J.M.Pullen
George Mason U.
M.Myjak
The Virtual Workshop
C.Bouwens
SAIC
14 September 1998

Limitations of Internet Protocol Suite for Distributed Simulation in the Large Multicast Environment

draft-ietf-lsma-limitations-03

Status of this Memo

Abstract

The Large-Scale Multicast Applications (LSMA) working group was chartered to produce Internet-Drafts aimed at a consensus-based development of the Internet protocols to support large scale multicast applications including real-time distributed simulation. This draft defines services that LSMA has found to be required, and aspects of the Internet protocols that LSMA has found to need further development in order to meet these requirements.

1. The Large Multicast Environment

The Large-Scale Multicast Applications working group (LSMA) was formed to create a consensus-based requirement for Internet Protocols to support Distributed Interactive Simulation (DIS) [DIS94], its successor the High Level Architecture for simulation (HLA) [DMSO96], and related applications. The applications are characterized by the need to distribute a real-time applications over a shared wide-area network in a scalable manner such that numbers of hosts from a few to tens of thousands are able to interchange state data with sufficient reliability and timeliness to sustain a three-dimensional virtual, visual environment containing large numbers of moving objects. The network supporting such an system necessarily will be capable of multicast [IEEE95a,IEEE95b].

Distributed Interactive Simulation is the name of a family of protocols used to exchange information about a virtual environment among hosts in a distributed system that are simulating the behavior of objects in that environment. The objects are capable of physical interactions and can sense each other by visual and other means (infrared, etc.). DIS was developed by the U.S. Department of Defense (DoD) to implement systems for military training, rehearsal, and other purposes. More information on DIS can be found in [SSM96].

The feature of distributed simulation that drives network requirements is that it is intended to work with output to and input from humans across distributed simulators in real time. This places tight limits on latency between hosts. It also means that any practical network will require multicasting to implement the required distribution of all data to all participating simulators. Large distributed simulation configurations are expected to group hosts on multicast groups based on sharing the same sensor inputs in the virtual environment. This can mean a need for thousands of multicast groups where objects may move between groups in large numbers at high rates. Because the number of simulators is known in advance and their maximum output rate in packets per second and bits per second is specified, the overall total data rate (the sum of all multicast groups) is bounded. However the required data rate in any particular group cannot be predicted, and may change quite rapidly during the simulation.

DIS real time flow consists of packets of length around 2000 bits at rates from .2 packets per second per simulator to 15 packets per second per simulator. This information is intentionally redundant and is normally transmitted with a best-effort transport protocol (UDP). In some cases it also is compressed. Required accuracy both of latency and of physical simulation varies with the intended purpose but generally must be at least sufficient to satisfy human perception. For example in tightly coupled simulations such as high performance aircraft maximum acceptable latency is 100 milliseconds between any two hosts. At relatively rare intervals events (*e.g.* collisions) may occur which require reliable transmission of some data, on a unicast basis, to any other host in the system.

The U.S. DoD has a goal to build distributed simulation systems with up to 100,000 simulated objects, many of them computer-generated forces that run with minimal human intervention, acting as opposing force or simulating friendly forces that are not available to participate. DoD would like to carry out such simulations using a shared WAN. Beyond DoD many people see a likelihood that distributed simulation capabilities may be commercialized as entertainment. The scope of such an entertainment system is hard to predict but conceivably could be larger than the DoD goal of 100,000.

The High Level Architecture (HLA) is a DoD development beyond DIS that aims at bringing DIS and other forms of distributed simulation into a unifying system paradigm. From a distributed systems standpoint HLA is considerably more sophisticated than DIS. For example attributes of distributed objects may be controlled by different simulators. From the standpoint of the supporting network the primary difference between HLA and DIS is that HLA does not call for redundant transmission of object attributes; instead it specifies a "Run Time Infrastructure" (RTI) that is responsible to transmit data reliably, and may choose to do so by various means including redundant transmission using best-effort protocols. It is reasonable to say that any network that can meet the needs of DIS can support HLA by DIS-like redundant transmission, however this approach ignores the possibility that under HLA some mixture of redundant and reliable transmission can make significantly better use of network resources than is possible using DIS. While HLA, like DIS, does not specify use of a multicasting network, it has similar requirements for many-to-many transmission of object attributes at rates in excess of one update per object per second that cannot be met without multicasting. Further, HLA calls for transmission of semantically organized data (for example, groups of objects with similar capabilities such as tanks or aircraft) in this many-to-many context.

One solution that has been employed to deal with these challenges is to aggregate the contents of many multicast groups into a single multicast transmission [PuWh95, CSTH95]. Termed “dual-mode” or “bi-level” multicast, this approach takes advantage of the fact that although the amount of traffic in any particular multicast group can vary greatly, the aggregate of all transmissions is bounded. If the traffic is all aggregated into one large flow, an underlying ATM network can create multicast SVCs with acceptable QoS to support the requirement. It also bounds the network control problem of group joins, in that the joins take place among dedicated collections of routers and across the dedicated SVCs, rather than contending with other LSMAs that may be sharing the same network. But it does this at the cost of adding to the network a new, nonstandard aggregation element that is a hybrid of the Internet and ATM protocols. We address below the requirement to achieve such a result using a purely IP network with aggregated reservation via RSVP.

The defense distributed simulation community has created a number of multicast-capable networks for various simulated exercises, ranging from tens to hundreds of simulated objects distributed across numbers of sites ranging from two to twenty. As the number of objects has increased they have found that building multicasting networks potentially supporting thousands of simultaneous multicast groups with large group change rates is a hard problem. This defense problem is the precursor of similar problems that can be expected in commercial networks. Therefore the following sections describe the services required and the shortcomings that have been found in using today’s Internet protocols in providing these services, with the intention of informing the IETF to enable it to produce protocols that meet the needs in these areas.

2. Distributed Simulation (DIS and HLA) network service requirements.

a. real-time packet delivery, with low packet loss (less than 2%), predictable latency on the order of a few hundred milliseconds, after buffering to account for jitter (variation of latency) such that less than 2% of packets fail to arrive within the specified latency, in a shared network

b. multicasting with thousands of multicast groups that can sustain host group join requiring less than one second, at rates of hundreds of joins per second

c. multicasting using a many-to-many paradigm in which 90% or more of the group members act as receivers and senders within any given multicast group

d. support for resource reservation; because of the impracticality of over-provisioning the WAN and the LAN for large distributed simulations, it is important to be able to reserve an overall capacity that can be dynamically allocated among the multicast groups

e. support for a mixture of best-effort and reliable low-latency multicast transport, where best-effort predominates in the mixture, and the participants in the reliable multicast may be distributed across any portion of the network

f. support for secure networking, in the form of per-packet encryption and authentication needed for classified military simulations

3. Internet Protocol Suite facilities needed and not yet available for large-scale distributed simulation in shared networks: These derive from the need for real-time multicast with established quality of service in a shared network. (Implementation questions are not included in this discussion. For example, it is not clear that implementations of IP multicast exist that will support the required scale of multicast group changes for LSMA, but that appears to be a question of implementation, not a limitation of IP multicast.)

3.1 Large-scale resource reservation in shared networks

The Resource reSerVation Protocol (RSVP) is aimed at providing setup and flow-based information for managing information flows at pre-committed performance levels. This capability is generally seen as needed in real-time systems such as the HLA RTI. Concerns have been raised about the scalability of RSVP, and also about its ability to support highly dynamic flow control changes. In terms of existing RTI capabilities, the requirement in LSMA is for rapid change of group membership, not for rapid change of group reservations. This is because in existing RTIs the aggregate requirement for all groups in a large scale distributed simulation is static. However the current RSVP draft standard for LSMA does not support aggregation of reservation resources for groups of flows and therefore does not meet the needs of existing RTIs. Moreover, there is at least one RTI development underway that intends to use individual, dynamic reservations for large numbers of groups, and therefore will require a dynamic resource reservation capability that scales to thousands of multicast groups.

Further, RSVP provides support only for communicating specifications of the required information flows between simulators and the network, and within the network. Distributing routing information among the routers within the network is a different function altogether, performed by routing protocols such as Multicast Open Shortest Path First (MOSPF). In order to provide effective resource reservation in a large shared network function, it may be necessary to have a routing protocol that determines paths through the network within the context of a quality of service requirement. An example is the proposed Quality Of Service Path First (QOSPF) routing protocol [ZSSC97]. Unfortunately the requirement for resource-sensitive routing will be difficult to define before LSMA networks are deployed with RSVP.

3.2 IP multicast that is capable of taking advantage of all common link layer protocols (in particular, ATM)

Multicast takes advantage of the efficiency obtained when the network can recognize and replicate information packets that are destined to a group of locations. Under these circumstances, the network can take on the job of providing duplicate copies to all destinations, thereby greatly reducing the amount of information flowing into and through the network.

When IP multicast operates over Ethernet, IP multicast packets are transmitted once and received by all receivers using Ethernet-layer multicast addressing, avoiding replication of packets. However, with wide-area technology Asynchronous Transfer Mode (ATM), the ability to take advantage of data link layer multicast capability is not yet available beyond a single Logical IP Subnet (LIS). This appears to be due to the fact that (1) the switching models of IP and ATM are sufficiently different that this capability will require a rather complex solution, and (2) there has been no clear application requirement for IP multicast over ATM multicast that provides for packet replication across multiple LIS. Distributed simulation is an application with such a requirement.

3.3 Hybrid transmission of best-effort and reliable multicast

In general the Internet protocol suite uses the Transmission Control Protocol (TCP) for reliable end-to-end transport, and the User Datagram Protocol (UDP) for best-effort end-to-end transport, including all multicast transport services. The design of TCP is only capable of unicast transmission.

Recently the IETF has seen proposals for several reliable multicast transport protocols (see [Mont97] for a summary). A general issue with reliable transport for multicast is the congestion problem associated with delivery acknowledgments, which has made real-time reliable multicast transport infeasible to date. Of the roughly 15 attempts to develop a reliable multicast transport, all have shown to have some problem relating to positive receipt acknowledgments (ACK) or negative acknowledgments (NAK). In any event, its seems clear that there is not likely to be a single solution for reliable multicast, but rather a number of solutions tailored to different application domains. Approaches involving distributed logging seem to hold particular promise for the distributed simulation application.

In the DIS/HLA environment, five different transmission needs can be identified:
(1) best-effort low-latency multicast of object attributes that often change continuously, for example position of mobile objects;
(2) low-latency reliable multicast of object attributes that do not change continuously but may change at arbitrary times during the simulation, for example object appearance (An important

characteristic of this category is that only the latest value of any attribute is needed by the receiver.);
(3) low-latency, reliable unicast of occasional data among arbitrary members of the multicast group (This form of transmission was specified for DIS "collisions"; it is not in the current HLA specification but might profitably be included there. The requirement is for occasional transaction-like exchange of data between two arbitrary hosts in the multicast group, with a low latency that makes TCP connection impractical.);
(4) reliable but not necessarily real-time multicast distribution of supporting bulk data such as terrain databases and object enumerations; and
(5) reliable unicast of control information between individual RTI components (this requirement is met by TCP).

All of these transmissions take place within the same large-scale multicasting environment. The value of integrating categories (1) and (2) into a single selectively reliable protocol was proposed by Cohen [Cohe94]. Pullen and Laviano implemented this concept [PuLa95] and demonstrated it within the HLA framework [PLM97] as the Selectively Reliable Transmission Protocol (SRTP) for categories (1) through (3). Category (4) could be supported by a reliable multicast protocol such as the commercial multicast FTP offering from Starburst [MRTW97], however adequate congestion control has not been demonstrated in any such protocol. There has been some discussion of using the Real-Time Streaming Protocol, RTSP, for this purpose, however as the databases must be transmitted reliably and RTSP uses a best-effort model, it does not appear to be applicable.

In summary, it is clear that a hybrid of best-effort and reliable multicast (not necessarily all in the same protocol) is needed to support DIS and HLA, and that the low-latency, reliable part of this hybrid is not available in the Internet protocol suite.

3.4 Network management for distributed simulation systems

Coordinated, integrated network management is one of the more difficult aspects of a large distributed simulation exercise. The network management techniques that have been used successfully to support the growth of the Internet for the past several years could be expanded to fill this need. The technique is based on a primitive called a Management Information Base (MIB) being polled periodically at very low data rates. The receiver of the poll is called an Agent and is collocated with the remote process being monitored. The agent is simple so as to not absorb very many resources. The requesting process is called a Manager, and is typically located elsewhere on a separate workstation. The Manager communicates to all of the agents in a given domain using the Simple Network Management Protocol (SNMP). It appears that SNMP is well adapted to the purpose of distributed simulation management, in addition to managing the underlying simulation network resources. Creating a standard distributed simulation MIB format would make it possible for the simulation community to make use of the collection of powerful, off-the-shelf network management tools that have been created around SNMP.

3.5 A session protocol to start, pause, and stop a distributed simulation exercise

Coordinating start, stop, and pause of large distributed exercises is a complex and difficult task. The Session Initiation Protocol (SIP) recently proposed by the Multiparty Multimedia Session Control (MMUSIC) working group serves a similar purpose for managing large scale multimedia conferences. As proposed, SIP appears to offer sufficient extensibility to be used for exercise session control, if standardized by the IETF.

3.6 An integrated security architecture

It appears that this requirement will be met by IPv6 deployment. A shortcoming of the current Internet Protocol (IPv4) implementation is the lack of integrated security. The new IPv6 protocol requires implementers to follow an integrated security architecture that provides the required integrity, authenticity, and confidentiality for use of the Internet by communities with stringent security demands, such as the financial community. The possibility that the IPv6 security architecture may meet military needs, when combined either with military cryptography or government-certified commercial cryptography, merits further study.

3.7 Low-latency multicast naming service

Name-to-address mapping in the Internet is performed by the Domain Name Service (DNS). DNS has a distributed architecture tuned to the needs of unicast networking with reliable transmission (TCP) that is not considered problematic if its latency is on the order of a second or more. The requirement of distributed simulation for agile movement among multicast groups implies a need for name-to-multicast-address mapping with latency of under one second for the name resolution and group join combined. This problem has been circumvented in military simulations by using group IP addresses rather than names. While military simulations may be satisfied to communicate using a known mapping from grid squares to multicast groups, growth of distributed simulation into commercial entertainment cannot be based on such a simple capability. The players in distributed entertainment simulations will want to be organized symbolically by virtual world and role. A low-latency multicast naming service will be required.

3.8 Inter-Domain Multicast Routing for LSMA

While military LSMAs typically take place within a single administrative domain, future entertainment LSMAs can be expected to involve heavy inter-domain multicast traffic so that players can be supported by multiple service providers. Standardized protocols able to support large numbers of multicast flows across domain boundaries will be needed for this purpose. Current work to create a Border Gateway Multicast Protocol (BGMP) shows promise of meeting this need.

4. References

[CSTH95] Calvin, J., J. Seeger, G. Troxel and D. Van Hook, "STOW Realtime Information Transfer and Networking Architecture," *12th DIS Workshop on Standards for the Interoperability of Distributed Simulations*, March 1995

[Cohe94] Cohen, D., “Back to Basics,” *Proceedings of the 11th Workshop on Standards for Distributed Interactive Simulation*, Orlando, FL, September 1994

[DIS94] DIS Steering Committee, “The DIS Vision,” Institute for Simulation and Training, University of Central Florida, May 1994

[DMSO96] Defense Modeling and Simulation Office, *High Level Architecture Rules Version 1.0*, U.S. Department of Defense, August 1996

[IEEE95a] IEEE 1278.1-1995, *Standard for Distributed Interactive Simulation - Application Protocols*

[IEEE95b] IEEE 1278.2-1995, *Standard for Distributed Interactive Simulation - Communication services and Profiles*

[MRTW97] Miller, K, K. Robertson, A. Tweedly, and M. White, “StarBurst Multicast File Transfer Protocol (MFTP) Specification,” IETF draft-miller-mftp-spec-02.txt, work in progress, January 1997

[Mont97] Montgomery, T., Reliable Multicast Links webpage, http://research.ivv.nasa.gov/RMP/links.html

[PuLa95] Pullen, J. and V. Laviano, "A Selectively Reliable Transport Protocol for Distributed Interactive Simulation", *Proceedings of the 13th Workshop on Standards for Distributed Interactive Simulation*, Orlando, FL, September 1995

[PuWh95] Pullen, J.M. and E. White, “Dual-Mode Multicast: A New Multicasting Architecture for Distributed Interactive Simulation,” *12th DIS Workshop on Standards for the Interoperability of Distributed Simulations*, Orlando, FL, March 1995

[PLM97] Pullen, J., V. Laviano and M. Moreau, “Creating A Light-Weight RTI As An Evolution Of Dual-Mode Multicast Using Selectively Reliable Transmission,” Proceedings of the Second Simulation Interoperability Workshop, Orlando, FL, September 1997

[SPW94] Symington, S., J.M.Pullen and D. Wood, “Modeling and Simulation Requirements for IPng,” IETF RFC1667, August 1994

[SSM96] Seidensticker, S., W. Smith and M. Myjak, “Scenarios and Appropriate Protocols for Distributed Interactive Simulation,” IETF draft-ietf-lsma-scenarios-00.txt, work in progress 13 June 1996 (companion Internet Draft)

[ZSSC97] Zhang, Z., C. Sanchez, W. Salkewicz, and E. Crawley, “Quality of Service Path First Routing Protocol,” IETF draft-zhang-qos-ospf-01.txt, work in progress September 1997

5. Authors' Addresses

J. Mark Pullen
mpullen@gmu.edu
Computer Science/C3I Center
MS 4A5
George Mason University
Fairfax, VA 22032

Michael Myjak
mmyjak@mail.imtinc.com
The Virtual Workshop
P.O. Box 98
Titusville, FL 32781

Christina Bouwens
christina.bouwens@cpmx.mail.saic.com
ASSET Group, SAIC Inc.
Orlando FL

Expiration: 14 March 1999