WWW Media Distribution via Hopwise Reliable Multicast[1]

James E. (Jed) Donnelley[2]

Abstract

Repeated access to WWW pages currently makes inefficient use of available network bandwidth. A Distribution Point Model is proposed where large and relatively static sets of pages (e.g. magazines or other such media) are distributed via bulk multicast to LAN distribution points for local access. Some access control issues are discussed.

Hopwise Reliable Multicast (HRM) is proposed to simplify reliable multicast of non real time bulk data between LANs. HRM uses TCP for reliability and flow control on a hop by hop basis throughout a multicast distribution tree created by today's Internet MBone.

Keywords: WWW, multicast, MBone, distributed file system, media

1 Introduction

The World Wide Web (WWW [1]), while still often associated with specific technologies such as server protocols (http, gopher, ftp, news, mail, etc.) and information browsers (Mosaic, Lynx, Netscape, etc.), has become generalized to the point where it is synonymous with on-line access to information. Coupling the point and click simplicity of the user interface in good browsers and their powerful and extendable information presentation capabilities with their access to the world of Internet information make them a good base for

1. This is a draft of a paper submitted to the Third International World-Wide Web Conference in Darmstadt, Germany, April 10-14, 1995. It can be viewed on the WWW at the URL: http://www-atp.llnl.gov/atp/papers/HRM/HRM.html
Postscript versiona are also available from the above location.

2. Lawrence Livermore National Laboratory, Livermore, California, 94550, USA.
Currently on a research leave to the Rechenzentrum, University of Stuttgart, Germany
http://www-atp.llnl.gov/atp/jed.html

just about any information manipulation task.

In the present paper we focus on access to non real time "mass" media publications. These include magazines, their briefer "zine" cousins, newspapers, journals, advertising catalogs, etc. The discussion is also applicable to distribution of many types of relatively static bulk data. Excluded from this discussion are real time media such as radio, television, telephone, etc. and dynamic WWW pages.

Magazines, newspapers, advertising catalogues, and other such media are beginning to appear on the WWW at an amazing rate. The author is currently supporting a set of WWW pages (the Computer and Communications set at: http://www-atp.llnl.gov/atp/comp-comm.html) that provide "One-stop shopping" for information about computers and communication. In this set there is a media page which at the time this paper was written contained URLs to about 50 magazines and newspapers relating to computers and communications.

Many of the publications beginning to appear on the Web should be considered experimental. Even in their current form, however, they are quite useful for those of us that use the Internet as an important source of information. The available searching facilities alone in some cases make Web access to such information considerably more useful than access to the equivalent printed page. To this advantage can be added the ability to reference material outside the article with a click, the ability to copy and paste relevant material, the ability to access such pages anywhere the Internet can be accessed, and others. The unique capabilities of Web access are beginning to incline some of us to read WWW media pages in preference to reading hardcopy, despite some continuing advantages to hardcopy such as portability, mass quick scanning, markability, etc.

As Internet users we must be concerned by the prospect of millions of readers beginning to access media material over the Web. With the current mechanisms for transporting this material, repeated transport of just the graphical page headings alone would constitute a significant additional load for the Internet.

To ease the load on Internet produced by WWW media requests this paper presents two suggestions:

1. The Distribution Point Model - a model where media is copied to distribution points on Local Area Networks and then accessed locally, and

2. Hopwise Reliable Multicast - a multicast mechanism that builds on the current Internet MBone and allows efficient transmission of bulk data from one point to many others.

2 The Distribution Point Model

The Distribution Point Model is analogous to current distribution of print media to distribution centers, libraries, periodical rooms, or even mail rooms. In these cases copies (not a single copy, but we will get to that later) are “staged” to these intermediate locations. They are then forwarded to readers. In the approach considered here, distribution points are set up on Local Area Networks (LANs) that support readers interested in media access. These distribution points must be willing to supply the needed computer resources to service and administer access to media over the local LAN. Media is multicasted to the distribution points and then accessed by readers over the local LAN. LANs are considered to be capable of supporting low enough latency, high enough reliability, and high enough bandwidth to deal effectively with local access to such media.

Before considering the Distribution Point Model in more detail, it is appropriate to consider some alternatives.

2.1 Won’t Caching solve the problem?

Caching has already been implemented at many sites on the Web. Caching can result in a reduced wide area network (WAN) load. Some limitations of caching are:

1. Caching reduces the WAN load for multiple accesses from the same site, but still requires as many WAN accesses as there are requesting cache sites.

2. The first access to a page and any access that results in a cache miss subjects readers to the vagaries of WAN access (higher latencies, service disruptions, etc.).

3. There is currently no effective cache notification for WWW page changes. This means that anyone using a caching service may get an old page from the caching service when a new page is available from the source server. This discourages use of caching services.

2.2 Why not use a Distributed File System?

This approach has technical appeal. Many of the performance and access control problems of using the Web to access files are similar to the problems faced by a distributed file system (DFS). In the author’s opinion, the currently available distributed file systems (e.g. see [2]) come with such a heavy administration and configuration burden, have so little current market penetration, and currently require such awkward name mapping support for WWW access that this approach isn’t ready to significantly help WWW performance. It is worthwhile to begin testing WWW access over distributed file systems to better understand this approach.

If the distributed file system approach does prove successful (whether with today’s systems or with further DFS development) there is still a difficult optimization problem that may benefit from some of the considerations in this paper.

When a file is remotely requested for the first time, a DFS must decide how to get it to the requesting site. One approach is to move it directly through the network from the source to a cache at the requesting site. This approach, while it may be appropriate in some instances, suffers from problems #1 and #2 discussed in the caching section above.

Another approach a DFS might take is to “stage” a requested file through multiple caches between the source and the requesting site. If caching sites are as dense as nodes in a multicast distribution tree, this approach reduces WAN load exactly like multicast distribution (e.g. with HRM). A difference is that with such a distributed caching approach copies of the file must be supported at the intermediate caching nodes for an extended period of time.

To eliminate the need for extended caching at intermediate nodes while still retaining the advantages of reduced WAN load due to multicast distribution, a DFS could use a reliable multicast transport mechanism (whether HRM or some other) to transport files to remote caches for local access. Doing such transport to likely access sites when a file is first created would be equivalent to the Distribution Point Model over multicast as proposed in this paper. Waiting until a file is first accessed and then using multicast distribution to a set of known likely access sites would have the advantage of not transporting any data until it is requested, but would still suffer from problem #2 above (slower first access).

Distributed file systems face a difficult challenge in providing configuration parameters that efficiently deal with the many common patterns of file access. A suggestion from this paper is that they consider multicast distribution to likely access points.

3 Control for Media Access at Distribution Points

Many magazines may be distributed for free as they are today on the Web and in printed form. Some may also be charged for. The issue arises as to how to control access to magazines with restricted access.

Mechanisms are under development to securely identify who is accessing a Web page (e.g. PGP [3] integration into WWW software[4], using the Secure Socket Layer [5] under Web software, etc.). We assume that such mechanisms will exist and can be used by the servers at the distribution points to identify users and restrict access. These identification mechanisms can be used directly with the current magazines on the Web serviced by a single server. In this paper we only consider the additional complications introduced by utilizing multiple distribution points as suggested by the Distribution Point Model.

With distributed access the publisher and the distribution point servers must together manage media access. This requires a distributed database for which a caching or distributed file system would be appropriate. There is little data transfer required so performance is not a concern. A publisher could simply send updates to the distribution points periodically.

As an aside, it seems reasonable to the author to allow some number of accesses to a magazine for free before beginning to charge for it. This sort of “lost leader” is in many ways analogous to the “first issue is free” policy that is often followed today by hardcopy magazines. This would suggest that Web servers should have somewhat more sophisticated access control policies that can be used to count accesses by users and notify them as access control restrictions are about to be (and ultimately are) imposed.

In the remainder of this paper we assume that we are going to use the Distribution Point Model to improve performance and reduce WAN load for WWW page access and focus on the technical problem of getting the bulk data to the distribution points.

4 Multicast

Having decided to move the bulk data to all the distribution points, we are now faced with the task of doing this efficiently. This is where multicast technology can be used effectively. By using multicast (more about the details later) we can move this bulk data to any number of sites around the world with at most one copy of the data passing over any given communication line. In the remainder of this paper we accept this “one time” transmission cost and consider how best to accomplish such a multicast.

4.1 Why not just use IP Multicast?

IP Multicast ([6]) could be used to accomplish the distribution that we seek. To accomplish this distribution, however, we need a reliable transfer. Particularly if we have compressed and perhaps even encrypted data, it won’t do to have pieces of it missing when it gets to the distribution points. There are quite a number of groups working on reliable multicast technologies ([7], [8], [9], [10], [11], [12], [13]). Unfortunately, it is a difficult technology to develop. To understand why and to see how significant the difference is between the needs of “real time” multicast (e.g. audio or video transmissions) and “non real time” bulk multicast, we look at multicast and particularly at the current Internet MBone in a little more detail.

4.2 The Internet MBone

The Internet MBone ([14], [15]- we consider a representative piece in figure 1) consists of hundreds of computers around the world. These computers are workstations or personal computers that are connected to the Internet and are running multicast routing software, the mrouted daemon. Each of the mrouted daemons is configured with information about “tunnels” to neighboring multicast routers. This software creates a virtual multicast network, the MBone, on top of the real Internet. The mrouted software understands the concept of a “multicast address”. It knows how to play its part in building the set of destinations for packets directed to any specific multicast address. Most importantly, it knows how to decide, when it receives an incoming packet, which “tunnels” to send it out over. By directing incoming packets out more than one tunnel, mrouted can effect multicast. By not directing packets to any but the correct lines of a distribution tree for the multicast group, mrouted can insure that packets don’t flow over communication lines (virtual ones = tunnels anyway) more than once.

The Internet MBone currently sends UDP datagrams over this multicast structure. The IP multicast structure does not concern itself with lost packets. Typically packet losses are due to congestion on the network rather than errors on the lines. Dealing with any such losses are left to higher level protocols (e.g. the reliable multicast transport protocols). For real time communication such as audio or video there is not too much one can do about lost packets. If lost packets are retransmitted, they will likely arrive too late to be useful in playout of real time media transmissions.

4.3 Reliable Multicast Transport of Non Real Time Bulk Data

There is a difficult issue to be dealt with by a multicast transport protocol when transmitting non real time bulk data over a multicast datagram network. Consider a host, e.g. node A in figure 1, that has bulk data to multicast. If A has a high speed line to send the data out over (T1), it’s first inclination might be to send the data out as quickly as this line can accept it. If it does so, however, it can easily create a congestion problem at nodes that are then required to receive and forward this data over relatively slow or congested links (e.g. forwarding from B to D or D to E in figure 1).

With real time data this specific form of congestion is not quite so readily generated. If the sender is to be received by everyone listening, it must not send data so quickly that such congestion is created. However, having selected a transmission rate (e.g. media quality, frame rate, compression, or other parameters) the sender cannot send data any faster than the data is generated.

In the bulk data case the sender would prefer to send the data as quickly as possible. In general the sender has no idea how bad the available bandwidth is on the worse case link that will ultimately receive the data. Of course senders can get reports from receivers (e.g. “Session” Packets [16]), but such reporting creates a rather complex and long distance feedback loop with possibly long latencies. Lost packets must be retransmitted to the whole multicast group, thereby requiring bandwidth over many links for a problem that may well be relatively isolated. An additional problem is the “Ack Implosion” problem (many acks from all the receivers arriving at the sender [16]). While programs like “wb,” the currently most popular “white board” tool on the MBone, and others can overcome these problems to various degrees and succeed in multicasting reliably, the focus of this paper is on an alternative approach.

4.4 Multicasting VERY Large “Datagrams” with Hopwise Reliability

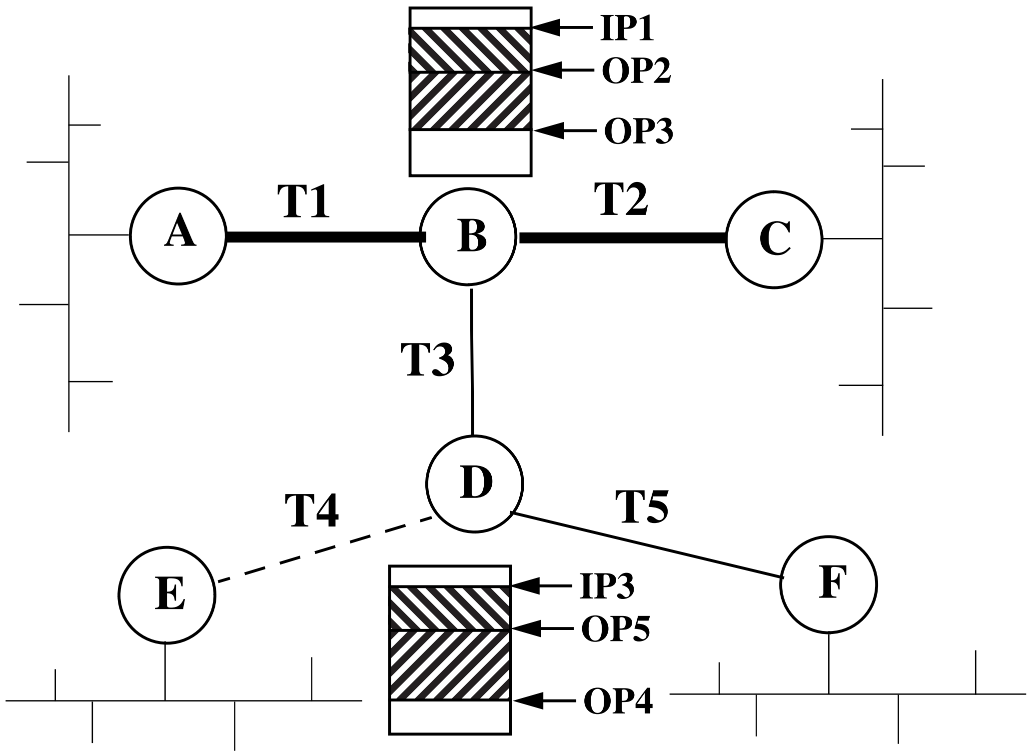

Figure 1: Multicast segment showing hopwise transfer buffers

The Hopwise Reliable Multicast (HRM) approach transmits bulk data of whatever size as a single block as if it were a single datagram for the purposes of multicast.

In more detail, the multicast routing mechanisms of mrouted are used as they operate now. However, to send a multicast "packet" (which we refer to here as a "block"), a TCP connection is opened to each of the links that are to receive the block and the data is transmitted at the best possible rate. Since very large blocks of data are being transmitted, it is important that each node begin to transmit the received data onward to the next appropriate links (tunnels) in the distribution tree before the data is completely received. This "cut through" approach is important both to allow buffer sizes in the nodes to be bounded and to cut down on the latency of the transmission.

With HRM the buffers in the intermediate nodes can be configured to optimize local performance. If the buffers are too small, small TCP transmissions will result in inefficient use of network resources. Beyond the point where buffers are adequate for smooth hopwise TCP transmission, larger buffers effectively allow more buffering of data in the network, thereby offloading the source more quickly and possibly allowing some nodes to receive completed blocks more quickly. Such larger buffers have no effect on the time to complete a multicast, i.e. the time until the last node receives the complete block.

An example of the buffering at intermediate nodes in a HRM transfer is illustrated in figure 1. The buffers are circular with an input pointer (IP<i> in the figure) and as many output pointers (OP<i>) as there are output links in the multicast distribution tree (bounded above by the fanout of the multicast node).

In the example shown, node B has a relatively fast output link T2 and a slower (or currently more congested) output link T3. In this case OP2 would tend to track the IP1, while OP3 would tend to lag behind. When IP1 wraps around and catches OP3, the buffer is full and the TCP window to node A is closed.

Naturally with a scheme of this sort it is important to only send data in blocks large enough for reasonable performance. Until a connection is closed a node always knows to expect more data and so need only transmit data out when its buffers are “reasonably” full.

4.5 HRM Error Considerations

It is suggested that errors in block transmissions with HRM, namely TCP errors such as time-outs, addressing errors, etc. be dealt with by silently terminating the transmission. This is the same as the handling of errors in transmitting UDP packets with IP multicast. Determining what destinations received the block transmission is then left to higher level protocols. In the case of an application that is distributing WWW page blocks to LAN distribution sites, it would seem appropriate to return a receipt notification to the sender. Such a receipt notification, probably best done with TCP, would create an “Ack Implosion” at a higher level. However, such a single set of acks for a large block transfer seems a relatively small cost, even with a rather large number of distribution sites.

If it is determined that one or more sites did not receive a distribution, it would be necessary to retransmit the whole block, either by unicast if there were very few such sites or again by HRM if there were many such sites. In either case communication with the remaining sites must again be established before initiating further distribution.

5 Whither Internet Multicasting?

Hopwise Reliable Multicasting of the type described here would fit nicely on today’s Internet MBone. The MBone of today, however, is imagined to be a temporary research and development expedient on the way to a truly routed multicast Internet. Indeed the transition to utilizing multicast routers is well underway. The Hopwise Reliable Multicast described in this paper can easily be added as an extension to today’s mrouted software, but it is not easy to see how it could appropriately be mapped onto routed Internet multicasting. HRM’s use of TCP and relatively large buffers do not fit into the network level protocols supported by routers.

It would be unfortunate if a technology like HRM caused the life of the current “virtual” Internet MBone to be extended beyond the time when routed IP multicasting is supported. There are some very awkward administrative features of the current MBone, most notably its requirement to “track” the topology of the actual Internet to avoid wasteful retranmissions of data. Perhaps one of the reliable transport protocols currently under development will work out well and come to be widely accepted. In this case such a protocol could be used across a routed Internet multicast facility to provide distribution of WWW media.

A limitation of HRM is that it is not appropriately used over LAN links with hardware supported multicast. In this case use of HRM would unnecessarily replicate data transmission over the LAN. This limitation is not significant for the Distribution Point Model where distribution is only needed to a single LAN site. In the larger context of multicast transport, the approach of HRM could be combined with a reliable “single hop” multicast transport over links where hardware multicasting is supported.

Despite it’s limitations, the author finds the mechanisms of the hop by hop approach to reliable multicast compellingly simple. Both congestion control and error control are significantly simplified by this approach. Perhaps it will be worthwhile to experiment with the HRM approach in parallel with the efforts to develop reliable transport protocols on top of routed multicast datagram services.

6 Comparable Technologies

There are several mechanisms in use in the Internet today that resemble the Distribution Point Model. Three examples are SMTP mail distribution, Usenet news distribution, and the “mirroring” of software and other such relatively static bulk data that is so commonly done on an ad hoc basis. These mechanisms are addressing the same basic need. It is suggested here that these mechanisms be replaced with one based on the maturing technology of multicast addressing and routing. This approach allows new distribution points to be easily added to “groups”, provides a common distribution mechanism for many types of data, and allows the topological configuration to be done once.

There are several groups currently investigating using reliable multicast transport for bulk data distribution. It is suggested that the current MBone with the extension described here (HRM) may be better suited to dealing with the distribution of such non real time bulk data, at least in the short term.

7 Conclusions

A Distribution Point Model is suggested to reduce the network load induced by many accesses to popular but relatively static WWW pages such as the various media being rapidly added to the WWW.

Reliable multicast technology is suggested to further reduce the network load induced by media distribution. Hopwise Reliable Multicast is suggested as an approach that can extend the current Internet MBone to simplify multicast transport in the case of non real time bulk data transport.

8 Acknowledgments

The author would like to thank his colleagues at the Rechenzentrum of the University of Stuttgart (RUS) who provided valuable feedback on the development of the ideas in this paper, particularly Claus-Dieter Schulz who provided the initial “seed” idea for media distribution. Discussions of buffering and flow control strategies for HRM were considerably refined in discussions with Peter Haas and Andreas Rozek from RUS, Mark Boolootian from LLNL, and Jon Knight from Loughborough University of Technology, UK.

“Work performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under contract number W-7405-ENG-48.”

9 References

[1] Berners-Lee, T.J., Cailliau, R., Groff, J-F, Pollermann, B.,
"World-Wide Web: The Information Universe,"
Electronic Networking: Research, Applications and Policy, Vol. 2 No. 1, pp. 52-58, Spring 1992, Meckler Publishing, Westport, CT, USA.
Preprint: http://info.cern.ch/pub/www/doc/ENRAP_9202.ps
copy: http://www-atp.llnl.gov/atp/papers/HRM/references/www-info-univ.ps

[2] Spasojevic, M., Bowman, M., Spector, A.,
"Using a Wide-Area File System Within the World-Wide Web,"
copy: http://www-atp.llnl.gov/atp/papers/HRM/references/dfs-www.ps

[3] PGP (tm) Pretty Good (tm) Privacy, Phil Zimmerman,
Users Guide: Vol. 1, Vol. 2, and Internal Data Structures
From Phils's Pretty Good Software,
http://www.pegasus.esprit.ec.org/people/arne/pgp.html
copy: http://www-atp.llnl.gov/atp/papers/HRM/references/pgp-refs.html

[4] NetMarket,
"PGP Help,"
Describes how to integrate PGP with NCSA XMosaic-2.4,
http://www.netmarket.com/pgp/bin/help
copy: http://www-atp.llnl.gov/atp/papers/HRM/references/netm-pgp.html

[5] Hickman, K.,
"The SSL Protocol",
Draft RFC submitted to the W3O working group on security,
http://mosaic.mcom.com/info/SSL.html
copy: http://www-atp.llnl.gov/atp/papers/HRM/references/ssl.html

[6] Deering, S.,
"Host Extensions for IP Multicasting,"
Network Working Group Internet RFC-1112, August 1989.
ftp://nic.ddn.mil/rfc/rfc1112.txt
copy: http://www-atp.llnl.gov/atp/papers/HRM/references/ip-multi-rfc.txt

[7] Armstrong, S., Freier, A. and Marzullo, K.,
"Multicast Transport Protocol,"
DARPA RFC 1301, February 1992,
ftp://src.doc.ic.ac.uk/rfc/rfc1301.txt
copy: http://www-atp.llnl.gov/atp/papers/HRM/references/multi-trans-rfc.txt

[8] Bormann, C., Ou, J., Gehrcke, H-C, Kerschat, T., and Seifert, N.,
"MTP-2: Towards Achieving the S.E.R.O. Properties for Multicast Transport,"
ICCCN '94, San Francisco, September 1994,
http://hill.lut.ac.uk/DS-Archive/sero.ps
copy: http://www-atp.llnl.gov/atp/papers/HRM/references/sero.ps

[9] Cheung, S. Y., Ammar, M. H.,
"Using Destination Set Grouping to Improve the Performance of Window-

controlled Multipoint Connections,"
GIT, College of Computing, Tech Report GIT-CC-94-32, August 1994,
ftp://ftp.cc.gatech.edu//pub/coc/tech_reports/1994/GIT-CC-94-32.ps.Z
copy: http://www-atp.llnl.gov/atp/papers/HRM/references/multi-conn.ps

[10] Crowcroft, J., and Paliwoda, K.,
"A Multicast Transport Protocol,"
ACM SIGCOMM 1988: Commun. Arch. Protocols, 18(4),
August 1988, p. 247-256.

[11] Dempsey, B. J., Strayer, T. and Weaver, A.,
"Issues in Providing a Reliable Multicast Facility,"
University of Virginia Department of Computer Science Technical Report
CS-89-15, November 1989.

[12] Kaplan, S. and Montgomery, T.,
"A High Performance Totally Ordered Multicast Protocol,"
submitted to INFOCOMM 95,
ftp://research.ivv.nasa.gov/pub/doc/RMP/info_s2.ps
copy: http://www-atp.llnl.gov/atp/papers/HRM/references/highp-multi.ps

[13] Maffeis, S., Bischofberger, W., and Maetzel, K.,
"GTS: A Generic Multicast Transport Service,"
University of Zurich Technical Report UBILAB 96.6.1, June 1994,
ftp://ftp.ifi.unizh.ch/pub/techreports/gts.ps.Z
copy: http://www-atp.llnl.gov/atp/papers/HRM/references/gts.ps

[14] Eriksson, H.,
"MBone: The Multicast Backbone,"
Communications of the ACM, August 1994, Vol.37, pp.54-60.

[15] Macedonia, M. R., Brutzman, D. P.,
"MBone Provides Audio and Video Across the Internet,"
IEEE Computer, Vol. 27 #4, April 1994, pp. 30-36.
ftp://taurus.cs.nps.navy.mil/pub/mbmg/mbone.html
copy: http://www-atp.llnl.gov/atp/papers/HRM/references/mbone-av.html

[16] Jacobson, V.
"Multimedia Conferencing on the Internet,"
Tutorial 4, SIGCOMM '94, University College, London, U.K.