Group communication seems to have re-opened the door to designing a set of protocols from the ground up, and the requirement for re-designing them from the services down. This represents an exciting challenge in communications research today.

6 Acknowledgments

The authors would like to thank first Mostafa Ammar for helping improve the paper and gaving direction for better integration in this special issue of JSAC. We would like to acknowledge useful comments also from Ken Carlberg, Tony Ballardie and Mark Handley.

- Multicast routing schemes may mean that it is more complex to determine which links are in use (or control this) than it is for the equivalent set of point-to-point communication. This further exacerbates the last point.

Group Security; Referenced papers

[Stubblebine93] S. Stubblebine. Security Services for Multimedia Conferencing. Proceedings of the 16th National Computer Security Conference. Baltimore, Ma. September 1993.

[Ballardie95] A. Ballardie and J. Crowcroft. Multicast-Specific Security Threats and Counter-Measures. In ISOC Symposium on Network and Distributed System Security, February 1995.

Group Security; Other readings

[Gong94] L. Gong and N. Shacham. Elements of Trusted Multicasting. International Conference on Network Protocols (ICNP '94). Boston. October 1994.

[Gong95] L. Gong and N. Shacham. Multicast Security and its Extension to a Mobile Environment. Wireless Networks, J.C. Baltzer Pub. 1995.

[Reiter92] M. K. Reiter, K. P. Birman http://www-mice.cs.ucl.ac.uk/mice/rat/and L. Gong. Integrating Security in a Group Oriented Distributed System. Research Report (TR92-1269). Cornell University. 1992.

[Tseung90] L. C. N. Tseung and K. C. Yu. Guaranteed, Reliable, Secure Broadcast networks. 9th Intl. Phoenix Conference on Computer and Communications. 1990.

5 Conclusion

Multipoint communication is a rich area of work, past, current, and future. In this paper, we have surveyed the whole range of communication functions and services, and presented the state of the art. It is clear that multipoint is an extremely valuable service for many new applications, but that a number of outstanding problems exist that require future research:

- The "receiver makes good" approach to reliability, congestion control, resource reservation and so on, has proved extremely powerful. This architecture has not reached the level of maturity that sender based schemes have for point-to-point applications and protocols.
- Many new applications and environments can tolerate loss, delay and out-of-sequence delivery, and group communication protocols should exploit this to the advantage (typically for performance gains). This needs further work.
- The Mbone has demonstrated that large group communication is possible, but it is not clear how well this will scale to very large groups, or numbers of groups, for example replacing cable TV distribution with millions receivers, or for new applications like multi-player games and simulations with thousands of groups changing very rapidly.
- Very fast group dynamics may place a stress on current membership and routing and addressing protocols. The DIS, with thousands of group members change per second, may cause the same problems for current multicast schemes as high signaling rates would to a current telephone exchange. Group aggregation schemes may be needed.
- Congestion Control for group communication clearly needs further research, both for network switch/router based techniques, and for end-to-end schemes. One area of promise is that of Active Networking, which, combined with distributed merge/branch points, multiple multicast group usage, and distributed intelligent filtering offers further scaling and flexibility.
- Fairness for multipoint traffic, whether on its own or combined with unicast traffic, seems to be a relatively under-researched area
- Billing is also a real problem that will be more complex to solve for multipoint than for point-to-point.

IEEE Infocom '96. San Francisco, pp. 232-239, April 1996.

[Clark95] R. Clark and M. H. Ammar. Providing Scalable Web Services Using multicast. Proceedings of the 2nd Intl. workshop on Services in Distributed and Networked Environments. June 1995.

[Crowcroft95] J. Crowcroft. Internet Videoconferencing. To be published in Video-Mediated Communication. S. Wilbur, A. Sellen, and K. Finn Editors. Published by Lawrence Erlbaum & Associates. 1995.

[Frederick92] R. Frederick, NV - X11 video-conferencing tool, Unix Manual Page, XEROX PARC, 1992.

[McCanne95] S. McCanne and V. Jacobson. Vic: A Flexible Framework for Packet Video. Proceedings of ACM Multimedia. November 1995.

[Rat96] Rat web server: http://www-mice.cs.ucl.ac.uk/mice/rat/.

[Turletti94] T. Turletti and J. C. Bolot. Issues with Multicast Video Distribution in Heterogeneous Packet networks. 6th International Workshop on Packet Video. Portland, Oregon. pp. F 301-304. Sept 26-27, 1994.

[Vat96] Vat web server: http://www-nrg.ee.lbl.gov/vat/.

[Wb96] White board software available through ftp://ftp.ee.lbl.gov/conferencing/wb/.

Internet Multimedia Applications; Other readings

[Aguilar86] L. Aguilar, J. Garcia-Luna-Aceves, D. Moran, E. Craighill, R. Brungardt, Architecture for a Multimedia Teleconferencing System, ACM 1986 Symposium on Communications Architectures and Protocols, Stowe,Vermont,pp 126-135, Aug. 1986.

[Almeroth96] K. Almeroth and M. Ammar. The Use of Multicast Communication to Provide Scalable and Interactive Video-on-Demand Service. IEEE JSAC Special Issue on Video Services to the home. 1996.

[Bolot94] J. Bolot, T. Turletti, and I. Wakeman. Scalable Feedback Control for Multicast Video distribution in the Internet. ACM SIGCOM'94. London. August 1994.

[Glicksman93] J. Glicksman and V. Kumar. A SHAREd Collaborative Environment for Mechanical Engineers. Proceedings of Groupware '93. pp. 335-347. 1993.

[Schulzrinne96] H. Schulzrinne. Dynamic Configuration of Conferencing Applications using Pattern-Matching Multicast. ACM/Springer Multimedia Systems Journal. January 1996.

4.3 Security

As stated in the introduction, traditional point-to-point communication is largely asymmetric. However, form the point of view of security, its requirements are also symmetric, in that secret key cryptography is an appropriate building block for authentication and privacy services.

However, group communication introduces some different problems[stubblebine93], and places different emphasis on some areas:

- An efficient key distribution scheme is needed for multiple recipients - this tends (as with electronic mail) to mean that asymmetric key systems such as Public Key Cryptography are more useful, especially for large groups [Ballardie95].
- The very fact that a group wishes to communicate rather than two single entities means that there are more opportunities for traffic analyses, denial of service and covert signaling type attacks. This is not a feature of multipoint, or multicast, though: it is inherent to the users' requirement!

Report (TR90-1141). Cornell University. 1990.

[Garcia91] H. Garcia-Molina, A. Spauster. Ordered and reliable Multicast Communication. ACM Transactions on Computer Systems. Vol. 9, No. 3, pp. 242-271. August 1991.

[Kalantar95] M. Kalantar. Issues in Ordered Multicast Performance: A Simulation Study. Research Report (TR95-1531). Cornell University. 1995.

[Peterson]: Larry L. Peterson, Nick C. Buchholz, Richard D. SchlichtingPreserving and using Context Information in Interprocess Communication, ACM Transactions on Computer Systems, vol 7, No. 3 pp 217-246, August 1989

[VanRenesse94] R. Van Renesse. Why Bother with CATOCS? Appeared in Operating Systems Review. 28 (1). 22-27. January, 1994.

[Verissimo89] P. Verissimo, L. Rodrigues, M. Baptista, AMp: A highly parallel atomic multicast protocol. ACM SIGCOMM. September 89.

4.2 Multimedia application software on the Internet

In this section, we briefly describe some multicast applications and discuss their requirements from the multicast viewpoint. The main design principle, common to all these applications, is that end-to-end control is provided at the application level according to ALF [Clark90], and not at a general purpose multicast transport protocol.

4.2.1. Audio and Video Conferencing

Many audio and video conferencing application have been developed since 1992, among them vic [McCanne95], vat [Vat96], nv [Frederick92], ivs [Turletti94], rat [Rat96], freephone [Bolot96]. These application do not require any reliability or transport level ordering: audio packets are reordered in the application play-out buffer. An application level congestion control is generally implemented (as described in section 4.4). Layered coding schemes are proposed to provide receiver controlled quality. On the other hand, the intensive use of these application on the Mbone showed the need for sparse mode multicast routing.

4.2.2. Shared Workspace

Wb (white board) [Wb96] is the most well-known shared workspace application on the Mbone. The communication system supporting wb provides reliable but not ordered multicast: messages are delivered as soon as they are received and application level recovery is performed if an out-of-order packet is received. Any participant having a copy of a requested data may retransmit it, according to the SRM model described above. A simplified version of NTP is used to estimate the distance to the other sites. Nt [Crowcroft95], a shared text editor, is another example of shared workspace application.

4.2.3. Session Directory

Sd is not a group application by itself, but it provides the possibility to perform multicast address allocation. Numbers are chosen randomly. If the number of already allocated addresses is lower than the square root of address space size, then the probability of collision is very small. For IPv4, about 2^{28} multicast addresses are available, therefore the limit is 16K. More room may be made either by taking the scope of the address into account or by expanding the address space for regions with already allocated addresses. This requires that the end of session be indicated explicitly.

4.2.4 Web Browser

Web Browsers are not multipoint applications. But multicasting could be seen as a way to optimize data distribution for a Web server. This direction is being investigated by [Clark95].

Internet Multimedia Applications; Referenced papers

[Bolot96] J-C. Bolot and A. Vega Garcia. Control mechanisms for packet audio in the Internet. Proc. of

The (numerous) duplicates are discarded. Some optimizations were then proposed in order to increase the protocol efficiency. The first optimization is to not piggyback messages to a site if they have already been sent to that site or after an ACK is received. Another optimization consists in piggybacking only on messages going directly to the destination sites, other sites are sent a descriptor. [VanRenesse94]

Transport level ordered and reliable multicast systems are the subject of controversy. In fact, guaranteeing order properties implies the system provides for atomic message delivery. This is implemented by having each station buffer each message is receives until it is stable, i.e. received by all other members of the group. However, most of the systems providing causally and totally ordered communication do not provide durable atomic message delivery: action changes do not survive failures. This is due to the transport-application levels boundary. A host may crash before messages received by the transport layer are delivered to the application. Cheriton and Skeen [Cheriton93] presented several limitations of these systems: the main idea is that ordering is an application problem, which is easier to control at the application level. Therefore, a transport level re-ordering does not simplify the application design. In addition, with transport protocol level ordering it is not possible to handle higher-level error conditions. Another drawback of transport level ordering is the increase of the response time due to data delivery delay. In addition, some applications do not require packet re-ordering at the transport level. Out of sequence packets can be processed by the application. In fact, several applications may be designed to allow this out of sequence packets processing according to the "Application Level Framing" concept, resulting in enhanced performance on heterogeneous networks.

Ordering and Synchronization; Referenced papers

[Birman87]K. Birman, T. Joseph, Reliable Communication in the Presence of Failures, ACM Transactions on Computer Systems, Vol.5, No1, February 1987, pp 47-76.

[Birman91] K. Birman, A. Schiper, P. Stephenson. Lightweight Causal and Atomic Group Multicast, ACM Transactions on Computer Systems, Vol.9, No 3, Aug. 1991, pp 272-314.

[Chang84] J. Chang and N. Maxemchuk. Reliable Broadcast Protocols. ACM Transactions on Computer Systems. Vol.2. No. 3. pp. 251-275. August 1984.

[Cheriton93] D. R. Cheriton and D. Skeen. Understanding the Limitations of Causally and Totally Ordered Communication. 14th Symposium on Operating System Principles. ACM. December 1993.

[Lamport78] L. Lamport. Time, Clocks, and the Ordering of Events in a Distributed System. Communication of the ACM 21, No. 7. pp. 558-565. July 1978.

[VanRenesse95] R. Van Renesse, K. P. Birman, B. B. Glade, K. Guo, M. Hayden, T. Hickey, D. Malki, A. Vaysburd and W. Vogels. Horus: A Flexible Group Communications System. Research Report (TR95-1500). Cornell University. 1995.

Ordering and synchronization; Other readings

[Aiello93] R. Aiello, E. Pagani, and G. P. Rossi. Causal Ordering in Reliable Group Communication. ACM SIGCOMM '93. September 1993.

[Agrawal94] D. Agrawal, P. Melliar-Smith, L. Moser. Reliable ordered delivery across interconnected Local-Area Networks.

[Birman93] K. Birman. A Response to Cheriton and Skeen’s Criticism of Causal and Totally Ordered Communication. Technical Report 93-1390. October, 1993.

[Cheriton89] D. R. Cheriton, C. L. Williamson, VMTP as the Transport Layer for High-Performance Distributed Systems, IEEE Communications Magazine, June 1989.

[Cooper94] R. Cooper. Experience with Causally and Totally Ordered Communication Support. Appeared in Operating Systems Review. 28, 1. 28-32. January, 1994.

[Freier90] A. O. Freier and K. Marzullo. MTP: An Atomic Multicast Transport Protocol. Research

[Ammar93] M. H. Ammar. Probabilistic Multicast: Generalizing the Multicast Paradigm to Improve Scalability. Proceedings of INFOCOM'94.

[Hoffmann96] M. Hoffmann. A Generic Concept for Large-Scale Multicast. Proceedings of International Zurich Seminar on Digital Communication (IZS'96), Springer Verlag, February 1996.

[Jones91] M. G. W. Jones, S. A Sorensen, and S. R. Wilbur. Protocol Design for Large Group Multicasting: The Message Distribution Protocol. Computer Communication. Vol. 14, No. 5, 1991.

[Rajagopalan92] B. Rajagopalan. Reliability and Scaling Issues in Multicast Communication. Proceedings of ACM SIGCOM 1992. pp. 188-198.

[Srinivasan95] S. Srinivasan and B. de Supinski. Multicasting in DIS: A unified solution. Research Report (CS-95-17). University of Virginia. 1995.

[Yavatkar93] R. Yavatkar and L. Manoj. Optimistic Strategies for Large-scale Dissemination of Multimedia Information. ACM Multimedia '93.

4 Applications

In this section, we analyze problems that are directly related to the application. Before describing some group applications that have been experimented with within the Internet, we will first discuss the problems of ordering and synchronization which are inherent to most multimedia and time constrained applications. This section will end with a short discussion of security in group applications.

4.1 Ordering and Synchronization

Packets may arrive out of sequence at their destination due to packet losses or changing datagram routing. For many distributed applications, an ordered reception of packets is required, because any misordering may give different view of the state of the group. Packets re-ordering is then necessary. Current solutions to solve this problem use either local or global sequence numbers and are based on a central sequencer or use token techniques [Chang84] [Birman87]. The packet order may be causal or total. With causal ordering, packets are delivered in accordance with the "happens before" relationship among the sending events defined in [Lamport78]. An event e_1 is said to "happen before" event e_2 if e_1 and e_2 occurs in the same process and e_1 "occurs" before e_2 or if e_1 is the sending of a message and e_2 is the reception of the same message. A message *a* is said to causally precede a message *b* (*a* -> *b*) if the sending of *a* happens before the sending of *b*. This causal order is not total: two messages (*a* and *c*) may be concurrent: i.e. both *a* -> *c* and *c* -> *a* are false. Total ordering means that the receivers will see all data units from all senders in the same order (even "concurrent" data units).

A protocol that ensures, in addition to reliability, a total ordering of the delivered messages is called an atomic protocol. Such a protocol may be used for reliable validation, atomic operations, group membership, etc., while a causal protocol (ensuring causal and not total ordering) may be used for ensuring consistency in updates to replicated data.

Birman proposed a distributed solution based on a two-phase protocol in [Birman87]. Each site maintains a pending queue, and when a new message is received it is sorted in the queue and marked as undeliverable. A proposed time-stamp is then sent back to the initiator, which collects the time-stamps and sends back the largest one. All sites assign this final time-stamp to the message and mark it as deliverable. The queue is then re-ordered in order of increasing time-stamps, and deliverable messages at the head of the queue are passed to the application. CBCAST [Birman91] implements causal ordering only. This reduces the latency problems of total ordering protocols, while still providing useful service for some distributed applications. Causal ordering was first implemented in an inefficient version described as follows. When a message *m* is sent from process P to Q, a (so-called piggyback) copy of all undelivered messages preceding *m* is also sent. Messages may be delivered as soon as they are received. In fact, when *m* arrives, copies of messages preceding *m* arrives with *m* or has arrived earlier.

1993.

[Ferrari92] D. Ferrari, J. Ramaekers, G. Ventre. Client-Network Interactions in Quality of Service Communication Environments. Proceedings of the fourth IFIP International Conference on High Performance Networking (HPN '92). Liège. December 1992.

[Ferrari95] D. Ferrari, A. Gupta, G. Ventre. Distributed Advanced Reservation of Real-time Connections. Proceedings of the 5th International workshop on Network and Operating System Support for Digital Audio and Video. Durham, NH. April 1995.

[Pingali94] S. Pingali, D. Towsley, and J. Kurose. A Comparison of Sender-Initiated and Receiver-Initiated Reliable Multicast Protocols. ACM SIGMETRICS. 1994.

[Szyperski93] C. Szyperski and G. Ventre. Efficient Group Communication with Guaranteed Quality of Service. Proceedings of the Fourth IEEE Workshop on Future Trends in Distributed Computing Systems, Lisboa, Portugal, September 1993.

3.3 Large groups and scalability

Distributed Interactive Simulation (DIS) [Ieee93] and distributed games may be one of the most stressful applications requiring multicast support: simulation networks with up to 100,000 dynamic entities are being developed. The multicast requirements of these applications are for a large number of multicast groups, and for a high number of group changes per second. A log-based receiver reliable protocol for distributed simulation is proposed in [Holbrook95].

Collaborative simulation systems, distributed Virtual Reality, games and multi-user environments can often generate quite low data rates, but require low latency, due to the high degree of interactivity (possibly incorporating interfaces to the real world), more so even than audio conferencing. This has meant that dense mode multicast distribution is typically appropriate. Simulation systems are part way between data systems, and multimedia ones, in that the information that is distributed is often that of object locations, ready reckoning or trajectory information, which is partly loss tolerant. It is also possible to adjust the rate of sending over quite a large range in some applications, and so congestion avoidance techniques are applicable.

The use of large numbers of groups stems from the DIS design, and is not necessarily the optimal way to proceed, but does provide a very elegant architecture. Essentially, objects in the DIS are in an hierarchy of groups, and updates on their location are multicast. Participants in the distributed simulation use the world view to decide which objects are visible or not, and so long as their world view is not changing too quickly, then join (and leave) groups appropriately, so that they receive information concerning the relevant (visible) objects. For a fast moving viewer (e.g. virtual plane pilot) it may be possible to ameliorate the rate of change of group membership by reducing the accuracy of view (and even creating new groups to which only a subset of updates for object location information is sent, in a similar fashion to that used for layered video and quality adjustment).

Nevertheless, the quantity and rate of change of group membership information for such systems is very large - it has lead to proposals for hierarchical group addressing (and routing) schemes in an attempt to reduce the amount of state stored per group or per source per group, but as yet, this is still research.

Large Group; Referenced papers

[Ieee93] IEEE Standard 1279-1993 Distributed Interactive Simulation Standard. For more information about DIS see http://www.sc.ist.ucf.edu/. Modeling, Simulation and Training Service Center.

[Holbrook95] H. W. Holbrook, S. K. Singhal, D. R. Cheriton. Log-Based Receiver-Reliable Multicast for Distributed Interactive Simulation. SIGCOMM ’95. Boston. August 1995.

Large Group; Other readings

[Towsley85] D. Towsley. On the Statistical Analysis of Point-to-Multipoint Channel Using Go-back-N Error Control protocol. IEEE Transactions on Communications. Vol. 33. No. 3. 1985.

[Yavatkar95] R. Yavatkar, J. Griffioen and M. Sudan. A Reliable Dissemination Protocol for Interactive Collaborative Applications. ACM Multimedia '95.

[Whetten95] B. Whetten, S. Kaplan, T. Montgomery. A High Performance Totally Ordered Multicast Protocol. Theory and Practice in Distributed Systems, Springer Verlag LCNS 938. 1995.

3.2 Receiver based transmission control and QoS

The problem of QoS is not yet completely solved even for unicast communication. Multicast applications may provide the motivation and the opportunity to resolve this for a broader environment. QoS is a typical example of a function where a unicast solution can be totally non-adaptive. Such QoS models cannot usually be applied to group communication. Each participant has its own constraints that are not necessarily acceptable to the other participants: group communication whose characteristics were based on consensus is not credible. There are three types of QoS that could be used in group communication:

- QoS is defined by the sender (no negotiation). If a potential member cannot accept this QoS, it cannot join the group.
- QoS is negotiated over the group members to be the minimum of each member's QoS.
- The senders send with the higher QoS and each receivers controls its own QoS. In that case, support from the network is usually required.

In case of receiver controlled QoS, the following facilities should be provided by the network:

- filtering at the node level would simplify route management, and allow bandwidth reservation,
- use hierarchical encoding at the source in order to allow each receiver to adapt the accepted information to its own capacity, and
- forget QoS negotiation on a point-to-point basis.

A guarantee usually implies resource reservation is used; best-effort implies adaptability. In a multicast session, each of the member has different characteristics. The consequence is group communication will be efficient if the communication is tailored to each participant. Resource reservation is complex to setup in a multipoint environment. QoS constrained multicast route design is a NP-complete problem. So what will be the efficiency of the routing algorithm if available resource has to be considered. Moreover, in the same session, the resource requirement may not be the same for all the participants.

Adaptability and reservation have to cohabit in a multimedia communication system (it has been proved that use of pure adaptation creates instability in the network). However, because of the nature of group communication, adaptability is highly desirable to tailor communication to the receiver requirements. The guarantee of minimum bandwidth on multicast route is not a problem that can be solved simply [Borella95, Zhang93].

Receiver based control and QoS; Referenced papers

[Borella95] M. Borella and B. Mukherjee. A Reservation-Based Multicasting Protocol for WDM Local Lightwave Networks. ICC 95. pp 1277-1281. June 1995.

[Zhang93] L. Zhang, S. Deering, D. Estrin, S. Shenker, D. Zappala. RSVP: A New Resource ReSerVation Protocol. IEEE Networks. September 1993.

Receiver based control and QoS; Other readings

[Cheung96] S. Y. Cheung, M. H. Ammar and X. Li. On the use of Destination Set Grouping to Improve fairness in Multicast Video distribution. Proceedings of INFOCOM 96. San Francisco. March 1996.

[Effelsberg93] W. Effelsburg and E. Mueller-Menrad. Dynamic Join and Leave for Real-Time Multicast. Technical Report TR-93-056, International Computer Science Institute, Berkeley, CA, October

[Strayer95] T. Strayer. XTP Home Page (includes link to XTP 4.0 spec). http://www.ca.sandia.gov/xtp/xtp.html. 1995.

End to End; Other readings

[Aiello93] R. Aiello, E. Pagani, and G. P. Rossi. Design of a Reliable Multicast Protocol. Proceedings of IEEE INFOCOM. pp. 75-81. 1993.

[Ammar92] M. H. Ammar and L. R. Wu. Improving the Performance of Point to Multi-Point ARQ Protocols through Destination Set Splitting. Proceedings of IEEE INFOCOM '92, Florence. Italy, May 1992, pp 262-271.

[Armstrong92] S. Armstrong, A. Freier, and K. Marzullo. Multicast Transport Protocol. RFC 1301. 1992.

[Birman87] K. Birman, T. Joseph. Reliable Communication in the Presence of Failures, ACM Transactions on Computer Systems, Vol.5, No1, February 1987, pp 47-76.

[Birman93] K. Birman. The Process Group Approach to Reliable Distributed Computing. Communications of the ACM. December 1993.

[Crowcroft88] J. Crowcroft and K. Paliwoda. A Multicast Transport protocol. Proceedings of ACM SIGCOMM. pp. 247-256. 1988.

[Erramilli87] A. Erramilli, R. P. Singh. A reliable and efficient multicast protocol for broadband broadcast networks, SIGCOMM'87 Workshop, ACM, August 1987.

[Grossglauser96] M. Grossglauser. Optimal Deterministic Timeouts for Reliable Scalable Multicast. Proc. IEEE Infocom '96. San Francisco. March 1996.

[Holbrook95] H. W. Holbrook, S. K. Singhal, D. R. Cheriton. Log-Based Receiver-Reliable Multicast for Distributed Interactive Simulation. SIGCOMM '95. Boston. August 1995.

[Mockapetris83] P. V. Mockapetris, Analysis of reliable Multicast Algorithms for Local networks. ACM SIGCOMM. 1983.

[Mohan88] S. Mohan, J. Quian, N. L. Rao. Efficient Point-To-Point and Point-To-Multipoint Selective-Repeat ARQ scheme with Multiple Retransmissions: A Throughput Analysis. ACM SIGCOMM 1988.

[Paliwoda88] K. Paliwoda, Transactions Involving Multicast. Computer Communication. No.11. 1988.

[Paul96] S. Paul, K. Sabnani, D. M. Kristol. Multicast Transport Protocol for High Speed networks. JSAC special issue on group communication. 1996.

[Ramakrishnan88] S. Ramakrishnan, and B. Jain. An Unbounded Protocol for Point to Multipoint Communication. 7th Intl. Phoenix Conference on Computer and Communications. 1988.

[Rezende96] Jose F. de Rezende, Andreas Mauthe, Serge Fdida and David Hutchison. Fully Reliable Multicast in Heterogeneous Environments. 5th IFIP workshop on protocols for High Speed Networks. Chapman&Hall editor. Sophia Antipolis (France). October 1996.

[Sabnani85] K. Sabnani and M. Schwartz. Multidestination Protocols for Satellite Broadcast Channels. IEEE Transactions on Communications. Vol. 33. 1985.

[Sabnani86] K. Sabnani and M. Schwartz. A New Connection Establishment Procedure for Multidestination Protocols. IEEE Transactions on Communications. Vol. 34. 1986.

[Segall83] A. Segall, B. Awerbuch. A Reliable Broadcast protocol. IEEE Transactions on Communications. Vol. 31. 1983.

[Talpade95] R. Talpade and M. H. Ammar. Single Connection Emulation: An Architecture for Reliable Multicast Transport Service. Proceedings of the IEEE Distributed Systems. June 1995.

the current time-stamp have been received by all destinations. The local buffer containing a copy of these messages may then be freed. If a message or an ACK is lost, a unicast NACK is sent to the presumed token site. The message is then retransmitted in unicast. The transport layer at each destination delivers the messages to the application according to the ACKs time-stamp. This delivery (or commit) may be delayed until the token rotates K times. This ensures K-resilient fault tolerance: up to K sites may crash without violating the atomicity property for the remaining sites. As the delivery to the application may be delayed if no new messages are transmitted, the token site sends a Null ACK if no messages are received for a period of time. This token site change reduces the "commit" delay but increases the number of messages transmitted on the network. Note that broadcasting the message to the token site, passing the token, and sending a NACK are point to point reliable operations. If a site does not receive response to a reliable operation, a reformation protocol is run and a new token list is constructed.

3.1.2. XTP

XTP provides a statistical reliable multicast data transfer [Chesson91][Strayer95]. In the 3.6 version, a specific algorithm, the bucket algorithm, is proposed to provide reliable multicast for a destination group. According to this algorithm, the retransmission strategy is based on buckets that collect acknowledgments. Senders regularly receive a status request from destinations. A bucket contains the information relative to an epoch (e.g. between two request for status emitted by the sender). The "oldest" bucket reflects the "best" view of the group status. However, there is no constraint on receiving all the messages from all the sources before delivering the information from the oldest bucket to the application. The content is in fact delivered when all buckets are full, and this bucket is used as the "newest" bucket. This allows a sender to "discard" slow receivers without degrading the performance of other "active" participants. The number of buckets defines a trade-off between response time and reliability.

Heuristics were defined in order to enhance the performance of transmission control. "Slotting" consists in forcing receivers to spread their reports in order to avoid an implosion of reports in case a problem (e.g. a packet loss) occurs. Receivers also apply "damping" i.e. they avoid sending their report if it is useless (e.g. a packet loss already declared by another receiver). Note that packets and reports are all sent in multicast to the group address. XTP v4.0 [Strayer95]proposes to use cloning in order to allow concentration (or many-to-one communication). N-by-M communication is supported using cloning and transport bridging.

3.1.3. SRM

A Scalable Reliable Multicast framework (SRM) is proposed in [Floyd95]. This protocol (designed to work over IP-Multicast), which is based on NACKs, guarantees reliable delivery of packets with no control on the order of sequence (application level mechanisms may be added to enforce a particular order). The most important issue is therefore the performance of the transmission control. The authors propose to use the slotting and damping techniques originally proposed by XTP. In order to reduce the response time, each host estimates the delay from every other sender. Closer hosts will choose a smaller randomization interval than distant hosts for both NACK and retransmission timers. A CSMA-like algorithm is used to avoid NACK and retransmission implosion. This architecture was tested in a well known Mbone application, wb.

End to End; Referenced papers

[Chang84b] J. Chang and N. Maxemchuk. Reliable Broadcast Protocols. ACM Transactions on Computer Systems. Vol.2. No. 3. pp. 251-275. August 1984.

[Chesson91] G. Chesson. The Evolution of XTP. Proceedings of IFIP Intl. Conference on High Speed networks. North Holland Ed. Berlin. 1991.

[Floyd95] S. Floyd, V. Jacobson, S. McCanne, C. G. Liu, and L. Zhang. A Reliable Framework for Light-Weight Sessions and Application Level Framing. ACM SIGCOMM '95. Boston. August 30-September 1, 1995.

cols. Proceedings of ACM SIGCOMM '90. September 1990. pp. 201-208.

[McCanne96] Receiver Driven Layered Multicast, Proc of Sigcomm 96, Stanford, September 1996.

[Turletti94] T. Turletti and J. C. Bolot. Issues with Multicast Video Distribution in Heterogeneous Packet networks. 6th International Workshop on Packet Video. Portland, Oregon. pp. F 301-304. Sept 26-27, 1994.

[Schulzrinne96] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson. RTP: A Transport Protocol for Real-Time Applications, RFC 1889, IETF, January 1996.

Traffic Control; Other readings

[Guerney95] Guerney et al. Multicast Flow Control on Local Area Networks. Research Report (TR95-1479). Cornell University. 1995.

[Shacham95] N. Shacham. Preemption-Based Admission Control in Multimedia Multiparty Communications. Proceedings of IEEE INFOCOM, Boston. April 1995.

3 End-to-End

End-to-end related tasks are those that are processed at end systems only (i.e. ignored by switching and routing nodes). This is usually the case for ordering, synchronization, and reliability. We have decided not to address ordering and synchronization in this section because some researchers propose providing these functions at the application level. However, more advanced topics such as Quality of Service and scalability will be discussed in this section.

3.1 Reliability vs. unreliability

Point to point transport protocols generally use positive acknowledgments (ACKs) sent by the destination to the source, in order to guarantee reliability. Extending this approach to multicast transmission means that the message is (re-)sent until ACKs from "all destinations" are received. This approach does not scale well because each destination has to send an ACK for each received packet (or group of packets). This may lead to a network congestion at the source level and/or a source overload due to the synchronization of ACK emission. Using a negative acknowledgment mechanism (NACK) with a semantic of a retransmission request is better suited for multicast transmission. This mechanism shifts the error control load from the source to the destination. A station transmit packets without waiting for ACKs. Error detection is performed by destinations using the packets sequence numbers. The philosophy is to avoid sending state messages (e.g. ACKs) when everything is normal, so to improve the protocol efficiency (i.e. the ratio of the number of "useful" data packets emitted over the total number of transmitted packets (including the retransmissions)).

Many reliable transport protocols have been developed. Most of these protocols satisfy the properties of: (1) atomicity: either all or none of the destinations receive and validate a multicast message and (2) termination: the result of the message multicast in known in finite time. We will now describe some classical mechanisms proposed by multicast protocols to guarantee a reliable transmission.

3.1.1. A Reliable Broadcast Protocol

Chang and Maxemchuk proposed a token ring based protocol in [Chang84]. This reliable broadcast protocol combines the ACK and NACK based approaches. All messages are stamped with a pair (host number, local sequence number), then broadcasted. They are handled by a primary receiver, the token site, which serializes the messages from all the senders. When a message is received at the token site, a time-stamp is added and an ACK is broadcast to the group in order to inform the sender that the token site has received the packet. The token rotates in order to balance the ACKs load. The new token site is required to have all time-stamped messages before accepting the token. This also solves an important problem concerning the reliability: after N rotations, messages with time-stamp at least N smaller than

-end.

As multicast routing has only recently been deployed in the Internet, and multicast switched circuits are rarely implemented in ATM or ISDN networks yet, it is too early for us to see congestion control schemes for multicast traffic within the network, although some early research has been done in this area. A key unsolved problem for congestion control schemes that operate within the network for multicast traffic is how to retain the capability for heterogeneity. Other problems include: defining fairness, and relative fairness between unicast and multicast traffic; timescales for congestion control; scaling of control traffic and techniques.

However, end-to-end congestion control is a mature topic. For example, since 1988, the TCP scheme and approach has been studied in its original and many modified forms in many papers. The same approach can be deployed in a limited fashion for group communication. In [Turletti94] an algorithm for multicast congestion control is proposed. Some design issues are discussed by the authors. First, a scalable feedback mechanism should be designed. Congestion detection should be performed by the receivers and not by the sources. Implosion is avoided by combining probabilistic query/reply schemes, random delay responses and expanding scope search. The goal of this scalable feedback mechanism is to keep the number of control packets a fixed proportion of data packets. Second, the feedback signal should be adapted to the payload type in order to optimize the "utility" of the delivered information. Third the question of who to satisfy should be resolved. If the source throughput is adapted to the slowest destination, fast destination may be unhappy. Otherwise, if some slow destination are "discarded" the links to these destinations may be overloaded. A possible solution for this problem is to use hierarchical coding (e.g. for video) and send only the most important informations towards congested destination, and all informations to the uncongested destinations.

A more general solution was proposed in [Schulzrinne96]. In this model, All sites quasi-periodically multicast session packets containing their identity, reception reports, packet loss, inter-arrival delay variation, and synchronization information. All other receivers hear the reports. Adaptive senders use these feedback informations. The detailed description of this mechanism is in the RTP protocol specification [Schulzrinne96]. RTP is a deliberately incomplete protocol framework following the Application Layer Framing (ALF[Clark90]) concept. It is malleable to provide information required by a particular application. This technique is also used to construct the Scalable Reliable Multicast protocol described in the next section. It is based on the observation that if the unit of recovery is the same as the unit of application semantics, functions from different protocol layers may be combined into single passes over the data.

End-to-end feedback signals for congestion control have limited effectiveness. The two approaches above constrain the feedback control traffic to a fixed percentage of the overall data transfer traffic; as the number of recipients grows, they sample the various distribution trees bottleneck links with higher probability, but less frequently. This means that they provide increasingly less timely information about short term conditions on links.

Receiver Driven Layered Multicast (RLM [McCanne96]) uses a different approach to this to provide a more rapid reaction to network conditions. Here, it assumed that the data stream can be divided into multiple layers of differing quality, and that receivers can subscribe to different distribution groups for the different layers. On detecting loss (through gaps in the send sequence number space), receivers dynamically adjust the number of groups they are subscribed to, using the typical "exponential backoff, linear increase" control algorithm typical of all the end-to-end schemes above, and TCP.

Finally, a feedback scheme proposed for multipoint traffic in ATM networks involves a mixture of end-to-end and network based congestion control - in the ATM ABR service, RM cells convey explicit rate feedback information to sources; on a multipoint call, these feedback messages are accumulated, and the worst case rate is the one returned to the source.

Traffic Control; Referenced papers

[Clark90] D. Clark and D. Tennenhouse. Architectural Considerations for a New Generation of Proto-

address is mapped to the ARP server's ATM address. A host wishing to join a group register its join in the MCS. The MCS retransmits it to all hosts by sending IGMP reports. Only active group senders process the reports sent by the MCS. The distribution tree may then be updated. A new sender requests the address list from the MCS before sending its data. However several issues are left open in the IP Multicast over ATM proposal [Armitage95]. Among them we cite the support of layer 3 broadcast and unicast as a special case of multicasting. In addition, the support of ATM group addresses and leaf initiated Join in the ATM Forum's UNI specification has not yet been addressed in the proposal.

LAN Emulation

The ATM Forum LAN Emulation (LANE) subworking group issued a document describing the specification of LAN Emulation Over ATM [Atmf95b]. In this document, the components of an Emulated LAN are described. A LE Client (LEC) is an entity that performs data forwarding, address resolution and other control functions in order to provide a IEEE 802.3 or IEEE 802.5 service interface to higher layers. A LE Server (LES) provides the facility for registering and resolving the MAC addresses and/or route descriptors to ATM addresses. All broadcast, multicast and unknown traffic to and from an LE Client passes through a single entity called the Broadcast and Unknown Server (BUS). Therefore, both the LES and BUS have multicast VCs with all registered LECs. This solution suffers like the MCS based approach from the single congestion point.

ATM Multicast; Referenced papers

[Armitage95] G. Armitage. IP Multicast over ATM Networks. JSAC special issue on multicast. 1996.

[Atmf95a] ATM User Network Interface (UNI) Specification Version 3.1. Prenctice Hall. 1995.

[Atmf95b] LAN Emulation Over ATM. Version 1.0 ATM Forum document af-lane-0021.000. January 1995.

[Cole95] R. G. Cole, D. H. Shur, and C. Villamizar. IP over ATM: A Framework Document. Internet Draft. draft-ietf-ipatm-framework-doc-03.txt. June 1995.

[LeBoudec] E.Gauthier, J-Y Le Boudec, Ph. Oechslin, "Many-to-Many ATM Multicast", Technical Report no. 96/168, March 1996, EPFL

[Mpoa95] Baseline text for MPOA. ATM Forum document atmf 95-0824. July 1995.

[Laubach94] M. Laubach. Classical IP and ARP over ATM. RFC1577. IETF. January 1994.

ATM Multicast; Other readings

[Bala93] K. Bala, K. Petropoulos and T. E.Stern. Multicasting in a Linear Lightwave Network. Proceedings of the IEEE Infocom '93, Vol 3. San Francisco, April, pp. 1350-1358.

[Sethi95] A. S. Sethi. A Model for Virtual Tree Bandwidth Allocation in ATM Networks. IEEE INFOCOM. Boston. April 1995.

[Grossglauser96] M. Grossglauser and K. K. Ramakrishnan. SEAM: Scalable and Efficient ATM Multipoint-to-Multipoint Communication [Extended Abstract]. Proc. 6h Intl. Workshop on Network and Operating System Support (NOSSDAV '96). Zushi. April 1996.

2.3 Traffic control

Group communication is by definition more greedy in bandwidth than point to point communication[1]. The design of a multicast congestion control algorithm is then an important and useful task. There are two potential approaches for congestion control: within the network (we might call this hop-by-hop-by hop, since it involved distribution trees rather than simple paths as the unicast case would), and end-to-

1. Of course, a set of users who wish to communicate may be less greedy if they use multicast than if they use multiple unicast communication facilities.

- The multicast VC mesh. In the multicast VC mesh, a point-to-multipoint VC originates from each sender to all members of the multicast group. If a member joins or leaves the group, the VC needs to be updated. In addition the ATM interface must terminate one VC for each active source in the cluster.
- In the multicast server (MCS) model, a server is chosen within each cluster. Each source establishes a point to point VC to the multicast server. The MCS establishes a point-to-multipoint VC to the desired destinations. During the data transmission phase, the MCS reassembles messages arriving on all incoming VCs and queues them for transmission on the VC. The side-effect is that some interfaces will receive "reflected" messages: sources that are also group members will get copies of their own cells as the multicast server sends the same informations to all group members; the message has a source identifier inserted in the AAL frame so that the source has the choice to drop the reflected cell(s).

Both VC meshes and multicast servers have advantages and drawbacks. If we focus on throughput, the VC mesh solution might be preferable as it lacks the traffic concentration point introduced by the MCS. Data transmission delays are likely to be lower in the VC mesh approach as the message reassembly at the MCS is avoided. However, the MCS approach is more adapted to dynamic sets of receivers because it has a more efficient group membership control. Concerning resource consumption, the MCS server is also better: only two VCs are needed per ATM interface compared to one VC termination per source in the VC mesh solution. Both VC meshes and MC use source based distribution trees.

A third model for multicast ATM called SMART has been devised by le Boudec [LeBoudec]. It entails the use of a multipoint-to-multipoint ATM VC. This introduces the problem for protocols that use messages larger than a single ATM cell size, that a receiver may have to re-sequence interleaved cells from different sources. One solution for this is to use an ATM Adaptation layer that includes a source multiplexing identifier such as AAL 3/4's MID. This introduces a significant packet header overhead, and in any case, AAL5 has largely been accepted as the appropriate framing protocol for data protocols such as IP and CLNP. In contrast, SMART uses a shared tree approach together with an access protocol based on the use of special RM cells to determine which source may send when. For sessions with large numbers of active sources and/or relatively small ADUs, multiple multipoint VCs can be used (the trade-off being somewhere between those of the MARS and MCS service approaches).

Another important issue which is not resolved today is the group membership notification. A group membership notification mechanism should allow the multicast router or center to receive join request from group members, to maintain the mapping between the layer 3 group address and group members ATM address list and to provide sufficient information for senders to set up their VC.

MARS

One of the interesting problems is how to map a high level group address to a point-to-multipoint VC. A solution based on a extension of the ATM ARP server [Laubach94] is proposed in [Armitage95]. A so-called Multicast Address Resolution Server (MARS) keeps extended tables of mapping between layer 3 group address and a list of ATM interfaces representing group members. A single MARS supports one cluster. A point-to-multipoint VC is maintained between the MARS and all ATM hosts desiring multicast support, in order to provide asynchronous group membership changes notification. Two MARS classes are defined: Class I allowing VC meshes to support layer 3 multicast traffic, and Class II allowing both VC meshes and MCS to be assigned for use on a per-group basis; this choice is at configuration time, and transparent to the MARS client.

IP multicast over ATM

Several solutions were proposed for IP over ATM [Cole95]. IP Multicast support is more or less difficult according to the chosen solution. The most advanced solution with regard the multicast support is the Classical IP Model [Laubach94]. The proposal for IP multicasting over UNI 3.1 based ATM networks is detailed in [Armitage95]. According to this proposal, the subnetwork ARP server (possibly implemented as a separate unit) acts as a multicast server (MCS). All end-systems maintain a VC with the MCS, and the MCS maintains a point-to-multipoint VC to all hosts. In addition, the all-hosts group

ously designed using a latitude/longitude based location). Other problems, which are typical to reliable multicast in mobile environments, have been identified by Acharya [Acharya93, Acharya95]:

- When the source is a mobile host, then a copy of the packet may not reach all group members if using source based routing.
- A mobile host may experience significant delays when it enters a new cell. It is easier if there is already a group member attached to that cell but if it is not the case, the delay may be high and have side effect on the transmission reliability. There also could be a time lap before connection to a new cell.
- TTL can limit reachability of cells. A mobile host can be in a situation where it was connected to a cell and while moving, it loses the connectivity because the next cell cannot be reached.

Mobile Multicast; Referenced papers

[Acharya93] A. Acharya and B. R. Badrinath. Delivering multicast messages in networks with mobile hosts. 13th International Conference on Distributed Computing Systems.Pittsburgh. 1993.

[Acharya95] A. Acharya, A. Bakre, and B. R. Badrinath. IP Multicast Extension for Mobile Networking. Rutgers DCS TechReport LCSR-TR-243. 1995.

[Cho94] K. Cho and K. Birman. A Group Communication Approach for Mobile Computing. Workshop on Mobile Computing Systems and Applications. Santa Cruz. 1994.

[Duchamp92] D. Duchamp. Issues in Wireless Mobile Computing. Third Workshop on Workstation Operating Systems. pp. 2-10. EEE Computer Society Press. Key Biscayne.1992.

Mobile Multicast; Other readings

[Bhagwat93] P. Bhagwat and C. E. Perkins. A Mobile Networking System Based on Internet Protocol ({IP}). USENIX Symposium on Mobile and Location Independent Computing. pp 69-82. Cambridge (MA). 1993.

[Caceres94] R. Caceres and L. Iftode. The Effect of Mobility on Reliable Transport protocols. 14th International Conference on Distributed Computer Systems. Poznan. 1994.

[Corson95] M. S. Corson and S. G. Batsell. A Reservation Based Multicast (RBM) Routing Protocol for Mobile Networks: Overview of Initial Route Construction. IEEE INFOCOM. Boston. April 1995.

[Dolev95] S. Dolev, D. K. Pradhan, J. L. Welch. Modified Structure for Location Management in Mobile Environments. IEEE INFOCOM. Boston. April 1995.

[Ioannidis93] J. Ioannidis and G. Q. Maguire. The Design and Implementation of a Mobile Internetworking Architecture. USENIX Winter 1993 Conference.

[Johnson94] D. B. Johnson. Scalable and Robust Interwork Routing for Mobile Hosts. 14th International Conference on Distributed Computer Systems. Poznan. 1994.

2.2.2.3 ATM architectures

The notion of point to multipoint Virtual Circuits was introduced in the UNI [Atmf95a] for audio conference purposes. A point-to-multipoint VC is a very narrow approach that doesn't provide an efficient solution to the requirements of most group applications in terms of flexibility and scalability. In the current standards for ATM, the multicast group address abstraction does not exist. The sender should be aware of all the members of the multicast group. Only the VC root node may add or remove leaf nodes. There is no receiver controlled group membership in UNI 3.1[Atmf95]. Multicast is supported through point to multipoint VCs. Multicast capable ATM interfaces are grouped into clusters. A cluster is a set of ATM interfaces able and willing to achieve AAL level multicasting. The introduction of this hierarchy divides the problem of multicast support in two parts: how to achieve intra-cluster and inter cluster multicasting. Two models were proposed for intra-cluster multicast in the baseline text for the ATM Forum MPOA (Multi Protocol Over ATM) sub-working group [Mpoa95].

[Deering94] S. Deering et al. An Architecture for Wide Area Multicast Routing. ACM SIGCOMM '94. London. August 1994.

[Deering 95] S. Deering, D. Estrin, D. Farinacci, V. Jacobson, C. G. Liu, L. Wei, Protocol Independant Multicast (PIM): Sparse Mode Protocol Specification. Internet Draft. March 1994.

[Moy91] J. Moy. OSFP version 2, Internet RFC 1247, 189 p., July 1991.

[Thyagarajan95] A. Thyagarajan, and S. Deering. Hierarchical Distance-Vector Multicast Routing for the MBone. Proceedings of ACM SIGCOMM'95. September 1995.

[Wall80] Mechanisms for Broadcast and Selective Broadcast, David W. Wall, PhD Thesis, Stanford University, Dept of Electrical Engineering, 1980.

Internet Multicast Routing Protocols; Other readings

[Aguilar84] L. Aguilar. Datagram Routing for Internet Multicasting. ACM SIGCOMM. 1984.

[Casner93] S. Casner. Frequently Asked Questions (FAQ) on the Multicast Backbone. File://venera.isi.edu/mbone/faq.txt. May 1993.

[Deering84] S. Deering, D. Cheriton, Host Groups: A Multicast Extension for Datagram Internetworks. ACM SIGCOMM. 1984.

[Deering89] S. Deering, Host Extensions for IP Multicasting. RFC 1112. 17. August 1989.

[Deering90] S. Deering, D. Cheriton, Multicast Routing in Datagram Internetworks and Extended LANs, ACM Transactions on Computer Systems, Vol.8, No 2, May 1990, pp 85-110.

[Eriksson94] H. Erikson. MBONE: The Multicast Backbone. Communication of the ACM. pp. 54-60. August 1994.

[Mah94] B. Mah. Measurements and Observations of IP Multicast Traffic. Technical Report UCB/CSD-94-858, University of California, Berkeley, CA, December 1994.

[Moy94] J. Moy. Multicast Routing Extensions for OSPF. Communication of the ACM. pp. 61-66. August 1994.

[Shacham92] N. Shacham. Multicast Routing of Hierarchical Data. Proceedings of ICC'92. Chicago, Il. June 1992.

2.2.2.2 Mobile hosts

Activities related to multicast and mobile networking are really in the early stages of development [Duchamps92, Cho94, Achayra93]. Mobile IP using PIM and RSVP is currently being researched. Two algorithms are used in mobile IP. When a mobile unit wants to send a packet, it uses a classic RPF based approach from the Mobile Support Router it is connected to. When a mobile unit has to receive a multicast packet, this packet is sent to the wired address of the mobile unit and then forwarded from this address to the mobile unit using a special tunnel[1].

There are two reasons to this approach. First, in the classic IP environment, it is expected that all the nodes connected to the same subnetwork, or LAN, are physically connected, or that if one of them receives a packet, all the nodes of the subnetwork will see this packet. This is not the case for mobile units that can move from a network to another. Second is the problem of correspondence between the address and identifier in the Internet environment. For mobile purpose, the name must be different from the address and independent from the wired location (such an address could be easily and unambigu-

1. Deering notes the possible synergy between multicast and mobility: multicast provides for a "level of indirection" through the use of logical addresses, and group membership is dynamic; a mobile host appears very like a multicast group of one single member, joining and leaving at different sites. However, the IP mobile approaches to date are not based on this type of mechanism at all, as we can see.

This technique would allow for incremental deployment of the hierarchical multicast routing[1]. This has not seen deployment, as work on successor protocols to DVMRP has taken priority.

MOSPF

The protocols described above are mainly based on the extension of a distance vector routing protocol. MOSPF [Moy91] is a multicast routing protocol based on OSPF V2, which takes advantage of the Link State data base. This allows routers to build efficient "source based trees" or a "shortest-path tree" without even flooding the first datagram of a group transmission. In addition, routers may use the TTL to immediately discard multicast datagrams that will never reach the receiver(s). Link efficiency is therefore higher than with DVMRP. MOSPF requires heavy computation for each source-group combination. MOSPF implementations carry out the computation on demand i.e. only when the first packet from a source S to a group G is received. There are, however, some concerns that while this solution works well for slight load: it becomes expensive in terms of CPU load when a large number of sources (start to) send to large numbers of groups.

PIM

The previous protocols are not suitable for use over the global Internet: they scale too poorly. In order to extend multicast support to wide area groups which may be either dense or sparse, to compliment DVMRP and CBT, the IETF Inter Domain Multicast Routing working group defined PIM (Protocol Independent Multicast) [Deering95]. PIM has two operating modes: sparse (SM) and dense (DM).

- A group is said to be dense if group membership is plentiful within a region of an internet. PIM-DM implements RPF and prunes. Dense mode PIM is essentially the same as DVMRP, except that the unicast routes are imported from existing unicast tables, rather than incorporating a unicast routing algorithm in the specification and implementation directly. This is why it is termed "Protocol Independent".
- A group is said to be sparse if group membership is spread out thinly across regions of an internet. In PIM Sparse Mode, explicit join messages are sent to Rendezvous Points (RPs) to meet new sources. It is based on the shared tree algorithm. These trees do not optimize the delay with respect to sources. However, it has been proven in [Wall80] that the maximum delay is bounded by two times the shortest path delay if the center location is optimal. The highest IP addressed router is chosen as Designated Router (DR) on a multiaccess network. The receiver's designated routers (DR) send join messages to the RP. The sender's DR sends register messages to the RP which send a join to sources. Data packets will follow the established RP-rooted shared tree. However, the receiver (or router nearest the receiver) may decide to switch to the source's shortest path tree. Even in this case, the source continue to send its data to the RP for other possible receivers. The possibility to switch from a RP-rooted shared tree to a source based tree is the main difference between PIM-SM and the Core Based Tree algorithm. However, several open issues still need to be addressed, including the interoperation DM/SM, the RP selection, the criteria for switching between a shared tree and a shortest path tree, the interaction with policy-based and QOS routing and the interaction with receiver-initiated reservation (such as RSVP). These issues are discussed in detail in [Deering94].

Routing protocols can be characterized by the state required at nodes in their distribution trees. In CBT and sparse-mode PIM, typically, we keep state per group. In dense mode distribution trees, nodes must keep per source, per group information.

Internet Multicast Routing Protocols; Referenced papers

[Deering88] S. Deering, C. Partridge, and D. Waitzman. Distance Vector Multicast Routing Protocol. RFC 1075. November 1988.

[Deering91] S. Deering. Multicast Routing in a Datagram Internetwork. PhD thesis. Stanford University. December 1991.

1. This encapsulation is merely another manifestation of the "tunneling" technique used to build the Mbone.

pp. 45-51. January-February 1992.

[Kompella93] V. P. Kompella et al. Multicast Routing for Multimedia Communication. IEEE/ACM Transactions on Networking. Vol. 1. No. 3. pp. 286-292. June 1993.

[Tode92] H. Tode, Y.Sakai, M.Yamamoto, H.Okada, Y.Tezuka, Multicast Routing Algorithm for Nodal Load Balancing, IEEE Infocom92, Florence, Italy, May 1992.

2.2.2. Multicast Routing Protocols

2.2.2.1 Internet

The first, and still predominant multicast routing protocol used in the Internet is based on the model established by Deering in his Ph.D. thesis [Deering91]. Before the IP multicast extensions, the Internet could be considered as a set of interconnected subnetworks with local multicast support. These subnetworks are either multicast capable LANs or point to point links, or switched networks. The IP multicast service model allowed then to provide Internet multicast support. This service model is based on an underlying unreliable *datagram* service. In the Internet service model, datagrams are delivered with "best effort" reliability to the group members. The forwarding of multicast datagrams between "islands" of multicast capable subnetworks is handled by "multicast routers" through tunnels[1]. Datagram delivery is done according to the truncated broadcast model: packets are forwarded on all non leaf "child" subnetworks in the tree and on all leaf "child" subnetworks where there are group members[2].

This multicast overlay on the Internet is called the MBone (Multicast backBONE). It is an "experimental" virtual network operating since 1992, which is becoming part of the operational infrastructure of the Internet at the time of writing. In July 1995, there were more than 2500 connected networks, 12000 routine users and 500 Kbps available bandwidth (bandwidth "reservation" is done by manual scheduling of usage of applications by the community of users). The Mbone is based on UDP for end-to-end transmission control, IGMP for group mangement, DVMRP for routing.

DVMRP (Distance-Vector Multicast Routing Protocol) [Deering88], is an extension of RIP. Multicast routers exchange reverse path distances in order to build the (source based) delivery tree for each group. Once the delivery tree is built, RPF is used to decide whether a packet should be forwarded on not to a router i.e. the datagram is forwarded if the receiving interface is on the shortest path to the source. Superfluous datagram copies are avoided by looking "one step further" i.e. by looking if the router is the next hop on its child's attached router shortest path to the source[3]. The "scope" of multicast delivery is limited by forwarding the datagrams if their TTL is higher than a given threshold defined at the tunnel set up. This forwarding technique was enhanced by the support of on demand "pruning" of tree branches not leading to group members.

Hierarchical DVMRP

The rapid growth of the MBone necessitated the revision of the DVMRP in order to introduce hierarchy. In [Thyagarajan95], the authors propose organizing the MBone in "regions" having address-independent identifiers. A two level hierarchy proposed: intra-region multicast where routers may run any protocol and inter-region multicast where boundary routers run DVMRP. Packets are tagged with region identifiers and boundary routers exchange routing information using region identifiers. The encapsulation header is stripped off in final destination region and local multicast routing is applied.

---

1. Tunnels are implemented by encapsulating IP packets destined for a multicast address within an IP packet with the unicast address of the next multicast capable router along the path. Once, they were implemented using Loose Source Routing, but this transpired to have poor implementations in many routers, and introduced high cost packet processing, or low priority forwarding compared with other router tasks such as route update handling.

2. We use the term leaf and child in the usual sense, where the root of a tree is the parent, the leafs are the ultimate descendants, and so forth.

3. A shared-access link between two routers might have replica packets delivered on to it, without the extra rules that the router with shortest path to source, or lowest address will win, and the other router desist from forwarding.

1993.

[Herzog95] S. Herzog, S. Shenker, D. Estrin. Sharing the "Cost" of Multicast Trees: An Axiomatic Analysis. ACM SIGCOMM. 1995.

[Hwang92] F. K. Hwang and D. S. Richards. Steiner Tree problems. Networks. Vol. 22. No. 1. pp. 55-89. January 1992.

[Kou81] L. Kou, G. Markowshy, and L. Berman. A Fast Algorithm for Steiner Trees. Acta Informatica 15. pp. 141-145. 1981.

[Noronha94] C. A. Noronha, F. A. Tobagi. Optimum Routing of Multicast Streams. IEEE INFOCOM '94. Toronto. June 1994.

[Perlman92] R. Perlman. Interconnection. Bridges and Routers. Addison-Wesley Professional Computing Series. 1992.

[Rayward86] V. J. Rayward-Smith and A. Clare. On finding Steiner Vertices. Networks. Vol. 16. No. 3. pp. 283-294. Fall 1986.

[Takahashi80] H. Takahashi and A. Matsuyama. An Approximate Solution for the Steiner Problem in Graphs. Math. Japonica 6, pp. 573-577. 1980.

[Thyagarajan95] A. Thyagarajan, and S. Deering. Hierarchical Distance-Vector Multicast Routing for the MBone. Proceedings of ACM SIGCOMM'95. September 1995.

[Waitzman88] D. Waitzman, C. Partridge, S. Deering, Distance Vector Multicast Routing Protocol. RFC 1075. IETF. November 1988.

[Waxman88] B. M. Waxman. Routing of Multipoint Connections. IEEE Journal on Selected Area in Communications. Vol. 6. No. 9. pp. 1617-1622. December 1988.

[Wei95] L. Wei and D. Estrin. The Trade-offs of Multicast Trees and algorithms. Internet Draft. draft-ietf-idmr-mtree-00.txt. March 1995.

[Winter87] P. Winter. Steiner problem in Networks: A Survey. Networks. Vol. 17, No. 2, pp. 129-167. 1987

[Zhu95] Q. Zhu, M. Parsa, and J. J. Garcia-Luna-Aceves. A Source-Based Algorithm for Delay-Constrained Minimum-Cost Multicasting. IEEE INFOCOM. Boston. April 1995.

Multicast Routing; Other readings

[Ammar93] M. H. Ammar, S. Y. Cheung, and C. M. Scoglio. Routing Multipoint Connections Using Virtual Paths in an ATM Network. IEEE INFOCOM. San Francisco. pp. 98-105. 1993.

[Bharath83] 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.

[Cheng88] C. Cheng, I. A. Cimet, S. P. R. Kumar. A Protocol to Maintain a Minimum Spanning Tree in a Dynamic Topology. ACM SIGCOMM. 1988.

[Chow91] C. H. Chow. On Multicast Path Finding Algorithms. IEEE INFOCOM. Bal Harbour. pp. 1274-1283. April 1991.

[Colombo90] G. Colombo, C. Scarati, and F. Settimo. Asynchronous Control Algorithms for Increasing the Efficiency of the Three-Stage Connecting Networks for Multipoint Services. IEEE Transactions on Communications. Vol. 38, No. 6. June 1990.

[Hwang95] R. H. Hwang. Adaptive Multicast Routing in Single Rate Loss Networks. IEEE INFOCOM. Boston. April 1995.

[Jiang92] X. Jiang. Routing Broadband Multicast Streams. Computer Communications. Vol. 15. No. 1.

Designing a center based tree is like designing a spanning tree per group. It has the advantage over RPF of only requiring a state information per group instead of a pair of information per group and source. The centered approach does, however, suffer from traffic concentration, as the traffic from all sources of a given group will converge to the center.

The principal advantage of CBT is to limit the expansion of multicast transmission to the set of receivers, and only to the set of receivers through the center. Choosing a center is a NP-complete problem. Locating optimally the tree center requires the complete knowledge of the network topology and of the group membership. There are various heuristic to locate the tree center [Wei95]. Solutions to this problem are proposed by Ballardie [Ballardie95].

2.2.1.4 Synthesis

Designing an efficient multicast route requires the knowledge of numerous parameters that are not easy to quantify, such as the topology of the network, the dynamicity of the group, the location of the group members, and other routing algorithms already used in the network. A multipoint session initiator should be able to chose its algorithm and to fix the parameters of the selected algorithm regarding to the previous parameters [Herzog95, Wei95].

Another problem of multicast route design is the problem of the dynamicity of the group [Effelsberg93]. Existing solution to group dynamics have been proved inefficient by [Waxman88] and [Doar93]. An efficient algorithm should be incremental (like CBT) instead of monolithic (like Steiner). It should also have the following characteristics:

- It should be transparent to the member that remain in the group.
- It should maintain the properties of the original route.
- It should not perturb on-going data transfers.
- It must be receiver driven.

It might seem unrealistic to ask a routing algorithm to behave that way. But the advent of ATM based network with guarantee of QoS and resource reservation make it indispensable. It will be impossible to make any guarantee if the routing algorithm is not capable to maintain a multicast route with chosen properties.

Multicast Routing; Referenced papers

[Ballardie93] T. Ballardie, P. Francis, J. Crowcroft. Core based Trees (CBT). An Architecture for Scalable Inter-Domain Multicast Routing, SIGCOM '93. September 13-17, 1993, San Francisco (USA).

[Ballardie95] A. J. Ballardie. A New Approach to Multicast Communication in a datagram Internetwork. PhD thesis. University College of London. May 1995.

[Bauer95] F. Bauer and A. Varma. Degree-Constrained Multicasting in Point-to-Point Networks. IEEE INFOCOM. Boston. April 1995.

[Cimet87] I. A. Cimet and S. P. R. Kumar. A Resilient Algorithm for Minimum Weight Spanning Trees. Intl. Conference on Parallel Processing. pp. 196-203. St Charles. August 1987.

[Dalal78] Y. K. Dalal and R. M. Metcalfe. Reverse Path Forwarding of Broadcast packets. Communication of the ACM. Vol. 21. No. 12. 1978.

[Deering89] S. Deering, Host Extensions for IP Multicasting. RFC 1112. 17. August 1989.

[Deering94]] S. Deering and alii. An Architecture for Wide Area Multicast Routing. ACM SIGCOMM '94. London. August 1994.

[Doar93] M. Doar and I. Leslie. How Bad is Naive Multicast Routing?. IEEE INFOCOM. San Francisco. pp. 82-89. 1993.

[Effelsberg93] W. Effelsburg and E. Mueller-Menrad. Dynamic Join and Leave for Real-Time Multicast. Technical Report TR-93-056, International Computer Science Institute, Berkeley, CA, October

acceptable when the number of groups and the number of sources per group is low. If the number of sources and/or groups grows too large, memory could be saturated in the routers. This point also applies to the flooding step of the algorithm. [Zhu95] and [Thyagarajan95] propose an alternate approach to a source based algorithm to make it more efficient in the context of WANs.

2.2.1.2 Steiner trees

The Steiner algorithm is a monolithic algorithm that designs a tree that spans the group of members with the minimal cost, according to a distance defined on the network edges (it globally optimizes the network resources). It is aimed at a centralized calculation (but heuristics can be distributed). It is very popular because of its mathematical complexity However, to our knowledge, it has never been implemented, only simulated; the reasons are:

- The Steiner problem is NP-complete. In other words, finding the minimum Steiner tree in a graph has an exponential cost for a result which is not necessarily optimal. It has been shown that the minimum cost of a Steiner tree algorithm is O (n log n), where n is the number of nodes in the network and with all distances equal to one on the links [Hwang92].
- The tree designed is un-directed. That means it can be applied to group communication only if all the links in the network are symmetric.
- It is a monolithic algorithm. It has to be run each time there is a change in the group membership or in the network topology. The inefficiency of a Steiner tree increases dramatically each time the group changes or the network changes [Doar93].

There are numerous heuristics that have been proposed to construct Steiner trees [Cimet87, Takahashi80, Rayward86, Winter87, Noronha94]. The one that is still considered the most optimal is the one described in [Kou81]. Among more recent research works, we note:

- [Doar93] shows that the most complex heuristic is not the best and in most of the cases, a sub-optimal tree can keep its properties after modification. The naive heuristic consists of designing a sub-optimal tree which is resilient to change. Member and node movement are achieved by joining or leaving the resilient tree.
- The cost function in the Steiner problem affects link. In the degree constrained problem [Bauer95], the design of the multicast tree is constrained by the multicast capabilities of nodes. This type of problem becomes important in ATM environments where the connectivity of switches is an important issue in the efficiency of the network technology. The degree constrained problem is then to find a minimum Steiner tree constrained by the multicast capability of nodes, which is also the number of interfaces on which a message can be duplicated. Simulations made by [Bauer95] show that a fanout of 2 per node (or 3 interfaces) is enough to find acceptable solutions. [Bauer95] converges with [Doar93] in the sense that they both confirm that the naive heuristic works quite well and produces trees that are more stable under network and group dynamicity.

2.2.1.3 Centered based trees

"Centered based trees" is the most recent routing approach. We can distinguish this from the Steiner problem by observing that this family of algorithms is aimed at multiple sender/multiple recipient, as opposed to single sender, effectively fixed recipient scenario that the previous section addressed. We will illustrate this solution with the Core Based Tree (CBT) algorithm [Ballardie93]. This is a totally receiver based approach that limits the diffusion of packets naturally to group members. It is suitable for sparsely distributed receivers, and does incur extra delay over the RPF tree approach. There are 3 steps in this algorithm:

- Choose a fixed node that will be the center of the group. Multiple cores can be used if higher fault tolerance and/or better delay characteristics are required.
- Then potential group members send a join message to the center. The role of each intermediate node is to mark the interface on which a multicast packet is received and to forward it to the center.
- Multicast packets from non-member senders are forwarded towards the tree center until they reach a node that already belongs to the tree.

- Minimizing state stored in routers for some types of multicast is an important goal, as otherwise, delivery to large group is not realistic.

We start this section with the theoretical framework for multicast tree construction, then we describe existing protocols. Most of these protocols are used in the Internet, but we will also show solutions proposed in the ATM environment and for mobile communication.

There are few basic algorithms for multicast routing that have been identified today. The simplest algorithm is flooding. It is used on broadcast topologies, but it can result in very low efficiency in terms of link utilization. The spanning tree algorithm [Perlman92] is a refinement of simple flooding to provide broadcast packets in a LAN in a more efficient manner. As these two techniques are dedicated to broadcast LANs, we will not detail them here. We will focus on algorithms and protocols that really design a multicast tree, applicable to an arbitrary network topology.

2.2.1. Theoretical Basis

All services for multiple destinations require some distribution tree rather than a path through the network graph. The simplest tree is a broadcast one, that reaches all destinations. In an heterogeneous network, the effect of flooding can be catastrophic. An ideal efficient routing algorithm will design a tree that spans the group members only, with the following characteristics:

- Evolve with group membership. The routing algorithm has to distinguish group members to specifically reach them, and only them.
- Minimize state information to be kept in nodes.
- Optimize the route considering cost functions.
- Avoid traffic concentration on a subset of the links and nodes.

There are three basic algorithm to construct multicast trees. They can be characterized as centralized or distributed, and are designed to support a dense or sparse distribution of membership amongst the potential receiver set.

2.2.1.1 Source based routing

The source based routing algorithm (also known as Reverse Path Forwarding) is due to Dalal and Metcalfe [Dalal78]. It has seen widespread use through IP multicast [Deering89, Waitzman88]. The RPF algorithm computes an implicit spanning tree per source which is minimum in terms of transit delay[1]. It is optimized for dense receiver distribution, and can be implemented in a distributed fashion, with local recovery.

There cannot be loops with RPF (there can be duplicate routes, but no loops). The tree designed is a directed graph. The main advantage of the RPF algorithm is that it does not require any other resource than the classic unicast routing tables. As soon as you know how to compute the route toward the source, then you can safely process multicast packets received from this source; however, this does not really take into consideration actual group membership. The pruning variant was proposed to solve this problem [Deering94]. The idea is to complete the basic RPF by recording the "group membership" and only forward a packet if there is a group member down the tree. The pruning variant is controlled by a timer. Periodically, state for a group at a router is cleared, and all the prune messages kept in the nodes are discarded; when a new data message is forwarded, it is flooded, which serves to trigger the emission of new prune messages and the construction of a new "pruned" multicast tree. A more natural name for this improved RPF algorithm might be "flood and prune". The main problem is that the pruning variant requires one piece of state information per source and per group to be kept in each node. This is quite

1. We will see later that ordering in group communication is especially complex: for example ordered delivery of a sequence of message from a given source, to each and every receiver may be required; or delivery of messages sent by all sources may need to be globally ordered; or else messages may only need to be delivered in a causally ordered fashion.

1. The reverse path is minimal in terms of hop count, if the reverse path is calculated on a unicast metric of hop counts. Note also, that RPF routing may fail badly if the underlying unicast routing is asymmetric.

initiated join" has been proposed, where receivers, through some directory, discover the existence of a multipoint circuit, and the "conversation identifier" of the circuit, and signal their wish to be added to the circuit. Currently, no virtual circuit networks support signaling for multipoint.

A key problem with virtual circuit network signaling for multicast support is that of state. For a sender to add receivers one at a time takes O(n) signaling steps and may require the mapping from the group to all the receivers' addresses to be held at one point. This is in contrast to the datagram multicast model, where the mapping between group address and recipient location is totally distributed, and O(1).

Group Addressing and Management; Referenced papers

[Atmf95] ATM User Network Interface (UNI) Specification Version 3.1. Prenctice Hall. 1995.

[Deering89] S. Deering, Host Extensions for IP Multicasting. RFC 1112. 17. August 1989.

[Deering91]S. Deering. Multicast Routing in a Datagram Internetwork. PhD thesis. Stanford University. December 1991.

[Huitema95] C. Huitema. Routing in the Internet. Prenctice Hall. 1995.

[Schulzrinne96] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson. RTP: A Transport Protocol for Real-Time Applications. Internet RFC 1889. January 1996.

Group Addressing and Management; Other readings

[Almeroth96] K. Almeroth and M. Ammar. Collecting and Modeling the Join/Leave Behavior of Multicast Group members in the MBone. IEEE HPDC 96. Syracuse. August 1996.

[Birman91] K. P. Birman, R. Cooper and B. Gleeson. Design Alternatives for Process Group Membership and Multicast. Research Report (TR91-1257). Cornell University. 1991.

[Chandra95a] T. D. Chandra, V. Hadzilacos and S. Toueg. Impossibility of Group Membership in Asynchronous Systems. Research Report (TR95-1533). Cornell University. 1995.

[Eleftheriadis95] A. Eleftheriadis, S. Pejhan, and D. Anastassiou. Address Mangement and Connection Control for Multicast Communication Applications. IEEE INFOCOM. Boston. April 1995.

[Pejhan95] S. Pejhan, A. Eleftheriadis, and D. Anastassiou. Distributed Multicast Address Management in the Global Internet. JSAC, Vol. 13, No. 8, pp. 1445-1456. October 1995.

2.2 Routing

Designing multicast routing algorithms is a complex problem. Group membership can change, network topology can also evolve (links and nodes can fail). The technical challenges of multicast routing are the following:

- Minimize the network load. Within the problem of optimizing the network resources, there are two sub-problems which are to avoid loops and to avoid traffic concentration on a link or a subnetwork.
- The routing algorithm should be able to design optimal routes taking into consideration different cost functions, including available resource, bandwidth, number of links (graph optimization), node connectivity, price to be paid and end-to-end delay. If designing an optimal route is a complex problem, *to maintain route optimality after changes in the group and network may be much more complex*. The problem is consequently to find a good compromise between the efficiency of the route and the dynamic of the group.
- Provide basic support for reliable transmission. Ideally, route changes should have no side-effects on the way data is delivered to group members that remain in the group. Link failure should not increase transmission delay or decrease resource availability. Time constraints are very important in multicast sessions as data may have to be delivered to the application in a synchronized fashion,[1]. Too high a rate of change of routes could degrade higher level reliability.

2.1 Group addressing and membership management

A group is a set of entities. We have identified two group levels: the ***social group*** and the ***network group***. Multipoint sessions are organized for social groups. The network group is an abstraction that has been defined to coordinate participation in the social group. All notions of group address, group identifier, group properties, and group management are mechanisms to control participation in a social group[1].Such group addresses are, of course, logical: unlike addresses used for unicast delivery which contain locality, and possibly end system identifier information, multicast addresses effectively act as an index to another table somewhere (potentially distributed or partitioned throughout routers, or even implicit in other routing information).

Group management procedures are used to advertise groups to potential members, to broadcast routing information to all multicast capable nodes, and to control various properties of the group.

2.1.1. Internet Group addressing and management

The most widely used group addressing and management today is the Internet one. The Class D portion of the IP address space is reserved for multicast delivery groups. The semantics of this addressing scheme [Deering91] are such that senders to a group do not need to be in the group or to know the group members. There is no hierarchy in the Internet group structure. An Internet group address is chosen by some group initiator. On a LAN, it is directly mapped to a group MAC address [Huitema95] to avoid costly table lookup. The Group management protocol associated to the Internet is called IGMP (Internet Group Management Protocol) [Deering89]. It is used to report host group memberships to neighboring multicast routers.

Internet multicast also makes heavy use of the capability to scope the lifetime of a packet by setting a time-to-live (TTL) field which limits how far (how many hops) a packet can traverse on the way to a destination. When the destination is a group, this scoping mechanism allows the application to determine how near or far the set of actual receivers are to a sender.

IGMP has gone through various stages of evolution. Initially, the protocol entailed host members simply joining groups by advertising their membership on the local net (to the nearest router and other members) and leaving is achieved by timing out entries for hosts that did not persist sending such messages. In fact, the router needs only to know if there is *any* group member in the subnetwork in order to forward the datagrams sent to the group. More recent versions of IGMP were designed to provide some enhancements. IGMP v2 adds a "low-latency leave" to IGMP. This allows a more prompt pruning a group after all members in the subnetwork leave. IGMP v3 adds selective source reception in order to avoid forwarding traffic from all sources to members of a group. The information on desired sources is used by the multicast routing protocol in order to reduce the amount of bandwidth for a multicast tree by source pruning[2].

2.1.2. Virtual Circuit Switched Network Group Management

Circuit networks may also support group addressing. Typically, a group of end points is named at circuit setup time. It may only be possible to do this once and for all; alternatively, it may be possible to signal the addition of new receivers or senders to a multipoint circuit. Most virtual circuit networks allow a master (perhaps a single source) to add recipients, one at a time [Atmf95]. More recently, "leaf

1. Some researchers identify two other types of groups, namely "process groups" and "host groups". We regard these as implementation specific - a member of a social group may use multiple hosts, or multiple processes; multiple members of a social group may use a single host or even a single process. A member of a host group may be a member of multiple network groups, or vice versa, in other words, there is a many-to-many relationship between social group and process group, and between host group and network group.

2. However, this results in group asymmetry which jeopardizes the stability of the RTP v2 reporting algorithm [Schulzrinne96]. Receivers need to estimate the number of the sources sending to the group in order to adjust their transmission control parameters. This is discussed further under congestion control for group communication.

tems are not well suited. Group communication introduces two new aspects in transmission control:

- Control of time dependencies (transmission delay, ordering, synchronization)
- Definition of session control parameters per participant.

The problems are consequently to provide:

- the set of services that application programmers have become used to for point to point applications, but now for multipoint communication, and
- new services, both for new multipoint applications, and to support new multipoint applications.

The proliferation of multimedia applications associated with new high speed networks, often based on ATM technology, is driving this need for reliable group communication mechanisms and protocols. In the Internet Protocol Suite, TCP is a point to point protocol; however, more recently, we have seen some modifications to the Internet Protocol (and to implementations of UDP) to support group communication (see section 2). Consequently, group applications generally use non reliable UDP multicast transmission over the Mbone (examples of public domain applications using such multicast support are given section 4).

In this paper, we have decided not to follow the classic layered architecture to describe group communication related problems. Instead, we have chosen to organize the paper following the natural modularisation of mechanisms and functionalities discussed. Thus node level mechanisms will be discussed in section 2, on the "hop-by-hop" regime, where we include addressing, group management. routing, and traffic control; end-to-end problems are discussed in section 3, where mechanisms for reliability, receiver based transmission control and Quality of Service, large groups and scalability are discussed. Section 4 is specific to application related problems including multimedia applications, security, ordering and synchronization. For each topic discussed in the paper, we describe the state of the art and we present the areas of current active research.

A bibliography is provided with each section. Each bibliography is organized in two parts: referenced papers and other readings, which is a selection of the most interesting papers among those we have reviewed[1]. Our objective is not to provide an exhaustive list of references (which is available through our web page: http://www.inria.fr/rodeo/group.communication/), but rather to provide the reader with filtered bibliography of the most relevant work in this very prolific area of research.

We have also tried to avoid the classic section on terminology. Instead, we use simple vocabulary that is more intuitive. For example, "multicast" will be used to characterize the routing tree, whereas "multipoint communication" will be used to describe any type of communication within a group (it is equivalent to "group communication")[2]. Any ambiguities will be clarified on-the-fly.

2 Hop-by-Hop

In this section we discuss all the aspects of group communication that are processed at the node level. This covers routing related tasks, including group addressing, group management, and traffic control. First of all, the style of group communication required must be defined.

---

1. We have omitted almost all the references to Internet and ATM forum drafts because they are transient documents. Important document that have not yet been published will be referenced.

2. In circuit based networks, we refer to point-to-point calls, point-to-multipoint calls, multipoint-to-point calls and multipoint-to-multipoint calls. In datagram networks, there is no long term association between source and destination, and so the models are reduced to unicast (or as we call it, "any-to-one") and multicast (or "any-to-some").This characterization naturally extends to the service known as "anycast", which is one-to-any.[partrdige, anycast RFC]

Multipoint Communication: A Survey of Protocols, Functions and Mechanisms.

Christophe Diot*, Walid Dabbous*, and Jon Crowcroft**

*INRIA, 2004 Route des Lucioles, BP 93, 06902 Sophia Antipolis, FRANCE
Ph. (33) 93 65 78 25; Fax. (33) 93 65 76 02; e.mail: [walid.dabbous|christophe.diot]@sophia.inria.fr
http://www.inria.fr/rodeo/

**Department of Computer Science, UCL, Gower Street, London, WC1E 6BT, UK
Ph. (44) 171 380 7298; Fax (44) 171 387 1397; e.mail: jon@cs.ucl.ac.uk

Abstract: Group communication supports information transfer between a set of participants. It is becoming more and more relevant in distributed environments. For distributed or replicated data, it provides efficient communication without overloading the network. For some types of multimedia applications, it is the only way to control data transmission to group members. This paper surveys protocol functions and mechanisms for data transmission within a group, from multicast routing problems up to end-to-end multipoint transmission control. We provide a bibliography which is organized by topic. This paper is intended to introduce this special issue with the necessary background on recent and ongoing research.

Keywords: Group communication, multipoint, multicast, routing, end-to-end control, ATM, Internet.

1 Introduction

One of the most pressing needs for enhanced communication protocols comes from multipoint (or group) applications. These involve more than two users (these users define a "group") that wish to exchange information; in contrast, point to point applications involve only two users. Such applications cover a very wide spectrum, including software distribution, replicated database update, command&control systems, audio/video conferencing, distributed games and Distributed Interactive Simulation.

Until recently, communication systems built around the OSI reference model or the Internet architecture were designed to support point to point services. Point-to-point communication often depends on implicit knowledge that is no longer valid for group communication. A point-to-point session is made of two participants, one of which is typically a client (or the active initiator or master participant), the other of which is a server (or passive or slave). Connection establishment/teardown, flow control, error recovery can be driven from one end (typically the active/sender end). In a multipoint session any participant can decide whether and when it wishes to join or leave the session. The join and leave operations have to be simple, with no side effect on the other participants, if they are to scale seamlessly from small to very large group membership. The session parameter negotiation, which is common in point-to-point protocols, is not always acceptable in group communication. A multipoint session has to be receiver controlled in order to allow dynamic join and leave. Window based flow control for example is not usable in the context of group communication. In general, we find that a shift from sender initiation, to a model of "receiver-makes-good" is needed for many protocol functions only some of which we have mentioned so far.

Existing protocols (e.g. IP, CLNP and UDP, CLTS) are sufficient for the applications which can be satisfied by a connectionless (non reliable) multicast communication support. New generation applications such as multimedia conferences, shared workspace, distributed games, and distributed simulation introduce new requirements in data transmission; this means that existing protocols and communication sys-