwarding (RPF) algorithm [Dalal-78]. The rule establishes that a multicast packet will only be forwarded if received on the interface used to send unicast packets to the source or RP. If the "RPF check" fails, the packet is silently discarded. If the "RPF check" succeeds, the packet is forwarded through the interfaces indicated in the outgoing list field for this group. If traffic is flowing using a shared tree, the "RPF check" uses the RP address. If traffic is flowing using a source tree, the "RPF check" uses the source address.

PIM-SM assumes that no hosts want the multicast traffic unless they specifically ask for it. This behavior is called **explicit join** (the feed will not appear to a host unless it asks to join a group). PIM-SM evolved from CBT and it uses the same idea of a core where sources meet receivers (the difference is that the core is called rendezvous point).

10.1.5 - How PIM-SM works

This section presents a summary of the operation of PIM-SM.

Joining

When a host in a LAN wants to join a group, it tells the designated router via an **"IGMP Report"** message. The DR creates a (*,G) entry that has the interface towards Receiver in the "Outgoing Interface List". Then, the DR forwards this request hop-by-hop towards the group's rendezvous point via a **"Join"** message. It knows the address of the RP with the **bootstrap** mechanism that maps a group id to RP's IP unicast address. For example, the DR creates a table that contains entries such as: for the group 224.2.3.4 use the RP 192.1.2.3.

When the RP receives the (*,G) Join from the DR, it creates a (*,G) entry that has the interface towards the DR in the "Outgoing Interface List". At this point, the host has joined the shared tree and it is ready to receive traffic from sources sending data to this group.

Registering

When a sender wants to send data, it just sends data to group G. The DR encapsulates the source's multicast packet in **"Register"** messages and unicasts them to the Rendezvous Point (RP). From the RP, the data stream flows to receivers. Figure 10. 1 shows joining and registering process in PIM-SM.

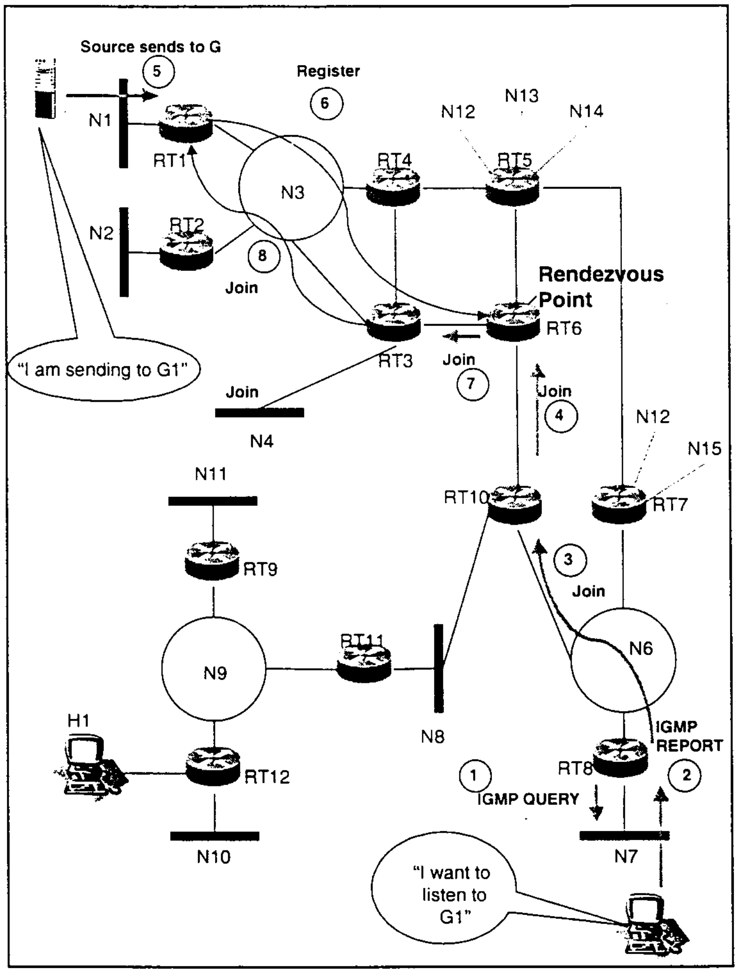


Figure 10. 1 - PIM-SM operation

Forwarding

When the RP de-encapsulates the **"Register"** messages coming from the DR, it realizes that these are packets for group "G". It can further forward the packet because it has (*,G) state. If it does not have state created for this group it silently drops the packet.

The incoming multicast packet for group "G" is forwarded out all interfaces that are in "outgoing interface list" (oilist) for this group. There is an "oilist" for each known group by the router. The multicast routing table is mainly composed of the tuple:

(group, oilist)

Where oilist is a list of interfaces in the form {Interface-1, Interface-2,..., Interface-N}. See Figure 10. 2.

```
rtr-c>sh ip mroute 224.1.1.1

     ***** SHARED TREE FOR 224.1.1.1 *********
(*, 224.1.1.1)  RP 171.68.28.140,
Incoming interface: Null, RPF nbr 0.0.0.0,
  Outgoing interface list:
     Serial0
     Serial1

     ***** SOURCE TREE FOR 224.1.1.1 *********
(171.68.37.121, 224.1.1.1)
Incoming interface: Serial3, RPF nbr 171.68.28.139,
  Outgoing interface list:
     Serial0
     Serial1

(123.122.12.12, 224.1.1.1)
Incoming interface: Serial3, RPF nbr 171.68.28.139,
  Outgoing interface list:
     Serial0
     Serial1
```

Figure 10. 2 - Typical PIM-SM Multicast Routing Table

If there is a source based tree created then this state has priority over the state of a shared tree. The state of source based tree is denoted as "(S,G) state". This means that there is a routing entry per each source 'S' sending to a group 'G'. The state of a core based tree is denoted as "(*,G) state". This means that there is only one entry for group 'G' and all senders to this group share it. For example, in Figure 10. 2 there is source based entry and shared based entry in the multicast routing table for group 244.1.1.1. A packet coming from 171.68.31.121 will use the source based entry and all other packets will use the shared based entry. Note that the source based entry does not need to keep track of the RP, since the root of the tree is the source. But, also note that there is an entry for each source sending to the group 224.1.1.1 for a shared based tree. In this case there is an entry for source 123.122.12.12. The RPF interface is used to check whether or not this multicast packet should be forwarded.

Sources Joining

Just after a RP receives a "**Register**" message, it sends a "Join" message towards the source, so that it can begin receiving "native" (i.e. un-encapsulated) packets from the source.

Once the data from the source starts arriving without encapsulation, the RP sends a "**Register-Stop**" message to notify the DR for this source that it no longer needs to encapsulate traffic in "**Register**" messages.

SPT-Switchover

PIM-SM has the option to switch to a shortest path tree if a network manager configures the router to do so. For example, a network manager could configure a bandwidth threshold. In this case, if the data rate of a source exceeds a pre-defined threshold, then a receiver will start the switch to a source based tree in order to remove any unnecessary hops and reduce the delay.

The DR starts the switching process to a Shortest Path Tree (SPT) by sending a "**(S,G) Join**" message towards the DR for this source. The DR for this source also sends a "**(S,G) RP-bit Prune**" message up the shared tree. This actions prunes traffic for this source from the shared tree.

When the DR receives the "**(S,G) RP-bit Prune**" message, it removes the interface from the "oilist" for this group. This action stops traffic for this group to be forwarded using the shared tree.

If the "oilist" is null, then the RP sends a "**(*,G) Prune**" towards the source. This stops the flow of this source traffic towards the RP.

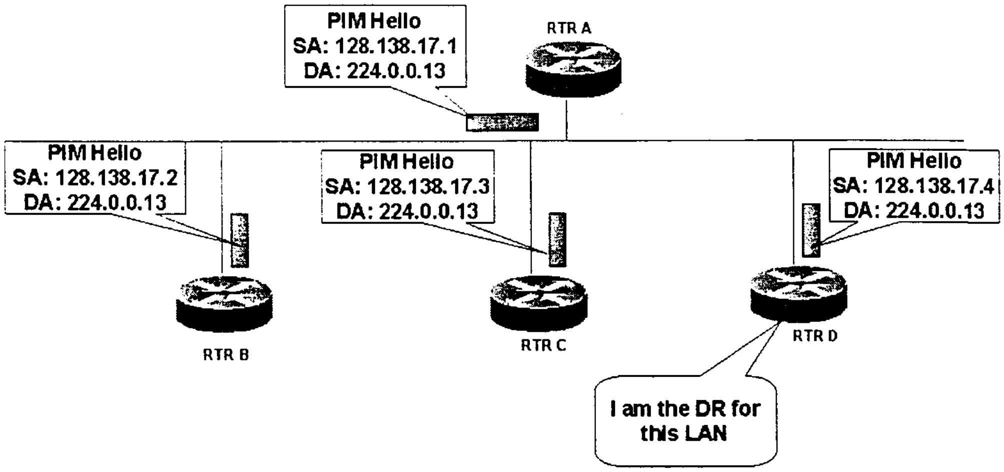
Pruning

When there are no more hosts in a LAN that belong to a group then the DR sends a "**Prune**" message to the RP. The state that forms the tree is deleted in all intermediate routers.

Neighbor Discovery

"Hello" Messages are sent periodically to discover the existence of other PIM routers on the network. Each LAN segment has a Designated Router (DR) that is also elected with the "Hello" messages. The router with the highest IP address is elected as the DR for the LAN. Hello messages are periodically multicast using address 224.0.0.13 (All-PIM-Routers group). The messages are sent every 30 seconds by default.

Each router keeps a table of the PIM neighbors present on each interface. If a PIM Hello message is not received from a neighbor in 105 seconds[1] then the PIM neighbor is no longer considered active and it is deleted from the PIM neighbor table. Figure 10. 3 shows the election process of the Designated Router (DR) on a LAN.



The router with the highest IP address interface
is elected as the Designated Router (DR) e.g. Router D is the DR.

Figure 10. 3 - PIM-SM DR election

State Maintenance

Once the tree is built the state kept in routers is refreshed periodically by sending "join" messages towards the RP.

[1] 105 seconds is the default for the [Hello-Holdtime]. This parameter is configurable.

10.1.6 - Message Types

There are eight message types defined in PIM-SM. The Hello message is used to detect neighboring PIM routers. Join, Prune and Assert messages are used to maintain the shared tree. Register and Register-Stop allow sources to send to a group. Finally, the bootstrap and Candidate-RP-Advertisement message are used to select the Rendezvous Point or RPs. Table 10. 1 summarizes PIM-SM messages.

Packet Name	Description
Hello	Detect neighboring PIM routers
Join	Join a multicast group by building a shared tree
Prune	No more interest in a group so leave tree
Assert	Election of a DR in a LAN
Register	Register a source for a group and switch to SPT
Register-Stop	The Source tree has been built
Bootstrap	Advertises set of possible RPs (RP-Set)
Candidate-RP Advertisement	This router is a possible RP

Table 10. 1 - PIM-SM Message Types

10.1.7 - PIM-SM Summary of key features

The most distinctive feature of PIM-SM is the switchover from shared trees to source-based trees. It initially builds a shared tree, which could be switched to a source based tree, if the data rate of the source exceeds a configured bandwidth threshold.

10.1.8 - More information on PIM-SM

Besides the protocol specifications and papers mentioned in this chapter, PIM-SM is well explained in these references [Maufer-98, Moy-98, Cisco10-98].

10.2 - Protocol Status

10.2.1 - Specification

The latest formal document describing PIM-SM was submitted to the IETF in June of 1998 as RFC-2362 [Estrin-98].

10.2.2 - Status

RFC-2362 is in experimental status.

10.2.3 - Availability

There are two known implementations of PIM-SM. These implementations are:

- **GateD version:** This version of PIM-SM was developed for GateD (5.0 alpha 0). This version of GateD requires a license for commercial use, academic and research licenses are free. The source code is available at:

  http://www.isi.edu/~eddy/pim/pim.html

- **USC's version:** The University of Southern California offers an implementation of PIM-SM at this web site:

  http://catarina.usc.edu/pim/pimd/

10.2.4 - Supported Platforms

The code is for FreeBSD-2.2.1, FreeBSD-2.2.5 and (untested) NetBSD-1.3 and SunOS-4.1.3.

10.2.5 - Management Information Base (MIB)

There is an internet-draft describing a MIB for PIM-SM [McCloghrie-98]

10.2.6 - Features tested

Not Available

10.2.7 - Operational experience

Some.

10.2.8 - Router Vendor Support

Cisco supports PIM-SM in some of its new routers.

10.3 - Basic Characteristics

10.3.1 - Tree Type

PIM-SM creates both types of trees. It starts creating a shared tree for the group. If a source exceeds a bandwidth threshold then the protocol switches to a source based tree to optimize the end-to-end delay. The use of this protocol is for applications that are delay sensitive, but increases the complexity of the protocol.

10.3.2 - Uni/Bi-directional Shared Trees

PIM-SM creates uni-directional trees. This is bad since a new source needs to join the RP which creates some control overhead traffic.

10.3.3 - Loop-free-ness

PIM-SM does not guarantee a loop free tree. It is also dependent how good the underlying unicast protocol is. For example, if RIP is used then PIM-SM inherits the problems of distance vector protocols (See section 3.4).

10.3.4 - RPF-Check

PIM-SM performs the RPF check. This creates more processing overhead for the CPU of the router. This is a disadvantage but it is not as important as other factors since CPU speeds are increasing very rapidly.

10.3.5 - Hard State vs. Soft State

PIM-SM uses "soft state". This means that join messages are repeated a regular intervals, the states are cached and simply "disappear" if the information is not refreshed. Soft state mechanisms require fewer control overhead packets. This makes PIM-SM more efficient since it uses less bandwidth to create the distribution tree.

10.3.6 - Protocol Independence

PIM-SM designers had in mind to separate the dependence of a multicast routing protocol from a unicast routing protocol. Any protocol could be running in a domain and PIM-SM would be able run correctly. This is a good property of the protocol, however it still requires that every domain in the Internet run PIM-SM. In other words, it does not support multiple multicast routing protocols. It is desirable that an inter-domain multicast routing protocol could support heterogeneous domains.

10.3.7 - RFC-1112 Compliant

PIM-SM supports RFC-1112. It uses IGMP to discover group membership in LANs. PIM-SM maintains the traditional IP multicast service model of receiver-initiated membership. This is good since it enables interoperability but it inherits the problems introduced by the IP Multicast model (See section 4.5).

10.4 - Technical Criteria

10.4.1 - Link bandwidth

The PIM-SM distribution tree contains only those parts of the network interested in receiving multicast traffic. In PIM-SM receivers explicitly join a shared multicast distribution tree. This is a good property that makes this protocol a better choice over "flood and prune" protocols such as DVMRP.

Senders unicast data packets to the RP, even though there is no one listening to a group. When this happens this wastes valuable bandwidth.

10.4.2 - CPU utilization

The CPU is used to generate, receive and interpret PIM-SM messages.

Receivers' outbound join traffic scales $O(G_{local})$ (number of inter-domain groups with local members). Inbound join traffic scales $O(S_{local})$ (number of local sources sending to interdomain groups) [Jacobson-94]. This gives an idea of the amount of processing a CPU would have to do to generate and process control packets. If number of groups increases without bound then this might negatively affect CPU performance. More quantitative research is needed in this area to make more appropriate conclusions. Ideally, measurements should be taken on real routers running in test networks. Simulations always have the problem that they are not real implementations and that is hard to simulate big groups.

10.4.3 - Router memory

In PIM-SM state is created explicitly by interested receivers. The state is maintained with a soft-state mechanism, i.e. it times out in routers unless refreshed periodically. The state scales between O(G) and O(S x G). This is because PIM-SM use both source and shared trees. Shared trees scale to O(G) while source-based trees scale O(S x G).

The big size of the routing tables created by PIM-SM is clearly a disadvantage of PIM-SM. CBT creates less state in routers since it uses shared trees only. Shared trees only create one entry per group in the multicast routing table. So, there will be as many routing entries as groups known by this router.

10.4.4 - End-to-end delay

PIM-SM initially builds a shared tree and it has the possibility to switch to a source based tree for sources that exceed a bandwidth threshold. This optimizes the end-to-end delay in this protocol since the path from the source to the receiver is the shortest possible. This is a good property of PIM-SM that enables it to be used for time-sensitive applications, however it adds complexity to the protocol.

10.4.5 - Join time

After the designated router of a LAN receives an "IGMP Host membership report", it sends a "Join" message towards the RP. Routers along the way to the RP create (*,G) state and when the "Join" message gets to the RP then the host has joined the shared tree. This is very similar to the "join" mechanism used in CBT. The only difference is that CBT acknowledges the Join message back to the designated router. This is a little advantage of PIM-SM since it creates less control overhead packets than CBT. But, again this parameter only affects end-user performance and it is of little importance in comparison with other parameters.

10.4.6 - Leave time

There is no previous work that points out this issue.

10.4.7 - Convergence time

There is no previous work pointing how much time it takes a network to discover that a change has occurred and propagate it across the entire domain.

10.4.8 - Traffic concentration

PIM-SM concentrates traffic around the RP. However, with the provision of source based trees, it provides mechanisms to distribute traffic across the Internet. In this respect PIM-SM is better than CBT since it provides a mechanism to provide distributed traffic in the Internet. However, it does this at the expense of more state in routers and added complexity to the protocol.

10.4.9 - Address allocation

There are no definitions of mechanisms for address allocation in PIM-SM. It is assumed that other protocols will handle the issue of address allocation. Newer IDMR proposals propose address allocations mechanisms as a part of the architecture [Kumar-98]. There is a need to extend CIDR [Fuller-93] to support multicast.

10.4.10 - Address aggregation

PIM-SM designers do not have an answer for this issue. They do not know how to aggregate addresses in multicast. This creates the problem of bigger routing tables and more control traffic overhead.

10.5 - Operational criteria

10.5.1 - Ease of configuration

In a Cisco router, it seems that is very easy to configure a router to run PIM-SM. There are 3 steps:

a) Add the "ip multicast-routing" global command to the configuration.

b) Add the "ip pim sparse-mode" interface command to each interface in the router configuration.

c) Configure the address of the Rendezvous Point (RP) using the "ip pim rp-address <addr>" global command.

The RP could be also automatically configured with the Auto-RP procedure. The RP router needs to be configured to broadcast its address to other PIM routers.

It seems that PIM-SM is easier to configure than CBT, since there is more information available and there is support from a commercial company. This is an advantage of PIM-SM over CBT.

10.5.2 - Ease of management

It is too early in the deployment of the protocol to really know how easy it will be to manage the protocol. Some freeware tools are available and are very well summarized in [Thaler-98b].

10.5.3 - Robustness

Even though PIM-SM defines its own set of mechanisms to deal with the failure of a RP router, this still doesn't seem to be well tested. The failure of RP could create serious problems in the protocol. However, since PIM-SM also creates source-trees this creates more redundancy and less dependence on a single router. More testing is needing in this area to discover how reliable the protocol is.

10.5.4 - Price

Every router in the network needs to be upgraded to support PIM-SM. In some cases, the upgrade will require just an upgrade in the operating system of the router, in other cases will require an upgrade of the memory of the router and finally in some other cases a new complete router would need to be bought.

10.5.5 - Interoperability

Interoperability for PIM-SM and other routing protocols is defined in [Thaler-98c]. This is an issue that requires further research. This factor would increase in importance if there are several multicast routing protocols running in the Internet. It would be better to have just one protocol but it seems that many protocols would be needed to support different types of applications. This means that interoperability is going to be a required goal for any multicast routing protocol of the future.

10.5.6 - Installation time

To multicast enable a complete campus could take months. There are few changes that need to be made in the router configuration, but the number of changes is in the order of routers on campus, which could be a considerable number.

10.5.7 - Billing capability

There is no proposal on the PIM-SM protocol on the issue of billing, which is a disadvantage of PIM-SM. ISPs need ways to account for multicast traffic. This is an area of open research at this time. For more on this issue see [Bellman-97].

10.5.8 - Impact on existing network infrastructure

The impact of deploying multicast in a network could be big. New traffic will be generated since new bandwidth intensive applications will be enabled. As the use of multicast applications increase, the number of groups present on the Internet at any time will increase considerably. This may create huge routing tables that routers will need to cope with, which may require memory upgrades in routers. PIM-SM requires more memory in routers than CBT and this is a factor that must be taken into account by networks managers deploying this protocol.

10.5.9 - Multipath routing

PIM-SM does not support many paths to a single destination. This is an improvement that may be needed for bandwidth intensive applications in order to load balance the traffic. It seems that support for this feature is not a near term goal in the IDMR working group of the IETF.

10.5.10 - Quality of Service (QoS) Support

Quality of Service is not supported by PIM-SM. This is a requirement that could become more important in the near future.

10.5.11 - Mobility support

There is no support for mobile hosts in the PIM-SM specification either. This should be included in future versions of the protocol, however for now, simpler things need to be solved first.

10.5.12 - Heterogeneity

If this protocol is to be used in the Internet, it is required that a network manager deploys PIM internally. PIM-SM does not have the concept of domains that are independently managed. The lack of support of heterogeneous domains is one of the main disadvantages of PIM-SM.

10.5.13 - Policy support

A manager cannot define policies using PIM-SM. This is a major disadvantage of the protocol, since it does not permit a manager to control multicast traffic.

10.5.14 - Security

There is no support of security mechanisms in PIM-SM. Group communications are not encrypted and can be intercepted; even worse a sender could send unwanted information to a group. It is assumed that application level protocols would encrypt the transmission. Privacy and authentication are issues that are currently studied by the IETF.

10.5.15 - Complexity

PIM-SM has the ability to switch to a source-tree, which makes the protocol more complicated. There are intermediate states while switching from one tree to another that are harder to understand and can create problems when troubleshooting the protocol. This is a bad characteristic of the protocol, ideally protocols should be as simple as possible.

10.5.16 - Third-party dependency

PIM-SM creates a third party dependency problem. The RP for a group could be located in a domain that is not under the control of an ISPs. If the RP fails then the ISP's customers will be affected and there is no way that the ISP manager could solve the problem since it depends on another ISP.

10.6 - Overall Assessment

10.6.1 - Scalability

Scalability was not one of the design goals of the PIM-SM protocol. However, since it is based in shared trees it scales better than previous multicast routing protocols, because it is not based on the "flood and prune" algorithm.

There is no previous work proving that PIM-SM could scale to hundreds of thousands of receivers. This is an important requirement that could only be proved if the protocol is widely deployed. Simulation studies are usually done over relatively small networks and set of receivers.

10.6.2 - Suitability of the protocol

Even though, the protocol was originally designed to be used for the global Internet, it is better suited to be used internally in a domain. The two main drawbacks of the protocol are the flooding of the RP set and the its inability to support heterogeneous domains.

10.6.3 - Advantages

- **Commercial Availability:** PIM-SM is supported by Cisco. This gives an advantage to the protocol over CBT, since commercial support means more testing, less bugs and customer support from a company.
- **Less end-to-end delay:** Source trees provide a minimal-delay path from a source to a group, which is better for multimedia applications.
- **Unicast independence:** PIM-SM does not depend on any unicast routing protocol. It can use whatever protocol is deployed on a domain.
- **Better robustness:** PIM-SM increases the robustness of the protocol by defining a set of available RPs.

- **Possibility of support QoS:** The fact that PIM-SM allows switching to a source tree allows specifying QoS criteria per source. This makes PIM-SM a better protocol if QoS is ever added to the protocol.
- **Better bandwidth utilization:** PIM-SM uses explicit joins which saves bandwidth. It does not use flood and prune strategies that waste bandwidth.
- **Soft State:** This implies that PIM-SM uses less bandwidth since control traffic does not need to be acknowledged.

10.6.4 - Disadvantages

- **Flooding of RP set:** The fact that PIM-SM needs to flood the RP set is a disadvantage in PIM-SM. This impedes this protocol to scale to the global Internet.
- **No support for policies:** PIM-SM does not allow a network manager to control multicast traffic. This is a big disadvantage of PIM-SM.
- **No support for heterogeneity:** PIM-SM requires that the protocol must be run in every single domain. Ideally, the protocol should support different multicast routing protocols in each domain.
- **No aggregation:** There is no proposal on how to aggregate routing entries and control traffic. This is an open issue at this point.
- **RP failure:** The heavy dependence of PIM-SM on the RP, makes the protocol a less robust protocol. There are mechanisms to recover from RP failure. However, the protocol is still not as reliable as desired.
- **Unidirectional state:** The forwarding state created by PIM-SM is unidirectional. Traffic can only flow away from the RP, not toward it. This is a disadvantage because if a receiver wants to become a sender then it

needs to create its own source tree. This is not an easy mechanism and PIM-SM may have trouble with applications such as conferencing and chat groups, in which end nodes could be senders or receivers.

- **Complex:** PIM-SM introduces complexity in the protocol with the "switchover" feature. This makes the protocol harder to implement for software developers and harder to debug for network managers.
- **More state:** It creates more state than CBT. It is in the order of O(S x G), if the protocol is using source based trees.
- **No support for billing:** ISPs like to have mechanisms to account for multicast traffic and this is not provided by PIM-SM
- **Third party dependency:** The fact that the RP could be managed by another ISP is something that it is not accepted by network managers.
- **No support for advanced features:** PIM-SM does not support QoS, multipath routing or mobile hosts. These are requirements that will probably gain more importance in the near future.
- **RP location:** In the mode of PIM-SM that uses shared trees there is the problem of how to choose the RP to obtain better multicast distribution trees. If the RP is located in an undesirable position then the end-to-end delay might be unacceptable.

10.7 - Chapter Summary

This chapter presented an analysis of PIM-SM. Five main criteria were considered: protocol status, basic characteristics, technical criteria, operational criteria and overall assessment.

PIM-SM is a protocol that defines initially a shared tree with the option of to switch to a source tree in order to optimize delay. It has the problem of the need of flooding the RP-Set to the entire Internet which makes the protocol a bad solution for inter-domain multicasting. Besides the protocol does not support policies or heterogeneous domains.

Chapter 11 - Summary and Conclusions

11.1 - Summary of work accomplished

This thesis has presented a comparison of the two most relevant proposals for inter-domain multicast routing (CBT and PIM-SM). The comparison was done based on the criteria outlined in chapter 8. There are a significant number of newer proposals, which are summarized in chapter 6.

A review of other comparisons of inter-domain multicast routing protocols was presented in chapter 7. The work presented in [Billhartz-XX][1] is the most similar to the work presented in this thesis. Chapters 9 and 10 present the protocol analysis for CBT and PIM-SM, respectively.

This thesis covered a number of review chapters. Specifically, the following areas were reviewed: components of today's Internet, unicast routing, multicast basics and intra-domain multicast routing protocols (See chapters 2 through 5). These chapters cover background that aids to better understand the comparative analysis presented in this thesis. The following sections in this chapter present a summary of the comparative analysis, and the conclusions of the comparison.

11.1.1 - Summary - Background

DVMRP [Waitzman-88] and PIM-DM [Estrin-96] periodically flood data packets throughout the network. Networks with no group members send prunes back to the source. This "Flood and Prune" mechanism exhibits poor scaling properties. It consumes bandwidth in links that do not lead to group members and requires every router in the network to keep state information for every active source/group pair. MOSPF [Moy2-94] floods group membership information to all routers in the network so that they can build multicast distribution trees. This flooding mechanism cannot be used in the global Internet either. In general, flooding control packets to the entire Internet does not scale. Any protocol that uses flooding of data packets or control packets is not a good solution for an Inter-Domain multicast protocol for the entire Internet.

Even though DVMRP is not a scalable solution it is the one that is used today in the Internet. The reason for that is that DVMRP was the only solution available in 1992. However, today it is clear that it is not a good solution (it does not scale because of flooding).

The next few paragraphs summarize the main protocols available for multicast routing, their characteristics and when they would be used. Table 11. 1 and Figure 11. 1 provide high level classification for multicast routing protocols.

Protocol	**Unicast Protocol Requirements**	**Network Size**
DVMRP	It provides its own.	Small
MOSPF	OSPF	Small
PIM-dense mode	Any	Small
PIM-sparse mode	Any	Large, but not global
CBT	Any	Large, but not global

Table 11. 1 - Comparison of Routing Protocols

Multicast routing protocols can be classified depending on the join type and the type of delivery tree they create. Figure 11. 1 describes these relationships. Explicit join trees are more efficient since they save bandwidth, i.e. multicast traffic only appears in networks that desire the traffic. Implicit join trees require flooding to

[1] Harris Corporation's work started around 1993 and their work is available in several

the entire network to discover new receivers. Source based trees provide low delay but use more memory in routers, while share trees create sub-optimal paths, but use less memory in routers.

Implicit Join	**Explicit Join**
- PIM-DM	- MOSPF
- DVMRP	- CBT
	- PIM-SM
Source Based Trees	**Shared Trees**
- DVMRP	- PIM-SM
- MOSPF	- CBT
- PIM-DM	
- PIM-SM	

Figure 11. 1 - Classification of routing protocols by join type and tree type.

11.2 - Summary of the comparison

11.2.1 - Summary - Protocol Status

PIM-SM is **commercially** available in Cisco routers while CBT is not supported by any vendor. This issue gives a stronger position to PIM-SM if deployment is required immediately. This may be a strong criterion on choosing which protocol to deploy in the enterprise or in an ISP network.

11.2.2 - Summary - Basic Characteristics

PIM-SM uses **explicitly joined shared trees** emanating from a “Rendezvous Point”. CBT also uses explicitly joined shared trees originated from a router that is called “core”. This is a more efficient mechanism than the "flood and prune"

places [Billhartz-95, Billhartz-95, Billhartz-96a, Billhartz-96, and Billhartz-97].

algorithm used by of other protocols such as PIM-DM and DVMRP. The core and the RP have the same function, which is to serve as the center of the shared tree.

The use of shared trees allows for all senders to use the same tree thus reducing considerably the amount of memory necessary in routers[2]. However, it introduces new problems such as increased end-to-end delay [Billhartz-97], center failure, center location [Thaler-97], address partitioning problems [Estrin-98b] and traffic concentration around the center [Wei-94]. These issues do not have a clear answer at this point, which means that there is room for improvement.

PIM-SM provides the option to switch to shortest path trees for sources that exceed a bandwidth threshold. The ability to switch to a source-based tree decreases end-to-end delay and allows specifying source specific policies. However, it makes the protocol more complicated to debug for network managers and also more complicated to develop for software developers.

PIM-SM **does not require having a specific unicast routing protocol**. The only thing that is needed is the presence of a routing table. Similarly, CBT does not depend on any specific unicast routing protocol either. This feature allows these two protocols to be developed independently from unicast routing protocols. This is an improvement over other protocols such as DVMRP, which has its own unicast routing protocol, or MOSPF, which is tied to OSPF.

CBT creates **bi-directional trees** that exhibit low bandwidth consumption. However, CBT cannot support unidirectional policies, neither can support source-

[2] The amount of memory needed in shared trees is in the order of the number of groups known by the router, while in source-based the amount of memory is in the order of the product of senders by groups. In mathematical terms, shared trees scale to O(G) while source trees scale to O(SxG), where S is the number of senders and G is the number of groups.

specific policies. PIM-SM offers source-based trees for certain sources. This means that it supports better source specific policies.

CBT uses "**hard states**", which means that messages are acknowledged and repeated after a time-out. PIM-SM uses "**soft state**" in which join messages are repeated a regular intervals, the states are cached and simply "disappear" if the information is not refreshed. Soft state mechanisms require fewer control overhead packets. This makes PIM-SM more efficient since it uses less bandwidth to create the distribution tree.

Both protocols are compliant with the **RFC-1112**. This enables interoperability, but then the protocols inherit the problems introduced by the IP Multicast Model. These problems are that any source can send to a group and that senders cannot account for who is listening to a multicast feed (see section 4.5).

All current multicast routing protocols rely on the **Reverse Path Forwarding** (RPF) algorithms (packets get dropped if not received on the shortest path to the source). Therefore, these protocols are not suited for asymmetric networks. Hodel proposed a multicast routing protocol called Policy Tree Multicast Routing (PTMR) that addresses this issue [Hodel-98].

11.2.3 - Summary - Technical Criteria

In general, **flooding** control packets to the entire Internet does not scale. Any protocol that has flooding of data packets or control packets is not a good solution for an Inter-Domain multicast protocol for the entire Internet. Both PIM-SM and CBT flood packets to the entire Internet. It is likely that neither of these solutions will be used in the Internet as the long-term solution because flooding does not scale.

Scalability of multicast routing protocols is directly affected by the amount of **forwarding state** that routers need to keep in order to forward multicast packets. That is, the protocol will scale better if it requires less memory in routers. Shared trees require less state but have more end-to-end delay and concentrate traffic around the center among other problems.

The **join time** is the time that elapses between a join request from a host and the reception of a multicast feed. PIM and CBT join times are low and about equal [Billhartz-97]. They both use a "join" mechanism towards the center of a tree.

The **end-to-end delay** that a multicast packet experiences by using a distribution tree built by PIM-SM using source-based trees is less than using a shared tree built CBT [Billhartz-97]. However, the delays are low and very similar. This criterion does not seem to have much impact on which protocol is the best. This factor only affects how much time a user has to wait to get packets from a group.

Shared trees do not provide minimum delay paths [Wei-94]. So, they are not a good solution for tele-conferencing that is a time-sensitive application. However, for applications with many senders, as in the case of Distributed Interactive Simulation (DIS), it is better to use shared trees because they consume less state in routers than source trees do.

Shared trees may not be optimal as Source-Based trees are. Shared-based trees have 10% worse latency that Source-Based Trees [Wei-95]. This implies that source-based trees are better suited for applications with real-time constraints such as video-conferencing. PIM-SM supports source-based trees, which makes it a better solution for multimedia applications with real-time constraints.

Neither PIM-SM nor CBT offer support for **Quality of Service (QoS)**. This is a limiting factor of these two protocols; it does not seem that in the near future, they will include support for QoS. Newer proposals such as YAM [Carlberg-97] and QoSMIC [Faloutsos-98] [Banerjea-98] propose alternatives that support quality of service.

Neither PIM-SM nor CBT offer support for **multipath routing**. That is, they only provide one route from each source to each recipient. In some applications, it is necessary to load balance traffic between links (e.g. high bandwidth applications). Multipath routing may increase considerably the size of the routing table. For example, if an implementation chooses to use source based trees and keep three routing entries for each source (S) sending to a group (G) present in the Internet then the state will be in the order of O(3 x S x G). This may create huge routing tables, but that may not be a big issue since memory prices are getting cheap everyday.

11.2.4 - Summary - Operational Criteria

It seems quite **easy to configure** multicast in a Cisco router. Just a global command and one command per interface. Support from a commercial vendor is a clear advantage of PIM-SM over CBT. However, PIM-SM is only supported in the latest versions of the Cisco's IOS. This implies that many existing routers may need to be upgraded which could be cumbersome for network managers in charge of large networks. Cost

Both PIM-SM and CBT are dependent on the multicast routing protocol that is running internally in a domain. This makes these protocols a bad choice for inter-domain multicast routing. Ideally, an inter-domain multicast routing protocol should be able to glue **heterogeneous domains** (each is running a different multicast routing protocol). MASC/BGMP [Kumar-98, Thaler-98, Estrin-98b] promises to be a solution that interconnects heterogeneous domains. Other old proposals tried to propose a hierarchy in the past but failed: HPIM [Handley-95] and HDVMRP [Thyagarajan-95].

None of the proposals offer the ability to do **billing**. This is one of the reasons ISPs are hesitant to deploy IP Multicast [Bellman-97].

There is also very little work on **security** and multicast. For example, in the current model a sender can send to a group without authorization. In the future, when multicast appears as a paid service, this issue will be of greater concern (see section 4.5).

Neither PIM-SM nor CBT offer the ability to control multicast traffic in a network, because none of them support **policies.** ISPs don't like this because they need to control from whom they receive multicast feed and also limit the amount of bandwidth used by multicast applications. This limitation is very crucial and it is a big obstacle to the deployment of IP multicast in the Internet. Hodel's recent work is offering an alternative for this issue but it is still too early in the development stage of his proposal [Hodel-98]. The BGMP protocol is another protocol that has a proposal on how to support policies for multicast [Thaler-98, Kumar-98].

Shared trees require traffic to travel through a core or RP. This creates the **third party dependency problem** [Meyer-97]. Autonomous Systems do not like to depend on a third party core. This is a main problem of PIM-SM and CBT that has slowed down their deployment.

Neither CBT nor PIM-SM support **Policy and QoS**. This is a big limitation of these protocols that makes them a less than an attractive solution for Inter-Domain Multicast Routing. There are two main recent proposals that offer ideas on how to solve these issues [Thaler-98, Hodel-98].

11.2.5 - Summary - Overall Assessment

Table 11. 2 is summary of the advantages and disadvantages of CBT and PIM-SM found by doing the comparative analysis done in this thesis.

	CBT	PIM-SM
Advantages	• Less State information • Better BW utilization • Better scalability • Unicast independence • It is free • Simple	• Commercial availability • Better bandwidth utilization • Less end-to-end delay • Unicast independence • Better robustness • Possibility of support of QoS • Soft state
Disadvantages	• Flooding of core set • No support of policies • Immaturity • Sub-optimal paths • Traffic Concentration • No heterogeneity • No aggregation • Core Failure • No support for billing • Third party dependency • No support of advanced features • Core Location • No security	• Flooding of core set • No support of policies • Unidirectional state • Complex • More state • No heterogeneity • No aggregation • RP failure • No support for billing • Third party dependency • No support of advanced features • RP location security?

Table 11. 2 - Advantages and Disadvantages of PIM-SM and CBT

11.2.6 - Summary - Table

Table 11. 3 is a summary of the comparative analysis presented in this thesis. Five sets of criteria were presented: Protocol Status, basic characteristics, technical criteria, operational criteria and overall assessment. For an explanation of the criteria meaning see chapter 8. The criteria are weighted so that the sum of the criteria is 100. Each protocol is "graded" from 1 to 10. The perfect protocol would obtain 100 as the "final grade".

It is shown that both protocols obtain low subjective ratings in many areas. This indicates that these two protocols require a review and that is not likely that they will become the definitive standard. Each criterion is weighted depending on its relative importance to a network manager. For example, the criteria considered most important for network managers is the support of policies. This is why it is given the highest weight of all the criteria used for the comparison.

For some of the criteria information was not available. For example, no information was found on the convergence time of the protocol. For this type of criteria a weight of zero was assigned to the criteria, so that it does not affect the comparison.

Protocol Status

#	Criteria	W	CBT	G/10	PIM-SM	G/10
1	Specification	1	RFC 2189 [Ballardie-97]	10	RFC 2362 [Estrin-98]	10
2	Status	2	Experimental	6	Experimental	6
3	Availability	2	Freeware	2	Commercial	8
4	Supported Platforms	2	FreeBSD 2.2.[67]	6	FreeBSD-2.2.1, FreeBSD-2.2.5, NetBSD-1.3 and SunOS-4.1.3, Cisco IOS.	8
5	MIB	1	Internet Draft	6	Internet Draft	6
6	Implementations	2	One	2	Two	8
7	Features Tested	1	N/A	2	N/A	2
8	Operational Experience	2	None	6	Some (e.g. UUNET)	8
9	Router Vendor Support	2	None	5	Cisco	10
		15		7.2		11.4

Basic Characteristics

#	Criteria	W	CBT	G/10	PIM-SM	G/10
1	Tree Type	1	Shared Tree	5	Both (Shared and Source)	7
2	Uni/Bi-directional	1	Bi-directional	9	Uni-directional	5
3	Loop-free-ness	1	No	1	No	1
4	RPF-Check	1	No	7	Yes	5
5	Hard/Soft State	1	Hard	3	Soft	7
6	Protocol Independence	1	Yes	10	Yes	10
7	RFC-1112 Compliant	1	Yes	10	Yes	10
		7		4.5		4.5

Technical Criteria

#	Criteria	W	CBT	G/10	PIM-SM	G/10
1	Link BW overhead	5	Moderate	5	Moderate	5
2	CPU Utilization	5	Moderate	5	Moderate	5
3	Router Memory	2	O(G)	7	From O(G) to O(SxG)	3
4	End-to-end Delay	2	Medium	5	Medium - Low	7
5	Join Time	1	Low	5	Low	5
6	Leave Time	0	N/A	0	N/A	0
7	Convergence Time	0	N/A	0	N/A	0
8	Traffic Characteristics	2	Concentrated around core.	3	Concentrated around RP, but could also distribute.	7
9	Address Allocation	2	No	1	No	1
10	Address Aggregation	1	No	1	No	1
		20		8.8		9.1

Operational Criteria

#	Criteria	W	CBT	G/10	PIM-SM	G/10
1	Ease of Configuration	2	Cumbersome	3	Easy	9
2	Ease of Management	1	Limited management tools for multicast	3	Limited management tools for multicast	3
3	Robustness	4	Yes	7	Yes	7
4	Price	2	Free	9	Moderate	5
5	Interoperability	1	Yes	7	Yes	7

6	Installation Time	2	Not commercial yet	4	Depends on size of the network	9
7	Billing Capacity	3	No	1	No	1
8	Impact on existing network	3	High	2	High	2
9	Multipath routing	1	No	1	No	1
10	QoS Support	1	No	1	No	1
11	Mobility Support	1	No	1	No	1
12	Heterogeneity	3	No	5	No	5
13	Policy Support	8	No	1	No	1
14	Security	6	No	3	No	3
15	Complexity	2	Less complex	7	More Complex	3
16	Third-Party dependence	5	Yes	3	Yes	3
		45		15.2		15.8

Overall Assessment

#	**Criteria**	**W**	**CBT**	**G/10**	**PIM-SM**	**G/10**
1	Scalability	4	No	5	No	5
2	Suitability	3	Intra-domain	5	Intra-domain	5
3	Advantages *	3	Less state	5	Source Tree	8
4	Disadvantages *	3	Flooding of Core set	3	Flooding of RP set	3
		13		5.9		6.8

Total **100** **41.6** **47.6**

* Only the main advantage and disadvantage are shown. For more see analysis in previous chapters

Table 11. 3 - Summary Table of the comparison

11.3 Conclusions

PIM-SM and CBT were designed with the aim to have a protocol that scales to the entire Internet. However, based on the comparative analysis performed, it is clear that many issues still need to be solved for IDMR protocols and that these two protocols are better suited to be used internally in a domain.

Neither PIM-SM nor CBT seem mature enough to become the final solution to become an Internet standard. If a protocol needs to be deployed immediately then PIM-SM is a better overall solution at this time. The main reason for this choice is that PIM-SM is supported by commercial routers, while CBT has not been implemented by any router vendor. In other terms, the two protocols are very similar (see Table 11. 3).

Many issues still need to be solved; however, there are four that are more important than other issues. These are: scalability, policies, heterogeneity and aggregation. The next paragraphs summarize these problems.

Scalability:

In PIM-SM and CBT, the mechanism to map group address to its core router requires **flooding** to the entire network the set of routers that are candidates to be a core. This flooding mechanism is called the "Bootstrap" method. This approach may work in small environments, but clearly it **does not scale** to the whole Internet. Recently, Otha et al. proposed a mechanism using DNS to advertise core routers associated with a group that does not require state in intermediate routers or control traffic overhead to map core to groups. This approach may scale better but it is still in draft status in the IETF and it may suffer considerable changes in the next few months [Ohta-98]. This is a flaw in the design of PIM-SM and CBT that clearly impedes its success as a good solution for Inter-Domain Multicast Routing (IDMR).

If this is not fixed in the next versions of these protocols, then it is likely that another proposal will become the standard for IDMR.

Policies:

Current proposals for inter-domain multicast routing do not allow Internet Service Providers to control multicast transit traffic, that is, they do not support **policies**. In other words, it is not possible for an ISP to control the multicast packets that travel its network. Multicast traffic could affect the bandwidth available for unicast applications and there is no way to control it at this point. New proposals should focus in providing an AS with the flexibility and the autonomy to control the receive path of multicast data traffic.

Aggregation:

In order for multicast routing protocols to achieve scalability for the entire Internet they need to support some form of **aggregation** so that the state in routers does not grow without bound. As November 1998, there is no clear solution for this problem. Neither CBT nor PIM-SM has an answer for this issue.

Heterogeneity:

PIM-SM and CBT work in a single routing layer - i.e. there is **no hierarchy** associated with these protocols. This means that they are mainly suited for intra-domain environments. This creates the problem that the protocol does not support heterogeneous domains. Ideally, each autonomous system in the Internet should be able to choose ~~which proto~~col a intra-domain multicast routing protocol independently.

There is an opportunity to design a protocol that addresses all the issues not solved by current proposals. The qualitative analysis presented in this thesis serves as a starting point for a new proposal for an IDMR protocol. The following **goals** must be achieved by a new proposal in order to improve current protocols:

a) Scalability: The protocol should use mechanisms that could be used globally in the Internet. For example, flooding should be avoided.

b) Policies: The protocol should be able to accommodate policies from network managers so that they can control multicast traffic.

c) Aggregation: There should be a way to aggregate routing table entries and control traffic so that they do not grow without bound.

d) Heterogeneity: The protocol should allow each domain to run a different multicast protocol that fits its needs.

d) Other features: The protocol should support multipath routing, billing, QoS, mobility. However, these characteristics are not as important as the above characteristics.

Another conclusion from this work is that the IP Multicast Model described in RFC-1112 [Deering-89] needs to be reviewed. Multicast routing protocols comply with the model proposed by Deering, however this model introduces problems that are inherited by other multicast protocols and applications (see section 4.5).

Future Perspective

The work performed for MASC/BGMP [Thaler-98, Kumar-98] has started to move in the direction to support the above requirements. However, this work is in early stages and it should closely monitored in the near future. Other proposals have also recognized these problems, but MASC/BGMP seems to be having the most attention at this point in time (see chapter 6).

Finally, it should be expected several years will pass before a solution that satisfies these requirements for IDMR is encountered. The process might be accelerated if a killer application for multicast appears. At this point it seems that TV-like Internet broadcasting may become a killer application for multicast (see section 4.15).

Chapter 12 - Recommendation s on Future work

There are many open issues in multicast technologies at the moment. The best place to find open research issues is a survey article written by Diot et al. [Diot-97]. This chapter attempts to summarize the state of multicast technologies as of November of 1998.

12.1 - Extend the comparison

This work could be expanded by comparing newer proposals for IDMR using as the basis the criteria presented in this thesis. For other alternatives for IDMR see Chapter 6.

12.2 - Simulation

A simulation would be helpful to better understand the proposals. Chapter 14 has an introduction to the trial simulations performed in this thesis.

12.3 - Review the IP Multicast model

The IP multicast model described in RFC-1112 is the basis of every other piece of work in IP multicast. This author argues that there are two design flaws in the model: a sender can send to any group it wants and a receiver can any join it desires. This may work in the research community but clearly it does not work in the commercial world. For more on this argument see section 4.5.

12.4 - Addresses management issues

There have been proposals to use DHCP to allocate multicast addresses. This technique is good for small groups however this technique is clearly not scalable for Internet scale networking. The Multicast Address-Set Claim (MASC) protocol is one of the newest proposals on the multicast address allocation issue [Estrin-98b].

12.5 - IP Multicast in IPv6

The future of multicast routing over Ipv6 is unclear at the time of this writing. It is highly unlikely that there will ever be a version of DVRMP for IPv6. OSPF is at the present time being defined by the IETF for IPv6, and M-OSPF will work similarly to the way it does with Ipv4. PIM (either of the versions) and CBT are easily modifiable to support Ipv6, because both only require that there be a unicast routing table present in the router.

12.6 - Interoperability

It is expected that there will be many different protocols deployed in the Internet. No single protocol would be able to meet the requirements for all the applications. This suggests that there is a need for inter-operability between all the existing multicast routing protocols. Since the MBONE core routing protocol is DVMRP, many vendors provide specifications on how to connect their routing protocol to DVMRP. For example, vendors providing support for MOSPF also provide a way to interoperate with DVMRP.

12.7 - Heterogeneity

It is desirable that a new protocol supports multiple domains running different multicast routing protocols. This will better serve the needs of the Internet. No single protocol has solved this issue as of today.

12.8 - Policy-based routing

Multicast policy and access-control do not exist in today's MBONE. Policy and access-control is done by packet filtering. A better solution is needed. Some form of policy-based hierarchical routing is required for Internet-scale operation. One of the possible alternatives is to have two levels of routing: one for the Intranet and another for the connection with the outside world. Just as the unicast routing protocol has an interior and an exterior routing protocol.

None of the current multicast routing protocols specify a mechanism to express and enforce multicast routing policies and forwarding policies among ISPs. The lack of policy control is one of the biggest reasons that multicast has not been deployed by many ISPs. They feel that they need a way to control the routing and forwarding of multicast traffic [Maufer-98]

However, the ISP's requirements are quite different from the Intranet managers. Current existing protocols such as DVMRP and MOSPF are perfect to be deployed in the internal network. However, they do not scale and have a lack of policy control mechanisms that ISPs need urgently in order to be able to deploy IP Multicast. Current multicast routing protocols do not allow ISPs to control how the routing and forwarding for multicast is done.

A recent piece of work by Hodel proposes a protocol that supports policies [Hodel-98].

12.9 - Congestion Control

How to do congestion control for multicast? Congestion control frames returning to a sender may create even more congestion.

12.10 - Convergence time

There is no previous work that focus on measuring how much time elapses between a change occurs in the network and the time that all the routers know about the change in multicast routing. Further research is needed in this area. It might be possible that the protocol behaves inadequately in the presence of link or router failures.

12.11 - Center Location

The placement of Cores in shared based trees is an area of open research at this time. Thaler presented a survey of current alternatives in [Thaler-97].

12.12 - ISP Billing and settlement issues

There is no mechanism to charge for multicast transmissions. The question on how to do billing for multicast is an open issue at this time. This is one of the reasons ISP have not deployed multicast yet [Bellman-97].

MBONE is free access and it does not generate any revenue. So, the people that maintain it do not have any urgency to solve its problems. The MBONE relies on the volunteer work of researchers and engineers around the world.

12.13 - Reliable Multicast Transport Protocols

A comparative analysis can be done on reliable multicast transport protocols. Several surveys can be found at [Diot-97] [Obraczka-98] [Tascnets-98]. One of the first proposals was presented in [Armstrong-92].

12.14 - Network Management for Multicast

There are very few tools to manage IP Multicast. The tools that exist are very rudimentary. There is room to create new improvements in this area. Thaler and Aboba presented a survey of existing tools in [Thaler-98b].

12.15 - Aggregation

Ways of bounding and aggregating control traffic in PIM is something with no answer at this point. This is an issue also for CBT [Ballardie-95]. Large-scale use of multicast may require some form of aggregation of IP level multicast tree indices (state in Mbone routers). As of November 1998, there is no clear alternative on how to aggregate multicast addresses and control traffic. In other words, there is a need to design the multicast version of CIDR.

12.16 - Multicast over different L2 technologies

It is not clear at this moment the mappings of IP multicast for many level 2 technologies. For example, work is underway to define the mappings of IP multicast over ATM, SONET, ADSL, Frame Relay, etc. The leading proposal for support of multicast in ATM is described in [Armitage-96].

12.17 - Chapter Summary

This chapter has presented some of the open areas of research in relation IP multicast technologies. Some of the issues presented in this chapter could serve as topics for a future thesis.

Chapter 13 - References

[Aggarwal-96] S. Aggarwal and S. Paul, "A Flexible Protocol Architecture for Multi-Party Conferencing," Proceedings of ICCN'96, pp. 81-91, October 1996. Available on-line at: http://remus.rutgers.edu/~psanjoy/ic3nfinal.www.ps.Z

[Aggarwal-98] S. Aggarwal, S. Paul, D. Massey, and D. Calderaru, "A Flexible Protocol for Multi-party Conferencing: from Design to Implementation," Bell Labs Technical Memorandum 11345-980424-03TM, submitted for publication, April 1998.

[Ammar-94] M. Ammar, S.Y. Cheung, and C. Scoglio, "Routing Multipoint Connections Using Virtual Paths in an ATM Network," Proceeding on IEEE INFOCOM 93, San Francisco, March 1994, pp. 98-105.

[Ammar-97] M. Ammar and D. Towsley, "Group (Multicast) Communication in Wide Area Networks," http://www.cc.gatech.edu/fac/Mostafa.Ammar/tutorial.html, Tutorial given on September 1997, Date of search: April 11, 1998.

[Armitage-96] G. Armitage, "Support for Multicast over UNI 3.0/3.1 based ATM Networks," RFC 2022, November 1996.

[Armstrong-92] S. Armstrong, A. Freier, and K. Marzullo, "Multicast Transport Protocol," RFC 1301, February 1992.

[Baker-95] F. Baker and R. Coltun, "OSPF Version 2 Management Information Base," RFC 1850, November 1995.

[Ballardie-93] A. Ballardie, P. Francis, and J. Crowcroft, "Core Based Trees (CBT) and Architecture for Scalable Inter-Domain Multicast Routing," in ACM SIGCOMM '93, pp. 85-95, ACM, September 1993.

[Ballardie-95] A. Ballardie, "A New Approach to Multicast Communication in a Datagram Internetwork," Ph.D. Thesis, University College London, May 1995. Available online at: ftp://cs.ucl.ac.uk/darpa/IDMR/ballardie-thesis.ps.Z

[Ballardie-97] A. Ballardie, "Core Based Trees (CBT version 2) Multicast Routing," RFC 2189, September 1997.

[Ballardie2-97] A. Ballardie "Core Based Trees (CBT) Multicast Routing Architecture," RFC 2201, September 1997.

[Ballardie-97c] A. Ballardie "Core Based Trees (CBT) Multicast Routing MIB," IETF Internet Draft, April 1997.

[Ballardie-98] A. Ballardie, B. Cain, and Z. Zhang, "Core Based Trees (CBT version 3) Multicast Routing," IETF Internet Draft, August 1998. Available on-line at: http://search.ietf.org/internet-drafts/draft-ietf-idmr-cbt-spec-v3-01.txt

[Banerjea-98] A. Banerjea, M. Faloutsos, and R. Pankaj, "Designing QoSMIC: a Quality of Service sensitive Multicast Internet protoCol," IETF Internet Draft:, April 1998. Available on-line at: http://search.ietf.org/internet-drafts/draft-banerjea-qosmic-00.ps

[Batsell-95] S. Batsell, "Multicast Requirements for Distributed Interactive Simulation," Slide Presentation, IDMR meeting of 32nd IETF, April 1995. Available on-line at: ftp://cs.ucl.ac.uk/darpa/IDMR/IETF-APR95/batsell-slides.ps

[Bauer-95] F. Bauer and A. Varma, "Degree-Constrained Multicasting in Point-to-Multipoint Networks," Proceedings of IEEE INFOCOM, April 1995, pp. 369-376.

[Bay-96] Bay Networks, "Exploiting Internetwork Multicast Services," http://www.baynetworks.com/Products/Reports/multicast.html, 1996.

[Bellman-97] R. Bellman, "The Push for IP Multicasting," Business Communications Review, June 1997.

[Billhartz-95] J. Billhartz, J. B. Cain, E. Farrey-Goudreau, and D. Fieg, "A comparison of CBT and PIM via simulation," IDMR Working Group presentation, April 1995, Available on line at: ftp://cs.ucl.ac.uk/darpa/IDMR/IETF-APR95/CBTvsPIM-cain.ps

[Billhartz2-95] J. Billhartz, J. B. Cain, E. Farrey-Goudreau, and D. Fieg, "Performance and Resource Cost Comparisons of Multicast Routing Algorithms," Report prepared by Harris Corporation for the Naval Research Laboratory under contract N00014-93-C-2186, 1995. Available on line at: ftp://taurus-littrow.itd.nrl.navy.mil/Pub/OpNet/

[Billhartz-96a] T. Billhartz, J. Cain, E. Farrey-Goudreau, D. Fieg and S. Batsell, "Simulation Comparison of CBT and PIM Multicasting for Distributed Interactive Simulation (DIS)," Proceedings of the 1996 Society Computer Simulation Western Multi-conference: Communication Networks

Modeling and Simulation Conference, January 14-17, 1996, pp. 246-251. Available on-line at: http://nrg.cind.ornl.gov/~sgb/mars/SCS.ps

[Billhartz-96] J. Billhartz, J. B. Cain, E. Farrey-Goudreau, D. Fieg, and S.G. Bastell, “Performance and resource cost comparisons for the CBT and PIM multicast routing protocols in DIS environments,” Proceedings IEEE INFOCOM '96, San Francisco, CA, USA, 24-28 March 1996. Available on-line at: http://nrg.cind.ornl.gov/~sgb/mars/Infocom.ps

[Billhartz-97] J. Billhartz, J. Bibb Cain, E. Farrey-Goudreau, D. Fieg, and S.G. Bastell, “Performance and Resource Cost Comparisons for the CBT and PIM multicast Routing Specification,” IEEE Journal on Selected Areas in Communications, Vol. 15, No. 3, April 1997, pp. 304-315. Available on-line at: http://nrg.cind.ornl.gov/~sgb/mars/bibb.ps

[Cain-97] B. Cain, A. Thyagarajan, and S. Deering, "Internet Group Management Protocol, Version 3," IETF Internet Draft, Expires November 21, 1997. Available-on-line: http://search.ietf.org/internet-drafts/draft-ietf-idmr-igmp-v3-00.txt

[Calvert-94] K. Calvert, R. Madhavan, and E. Zegura, "A Comparison of Two Practical Multicast Routing Schemes," Georgia Institute of Technology College of Computing Technical Report GIT-CC-94/25, February 1994. Available on-line at: ftp://ftp.cc.gatech.edu//pub/coc/tech_reports/1994/GIT-CC-94-25.ps.Z

[Carlberg-97] K. Carlberg and J. Crowcroft, "Building Shared Trees using a one-to-many joining mechanism," ACM SIGCOMM Computer Communication Review, pages 5-11, January 1997. Available on-line at: http://www.acm.org/sigcomm/ccr/archive/1997/jan97/ccr-9701-carlberg.ps

[Casner-92] S. Casner and S. Deering, “First IETF Audiocast,” ACM SIGCOMM Computer Communications Review, Vol. 22, No. 3, July 1992.

[Casner-93] S. Casner, “Frequently Asked Questions (FAQ) on the Multicast Backbone,” May 1993. Available on-line at: http://www.mbone.com/mbone/mbone.faq.html

[Casner-94] S. Casner, "Major MBONE Routers and links," Available on-line at: ftp://ftp.isi.edu/mbone/mbone-topology.gif

[Cisco10-98] Cisco, "IP Multicast training material," ftp://ftpeng.cisco.com/ipmulticast/multicast_training.html, Date of Search: August 13, 1998.

[Coltun-98] R. Coltun, S. Deering, T. Pusateri, R. Shekhar, "DVMRPv1 Applicability Statement for Historic Status," IETF Internet Draft, July 1998. Available on-line at: http://search.ietf.org/internet-drafts/draft-ietf-idmr-dvmrp-v1-as-00.txt

[Dalal-78] Y.K. Dalal and R.M. Metcalfe, “Reverse Path Forwarding of Broadcast Packets,” Commun. of the ACM, Vol. 21, No. 12, 1978.

[Deering] S. Deering, A. Thyagarajan, and W. Fenner. "mrouted UNIX manual page," mrouted(8).

[Deering-85] S.E. Deering and D. Cheriton, “Host groups: A Multicast Extension to the Internet Protocol,” RFC 966, December 1985.

[Deering-86] S.E. Deering, “Host extensions for IP multicasting,” RFC 988, July1986.

[Deering-88] S.E. Deering, “Multicast Routing in Internetworks and Extended LANs,” Computer Communications Review (Proc. SIGCOMM ’88), Vol18, No. 4, August 1988.

[Deering2-88] S.E. Deering, “Host extensions for IP multicasting,” RFC 1054, May, 1988.

[Deering-89] S.E. Deering, “Host extensions for IP multicasting,” RFC 1112, August 1989.

[Deering-90] S.E. Deering and D. Cheriton, “Multicast Routing in Datagram Internetworks and Extended LANs,” ACM Trans. on Computer Systems,Vol. 8, No. 2, pp. 85-110, May 1990.

[Deering-91] S.E. Deering, “Multicast Routing in a datagram Internetwork,” Ph.D. thesis, Stanford University, December 1991. Available on-line:
ftp://gregorio.stanford.edu/vmtp-ip/sdthesis.part1.ps.Z;
ftp://gregorio.stanford.edu/vmtp-ip/sdthesis.part2.ps.Z;
ftp://gregorio.stanford.edu/vmtp-ip/sdthesis.part3.ps.Z

[Deering-94] S.E. Deering, D. Estrin, D. Farinacci, V. Jacobson, C.G. Liu , and L. Wei, “An Architecture for Wide-Area Multicast Routing,” in Proc. ACM SIGCOMM’94, London, 1994, pp. 126-135.

[Deering-96] Deering, S.; Estrin, D.L.; Farinacci, D.; Jacobson, V.; Ching-Gung Liu; Liming Wei, "The PIM architecture for wide-area multicast routing," IEEE/ACM Transactions on Networking, vol.4, no.2, p. 153-62, April 1996.

[Deering-98] S.E. Deering and R. Perlman "Preliminary Report on the IAB Workshop on Routing and Addressing," March 23-25, 1998, Santa Clara, CA.

[Deering-98b] S. Deering, D. Estrin, D. Farinacci, V. Jacobson, A. Helmy, D. Meyer, and L. Wei "Protocol Independent Multicast Version 2 Dense Mode Specification," November 3, 1998. Available on-line at: http://www.ietf.org/internet-drafts/draft-ietf-pim-v2-dm-01.txt

[Deering-98c] S. Deering, S. Hares, C. Perkins and R. Perlman, "Report on the 1998 IAB Routing Workshop," IETF Internet Draft, November 15, 1998. Available on-line at: ftp://ftp.ietf.org/internet-drafts/draft-iab-rtr-workshop-00.txt

[Diot-97] C. Diot; W. Dabbous, and J. Crowcroft, "Multipoint Communication: A survey of protocols, functions, and mechanisms," IEEE Journal on selected areas in communications, Vol. 15, No. 3, April 1997.

[Dubray-98] K. Dubray, "Terminology for IP Multicast Benchmarking" RFC-2432, October 1998.

[Eriksson-94] H. Eriksson "MBONE: The Multicast Backbone," Communications of the ACM, Vol. 37, No. 8, August 1994.

[Estrin-96] S. Deering, D. Estrin, D. Farinacci, V. Jacobson, A. Helmy, D. Meyer, and L. Wei " Protocol Independent Multicast Version 2 Dense Mode Specification," November 3, 1998. Available on-line at: http://www.ietf.org/internet-drafts/draft-ietf-pim-v2-dm-01.txt

[Estrin-97] D. Estrin, D. Farinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, V. Jacobson, C. Liu, P. Sharma, and L. Wei, "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification," RFC 2117, June 1997.

[Estrin-98] D. Estrin, D. Farinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, V. Jacobson, C. Liu, P. Sharma, and L. Wei, "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification," RFC 2362, June 1998.

[Estrin-98b] D. Estrin, R. Govindan, M. Handley, S. Kumar, P. Radoslavov, D. Thaler, "The Multicast Address-Set Claim

(MASC) Protocol, IETF Internet Drafts, August 1998. Available on-line at: http://search.ietf.org/internet-drafts/draft-ietf-malloc-masc-01.txt

[Faloutsos-98] M. Faloutsos and A. Banerjea and R. Pankaj, "QoSMIC: Quality of Service sensitive Multicast Internet protoCol," SIGCOMM, Sep 2-4, Vancouver BC,1998. Available online at: http://www.acm.org/sigcomm/sigcomm98/tp/paper12.ps http://www.acm.org/sigcomm/sigcomm98/slides/slides_12.ppt

[Farinacci-98] D. Farinacci, Y.Rekter, P.Lothberg, H. Kilmer, J. Hall, "Multicast Source Discovery Protocol (MSDP)," IETF Internet Draft, August 1998. Available on-line at: http://search.ietf.org/internet-drafts/draft-farinacci-msdp-00.txt

[Fenner-95] B. Fenner and A. Ballardie, "IDMR Working Group minute," April 1995. Available on-line at: ftp://cs.ucl.ac.uk/darpa/IDMR/IETF-APR95/idmr-apr95-minutes.txt

[Fenner-97] W. Fenner, "Internet Group Management Protocol," RFC 2236, November 1997.

[Floyd-93] S. Floyd and V. Jacobson, "The Synchronization of Periodic Routing Messages," ACM SIGCOMM '93 symposium, September 1993.

[Fuller-93] V. Fuller, T. Li, J. Yu, and K. Varadhan, "Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy, " RFC 1519, September 1993.

[Gross-92] P. Gross, "Choosing a Common IGP for the IP Internet," RFC 1371, October 1992.

[Halabi-97] B. Halabi, "Internet Routing Architectures," Cisco Press, 1997.

[Handley-95] M. Handley, J. Crowcroft, I. Wakeman, "Hierarchical Protocol Independent Multicast (HPIM)," University of College London, Oct 1995. Available on-line at: ftp://cs.ucl.ac.uk/darpa/IDMR/hpim.ps

[Hanks1-94] S. Hanks, T. Li, D. Farinacci, and P. Traina, "Generic Routing Encapsulation," RFC-1701, October 1994.

[Hanks2-94] S. Hanks, T. Li, D. Farinacci, and P. Traina, "Generic Routing Encapsulation over IPv4 networks," RFC-1702, October 1994.

[Haskin-95] D. Haskin, "A BGP/IDRP Route Server alternative to a full mesh routing," RFC 1863, October 1995.

[Hedrick-88] C.L. Hedrick, "Routing Information Protocol," RFC 1058, June 1988.

[Helmy-97] A. Helmy, "STRESS: Testing Applied to a Multicast Routing Protocol," University of Southern California, July 22, 1997. Available on-line at: http://catarina.usc.edu/ahelmy/stress/mascots.ps.gz

[Hinden-97] R. Hinden and S. Deering, "IPv6 Multicast Address Assignments" http://www.ietf.org/internet-drafts/draft-ietf-ipngwg-multicast-assgn-04.txt, July 1997.

[Hodel -98] H. Hodel, "Policy Tree Multicast Routing: An Extension to Sparse Mode Source Tree Delivery," in Proc. ACM SIGCOM' 98, Volume 28, Number 2, April 1998. Available on-line at: http://www.acm.org/sigcomm/ccr/archive/1998/apr98/ccr-9804-hodel.html

[Holbrook-98] H. Holbrook, D. Cheriton, "Single-Source Multicast (Multicast)," Work in progress for Ph.D. Thesis, Stanford University, March 1998. Available on-line at: http://www.dsg.stanford.edu/holbrook/express/ http://www.dsg.stanford.edu/holbrook/express.ps

[Huitema-93] C. Huitema, "Routing in the Internet," Prentice Hall, 1993, Chp. 11, pag. 235-259, ISBN 0-13-132192-7

[Hwang-92] F.K. Hwang, and D.S. Richards, "Steiner Tree Problems," IEEE Networks,Vol. 22, pp. 55-89, January 1992.

[IDMR-email] IDMR Mailing List Archives, http://www3.juniper.net/~pusateri/idmr/

[Jacobson-94] V. Jacobson, "Some Notes on Multicast Scaling and PIM," Slide Presentation of IDMR working group meeting, July 1994. Available on-line at: ftp://cs.ucl.ac.uk/darpa/IDMR/IETF-JUL94/van-pim-scaling-slides.ps

[Keshav-98] S. Keshav, S. Paul, "Centralized Multicast," submitted for publication, April 1998. Available on line at: http://www.cs.cornell.edu/skeshav/papers/cm.ps

[Komandur-98] S. Komandur, M. Doar, D. Mosse, "The Domainserver Hierarchy for Multicast Routing in ATM Networks," Sixth IFIP Workshop on Performance Modeling and Evaluation of ATM Networks (IF IP ATM'98), West Yorkshire, UK, July 1998. Available on-line at: ftp://speedy.cs.pitt.edu/komandur/published/ifip98.ps

[Kou-81] L. Kou, G. Markowshy, and L. Berman, "A Fast Algorithm for Steiner Trees," Acta Informatica 15, pp. 141-145, 1981.

[Kumar-98] S. Kumar, P. Radoslavov, D. Thaler, C. Alaettinoglu, D. Estrin, and M. Handley, "The MASC/ BGMP Architecture for Inter-domain Multicast Routing," in Proc. ACM SIGCOMM 98, September 1998, Vancouver, Canada. Available on-line at:
http://www.acm.org/sigcomm/sigcomm98/tp/paper08.ps
http://www.acm.org/sigcomm/sigcomm98/slides/slides_08.ppt

[Lougheed-89] K. Lougheed and Y. Rekhter "A Border Gateway Protocol (BGP)," RFC 1105, June 1989.

[Lougheed-90] K. Lougheed and Y. Rekhter "A Border Gateway Protocol (BGP-2)," RFC 1163, June 1990.

[Lougheed-91] K. Lougheed and Y. Rekhter "Border Gateway Protocol 3 (BGP-3)," RFC 1267, October 1991.

[Malkin-93] G. Malkin, "RIP Version 2 Carrying Additional Information," RFC 1388, January 1993.

[Malkin2-94] G. Malkin, "RIP Version 2 – Protocol Applicability Statement," RFC 1722, November 1994.

[Malkin-94] G. Malkin, "RIP Version 2 - Carrying Additional Information," RFC 1723, November 1994.

[Malkin-97] G. Malkin, R. Minnear, "RIPng for IPv6," RFC 2080, January 1997.

[Maufer-98] T. Maufer, "Deploying IP Multicast in the enterprise," Prentice Hall, Upper Saddle River, 1998, 275 pages. ISBN 0-13-8997687-2

[McCloghrie-98] K. McCloghrie, D, Farinacci, D. Thaler, "Protocol Independent Multicast MIB," July 1998, Available on-line at:
http://search.ietf.org/internet-drafts/draft-ietf-idmr-pim-mib-05.txt

[Meyer-94] G. Meyer, “Extensions to RIP to Support Demand Circuits,” RFC 1582, February 1994.

[Meyer-97] D. Meyer, "Some Issues for an Inter-domain Multicast Routing Protocol," IETF Internet Draft, March 1997. Available on-line at: http://info.internet.isi.edu/0/in-drafts/files/draft-ietf-idmr-membership-reports-01.txt

[Meyer-98] D. Meyer, "Administratively Scoped IP Multicast," RFC-2365, July 1998. Available on-line at: http://info.internet.isi.edu:80/in-notes/rfc/files/rfc2365.txt

[Mills-84] D.L. Mills “Exterior Gateway Protocol formal specification,” RFC 904, April 1984.

[Moy-91] J. Moy, “OSPF Version 2,” RFC 1247, July 1991.

[Moy-94] J. Moy, “OSPF Version 2,” RFC 1583, March 1994.

[Moy2-94] J. Moy, "Multicast Extensions to OSPF ," RFC 1584, March 1994.

[Moy3-94] J. Moy, "Multicast routing extensions for OSPF," Communications of the ACM, Vol. 37, No. 8, pp. 61-66, August 1994.

[Moy4-94] J. Moy, "MOSPF: Analysis and Experience," RFC-1585, March 1994.

[Moy-98] J. Moy, “OSPF Version 2,” RFC 2328, April 1998.

[Moy-98c] J. Moy, "OSPF: Anatomy of an Internet Routing Protocol," Addison-Wesley, January 1998. ISBN 0-201-63472-4

[Nerney-97] C. Nerney, “The spreading of IP Multicast,” Network World, October 20, 1997, page 48.

[Noronha-94] C.A. Noronha and F.A. Tobagi, “Optimum Routing of Multicast Streams,” IEEE INFOCOM ‘94, Vol. 2, Toronto, pp. 865-873, June 1994

[Obraczka-98] Obraczka, K., "Multicast Transport Mechanisms: A Survey and Taxonomy,” to appear in IEEE Communications, 1998.

[Ohta-98] M. Ohta, J. Crowcroft, "Static Multicast," IETF Internet Draft, March 1998. Available on line at: http://search.ietf.org/internet-drafts/draft-ohta-static-multicast-00.txt

[Perkins-96] C. Perkins, "IP Encapsulation within IP," RFC-2003, October 1996.

[Perlman-98] R. Perlman, C-Y. Lee, A. Ballardie, J. Crowcroft, "A Design for Simple, Low-Overhead Multicast," IETF Internet draft, August 1998. Available on-line at: http://search.ietf.org/internet-drafts/draft-perlman-simple-multicast-00.txt

[Peterson-96] L. Peterson, and D. Bruce, "Computer Networks," Morgan Kaufmann Publishers, Inc., 1996, p. 4, 14, 125, 215-216, 254, 262-267, 460-461, 502.

[Parsa-97] M. Parsa and J.J Garcia-Luna-Aceves, "A protocol for scalable loop-free multicast routing," IEEE Journal on Selected Areas in Communications, vol. 15, no. 3, pp. 316-331, April 1997.

[Petitt-96] D. Petitt, "Solutions for Reliable multicasting," M.S. Thesis, Naval Postgraduate School, September 1996. Available on-line at: http://web.nps.navy.mil/~seanet/mcast/Thesis.htm

[Postel-92] J. Postel, "Introduction to the STD Notes," RFC 1311, March 1992.

[Pullen-96] M. Pullen, "QoS IP Network Simulation Ipmc, RSVP, QOSPF in OPNET," Slide presentation at IETF San Jose Meeting, December 1996.

[Pullen-98] M. Pullen, R. Malghan, L. Lavu, G. Duan, J. Ma, H. Nah, "A Simulation Model for IP Multicast with RSVP," IETF Internet draft, July 1998. Available on-line at: http://search.ietf.org/internet-drafts/draft-pullen-ipv4-rsvp-04.ps

[Pusateri-93] T. Pusateri, "IP Multicast over Token-Ring Local Area Networks," RFC 1469, June 1993.

[Pusateri-98] T. Pusateri, "Distance Vector Multicast Routing Protocol," Internet Draft from IDMR WG, March 1998. Available online at: http://www.ietf.org/internet-drafts/draft-ietf-idmr-dvmrp-v3-06.txt

[Rekhter-95] Y. Rekhter and T. Li, "A Border Gateway Protocol 4 (BGP-4), " RFC 1771, March 1995.

[Reynolds-94] J. Reynolds and J. Postel, "ASSIGNED NUMBERS," RFC 1700, October 1994.

[Rivest-92] R. Rivest, "The MD5 Message-Digest Algorithm," RFC 1321, April 1992.

[Semeria-96] C. Semeria, and T. Maufer, “Introduction to IP Multicasting Routing,” http://www.ipmulticast.com/community/semeria.html March 1996, Date of search on the WWW: May 3, 1998.

[Semeria-98] C. Semeria, and T. Maufer, “Introduction to IP Multicasting Routing,” http://www.3com.com/nsc/501303.html, Date of search: March 21, 1998.

[Schulzrinne-96] H. Schulzrinne, S. Casner, R. Frederick and V. Jacobson, “RTP: A Transport Protocol for Real-Time Applications. Audio-Video Transport Working Group,” RFC 1889, January 1996.

[Schulzrinne2-96] H. Schulzrinne, “RTP Profile for Audio and Video Conferences with Minimal Control. Audio-Video Transport Working Group,” RFC 1890, January 1996.

[Shields-97] C. Shields and J.J. Garcia-Luna-Aceves, "The ordered core based tree protocol,” in Proc. IEEE INFOCOM' 97, Kobe, Japan, pp. 884-91, April 1997.

[Shields-98] C. Shields and J.J. Garcia-Luna-Aceves, "Hierarchical Multicast Routing,” in Proc. Seventeenth Annual ACM SIGACT-SIGOPS Symposium on principles of distributed computing (PODC 98), Puerto Vallarta, Mexico, June 28-July 2 1998. Available on-line at: http://www.cse.ucsc.edu/research/ccrg/publications/clay.podc98.ps.gz

[Shulka-94] S. Shulka, E. Boyer, J. Klinker, "Multicast Tree Construction in Network Topologies with Asymmetric Link Loads,” Naval Postgraduate School NPS-EC-94-012, September 20, 1994. Available on-line at: ftp://ftp.nps.navy.mil at /pub/ece/shulka/nps-mltcst-asym-linksvl.ps

[Sola-98] M. Sola, M. Ohta, T. Maeno "Scalability of Internet Multicast Protocols,” in Proc. INET'98, July 1998. Available on-line at: http://www.isoc.org/inet98/proceedings/6d/6d_3.htm

[Sola-98b] M. Sola, M. Ohta, "Modifications to PIM-SM for Static Multicast,” IETF Internet Draft, August 1998. Available on-line at: http://search.ietf.org/internet-drafts/draft-sola-pim-static-multicast-00.txt

[Sola-98c] M. Sola, M. Ohta, "Modifications to OCBT for Static Multicast," IETF Internet Draft, August 1998. Available on-line at: http://search.ietf.org/internet-drafts/draft-sola-ocbt-static-multicast-00.txt

[Stark-98] T. Stark, "Multicast your fate to the wind," Boardwatch Magazine, May 1998. Available on line at: http://boardwatch.internet.com/mag/98/may/bwm47.html

[Stein-98] B. Stein "268,000 channels and still nothing will be on Internet Multicasting and the multicast backbone," Boardwatch Magazine, April 1997. Available on-line at: http://boardwatch.internet.com/mag/97/aug/bwm32.html

[Talpade-98] R. Talpade, E. Bommaiah, L. Mingyan, A. McAuley, "AMRoute: Adhoc Multicast Routing Protocol," IETF Internet Draft, August 1998. Available on-line at: http://search.ietf.org/internet-drafts/draft-talpade-manet-amroute-00.txt

[Tanenbaum-96] A. Tanenbaum, "Computer Networks," Prentice Hall, 1996

[Tanenbaum2-96] Tanenbaum, Todd "IP Multicasting: Diving through the layers," Network Computing, November 15, 1996.

[Takahashi-80] H. Takahashi and A. Matsuya, "An Approximate Solution for the Steiner Problem in Graphs," Math Japonica 6, pp. 573-577, 1980.

[Tascnets-98] "Reliable Multicast Transport Protocols Comparison" http://www.tascnets.com/mist/doc/mcpCompare.html

[Thaler-97] D. Thaler, C. Ravishankar, "Distributed center location algorithms," IEEE Journal of Selected Areas in Communications, 15(13):291-203, April 1997.

[Thaler-98] D. Thaler, D. Estrin, D. Meyer, "Border Gateway Multicast Protocol (BGMP)," IETF Internet Draft, August 5, 1998. Available on-line at: http://search.ietf.org/internet-drafts/draft-ietf-idmr-gum-03.txt

[Thaler-98b] D. Thaler and B. Aboba, "Multicast Debugging Handbook," IETF Internet Draft, October 14, 1998, Available on-line at: http://search.ietf.org/internet-drafts/draft-ietf-mboned-mdh-01.txt

[Thaler-98c] D. Thaler, "Interoperability Rules for Multicast Routing Protocols," IETF Internet Draft, July 31, 1998, Available on-line at:

http://search.ietf.org/internet-drafts/draft-thaler-multicast-interop-03.txt

[Thyagarajan-95] A.S. Thyagarajan and S. Deering, “Hierarchical Distance-Vector Multicast Routing for the Mbone,” Computer Communications Review (Proc. SIGCOMM ´95), Cambridge, MA, 1995.

[Waitzman-88] D. Waitzman, C. Patridge, and S. Deering "Distance Vector Multicast Routing Protocol,” RFC 1075, November 1988., ftp://ftp.isi.edu/in-notes/rfc1075.txt

[Wall-80] D. Wall, "Mechanisms for Broadcast and Selective Broadcast,” Technical Report 190, Stanford University, June 1980.

[Waxman-88] B.M. Waxman, “Routing of Multipoint Connections,” IEEE Journal in Selected Areas Communication, Vol. 6, pp. 1617-1622, December 1988.

[Waxman-93] B. Waxman, "Performance Evaluation of multipoint routing algorithms,” Proc. IEEE INFOCOM, pages 980-986, 1993.

[Wei-94] L. Wei, D. Estrin, "The tradeoffs of multicast trees and algorithms,” In Proceedings of the 1994 International Conference on Computer Communications and Networks (ICCCN'94), San Francisco, September 1994. Available on-line at: 93-560 USC Technical Report, http://www.usc.edu/dept/cs/tech.html

[Wei-95] L. Wei, D. Estrin, "Multicast routing in dense and sparse modes: simulation study of tradeoffs and dynamics,” Proceedings Fourth International Conference on Computer Communications and Networks (ICCCN'95), Las Vegas, NV, USA; 20-23 Sept. 1995
Available on-line at: 95-613 USC Technical report, http://www.usc.edu/dept/cs/tech.html

[Zappala-97] D. Zappala, D. Estrin, and S. Shenker, "Alternate path routing and pinning for interdomain multicast routing. Technical Report USC CS TR 97-655, U. of Southern California, 1997.

13.1 - OTHER READINGS

A. Ballardie, “Scalable Multicast Key Distribution,” RFC-1949, May 1996.

A. Ballardie, B. Cain, and Z. Zhang "Core Based Tree (CBT) Multicast Border Router Specification," IETF Internet Draft, March 1998. Available on-line at: http://search.ietf.org/internet-drafts/draft-ietf-idmr-cbt-br-spec-02.txt

K. Bharath-Kumar and J.M. Jaffe, "Routing to Multiple Destinations in Computer Networks," IEEE Transaction on Communication. Vol. COM-31. pp. 343-351, March 1983.

S. Bradner, "Internet Protocol Multicast problem statement," IETF Internet Draft, September 1997. http://www.ietf.org/internet-drafts/draft-bradner-multicast-problem-00.txt

R. Braudes, and S. Zabele, "Requirements for Multicast Protocols," RFC 1458, May 1993.

K. Calberg, "Comparison of CBT and PIM via simulation," IDMR Working Group presentation, July 18, 1995. Available on-line at: ftp://cs.ucl.ac.uk/darpa/IDMR/IETF-JUL95/carlberg-slides.tar

R. Chandra, P. Traina, and T. Li, "BGP Communities Attribute," RFC 1997, August 1996.

L. Chapin, "Applicability Statement for OSPF," RFC 1370, October 1992.

E. Chen and T. Bates, "An Application of the BGP Community Attribute in Multi-home Routing," RFC 1998, August 1996.

C. Cheng, I.A. Cimet, and S. Kumar, "A Protocol to Maintain a Minimum Spanning Tree in a Dynamic Topology," in ACM SIGCOMM August 1988, pp. 330-337.

C.H. Chow, "On Multicast Path Finding Algorithms," in IEEE INFOCOM, Bal Harbour, FL, April 1991, pp. 1274-1283.

I.A. Cimet and S. Kumar, " A Resilient Algorithm for Minimum Weight Spanning Trees," in Int. Conf. Parallel Processing St. Charles, August 1987, pp. 196-203.

G. Colombo, C. Scarati, and F. Settimo, "Asynchronous Control Algorithms for Increasing the Efficiency of the Three-stage Connecting Networks for Multipoint Services," IEEE Trans. Commun., Vol. 38, pp. 898-905, June 1990.

D. Comer, "Internetworking with TCP/IP, Volume I, Principles, Protocols and Architecture," Prentice Hall, 1995, Chp. 17, pp. 289 – 302, ISBN 0-13-216987-8

J. Crowcroft, I. Wakeman, M. Handley, S. Clayman and P. White "Internetworking Multimedia," UCL Press, http://www.cs.ucl.ac.uk/staff/J.Crowcroft/mmbook/book/book.html, 1996.

M. Doar and I. Leslie, "How bad is naive multicast routing," Proceedings of IEEE INFOCOM 93, pp. 82-89, April 1993.

J. Halpern and S. Bradner, "RIPv1 Applicability Statement for Historic Status," RFC 1923, March 1996.

M. Handley, "On Scalable Internet Multimedia Conferencing Systems ," Ph.D. Thesis, University of London, November 1997. Available on-line at: http://north.east.isi.edu/~mjh/thesis.ps.gz

S. Herzog, S. Shenker, and D. Estrin, "Sharing the 'Cost' of Multicast Trees: an Axiomatic Analysis," in Proc. ACM SIGCOMM'95, Cambridge, September 1995, pp. 315-327.

M. Hurwicz, "Multicast to the masses," Byte, June 1997, page 93.

R.H. Hwang, "Adaptive Multicast Routing in Single Rate Loss Networks," IEEE INFOCOM, Boston, MA, April 2-6, 1995, pp. 571-578.

V. Jacobson, "Multimedia Conferencing on the Internet," ftp://cs.ucl.ac.uk/darpa/vjtut.ps , August 1994.

V. Johnson and M. Johnson, "IP Multicast Backgrounder," http://www.ipmulticast.com/community/whitepapers/backgrounder.html, Date of search: April 3, 1998.

V. Johnson, and M. Johnson, "How IP multicast works," http://www.ipmulticast.com/community/whitepapers/howipmcworks.html, Date of Search: March 24, 1998

V. Johnson, and M. Johnson, "Introduction to IP Multicast Routing," http://www.ipmulticast.com/community/whitepapers/introrouting.html, Date of Search: March 24, 1998

V. Johnson, and M. Johnson, "IP Multicast Glossary of Terms," http://www.ipmulticast.com/community/whitepapers/glossary.html, Date of Search: March 24, 1998

V. Johnson, and M. Johnson, "IP Multicast Making it happen," Data Communications, May 21, 1997.

V. Johnson, and M. Johnson, "Higher level protocols used with IP Multicast," http://www.ipmulticast.com/community/whitepapers/highprot.html, Date of Search: March 24, 1998

V. Johnson, and M. Johnson, "Implementing IP Multicast in different network infrastructures," http://www.ipmulticast.com/community/whitepapers/netinfra.html, Date of Search: March 24, 1998

V.P. Kompella, J.C. Pasquale, and G.C. Polyzos, "Multicasting for Multimedia Applications," Proceedings of INFOCOM '92, pp. 2078-2085, IEEE Computer Society, 1992.

V.P. Kompella, J.C. Pasquale, and G.C. Polyzos, "Multicasting Routing for Multimedia Communication," IEEE/ACM Trans. Networking, Vol. 1, pp. 286-292, June 1993.

V. Kumar, "MBONE: Interactive Media on the Internet," New Riders, 1996

X. Jiang, "Routing Broadband Multicast Streams," Comput. Commun., Vol. 15, No.1, pp. 45-51, January/February 1992.

M. Macedonia and D. Brutzman, "MBONE provides audio and video across the Internet," Computer, pp. 30-36, April 1994.

G. Malkin, F. Baker, "RIP Version 2 – MIB Extension," RFC 1724, November 1994.

G. Malkin, "RIPng Protocol Applicability Statement," RFC 2081, January 1997.

V. Mallela and M. Shand, "IP Multicast Protocols and Applications," http://www.networks.digital.com/dr/techart/ipmul-mn.html, April 1997.

B. Manning, "Registering New BGP Attribute Types," RFC 2042, January 1997.

J. Mascavage, "Multicasting and Enhanced Broadcast Delivery Service," Andersen Consulting presentation to ITP, March 1998.

D. Marlow, "Host Group Extensions for CLNP Multicasting," RFC 1768, May 1995.

K. Miller, "Multicast Services: The Medium is the message," Data Communications, March 21, 1995.

K. Milne, "Better Data Delivery for the net," Byte, April 1997, page 40.

J. Moy, "Experience with the OSPF protocol," RFC 1246, July 1991.

J. Moy, "OSPF Standardization Report," RFC 2329, April 1998.

S. Nightingale, "Multicast Study," http://snad.ncsl.nist.gov/snad-staff/night/multicast/study.html, September 1, 1995.

L.H. Ngoh, "Multicast support for group communications," Computer Networks and ISDN System 22, p. 165-178, 1991.

B. Quinn, "Internet Multicasting," Dr. Dobb's Journal, October 1997.

B. Rajagopalan and M. Faiman, "A New Responsive Distributed Shortest-Path Routing Algorithm," ACM SIGCOMM '89 symposium, September 1989.

R. Ramanathan, "Multicast Support for Nimrod : Requirements and Solution Approaches," RFC 2102, February 1997.

V.J. Rayward-Smith and A. Clare, "On Finding Steiner Vertices," Networks, Vol. 16, No. 3, pp. 283-294, 1986.

Y. Rekhter, "BGP Protocol Analysis, " RFC 1265, October 1991.

Y. Rekhter and P. Gross, "Application of the Border Gateway Protocol in the Internet, " RFC 1772, March 1995.

K. Savetz, N. Randall and Y. Lepage, "MBONE: Multicast Tomorrow's Internet," IDG Books Worldwide, 1996.

H. Schulzrine, S. Casner, R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real Time Applications," ftp://ftp.isi.edu/in-notes/rfc1889.txt, January 1996.

T. Shafron and P. Loshin, "Multicast offers bandwidth salvation," Byte, March 1998.

W. Stallings, "IPv6: The New Internet Protocol," IEEE Communications Magazine, pp. 96-108, July 1996.

Stardust Forums, Inc. "IETF 38, Memphis TN, Multicast Technologies Report," http://www.ipmulticast.com/community/ietf_38.html, Date of Search: March 26, 1998.

Stardust Forums Inc. "Writing IP Multicast-enabled Applications," http://www.ipmulticast.com/community/whitepapers/ipmcapps.html, Date of Search: March 24, 1998.

Stardust Forums, Inc. "IETF 40, Memphis TN, Multicast Technologies Report," http://www.ipmulticast.com/community/ietf_40.html, December 8, 1997, Date of Search: March 26, 1998.

M. Steenstrup, "IDRP as a Proposed Standard," RFC 1477, July 1993.

R. Stevens, "TCP/IP Illustrated: Volume 1, The Protocols," Addison Wesley Publishing Company Reading, MA, 1994.

R. Talpade, M. Ammar, "Multicast Server Architectures for MARS-based ATM multicasting," RFC 2149, May 1997.

P. Traina, "BGP-4 Protocol Document Roadmap and Implementation Experience," RFC 1656, July 1994.

P. Traina, "Experience with the BGP-4 Protocol," RFC 1773, March 1995.

P. Traina, "BGP-4 Protocol Analysis," RFC 1774, March 1995.

R. Voigt, "Distribution Center Location for Multicast Trees," IDMR Working Group presentation, April 5, 1995. Available on-line at: ftp://cs.ucl.ac.uk/darpa/IDMR/IETF-APR95/voigt-slides.ps

R. Voigt, "A Hierarchical approach to multicast in a datagram internetwork," Ph.D. thesis, Naval Postgraduate School, March 1996. Available on-line at: ftp://ftp.nps.navy.mil at /pub/ece/rjvoigt/CHARM.ps.Z

P. Winter, "Steiner Problem in Networks: A Survey," IEEE Networks, Vol. 17, No. 2, pp. 129-167, 1987.

R.T. Wong, "A Dual Approach for Steiner Tree Problems on a Directed Graph," Mathematic Programming, Vol. 28, pp. 271-287, 1984.

G. Wright and R. Stevens, "TCP/IP Illustrated: Volume 2, The implementation," Addison Wesley Publishing Company, Reading MA, 1995.

G. Xylomeno and G. Polizos, "IP Multicast for Mobile Hosts," IEEE Communications Magazine, January 1997.

Chapter 14 - Appendix

14.1 - Appendix A - OPNET Simulation

Simulations were tried in order to complete the comparison presented in this thesis. This appendix presents the problems encountered while trying to use the OPNET program.

OPNET is a commercial network simulation package developed by MIL3[1]. OPNET allows simulating network topologies, node architectures and process behaviors by using a graphical user interface. The package allows measuring parameters on a network topology such as CPU utilization, end-to-end delay, etc. Figure 14. 1 illustrates the look and feeling of the simulator.

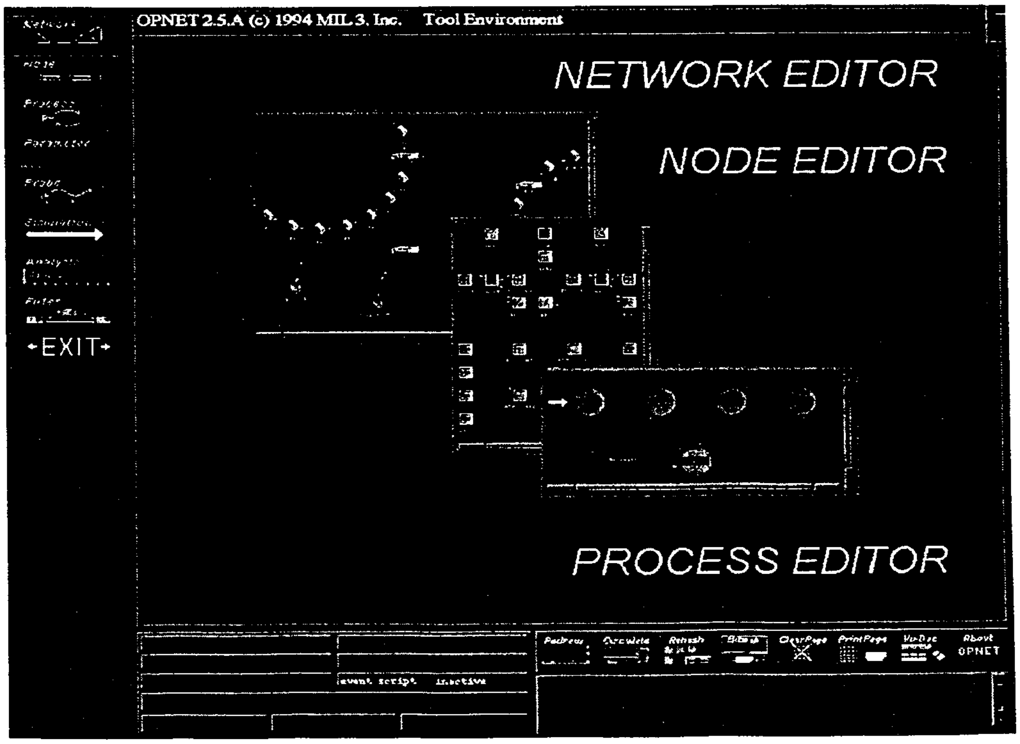


Figure 14. 1 - OPNET simulation environment.

[1] More information on OPNET is available at http://www.mil3.com

OPNET has two issues that make hard to run a simulation:

- Simulation with OPNET requires creating source code in C for the two protocols beign analyzed in this thesis (CBT and PIM-SM). IGMP is not supported in the standard libraries either so it also needed to be modeled. This requires to define a finite state machine for the protocol, and write C code for each state. In essence, the complete protocol needs to be implemented from scratch. However, I was able to find the models developed by Harris Corporation[2], but they were developed in a previous version of OPNET and compatibility version impede to use them[3].
- Simulation times are very slow. For example, a recent simulation performed by a team at George Mason University took 7 hours of real running time for a simulation of 100 seconds of a small network (11 routers and 12 hosts). This simulation was done using a SGI Octane Station with 512 MB main memory and MIPS R10000 CPU [Pullen-98]. The University of Colorado lab does not have this type of machines, so simulations are even slower. It was suggested that one complete week could be spent running a similar simulation on a Sparc 20 workstation [Pullen-96].

[2] The models (implemented in OPNET 3.0) can be obtained by anonymous ftp to the following machine:

taurus-littrow.itd.nrl.navy.mil

The models and documents are in the directory Pub/OpNet.

[3] The command "op_cvmod -all" was used to upgrade the models, but even then the models had problems.

14.2 Appendix B - Useful Web Sites

- **Multicast Resources by StarBurst Communications** http://www.starburstcom.com/patches/mcastres.htm. It contains links to the main resources available on line related to IP Multicast.

- **The IP Multicast Initiative (IPMI)** http://www.ipmulticast.com/. This web site contains several good white papers about IP Multicast that are very good introductory material.

- **The Mbone web site** http://www.mbone.com. It is a general reference for the MBONE. It contains a page with useful references and also has an archive with IP multicast related email lists.

- **Inter-Domain Multicast Routing (IDMR) IETF Working Group** http://www.ietf.org/html.charters/idmr-charter.html The IDMR WG is part of the Routing Area and it is in charge of defining the routing protocols that make multicast a reality.

- **Mbone Deployment (MboneD) IETF Working Group,** http://www.ietf.org/html.charters/mboned-charter.html. This groups is part of the Operations Area and it is in charge of the technical and engineering details of deploying IP multicast to the enterprise.

- **Cisco IOS Software Solutions** http://www.cisco.com/warp/public/732/Multi/index.html. It is a web site that focuses on Multimedia networking. It has links to white papers, vendors, press releases, etc.

- **The Networked Multimedia Connection** http://www.nmc-info.org/ it is an alliance by Cisco, Intel and Microsoft. It is worthwhile to keep an eye to this web site.

- **Internet Technical Resources** http://www.cs.columbia.edu/~hgs/internet/. This is a web site put together by Henning Schulzrine, a professor a Columbia University, which has a wealth of information on Networked Multimedia technologies.

- **IP multicast Cisco Systems early field test (EFT) and beta customers** ftp://ftpeng.cisco.com/ipmulticast.html. This web site has information for Cisco customers that are deploying IP multicast currently.

- **Protocol Independent Multicast-Sparse Mode (PIM-SM)** http://netweb.usc.edu/pim/ it is a web site dedicated to the PIM protocol.

- **IP Multicast deployment guide** http://www/ipmulticast.com/deployment_guide/deployment_guide.html

14.3 - Appendix C - Glossary

BGMP	Border Gateway Multicast Protocol
BGP	Border Gateway Protocol
CBT	Core Based Trees
CIDR	Classless Inter-Domain Routing
DVMRP	Distance Vector Multicast Routing Protocol
IGMP	Internet Group Management Protocol
ISP	Internet Service Provider
MBONE	Multicast Backbone
MOSPF	Multicast Open Shortest Path First
OSPF	Open Shortest Path First
PIM-DM	Protocol Independent Multicast-Dense Mode
PIM-SM	Protocol Independent Multicast-Sparse Mode
RIP	Routing Information Protocol
RP	Rendezvous Point
RSVP	Resource Reservation Protocol
RTP	Real-Time Transport Protocol
RTCP	Real-Time Transport Control Protocol
SDP	Session Directory Protocol
VLSM	Variable Length Subnet Mask