A Survivability Framework for Connection-Oriented Group Communications*

William Yurcik**

Illinois State University
Dept. of Applied Computer Science
wjyurci@ilstu.edu

David Tipper

University of Pittsburgh
Dept. of Info. Science and Telecommunications
tipper@tele.pitt.edu

Abstract

This paper presents a framework for providing survivability to group communications where part of the underlying traffic layer infrastructure is connection-oriented. The framework is multi-layered to express the virtual overlays inherent to networked systems, describes survivability issues unique to group communications, and outlines tradeoffs between restoration techniques that can be extended from circuit-switched communications. The conclusion is disjoint dedicated backup route sets to provide survivability for connection-oriented group communications is preferable but more research needs to be done on coordination between layers and scalable techniques within the survivability framework presented.

1. Introduction

The importance of group communications and the need to efficiently and reliably support it across a network is a very important issue for the next decade. New group communication services are emerging such as multimedia conferencing/groupware, distributed interactive simulations, sensor fusion systems, command and control centers, and entertainment-on-demand. While a succession of point-to-point unicast routes could provide group communications, this approach is inherently inefficient and unlikely to support the increased resource requirements of these new services.

Multicast mechanisms provide group communications by reducing the amount of duplicate traffic in the network to conserve bandwidth and switch resources. When a sender transmits data to multiple receivers, some delivery paths may have common links. These common links can be shared such that data is sent only once over the common links and replicated at branch points. Thus multicast is efficient because it minimizes aggregate network load by decreasing the number of links and nodes utilized in a conversation between group members. This efficiency increases with the size of the network and the number of group members.

Thus to manage large bandwidth flows and slow congestion by eliminating waste, multicasting concentrates resources to gain efficiency. However, this concentration amplifies the impact of failures such that a random fault or malicious attack will impact a larger number of users.

The connectionless IP multicast model is based on the abstract concept of multicast group addresses. A sender transmits a packet to a multicast group address and underlying layers route it transparently. Establishing a multicast tree spanning group members has emerged from the Internet community as one efficient solution for IP multicast [2,3]. There have been several types of delivery trees proposed for different IP multicast algorithms. These trees are either source-based trees, whereby the source tree is built from an active sender, or shared trees, which disseminate multicast traffic using one delivery tree spanning all group members.

As more Internet Service Providers (ISPs) implement IP multicast, there will be increasing pressure on peer networks to support IP multicast as well as supporting native connection-oriented multicast on their backbones for internal/transit/originating/terminating flows requiring Quality-of-Service (QoS) guarantees. ISPs realize that without a multicasting capability that their core backbone could become overcongested at any moment. For example, a single IP stream in one ISP's network could explode into 40,000 streams upon entering a peer's ISP network, which does not support multicasting.

Supporting connection-oriented multicast is more difficult than connectionless multicast because: (1) connectionless tree heuristics for bidirected networks do not distinguish sending nodes from receiving nodes, and (2) connection-oriented communications require setup and release which is not compatible with group dynamics of participants joining and leaving a session.

The search for a connection-oriented group communications solution began with the introduction of

* This work was supported in part by NASA Earth Systems Science grant # NGT-30019, Defense Advanced Research Projects Agency grant # F30602-97-1-0257, SAE International – The Engineering Society for Advancing Mobility Land Sea Air and Space, and State Farm Insurance.
** corresponding author; additional contact information: voice/fax 309-438-8016/5113, hard copy: Campus Box 5150, 202 Old Union, Normal IL 61790 USA.

0-7695-0975-4/00 $10.00 © 2000 IEEE

Asynchronous Transfer Mode (ATM) multicast routing capability (point-to-multipoint) in the ATM Forum's User Network Interface (UNI) specification 3.1. The ATM Forum MPOA (Multi-Protocol Over ATM) subworking group proposes two current approaches for ATM intracluster group communications: (1) the VC Mesh Model and (2) the Multicast Server Model (MCS). As a result of the perceived complexity and inefficiency of these two approaches, other techniques have been proposed such as shared trees and rings.

ATM is connection-oriented, unidirectional, cell-switched, standards-based network technology designed to handle all traffic types (i.e., multimedia, and LAN traffic) homogeneously and with guaranteed QoS in terms of delay, jitter, and loss. ATM is inherently connection-oriented, a circuit must be established (or released) through network signaling before communications can begin (or terminate). Proposed Internet schemes to provide QoS guarantees based on the end-to-end allocation of resources (e.g., RSVP) are converging connectionless IP multicast solutions into connection-oriented group communications solutions such as described in this paper.

System survivability is important due to the increasing number of information systems and society's increasing dependence on these systems for dependable service. While ATM network survivability has focused on point-to-point restoration, there has been very little work on the survivability of ATM group communications.

We focus investigation on the extension of circuit-switched restoration techniques to the survivability of connection-oriented group communications. The remainder of this paper is organized as follows: Section 2 presents a survivability framework for connection-oriented group communications. Section 3 highlights the multi-layer survivability approach for networked systems within this framework and Section 4 outlines the formulation of the survivability optimization problem specific to group communications. Section 5 introduces different survivability techniques that can be extended from circuit-switched communications for group communications and discusses empirical results from simulation experiments. We close with conclusions and future research directions in Section 7.

2. A Survivability Framework for Connection - Oriented Group Communications

Figure 1 shows a framework for evaluating overall survivability of a networked system. Given a system architecture, specified layers are designed for fault-tolerance while minimizing redundancy according to an optimization formulation. The multi-layer approach to designing survivable networked systems and the survivability optimization problem are described in detail in Sections 3 and 4 respectively. This survivability framework can be used to assess overall system survivability in terms of incremental strategies on a given system or to provide sensitivity analysis of a particular system characteristics. In general, survivability strategies can be compared on the basis of their feasibility (percentage of restorable sessions), cost (in terms of switch/bandwidth/memory resources used), and speed (restoration speed due to signaling and computational complexity). The solid arrow lines in Figure 1 refer to escalation strategies for a fault (marked by an **X**) restoration at a specific layer. The dotted arrow lines in Figure 1 refer to hierarchical downward coverage of restoration mechanisms at different layers. Networked system can be abstracted into a format suitable for simulation (matrix geometric representation of topology/demand/capacity, mathematical distributions or empirical traces for traffic generation). The state-of-the-

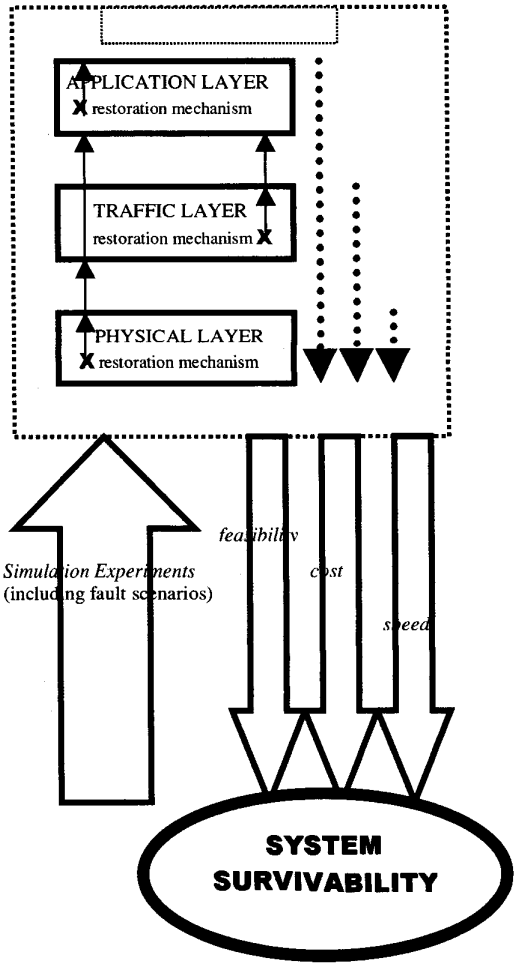

Figure 1. Survivability Framework

art in handling fault scenarios is single fault restoration; injecting multiple single faults in succession can approximate multiple fault scenarios.

3. A Multi-Layer Approach to Survivability

The survivability framework uses a multi-layered approach. Restoration mechanisms at each layer work either independently or cooperatively to handle fault scenarios as they occur.

3.1 Coordinating Layers

Coordination between restoration at the different layers will provide effective survivability by optimizing flexibility, speed, and cost. To understand the benefits of coordinating layers, tradeoffs can be characterized as follows: the physical layer has the fastest restoration; higher layers can provide restoration at lower layers; lower layers cannot restore faults at higher layers; higher layers can provide restoration with finer granularity so less redundancy is needed; layers in combination can provide the same restoration as a single layer; and different layer(s) can provide the same restoration as another layer for a lower cost (in terms of resources).

If one restoration mechanism performs well in a single layer, this does not mean it should be applied to every layer. In fact, if all the restoration mechanisms are similar at each layer, there is little to be gained and an increased system vulnerability to the potential faults of this replicated restoration mechanism. For example, if all layers used preplanned mechanisms then each layer, and the system as a whole, will not be able to handle unexpected fault events or if all layers used real-time search mechanisms then system behavior will be hard to predict. Instead it is better to provide complementary restoration mechanisms at the different layers, drawing on the benefits of each. For example, a preplanned mechanism first at a lower layer (for speed), followed by a real-time search mechanism at a higher layer to handle the unexpected faults that lower layers have been unable to restore.

3.2 Escalation Strategies

The set of rules used to decide which mechanisms to activate at which layers and when to halt mechanisms and activate others is referred to as the escalation strategy [6]. Within this context, an escalation strategy defines the coordination of restoration mechanisms in different layers to avoid contention, promote cooperation, and increase overall survivability.

There are three general classes of escalation strategies: (1) sequential; (2) parallel; and (3) managed. In the sequential escalation strategy, each layer acts in turn starting with the lowest possible layer and escalating upwards. In the simplest case, restoration is passed from one layer to another when restoration at the originating layer is exhausted and unsuccessful and/or a predefined time interval has passed. Sequential escalation is easiest to control and minimizes possible contention but may lead to the longest overall system restoration as each layer's restoration time adds to a system time sum.

In the parallel escalation strategy, each layer independently activates its own restoration mechanism when notified such that several layers will have separate restoration mechanisms activated simultaneously. When one mechanism at a layer succeeds in restoring a fault before all other layers, all other layers are signaled. Parallel escalation may be the fastest by reducing unnecessary delays but if left uncoordinated will result in contention for spare resources and possibly multiple restorations of the same fault event dependent upon the speed of signaling a successful restoration.

In the managed escalation strategy, restoration mechanisms at different layers are supervised under an integrated network management system. While possible in theory, a managed escalation strategy is likely to be slow due to the complex interpretation of distributed alarms and vulnerable to another layer of possible faults in network management software [5,6].

4. Optimization Problem

Providing survivability to group communications can be quantitatively expressed as a multidimensional optimization problem [9]. The overall goal is to make failures imperceptible to group communication users by providing adequate service continuity while minimizing the use of network resources (cost metric) and user transparency given congestion constraints, reliability threshold constraints, and restoration time constraints.

Other factors specific to group communications that must also be considered include consideration of group dynamics and scalable reliability. Many of the applications for group communications envision highly dynamic group membership in which members can join or leave a group session frequently. Protocols are needed that can handle rapid changes in connection topology and characteristics while ensuring a consistent view of the connection state at all times. As already mentioned, in contrast to connectionless IP group communications based on group addresses, connection-oriented group communications require a connection setup/release with every join/leave operation. In [8], Waxman shows that multicast trees can become increasingly suboptimal in terms of cost after a series of additions and deletions of members to a multicast group.

Reliable multicasting is concerned with providing a level of delivery assurance to multicast data streams

within an internetwork. While point-to-point protocols generally use acknowledgment packets (ACKs) sent by receivers to senders in order to guarantee reliability, extending this approach to multicast transmission means that the message is (re-)sent until ACKs from all group members are received. Receiving multiple simultaneous ACKs from group members is the "ACK implosion effect" of congestion at the source. This "ACK implosion effect" increases with the group size in terms of reverse traffic congestion at the source, source memory requirements for state information for each group/group member, and source processing to retrieve and derive state information about each group/group member. Thus the "ACK implosion effect" is the major limiting factor to providing reliable group communications on a large scale.

5. Survivable ATM Group Communications

Figure 2 shows four different implementation models for connection-oriented ATM group communications: (A) VC Mesh (VCMESH), (B) Multicast Server (MCS), (C) Shared Tree (ST), and (D) Virtual Ring (RING).

In the VCMESH, a unidirectional virtual circuit (VC) originates from each sender to all members of a group – each group member must terminate one VC for each active source in the group session. This criss-crossing of VCs across a network gives rise to the name "VC Mesh".

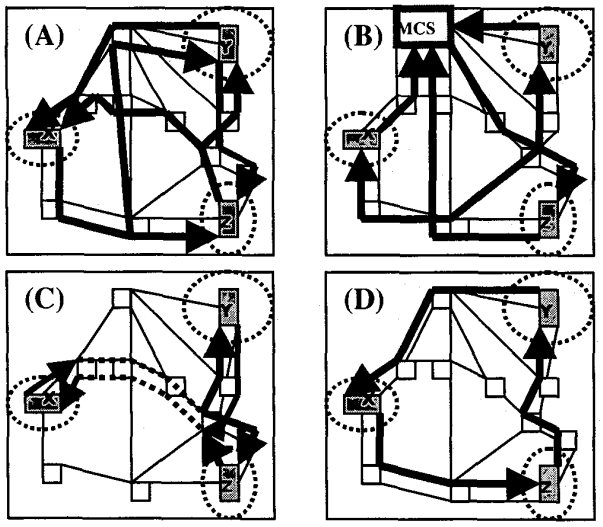

Figure 2. ATM Group Communication Models

In the MCS model, a server is chosen within each cluster to serve as a proxy for all senders in order to relieve senders from VC connection setup and release operations resulting from group dynamics. Conceptually the MCS serves as a centralized connection manager mediating and disassociating senders and receivers within a group. The MCS model is mentioned here for completeness but not included in our quantitative discussions because of the dominant single-point-of-failure vulnerability which requires a different set of restoration techniques [9].

In the ST approach, resources are reserved in both directions on all of the VC links of a shared multicast tree until the connections are released. Different ST schemes are similar in that they aim to provide a general-purpose control architecture by modifying in-band control mechanisms of ATM switches.

The RING is a unidirectional circular overlay of routes that join all desired members for group communications. Using a RING provides inherent feedback for reliability since a message send from one member of the group travels around th ring and back to the sender. In the application of a RING to ATM networks, each unicast connection between adjacent nodes on the RING can be implemented as a separate point-to-point VC.

A basic survivability concept is each group communication requiring survivability will establish a working set of routes and a disjoint set of backup routes with reserved bandwidth dedicated for restoration. When a single link or node fault occurs within a "working" set of routes, traffic flow is rerouted to the corresponding "backup" set of routes. In this preplanned method, each group communication session has guaranteed protection because when initially assigned it has a prearranged backup with bandwidth dedicated for restoration purposes.

Another approach would be for senders and receivers to dynamically search for backup routes with adequate spare capacity in real-time after being notified of a network fault. It turns out that this approach is not satisfactory for two reasons: (1) restoration is not guaranteed since no bandwidth is reserved for backup routes and residual bandwidth may not be available when a fault occurs; and (2) it is an unsolved problem that fault alarms cannot currently be processed such that signaling for detection, notification, identification, and recovery from specific fault events occurs under the required restoration time thresholds necessary for broadband ATM and IP [7].

The problem of finding disjoint end-to-end route sets is established in the point-to-point context but there has been very little work on finding disjoint route sets in the multipoint context. For the case of the VCMESH and ST, the problem is a search for a backup set of routes that is link and node disjoint such that a single fault in the "working" group will not affect the "backup" group. For the RING, bandwidth allocation for counter-rotating rings allows the upstream node of any single failure in a working ring to loop-back onto the counter-rotating backup ring and maintain connectivity (similar to SONET restoration at the physical layer).

The VCMESH specifies that each sender establishes circuit(s) to connect with all other group members but does not specify the mechanism so we have assumed the best case scenario such that each sender optimally selects the set of point-to-point/point-to-multipoint circuits that minimizes cost [11]. The ST does not explicitly state how to build the shared tree so we have assumed the best case scenario by identifying the minimum cost (Steiner) tree. Both VCMESH and ST are calculated using an implementation variant of a Steiner Tree procedure due to Lawler known as the spanning tree enumeration algorithm [1]. The minimum cost working and backup rings are calculated by an implementation of the "Disjoint Steiner Ring" (DSR) formulation [10].

Using actual network topologies, a simulation was developed in MATLAB to test the feasibility and cost of providing survivability to these different models. It is well-known that finding Steiner trees and Steiner rings is NP-complete but, depending on topology, the use of a hop-limit constraint may make an intractable problem solvable by restricting the exponential rate of growth due to combinatorics.

Results from [9,10,11] show that feasibility of restoration using the different approaches is not equal. While the RING displays 100% feasibility on the networks tested, the feasibility of the ST and VCMESH approaches is less (statistically significant decrease by 12% and 42% respectively) and rapidly decreases as the group size increases. The explanation for this difference in feasibility is the existance of "multipoint traps". The term "trap" was first introduced in the point-to-point context in [4]. A "trap" is a topology where a corresponding set of backup routes is not available due to the disjointness constraint although disjoint working and backup routes may be available if selected simultaneously and not sequentially (working then disjoint backup).

Although not always feasible on an arbitrary network, there still exist densely connected networks where VCMESH and ST can provide guaranteed single fault tolerance. In this case the question of cost in terms of link and node resources becomes important. Results from [9,10,11] show that cost of restoration using the different approaches is also not equal. The cost of the RING is lower, with statistical significance, than either the ST or VCMESH approaches with a cost differential that increases as the group size increases. The explanation for this difference in cost is that although the Steiner tree is the "best" solution for a working set of routes (minimum cost), it uses links and nodes that become unavailable for disjoint backup route sets.

These experiments simulate restoration which occurs exclusively at the traffic layer (rerouting). We are further developing these restoration mechanisms to handle multiple fault scenarios and coordination with the application layer using the framework depicted in Figure 1. Preplanned restoration appears to have addressed the need for speed to satisfy timing threshold constraints but has also introduced computational complexity requiring advanced search techniques.

6. Summary

We have presented a framework for survivable connection-oriented group communications based on multi-layered approach, the formulation of an optimization problem, and general survivability techniques extended from circuit-switched communications. It was pointed out that the these techniques are important not only to connection-oriented ATM group communications but also to connectionless IP group communications as solutions to provide QoS over the Internet are converging toward end-to-end allocation schemes very similar to connection-oriented communications.

Future research is needed to coordinate restoration mechanisms between different layers and to develop simple heuristics for disjoint routing which scale to large networked systems.

7.0 Acknowledgments

The authors would like to acknowledge the significant intellectual contributions of Deep Medhi/University of Missouri – Kansas City to this body of survivability research on group communications. We would also like to thank the anonymous reviewers whose insightful comments contributed directly to an improved paper and presentation.

8. References

[1] V.K. Balakrishnan, *Network Optimization*, Chapman & Hall, London U.K., 1995.

[2] S. Deering et. al., "An Architecture for Wide-Area Multicast Routing", *ACM Computer Communications*, Vol. 24, No. 4, pp. 127-135.

[3] S. Deering and D. Cheriton, "Multicast Routing in Datagram Internetworks and Extended LANs", *ACM Trans. on Computer Systems*, Vol. 8, No. 2, pp. 85-110.

[4] D. Dunn et. al., "Comparison of k-Shortest Paths and Maximum Flow Routing for Network Facility Restoration", *IEEE J. on Sel. Areas in Comm.*, Vol. 12, No. 1, pp. 88-99.

[5] D. Johnson, "Survivability Strategies for Broadband Networks", *Proceedings of IEEE Globecom*, 1996.

[6] L. Nederlof, et. al., "End-to-End Survivable Broadband Networks", *IEEE Comm. Mag.*, Sept. 1995, pp. 63-70.

[7] J. Sosnosky, "Service Applications for SONET DCS Distributed Restoration", *IEEE J. on Sel. Areas in Comm.*, Vol. 12, No. 1, pp. 59-68.

[8] B. Waxman, "Routing of Multipoint Connections", *IEEE J. on Sel. Areas in Comm.*, Vol. 6, No. 9, pp. 1617-1622.

[9] W. Yurcik and D. Tipper, "Survivable ATM Group Communications: Issues and Techniques", *8th Intl. Conf. on Telecomm. Systems*, 2000, pp. 518-537.

[10] W. Yurcik and D. Tipper, "Providing Network Survivability to ATM Group Communications Via Self-Healing Survivable Rings", *7th Intl. Conf. on Telecomm. Systems*, 1999, pp. 501-518.

[11] W. Yurcik, "Providing ATM Multipoint Survivability Via Disjoint VC Mesh Backup Groups", *7th IEEE Intl. Conf. on Computer Comm. and Networks (IC3N'98)*, 1998, pp. 129-136.