Reliable Ordered Multicast Service (ROMS) over NBMA Networks

Stefan Dresler

Institute of Telematics, University of Karlsruhe
Zirkel 2, D–76128 Karlsruhe, Germany
Phone: (+49) 721/608-6397, Fax: (+49) 721/388097
E-Mail: dresler@telematik.informatik.uni-karlsruhe.de

Georg Carle

Institut Eurecom
2229 Route des Cretes, F–06904 Sophia Antipolis Cedex, France
Phone: (+33) 4.93.00.26.91, Fax: (+33) 4.93.00.26.27
E-Mail: carle@eurecom.fr

Abstract: The programming of applications like distributed database systems or distributed simulations can often be simplified if the programmer can rely on a network service offering certain ordering semantics of messages. There have been numerous suggestions for systems providing such an ordering service over the Internet Protocol (IP), some taking into account NBMA (Non-Broadcast Multiple Access) networks like ATM (Asynchronous Transfer Mode) and ISDN.
This paper discusses the difference of the services offered by IP and ATM and its impact on the provision of a reliable ordered multicast service (ROMS). Two different solutions are presented to realize a ROMS on top of ATM: First, an intermediate service adaptation layer according to the IP-over-ATM and the MARS models can be introduced between an ordering service on top and ATM below. Second, a native ATM solution can be designed that takes advantage of the properties of ATM. This approach again can be realized by two different layers (reliable transport service and ordering service), or by an integrated layer.

1. Introduction

Several types of applications rely on a communication service offering reliable ordered communication in a group scenario. With appropriate support not being available at the beginning, these applications realized the service themselves by embedding appropriate functionality. This led to an increased complexity of the applications and more complicated programming. In a first conceptual step, the required functionality can be separated into a communication subsystem layered between the application and the network. Some libraries, especially for IP (Internet Protocol) environments, were developed which offer the desired services, e.g. the Transis [DoMa96], Isis [Birm93] and Trans/Total [MeMA90] systems.

The semantics of the ordering provided by the protocols often differ. The Transis system e.g. offers the following ordering semantics (in increasing strictness):

1. *FIFO* just assures the in-order delivery of messages of each sender, with no relation to messages of other senders.
2. *Causally ordered delivery* preserves a potential causal order among messages (compare [Lamp78]).

3. *Agreed delivery* ensures a unique order among every pair of messages in all of their destinations.
4. *Safe multicast* guarantees a unique order of message delivery. Furthermore, it delays the delivery of a message until it is acknowledged by the transport layers in all of the machines in the group. This ensures atomic delivery in case of communication failures.

With the advent of ATM, the service offered by the underlying network technology is different from that known from IP. ATM is connection-oriented, uses a fixed-length packet size (i.e., ATM cells) for multiplexing, and does not support broadcast, i.e., one-to-all communication. For the latter reason, ATM, like ISDN, is considered an NBMA (Non-Broadcast Multiple Access) technology. Furthermore, ATM is characterized by in-order delivery.
IP, on the other hand, works connection-less, uses variable-size packets and supports broadcasts (which is especially efficient in combination with LAN technologies like Ethernet which provide broadcast). IP packets may experience out-of-order delivery. The error characteristics of IP depends both on the underlying media and also on errors (mainly packet loss) in IP routers. Finally, the address format of ATM and IP is different.
These differences of the network service offered by IP networks and ATM networks lead to the following differences for protocols that offer reliable ordered multicast services on top of the respective network services:
1. Protocols that rely on IP multicasting or broadcasting have to emulate that functionality.
2. For point-to-point and point-to-multipoint communication operations it is necessary to establish a connection first. The connection establishment may be "hidden" by a sublayer realizing IP-over-ATM.
3. In order to not experience high packet loss rates in situations of buffer overflow in ATM switches, forward error correction (FEC) encoding can be used, as explained in section 4.3.

Since the area of reliable ordered multicast communication on top of ATM has not been covered extensively in literature yet, it is desirable to first state requirements to the service and to give guidelines for a high performance realization. This is goal of the paper.

Section 2 presents criteria for the evaluation of systems offering a reliable ordered multicast service (ROMS). Section 3 gives an overview on the deployment of reliable ordered multicast systems based on IP when deployed over ATM. Section 4 discusses procotols that are directly based on ATM. The paper concludes with remarks on issues that have not been solved yet.

2. Comparison and Evaluation Criteria

There are a number of criteria to compare and evaluate reliable ordered multicast protocols. These include criteria known from the evaluation of protocols in general as well as criteria specific to reliable ordered multicast protocols.

- Average and maximum delay. The following mechanisms contribute to the delay of a reliable ordered multicast service:
  (1) address resolution, e.g. by IP over ATM and MARS (see below),
  (2) connection establishment, if applicable,
  (3) packet copying and distribution, which may be performed in software even in ATM switches,
  (4) ordering of the packets, and
  (5) group management procedures.
- Number of messages. Especially for multicast services this metric can limit the efficiency of the protocol used.

- Number of connections. This is important for ATM networks, which can only support a finite number of connections. It also relates to the delay incurred by connection management.
- Scalability with respect to the number of participants (application processes) involved.
- Scalability with respect to the number of groups using the procotol at a time.
- Scalability with respect to the size of the underlying network (from LANs to WANs).
- Suitability for heterogeneous environments. If powerful workstations on high-speed network connections are in the same group as Personal Digital Assistants (PDAs) attached to the network via a wireless link, performance may be degraded to the speed of the least powerful participant.
- Interoperability with existing hardware and software, or modifications necessary to the network, if any.
- Ordering semantics implemented.
- Incorporation of flow control. Multicast protocols that use message duplication in intermediate systems can easily create a large amount of traffic disturbing other communications.
- Assignment of unique identifiers. In general it is desirable to be able to derive the sender from a packet received. To this aim, it is necessary to (explicitly or implicitly) assign an identifier to each sender participating.
- Robustness. Since IP does not have a connection concept, it cannot itself detect a link failure or an application crash. The lower layers of ATM are capable of detecting such an error condition. Fast detection of network problems is desirable.
- Two-layer or integrated approach. Typically, reliable ordered multicast protocols use a two-layer approach. The upper layer takes care of the ordering of messages, while the lower layer provides a reliable transport service. An integrated protocol combines both functionalities.

3. Reliable Ordered Multicast Systems Based on IP, Deployed over ATM

This section focuses on the provision of a reliable ordered multicast service by protocols designed for IP, which are used on top of ATM. To this aim, adaptation mechanisms between IP and ATM are presented first. Without referring to a specific library, the characteristics of such an adaptation are discussed.

3.1 Existing Adaptation Mechanisms

There is a number of proposals for the adaptation of IP to ATM environments:

- The classical IP over ATM approach [Pere95] is tailored towards Logical IP Subnets (LIS) only and does not support multicast.
- Combining [Pere95] with the MARS (Multicast Address Resolution Service) [Armi96] provides for the resolution of both unicast and multicast addresses in LIS environments.
- LAN Emulation Version 2 [ATMF97a] features a Broadcast and Unknown Server (BUS), which can be used to distribute data to all end systems in a LAN.
- The MPOA proposal [ATMF97b], which combines the LANE solution with the Next Hop Resolution Protocol (NHRP), only supports unicast communication.

None of these mechanisms supports a reliable delivery of data, has any flow or congestion control suitable for multipeer communication, or in itself ensures a certain ordering of messages in multipeer scenarios.

3.2 Performance and Resource Utilization

The MARS approach, which is combined with the IP-over-ATM address resolution, can serve as a basis for the adaptation from IP to ATM in one of three ways:

- The simplest way (which is not recommended in multipeer communication) is to use M*N 1:1 connections. This is clearly inefficient with respect to connection setup times and VC space usage.
- A lot less resources are wasted by establishing M 1:N connections, forming a mesh.
- The third way is to use a Multicast Server (MCS). It only requires for each sender a 1:1 connection to the MCS (for a total of M 1:1 connections) and from there a single 1:N VC to all receivers.

The MARS concept was modeled such that for senders and receivers no explicit difference is made whether they are a part of a mesh, an MCS connection, or possibly a hybrid of both. Any of the three ways can be deployed by the hidden sublayer between ATM and IP without IP or the application being able to distinguish them.

With Transis, which is optimzed for broadcast media in a LAN, it is necessary to establish and maintain one of these three ways to communicate. If a VC mesh is used, every participating IP over ATM subsystem that receives messages has to maintain M VCs, one from each sender. For the MCS solution, a lot of adaptation work is still covered by the sublayer.

Again, none of the solutions features flow control or ensures reliable delivery of data, and none spans a larger area than a LAN. The MCS does, however, sequentialize incoming frames before forwarding them. It is not sure how the system reacts to certain connection losses.

4. Reliable Ordered Multicast Service over NBMA Networks

Some of the protocols mentioned above to realize the ROMS using the Internet Protocol are implemented and ready to use. Besides those, there are also some suggestions on how to realize the reliable ordered multicast service directly on top of ATM. An advantage of this approach is the possibility to make use of the in-order delivery of ATM cells on unicast and multicast connections. The following lists key requirements to be met by protocols realizing a ROMS on top of NBMA networks.

4.1 Basic Principles of Multipeer Communication over ATM

A reliable ordered protocol over ATM can be thought of as providing an M:N connection that assures certain properties. Such a service cannot be achieved by simply merging data from M senders and distributing it to N receivers, because of ATM’s cell-based and connection-oriented nature.

If it should be possible for receivers to find out the originator of a packet, a unique identifier has to be assigned to each sender. This was already the case with IP-based protocols (using the sender’s IP address as a discriminator does not help if more than one application on a system belongs to the group). Such an identifier is created implicitly if for each sender a 1:M connection is established to all receivers (or even for each sender-receiver pair a 1:1 connection), because in this case receivers know the Virtual Path Identifier (VPI)/Virtual Channel Identifier (VCI) of an incoming packet, uniqely identifying a sender. If a common VC (either directly by the ATM system or by the means of a MCS) is to be used, the sender identifier has to be integrated into every ATM cell or AAL frame. The former idea is realized by ATM Adaptation Layer (AAL) 3/4, which provides a Message Identifier (MID) field of size 10 bits in each cell, which can be used to carry the sender identifier and allows for a convenient interleaving of cells of different senders on the VC. It limits the number of senders in a group to 1024 and — more importantly — the amount of data per ATM cell to 44 bytes, however. Integrating the identifier into the payload field of a, say, AAL5 frame is more

efficient with respect to bandwidth usage, but disallows interleaved sending of two or more frames at a time over the VC, because the receivers cannot properly assign the incoming cells to frames anymore. Thus, intermediate systems at merge points of a single VC have to buffer incoming frames and forward them one after the other. In both cases a method has to be provided to supply unique identifiers to participants.

4.2 Existing Protocols

In [VRCK97] it is suggested to use the VCI value as a sender identifier, thus supporting cell interleaving. For efficient usage of the protocol a modification of the treatment of VP/VC values in switches is needed, however. SMART [GaBO97] uses a token (here called grant) based approach which requires the introduction of at least one SMART-capable switch into the system. Thus, both protocols require at least some change to the existing switch implementations. This requisite does not seem to be desirable, though.

4.3 Possibilities of enhancing the service provision

The discussion above showed that there is still potential for improvement of the provisioning of a reliable ordered multicast service.

Loss of a single ATM cell results in the loss of the whole AAL SDU (or IP packet, if IP over ATM is deployed), so that an error handling algorithm should always be used. Unless Constant Bit Rate (CBR) is used, for large multicast scenarios it becomes increasingly likely that every single ATM cell of an SDU does not arrive at least at one receiver. Selective repeat is thus not an option in multipeer communication to handle the majority of cell losses. It has been shown that the introduction of forward error correction (FEC) allows for the recovery of most of the lost cells under typical error conditions [CaDr96].

At least for the distribution of the user data, redundancy based on error correcting coding should be used. In addition to that, dispersing ATM traffic between two systems onto several VC connections (VCCs) routed over different paths reduces burstiness on each of the VCCs and — in combination with FEC — lowers the resulting frame loss rate [DiLi95].

Furthermore, increasing the robustness of the communication system seems a necessity so that a crash of an application or a connection breakdown does not affect the whole group (but might be reported to all participants). A proposal for signalling procedures for fault tolerant connections (for unicast), which can be considered a step into that direction, can be found in [KuSp97].

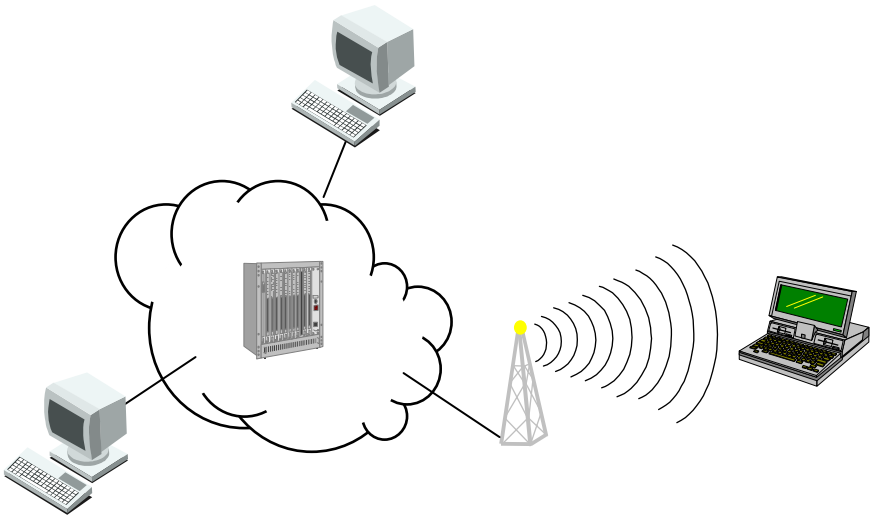

Figure 1: Heterogeneous Environment for Reliable Ordered Multicast Service

If less powerful machines like PDAs connected by a wireless link participate in a group, they tend to slow down overall communication progress. Accordingly, the protocol should be able to make progress unimpeded by slow machines or links, and it should be built in a robust way so that a failing link does not affect the overall communication more than necessary for the correctness of the protocol (compare Figure 1). This might also be achieved by (possibly dedicated) stations serving as proxies to slow stations.
Furthermore, the number of messages handled is often a limiting factor in communication systems. The ability to concatenate messages to form a larger SDU in order to reduce the number of messages a participating application has to process can be built into the protocol and the stations to this aim.
A potential gain to a protocol realizing reliable ordered multicast service might finally be achieved by permanently measuring delays in a group in order to adapt sending of messages to the current situation, forming a schedule.

4. Conclusion

The paper first presented criteria for the comparison and evaluation of protocols offering a reliable ordered multicast service. It then showed how existing protocols tailor-made for the Internet Protocol can be set on top of ATM and gave reasons why this approach is not optimal. Subsequently, implementing the reliable ordered multicast service directly on top of the ATM transport service was discussed, together with some protocols proposed for this purpose.
From the analysis given for each approach it can be concluded that no protocol has been suggested yet which meets all of the desired characteristics. Especially provision for communications over unreliable networks (due to cell loss or connection breakdown) and support for inhomogeneous groups of participants is often not included in the protocols.

5. References

[ATMF97a] ATM Forum: *LAN Emulation over ATM, Version 2, LUNI Specification*. af-lane-0084.000, July 1997.

[ATMF97b] ATM Forum: *Multi-Protocol over ATM, Version 1.0*. af-mpoa-0087.000, July 1997.

[Birm93] K. P. Birman: *The Process Group Approach to Reliable Distributed Computing*. CACM, Vol. 36, No. 12, December 1993, pp. 36–53.

[CaDr96] Georg Carle, Stefan Dresler: *High Performance Group Communication Services in ATM Networks*. Chapter in Book: "High-Speed Networks for Multimedia Applications", W. Effelsberg, O. Spaniol, A. Danthine, D. Ferrari (Eds.), Kluwer Academic Publishers, Boston/Dordrecht/London, 1996.

[DiLi95] Quan-long Ding, Soung C. Liew: *A Performance Analysis of a Parallel Communications Scheme for ATM Networks*. Proceedings of IEEE Globecom'95, pp. 898–902.

[DoMa96] Danny Dolev, Dalia Malki: *The Transis Approach to High Availability Cluster Communication*. CACM, Vol. 39, No. 4, April 1996, pp. 64-70.

[GaBO97] Eric Gauthier, Jean-Yves Le Boudec, Philippe Oechslin: *SMART: A Many-to-Many Multicast Protocol for ATM*. IEEE JSAC, Vol. 15, No. 3, April 1997.

[GrRa96] M. Grossglauser, K. K. Ramakrishnan: *SEAM – A Scheme for Scalable and Efficient ATM Multipoint-to-Multipoint Communication*. ATM Forum Draft 96-1142, 1996.

[KuSp97] David M. Kushi, Ethan M. Spiegel: *Signalling Procedures for Fault Tolerant Connections*. ATM Forum Draft 97-0391R1, 1997.

[Lamp78] Leslie Lamport: *Time, Clocks, and the Ordering of Events in a Distributed System*. CACM, Vol. 21, No. 7, July 1978, pp. 558–565.

[MeMA90] P. M. Melliar-Smith, L. E. Moser, V. Agrawala: *Broadcast Protocols for Distributed Systems*. IEEE Transactions on Parallel and Distributed Systems, Vol. 1, No. 1, January 1990, pp. 17–25.

[Pere95] M. Perez et al.: *ATM Signaling Support for IP over ATM*. Request for Comments 1755, IETF, February 1995.

[Armi96] G. Armitage: *Support for Multicast over UNI 3.0/3.1 based ATM Networks*. Request for Comments 2022, IETF, November 1996.

[VRCK97] R. Venkateswaran, C. S. Raghavendra, Xiaoqiang Chen, Vijay P. Kumar: *Support for Multiway Communications in ATM Networks*. ATM Forum Draft 97-0316, 1997.