Copyright ©2000, all rights reserved. Proceedings of the Eighth International Conference on Telecommunication Systems, Nashville TN. USA, March 2000, pp. 518-537.1

# Survivable ATM Group Communications: Issues and Techniques

William Yurcik *

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

David Tipper †

Department of Information Science
and Telecommunications
University of Pittsburgh
*tipper@tele.pitt.edu*

**Abstract**

*This paper presents the survivability issues relevant to ATM group communications and outlines the tradeoffs between restoration techniques that can be extended from circuit-switched communications. Current models proposed by the ATM Forum for group communications (VC Mesh and Multicast Server) are compared and restoration techniques for all possible single fault scenarios are examined. It is concluded that different restoration techniques are preferable for each proposed model depending on availability and connectivity to backup servers (adjacent or remote), guarantees of single fault tolerance, restoration time tolerance, level of group dynamics, and cost. Further research is needed to quantify the tradeoffs between different restoration techniques.*

*author for correspondence; additional contact information: telephone (309) 438-8016, fax (309) 438-5113, hardcopy mail Campus Box 5150, 111 Old Union, Normal IL. 61790 USA; supported in part by NASA Earth Systems Science grant # NGT-30019, Defense Advanced Research Projects Agency grant # F30602-97-1-0257, and SAE International - The Engineering Society for Advancing Mobility Land Sea Air and Space

†supported in part by Defense Advanced Research Projects Agency Grant # F30602-97-1-0257

## 1 Introduction

The importance of supporting group communications efficiently within a B-ISDN network based on ATM has atttracted a great deal of attention. Multicast mechanisms provide group communications while seeking to reduce the amount of duplicate traffic in the network in order to conserve bandwidth and switch resources. A variety of multicast schemes have been proposed and are still under study. One issue in ATM group communications which has not fully be addressed in the literature is how to make the communications survivable in the face of network failures and attacks.

The majority of work on providing survivability for ATM networks focuses on extending circuit-switched facility restoration techniques to ATM using Virtual Paths (VPs). VPs can be viewed as semi-permanent routes through the network with dedicated bandwidth onto which Virtual Circuits (VCs) are grouped. VP restoration is simpler and faster than VC restoration and has been the focus of ATM restoration systems currently proposed in the literature. Different algorithms for ATM network survivability have been proposed using either a pre-defined backup VP for each working VP and/or dynamically searching for new paths(s) after a single link failure [15, 13, 10]. Note that the bandwidth in the predefined backup VP is reserved and not used until a fault occurs and traffic is rerouted. Several different algorithms to efficiently provision network capacity to provide the necessary reserved backup capacity at the VP layer and lower layers (e.g., transmission layer) have been proposed [21, 8, 12].

The potential risk of vulnerabilities in ATM group communications is considerably greater than point-to-point or unicast communications. Traffic is concentrated due to underlying multicast mechanisms thus increasing the impact of a failure. The probability of encountering a failure also increases since traffic traverses potentially many more physical links and nodes than typical unicast communication. Protecting against these vulnerabilities is complex since determining all the links and nodes used in communications is nonscalable due to the combinatorics of possible groups. Lastly, higher layers may require all-or-nothing restoration where a threshold quality-of-service must be achieved for all members or the group communications will be abandoned (rather than partioned).

This paper presents a comprehensive structure to describe the survivability issues and restoration techniques relevant to ATM group communications. It should be noted that many of the proposed schemes for providing Quality of Service (QoS) guarantees in next generation Internet networks (e.g., Intserv, MPLS, RSVP) are converging connectionless IP multicast mechanisms into connection oriented multicast solutions such as discussed here. The remainder of this paper is organized as follows: Section 2 states the objectives of survivable group communications; Section 3 introduces ATM group communication models proposed by the ATM Forum; Section 4 discusses general survivability issues relevant from circuit-switched communications along with restoration techniques to address these issues and their corresponding tradeoffs; Section 5 examines specific restoration for the proposed ATM group communication models for all possible fault scenarios; and Section 6 closes with a summary of conclusions and future research directions.

## 2 Objectives of Survivable Group Communications

The objectives of ATM multipoint survivability are the same as in the ATM point-to-point scenario: to make failures imperceptible to group communication by providing adequate service continuity while minimizing the use of network resources and user transparency given congestion constraints, reliability threshold constraints, and restoration time constraints. These issues are summarized in Table 1.

Minimizing the use of network resources and other financial costs occurs when the number of physical links and switches needed for survivability is minimized and the aggregate capacity provisioned for both working and backup bandwidth demand is minimized. While a robust network can be built with a fully-connected topology and enough spare capacity to handle almost any arbitrary failure, the challenge is to provide an acceptable level of service for the least cost.

Adequate service continuity means that connectivity, a minimum required bandwidth, and guaranteed QoS is continuously available to specific services despite a failure that has left part of the network unavailable. These specified services may be clearly identified essential services (i.e., E911) or any service that may be willing to pay for survivability priority. The network-wide issue of providing service continuity to equal-priority users is addressed by techniques that either maximize the amount of demand restored or minimize the amount of interrupted connections that are not able to be restored (e.g., connections that cannot be rerouted around a failure and are thus dropped or blocked).

User transparency is ensuring that the failure of any network component affects only those network resources directly associated with that component. The reassignment of established connections in order to optimize network resources, accomplish restoration, or accommodate group dynamics should only be performed if it can be transparently accomplished such that is imperceptible to established users. In point-to-point circuit-switched communications, user transparency is often referred to as a "hit" (a temporary reconfiguration of service due to reframing which is negligible).

Congestion constraints refer to a congestion magnitude and duration below relevant service/network outage thresholds and a region of congestion contained to an acceptable area. A major factor of network performance after a failure is the transient congestion period that results from restored circuits attempting to send out a backlog of traffic accumulated for retransmission since the failure [22, 6]. In [22], the authors study alternative routing schemes to control such congestion after a failure.

Reliability probability threshold constraints refer to the calculated probability of specific services being non-disrupted over a given time interval or the specific connectivity of the network as a whole is above the required reliability threshold given probability of failure distribution information.

Although 100% survivability is the optimal goal, a lower level of survivability may be acceptable if the reliability probability is acceptable.[4]

Restoration time constraints refer to the total time for the network to restore adequate service continuity below service/network outage duration thresholds. In [20], Sosnosky provides a summary of outage duration impacts on services/networks. This total time of restoration includes the following timing stages: detection that a failure has occurred; notification of the failure to pertinent network elements (destination/source nodes); identification of the underlying failure event given multiple layers of failure alarms; and the time for rerouting connections around the failure. Each of these timing stages represents a significant challenge.

Other issues specific to group communications that must also be considered include consideration of group dynamics and scalable reliability. Many of the applications for multipoint connections envision highly dynamic connections in which nodes are added and deleted from a group 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. In [23], Waxman shows that multicast routing 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 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.

[4]reliability is defined as the ability of a system to perform its required functions under given conditions for a given time interval; survivability is defined as the ability of a system to perform required functions at a given instant in time after a subset of components of the system becomes unavailable.

## 3 Proposed ATM Group Communication Schemes

The two current models proposed for intracluster multicast by the ATM Forum MPOA (Multi Protocol Over ATM) subworking Group are: (1) the VC Mesh, and (2) the Multicast Server (MCS). A comparison of VC Mesh and MCS is summarized in Table 2. The choice of either model may be made on a per-group basis and is transparent to ATM users.[5]

In the VC Mesh Model, unidirectional VCs originate from each sender to all members of a group. Each group member must terminate one VC for each active source in the cluster. Each sender is essentially the root of its own source-based tree. If a member joins or leaves the group, each point-to-multipoint VC for each sender in the group must be modified (a signaling load that scales with the number of senders in a group). In the ATM Forum UNI 3.1 specification, the sender keeps track of all members in the connection in all cases, only the sender can add a new group member and either the sender can delete a member or the member can delete itself from the point-to-multipoint VC.[3] The ATM Forum UNI 4.0 specification allows a new member to join a point-to-multipoint VC without the intervention of the sender by using a Leaf-Initiated Join (LIJ) capability.[2] Receivers discover the existence of point-to-multipoint VCs through a directory and are able to signal a request to join a particular point-to-multipoint VC. LIJ does not specify where the new member should be joined to the existing point-to-multipoint VC. For example, the new member (leaf) may join at the nearest node which is already part of the point-to-multipoint VC or it may join via the shortest path to the sender (root).

Figure 1 shows an example of group communications using the VC Mesh Model. For a single group containing multiple senders, unidirectional point-to-multipoint/point-to-point VCs are established from each sender to all receivers with no coordination between VCs except for capacity allocation on shared links between senders. For multiple groups on the same network, the VCs for each sender are handled independently except for capacity allocation on shared links between groups. This criss-crossing of VCs across a network gives rise to the name - "VC Mesh".[1]

In the MCS Model, a server is chosen within each cluster to serve as a

[5]An ATM user is broadly defined as any entity that establishes and manages VCs and underlying ATM Adaption Layer (AAL) services to exchange data via a User-Network Interface (UNI).

proxy for all senders in order to relieve senders from direct 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. [25] All senders establish a unidirectional point-to-point VC to the MCS. The MCS establishes a unidirectional point-to-multipoint VC to the rest of the multicast group. During transmission, the MCS reassembles cells into packets arriving on all incoming VCs (avoiding interleaving cells from different packets) and queues cells belonging to particular packets for transmission on an outgoing point-to-multipoint VC.[6] The basic MCS architecture only supports one group at a time, reflecting all incoming cells to the outbound point-to-multipoint VC. A more complex MCS architecture can support multiple groups simultaneously with the addition of minimal layer 3 specific processing which allows the MCS to look inside received cells to determine which group and thus which point-to-multipoint VC to send the cell. All requests to create a group, delete a group, to add a member to a group, or delete a member from a group are sent to the MCS which maintains state information for all existing groups.

Figure 2 shows an example of a group with a single sender using the MCS Model. For a single group containing multiple senders, a point-to-point VC is established from each sender to the MCS and the same point-to-multipoint VC from the MCS to the group's receivers is shared (with capacity allocation on shared links between senders on the sender-side of the MCS). For multiple groups on a network sharing the same MCS, each group will have independent point-to-multipoint VCs from the MCS to their group's receivers and senders will have independent point-to-point VCs to the MCS (with capacity allocation on shared links between senders on the sender-side of the MCS and capacity allocation on shared links between groups of receivers on the receiver-side of the MCS).

## 4 Survivability Issues and Tradeoffs

There are general survivability issues relevant from circuit-switched communications which are applicable for ATM group communications. Choices between techniques to address these issues represent the following tradeoffs:

[6]Reassembly of incoming AAL_SDUs is required at the multicast server as AAL5 does not support cell level multiplexing of different AAL_SDUs on a single outgoing VC. [1]

**location** Two locations for initiating restoration are considered: (1) end-to-end restoration initiated between sender and receivers, or (2) local restoration initiated between the upstream and downstream nodes next to a fault. In general, the closer restoration is initiated to a fault offers greater speed and ease of implementation due to a smaller search space but may also result in blocking (non-restoration) due to the smaller set of possible routes. Local restoration may also result in a back-haul situation where an optimal route is identified between upstream and downstream nodes next to a failure, but this route may not be disjoint with the remaining end-to-end route beyond the nodes next to the fault and thus result in a non-optimal global solution. In general, the further restoration is initiated away from a fault results in slower restoration due to a larger search space (more possible routes), globally optimal solutions, and avoidance of back-haul situations. Results in [7] verify that end-to-end restoration methods are always more capacity efficient than local restoration methods but that the significance of this advantage depends upon network load (low network load decreases the capacity efficiency advantage of end-to-end over local restoration). In general, end-to-end restoration is well-suited for networks with higher reliability constraints, higher network loads, and lower connectivity while local restoration is well-suited for networks with lower reliability constraints, lower network loads, and higher connectivity. In the ATM context, local restoration is available only if VPs/VCs are concatenated together to form an end-to-end route such that an individual VP/VC within a route can perform restoration which appears as a local restoration to the entire route. A single VP or VC that extends over multiple hops will perform end-to-end restoration between sender and receiver unless a link/node fault and subsequent recovery is handled entirely at a lower layer transparent to ATM (i.e., SONET ring rerouting at the transmission layer).

**backup route** There are two basic methods of identifying a backup route for restoration: (1) a preplanned route can be computed in advance or (2) a dynamic search for a backup route can be computed in real-time. In the preplanned method, each working VP/VC has a prearranged backup VP/VC when initially assigned. In the dynamic search method, senders and receivers broadcast signaling messages after being notified of a network failure and backup VPs/VCs are selected via a specified algorithm. In the preplanned method, the backup VP/VC has reserved bandwidth that remains dedicated for restoration purposes. In dynamic search, no bandwidth is reserved for backup VPs/VCs so restoration is not guaranteed (i.e., residual bandwidth may not be available). The preplanned method is faster and dependent only on the number of VPs/VCs to be rerouted while the dynamic search method is slower and is dependent on network conditions (i.e., traffic loads, network topology, fault location, ATM signaling) in addition to the number of VPs/VCs to be rerouted. The preplanned method incurs a significant setup delay for reassignment of working and backup VPs/VCs as members join and leave a group. The dynamic search method does not need to update for group membership dynamics since the only group membership to be considered is when a fault occurs.

**circuit type** The two circuit types under consideration for restoration are (1) point-to-point circuits and (2) point-to-multipoint circuits. The setup of a point-to-point circuit is typically faster: 75-100 ms/hop in a large WAN environment.[7] The set-up of a point-to-multipoint circuit is typically slower: 200-300 ms/hop/receiver in a large WAN environment.[8] A point-to-point restoration usually involves only two end-nodes (the sender and the receiver) and a single route. A point-to-multipoint restoration by definition involves more than two end-nodes (a sender and multiple receivers) and multiple routes. A point-to-point restoration may preserve user transparency because the only receiver inherently affected is the disconnected node but other receivers may have to throttle while the disconnected receiver is restored. A point-to-multipoint restoration may not preserve transparency because the subset of disconnected receivers may not be the only receivers affected; at some point receivers that have continuous connections might need to be interrupted for restoration. To provide guaranteed single fault tolerance with point-to-point circuits requires more network resources: multiple backup point-to-point circuits with no sharing of link capacity. To provide guaranteed single fault tolerance with point-to-multipoint circuits requires less network resources: a single backup point-to-multipoint circuit.

[7]these setup times are for switched VCs measured as the elapsed time between the request or establishment of a connection at either the application or signaling software levels (both levels are considered in [14]).

[8]The reason for the wide difference in point-to-multipoint versus point-to-point set-up time is not yet determined but may be due to complex routing, extra indirection in UNI, and serialization of calls to multiple receivers. [14]

**buffering** Buffering is necessary in group communications for three reasons:

1. all cells belonging to a packet must be stored until the entire packet is received
2. reducing delay variation between widely dispersed group members
3. ensuring reliability of group communications by providing a re-transmission capability upon feedback from receivers

In [19], Rouskas examines the tradeoffs of buffering at the sender, within the network, and at the receivers. Buffering at the sender and within the network defeats the purpose of multicast mechanisms because cells can be buffered for different amounts of time at each location resulting in multiple unicast connections from a single multicast connection [19]. Buffering at receivers will allow all cells to be stored until an entire packet is received, can counterbalance delay variations resulting from receivers with routes of different delay characteristics, and retransmission capabilities for reliability can be directed toward other receivers (which also employ buffering) rather than the sender. The amount of buffering required is proportional to the larger of the maximum delay variation among the receivers or the retransmission window for reliability. There is a danger in relying on receivers to buffer information that may provide a competitive advantage so the development of a synchronized cryptographic mechanism based on public key infrastructure is needed.

**priority** Fault management signals are communicated in ATM using cells referred to as Operation-And-Management (OAM) cells. The entire restoration process builds upon the information relayed by these ATM OAM cells.[10] OAM cells are given no special priority and may be queued or dropped behind cells of an application or service that has been disrupted by the very same fault which initiated transmission of

[10]For example, a Loss-Of-Signal (LOS) alarm is declared by a receiving switch that no longer detects any signal transmission. The LOS alarm continues until it is manually suppressed or the problem is fixed. Detection of an LOS alarm prompts Alarm-Indication-Signal (AIS) OAM cells to be transmitted downstream toward the destination receivers. The AIS OAM cells replace the normal traffic signal and preserve transmission continuity while notifying the rest of the circuit that there is an alarm situation. Once notified, the destination receivers transmit Remote-Defect-Indication (RDI) OAM cells upstream toward the sender. The RDI OAM cells serve as notification of a fault but do not specifically identify the link where the fault occurred. [5]

the OAM cells in the first place (e.g., a deadlock situation). The only indication of priority in the ATM architecture is the cell-loss-priority (CLP) bit which is carried within the cell header. While this bit may be set by end systems, it is more likely to be set by the network. However, fault signaling is not the only signaling with a need for priority. In [14], the authors state; "The most significant contribution to the call setup latency is caused by a lack of prioritizing signaling message processing." ATM switches typically implement different priority levels for different service categories in order to effectively support multiple traffic categories. In [9], the authors propose a restoration priority scheme which aims to minimize the impact of a failure on the network while providing users the best possible service. The scheme involves both a priority for reconnection among ATM service classes and a rule for ordering and routing VCs within a service class. Schemes are also needed to specify how failure information should be transmitted (with priority) within an ATM network and across ATM intercarrier interfaces [4]. Of course, priority schemes are pure tradeoffs and any performance enhancement from their use in certain areas will result in a performance degradation in other areas.

**layers** In [11, 16] the authors present multi-layer restoration strategies and their interrelationships. The authors in [11] divide an ATM network into two layers: (1) the transmission layer and (2) the traffic layer. At the transmission layer restoration is provided by protection switching and self-healing SONET rings. At the traffic layer, restoration is provided by preplanned backup VPs and dynamic searching for backup VC routes. The authors also point out the tradeoffs between restoring at different layers: the transmission layer has the fastest restoration; higher layers can restore faults at lower layers; lower layers cannot restore faults at higher layers; 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. Providing ATM group communications introduces the need to consider an additional layer for restoration above the traffic layer: the application layer. With sender and receivers coordinating VC or VP multicast connections at the traffic layer, there is an opportunity to coordinate restoration at a higher and thus more flexible layer. While ATM multicast functionality at the traffic layer can take several forms (i.e., multicast trees or rings), users at the application layer of a group communication have the independent capability to join and leave different groups, split groups, form redundant groups, form entirely new groups, delete current groups, merge current groups, and/or a multitude of other group dynamics. These group dynamics at the application layer provide an opportunity for restoration that can be coordinated with other layers to provide tradeoffs in terms of cost and flexibility.

### 4.1 Survivability of Proposed Models: VC Mesh and and Multicast Server

The VC Mesh Model has a number of survivability issues that need to be addressed. In the VC Mesh Model, group dynamics drive the establishment of new VCs (combination of point-to-point and point-to-multipoint) for each sender in a group to reflect updated membership after each join or leave operation.[11] Also in the VC Mesh Model, multiple senders transmitting to the same group members use separate point-to-multipoint VCs while in the MCS Model multiple senders each establish point-to-point VCs to the MCS and then the MCS has the capability to integrate point-to-multipoint transmissions to the same group members. For both of these reasons, the number of VCs utilized in the VC Mesh Model increases with the number of senders at a faster rate than the MCS Model. If all group members are active senders at some point over the lifetime of a particular group, then the resulting paths will be a fully-connected VC mesh. As a result, there is potentially a larger number of routes over which a failure may occur for the VC Mesh which means an increased exposure to link and node failures.

Other survivability issues specific to the VC Mesh Model include increased signaling traffic due to more VCs and increased termination requirements (end hosts need one VC termination per source per each ATM interface). These issues both represent performance limitations due to lack of scalability. As group sizes grow, providing survivability becomes more difficult due to the increased capacity required for signaling and the exhaustion of end host ports due to previous sender transmissions.

The MCS Model has a number of survivability issues which make it an unattractive solution if not addressed. Table 2 highlights that although

[11] applies to ATM Forum UNI 3.1 specification without LIJ capability; the intention of the LIJ capability is to avoid all parties being aware of group dynamics, but LIJ may have a similar property depending on implementation

MCS constrains the number of potential paths over which a failure may occur, the MCS accomplishes this by concentrating traffic on a server which is both a performance bottleneck and single point-of-failure for a group's communication. If an MCS is disabled, all group sessions using that MCS terminate.

Other survivability issues specific to the MCS Model include placement of the MCS within a cluster topology; MCS scalability (buffering and processing constraints); QoS guarantees (i.e., delay, delay jitter, cell loss); and dependent faults (i.e., a single fault affecting more than one MCS).

Survivability issues common to both the VC Mesh and MCS Models: (1) due to group dynamics, it can be complex to identify and thus protect links/paths that will be used for group communications; (2) since it is complex to identify utilized links/paths in advance, this makes provisioning capacity for multipoint restoration difficult especially given a network of multiple groups with different bandwidth requirements; and (3) group access control such that only members of a group can send-to or receive-from the group, both the VC Mesh and MCS Models require external mechanisms to identify group members. The failure of individual senders (or receivers) within a group will degrade but not terminate group communications if there are other senders (or receivers) still operating. This paper will not be addressing strategies to recover from individual sender or receiver failures but rather focus on the survivability of a group's communication.

# 5 Restoration Techniques

## 5.1 VC Mesh Restoration Techniques

In the first paper on the survivability of ATM group communications [24], Wu et. al., focus exclusively on the VC Mesh Model and describe several pre-defined link and path restoration schemes to reconnect nodes to an established point-to-multipoint circuit. In one scheme, every working link in a multicast tree has a predefined backup path from its upstream node to provide restoration. In a second scheme, every working link in a multicast tree considers a predefined backup path from other potential nodes of the multicast tree (in addition to the immediate upstream node) to provide restoration. Other path restoration schemes are also described where point-to-multipoint VPs are treated as a combination of point-to-point working VPs that have backup VPs in which a backup VP may be disjoint from its corresponding working VP and independent backup VPs may share link capacity. In a comparison using bandwidth and node involvement performance metrics, the authors show that the link restoration scheme in which every working link in a multicast tree considers a predefined backup path from all nodes of the multicast tree provides the best performance from among these schemes.

Figure 3 shows a concise summary of single fault restoration scenarios of the VC Mesh Model. The restoration techniques can be categorized by location (end-to-end or local restoration), circuit type (point-to-point or point-to-multipoint restoration), and backup route (preplanned backup or dynamic search). Link capacity sharing between working and backup VCs or among backup VCs from disjoint working VCs is discussed separately.

Figure 3A shows an end-to-end, point-to-point, preplanned restoration scheme. This scheme requires an end-to-end disjoint backup VC be identified and reserved when each receiver joins the group. Restoration can be fast because it is preplanned but it may be slowed due to fault signaling that needs to travel end-to-end (receiver to sender). Backup VCs for different receivers can share link capacity but this may lead to receivers contending for the same backup link capacity depending on the location of the fault. If a single fault occurs beyond a branch point to a single receiver then link capacity sharing between multiple backup VCs can provide protection to different receivers. If multiple receivers are disconnected then backup VCs sharing link capacity may result in non-restoration (blocking) of some receivers. Although multiple receivers can be disconnected by multiple faults, this paper focuses on the single fault case where a single fault occurs in the "trunk"[13] part of the point-to-multipoint VC. If the fault occurs in the trunk then every receiver will simultaneously need their own disjoint point-to-point VC for restoration. In summary, the end-to-end, point-to-point, preplanned restoration scheme is fast, suited for infrequent single receiver disconnection events, and costly to provide single fault tolerance (due to the possibility of "trunk" faults).

Figure 3B shows a local, point-to-point, preplanned restoration scheme. This scheme requires additional backup point-to-point VC(s) with disjoint

[13]For the purposes of this paper, the term "trunk" is borrowed from multicast tree terminology to refer to the part of the point-to-multipoint VC which is shared between different receivers

route(s) be reserved between an upstream node (Figure 3B shows an upstream "branch node") and the receiver when each receiver joins the group. Restoration can be fast because it is preplanned and fault signaling is local (between the receiver and an upstream node) but it may be slowed if multiple backup VCs are attempted before one can be established upstream of the fault location. Backup VCs for different receivers can share link capacity; however, this scheme does not provide protection for a single fault event unless each receiver has a disjoint backup point-to-point VC upstream of all potential fault locations. Having a separate backup VC corresponding to each potential fault location is possible but very expensive. In summary, the local, point-to-point, preplanned restoration scheme is fast, suited for single receiver disconnection events, and allows link capacity sharing, but can only provide guaranteed restoration from a single fault at a high cost.

Figure 3C shows a local, point-to-point, dynamic search scheme. In this scheme when a receiver is disconnected, a dynamic search occurs to identify a point-to-point VC to the nearest non-disconnected node in the point-to-multipoint circuit. It may be the case that the nearest node in the point-to-multipoint circuit will be either disconnected itself by the same fault event or a route with the necessary residual bandwidth for a VC may not be available. In this case the next nearest node in the point-to-multipoint circuit is considered and so forth. Dynamic search is necessary because the location of the fault determines eligible reconnection nodes and it would be very costly to reserve preplanned backups for the entire multitude of fault location possibilities. Since no preplanned VCs must be identified and reserved when a receiver joins or leaves a group, this scheme has better performance under increasing group dynamics. This scheme is slow because of the signaling involved in a dynamic search and the worst case scenario of a trunk fault near the sender. If a fault occurs in the trunk then this scheme tends toward the end-to-end, point-to-point. preplanned VC restoration scheme in Figure 3A (as the fault occurs closer to the sender). This scheme also uses significant network resources due to signaling and may not preserve transparency because other receivers can be disrupted due to the dynamic search processing load. In summary, the local, point-to-point, dynamic search scheme is slow, relatively costly, well-suited for frequent group dynamics, and cannot guarantee single fault tolerance since the residual bandwidth for a restoration VC is not reserved and may not be available when needed.

Figure 3D shows an end-to-end, point-to-multipoint, preplanned restoration scheme. This scheme requires a new end-to-end disjoint point-to-multipoint backup VC be identified and reserved from the sender to all receivers when each receiver joins or leaves the group. Restoration can be fast because it is preplanned but it may be slowed due to fault signaling that needs to travel end-to-end (receiver to sender). Since a new disjoint point-to-multipoint VC must be identified and reserved when a receiver joins or leaves a group, this scheme has significantly worse performance under increasing group dynamics given the slow setup times for point-to-multipoint circuits. This scheme has three unique properties: (1) it allows for maximum sharing of link capacity among all receivers while still maintaining single fault tolerance; (2) its comparative efficiency to other schemes improves as the location of a fault tends closer to the sender; and (3) it can provide restoration for multiple receivers simultaneously regardless of single fault location. In summary, this end-to-end, point-to-multipoint, preplanned restoration scheme is fast and performance degrades with increasing group dynamics but it is guaranteed to provide single fault tolerance, well-suited for trunk faults and other multiple-receiver disconnection scenarios, and cost-effective due to the cost-decreasing effects of link capacity sharing.

Figure 3E shows an end-to-end, point-to-multipoint, preplanned restoration scheme very similar to Figure 3D but with one difference: the working VC and the preplanned backup VC are not disjoint. Figure 3E shows a link fault in the trunk which is restored by a point-to-multipoint backup VC which has every link in common with the working VC except for the link where the fault occurs. To provide such single link fault tolerance, there needs to be a separate point-to-multipoint backup VC for every link in the working point-to-multipoint VC. Each of these separate point-to-multipoint backup VCs will have every link in common with the working VC except for a disjoint route around the one link the backup VC is protecting. Any decrease in network resource costs due to sharing of link capacity between working and backup VCs is counteracted by the number of backup VCs required to provide single link fault tolerance. Single node fault tolerance can also be provided with this technique but not in all cases. For single node fault tolerance there needs to be a separate point-to-multipoint backup VC for every node in the working point-to-multipoint VC. Figure 3E is an example where single node fault tolerance cannot be provided with this technique because a failure of the branch node defaults the restoration to the point-to-point technique of Figure 3B. Thus this scheme can provide a single link fault guarantee and not a single node fault guarantee. This scheme will need to allocate exactly twice the working capacity for single link fault tolerance (in the simplest case every link shares a capacity allocation of 50% working capacity and 50% backup capacity). Restoration can be fast because it is preplanned but it may be slowed due to fault signaling that needs to travel end-to-end (receiver to sender). Since a new end-to-end, non-disjoint, point-to-multipoint VC must be identified and reserved for each new link in the working VC as receivers join and leave a group, this scheme has significantly worse performance under increasing group dynamics. In summary, this end-to-end, point-to-multipoint, preplanned, non-disjoint restoration scheme is fast, moderately costly, performance degrades with increasing group dynamics, and is only guaranteed to provide only single link fault tolerance (not single node fault tolerance).

## 5.2 MCS Restoration Techniques

The dominant survivability issue in the MCS Model is the single-point-of-failure vulnerability of the MCS. ATM uses statically-configured routes to an identified MCS without a dynamic discovery mechanism for other MCSs. Thus the loss of a primary MCS results in a catastrophic event isolating all senders and receivers. Three techniques to addressing this vulnerability are: (1) adjacent server replication at the traffic layer, (2) a backup scheme of primary and secondary servers at the traffic layer, and (3) redundant groups at the application layer. Different survivability techniques are appropriate depending on the availability and connectivity of redundant servers.

Replication of a primary MCS transparent to group members can be accomplished to a different environment (i.e., different physical locations and/or software environments) and involves the corresponding overhead necessary to keep all state information current and consistent as well as the traffic transmission requirements between the primary MCS and its replication (mirror MCS). Unlike the rerouting of group members to a secondary MCS in a remote backup scheme approach, a mirror MCS may assume the role of a failed primary MCS without any VC disconnects or performance degradation.

In the remote backup scheme approach, the failure of a MCS is handled by a secondary MCS taking over the role of the primary MCS. After a failure in the primary MCS occurs, all group members reroute new connections via preplanned disjoint VCs to a secondary MCS and proceed with connection establishment to rebuild the group. Unlike the mirror MCS(s) in the adjacent backup server approach, in the remote backup scheme the secondary MCS does not need to maintain state information on groups or group members.

At the application layer the same group can associate with multiple MCSs such that traffic from one MCS is processed at receivers while traffic from other MCSs are dropped. However, creating redundant groups on different MCSs will multiply the use of finite network resources (MCS processing and ports, intermediate switch processing, receiver processing, and link bandwith). On the other hand, this technique is flexible because a group can continue despite lower layer faults if there exists one redundant group connected to an operational MCS. This paper will focus on MCS restoration at the traffic layer.

### 5.2.1 MCS Restoration with an Adjacent Backup Server

Subject to the availability of appropriate backup servers at the network and the availability of appropriate customer-provided connections between servers, a backup server can be directly connected to the primary server via a dedicated connection isolated from the network. Given adjacent backup servers directly connected to the MCS, two survivability techniques for a primary MCS are multihoming and replication for the following reasons: (1) a multihomed MCS has a only a limited number of network access nodes to which to connect, thus allowing a mirror MCS to be created with a feasible number of additional links to the network access nodes and one link to the primary MCS; (2) traffic between the primary MCS and a mirror MCS does not affect network traffic or the bottleneck network access points because it does not share link capacity on the network but rather is transmitted over an dedicated link isolated from the network; and (3) all routes to and from an MCS must use the multihomed network access nodes which limits the number of routes and state information to be shared between the primary MCS and mirror MCS. The concept of "adjacency" between a primary MCS and backup MCSs refers only to the direct dedicated connection between them for the exchange of state information and does not necessarily imply co-location. An adjacent backup MCS can be located at a distant location as long as there exists a direct dedicated connection with the required available capacity and QoS guarantees. In fact, for purposes of security it is desirable that adjacent backup servers not be co-located to protect from a physical attack as well as adjacent backup servers having diverse software environments to protect from exploitation of the same software flaws.

A sender that wants to multicast to a group establishes a point-to-point VC to the MCS. Installing an MCS provides customer premises control and user simplicity since the MCS conceptually is a virtual meeting point. When considering performance, however, incoming/outgoing VCs may congest the bottleneck network access points (links/nodes) which attach the MCS to the network. This bottleneck can be alleviated with "multihoming". The term "multihoming" refers to the attachment of an MCS device to more than one access node of a network. In addition to providing an automatic backup route if there is a link or node fault on either attachment to the network, multihoming improves a number of other survivability characteristics. Multihoming an MCS at the edge of a network can protect against network partitioning, provide load balancing capabilities, and simplify multicast routing. [17] On the other hand, even a multihomed MCS attached at the network edge will have decreased survivability due to traffic concentration at the access nodes to the network and the limited number of disjoint routes which must include the access nodes. Another common use of the term "multihoming" refers to associating (or binding) multiple addresses and names to a single interface such that a server with a single interface can host multiple independent sessions. Multiple-binding can provide user transparency and network management flexibility but there is a direct tradeoff with network survivability such that an upper limit on the number of network addresses/names per interface should be set.

Replication is the maintenance of multiple copies of primary MCS state information at "mirror" MCS(s). The mirror MCS(s) should have independent failure modes from the primary MCS. One method to accomplish replication is to directly connect a mirror MCS in standby mode (aka hot standby) to the primary MCS such that if the primary MCS fails, a mirror MCS may assume the role of the failed primary MCS without any VC disconnects or performance degradation. This method is known as "passive" replication. The standby MCS can process all requests to the primary MCS in parallel to update itself or the primary MCS can perform all processing and update the standby mirror MCS via messages. Update messages between the primary MCS and mirror MCS(s) can take the form of a complete refresh (all state information) or a quick refresh (only changed state information). Because messaging and updating state information in the mirror MCS(s) takes a finite amount of time, this form of replication is inherently asynchronous.

Another method of replication is "active" replication where each mirror MCS receives and processes (in parallel) the same sequence of requests as the primary MCS and compare their request processing results with the primary MCS to vote on the result. In this way active replication protects against Byzantine failure scenarios[14] since mirror MCSs provide multiple results from which to discard random or malicious results and agree on the same majority result. Active replication may be implemented such that request processing can be shared by mirror MCSs which will improve performance in terms of response time.

The main requirement for both active and passive replication is replication transparency such that group members are not aware of multiple MCSs. There are also issues of data consistency between the primary MCS and mirror MCS(s) which have been addressed by a large number of replica control algorithms. The system of primary and mirror MCSs must ensure *single-copy serializability*, which means that requests execute as if there were a single copy of the data and the requests appear to execute in some serial order. Ensuring all mirror MCS(s) receive the same sequence of requests as the primary MCS can be addressed by multihoming each mirror MCS to the same network attachments as the primary MCS and comparisons with the direct link between the primary MCS and mirror MCS(s). While increasing the degree of replication (i.e., number of mirror MCSs) will increase survivability, the accompanying increase in messaging and/or processing overhead will decrease performance such that an optimum degree of replication can be determined for specific scenarios.

Figures 4 shows single fault restoration scenarios for a multihomed MCS device with an adjacent backup MCS. A failure which disrupts incoming VCs using one network attachment will be restored via a disjoint attachment to another network access point since each MCS is multihomed. Figure 4 also shows a fault in the primary MCS which is restored by an adjacent backup MCS (mirror MCS) with the same multihomed network at-

[14] "Byzantine failures" exhibit arbitrary error states and/or malicious error states which may involve deliberate deception such as sending contradictory messages to different recipients or impersonating one another; this contrasts with "fail-stop failures" in which a component stops after changing to a state in which its failure can be unambiguously detected

tachments. The adjacent backup MCS has all state information from the primary MCS and initiates recovery of all groups by either maintaining established VCs from itself to each group member (if possible) or restoring failed VCs by attempting to setup another VC from itself to each group member using the same previously established route.

#### 5.2.2 MCS Restoration with a Remote Backup Server

It may not be possible to directly connect a backup MCS to a primary MCS via a dedicated connection isolated from the network; there may be no available dedicated connections between the backup MCS and a primary MCS. In this case, a backup MCS available via the network may serve as a remote backup MCS as shown in Figure 5.

The appropriate survivability technique for a primary MCS and a remote MCS available over a network is a remote backup scheme of primary and secondary MCS(s). Multihoming and replication are not appropriate survivability techniques for a remote backup scheme with a backup servers accessed over the network for the following reasons: (1) maintaining current and consistent state information in two edge devices connected over a network is difficult due to the inherently larger delay (due to queueing at intermediate switches), delay jitter variation, increased cell loss ratio, and security vulnerabilities; (2) traffic between the primary MCS and the remote backup MCS has to share capacity on the network adding to network traffic congestion; (3) depending on the relative locations of the primary MCS and its remote backup MCS within a network, route information from the primary MCS may not be accurate route information for the remote backup MCS due to its different location within the network.

In a remote backup scheme approach, the failure of a MCS is handled by a secondary MCS taking over the role of the primary MCS as shown in Figure 5. After a failure in the primary MCS occurs, all group members reroute new connections via preplanned disjoint VCs to a secondary MCS and proceed with connection establishment to rebuild the group. Unlike the active replication approach, in the remote backup scheme the secondary MCS does not need to maintain state information on routes, group members, or groups. The remote backup scheme approach is sender/receiver driven - in the event of a fault to the primary MCS, all senders and receivers react by independently initiating rerouting over preplanned routes. This distributed sender/receiver driven aspect of the remote backup scheme approach is in direct contrast to the centralized adjacent backup MCS approach which relies solely on the adjacent MCS to maintain connections in the case of a primary MCS fault.

The first step in the remote backup scheme is for each group member to identify eligible secondary MCS(s). One way to implement this is for each member to consider all potential secondary MCSs known to the primary MCS at setup time when a new member requests to join a group or when an established member leaves a group. Each group member computes whether each potential secondary MCS is eligible as a backup given the requirements of a disjoint path to establish a VC, bandwidth, and QoS guarantees such as a delay bound. With a specified hop count limit to guarantee a delay bound, each group member will need to compute the residual bandwidth to each secondary MCS and the residual bandwidth from each secondary MCS to the most distant group member (i.e., the group member which has the maximum delay from the particular secondary MCS under consideration). If the residual bandwidth to a secondary MCS is available and the computed delay using the secondary MCS provides the guaranteed QoS then the group member will be said to find this secondary MCS *"eligible"*. Thus at setup time, each group member will identify eligible secondary MCSs.

It is not sufficient that each group member independently identify eligible secondary MCS(s); the second step is that all group members must decide upon the <u>same</u> secondary MCS for their group communications. A checkpoint (or recovery point) is established at specific operation points during normal group dynamics: (1) after the setup time when a new member joins a group or (2) after an established member leaves a group. At these checkpoints all information needed for recovery is processed and a single secondary MCS is selected and communicated to all group members (replacing any previously selected secondary MCS). If the primary MCS fails, all group members will be routed to the secondary MCS identified at the last checkpoint. This global synchronous checkpoint approach is a "pessimistic" recovery mechanism as opposed to the "optimistic" recovery mechanism of asynchronous checkpointing (each receiver checkpointing independently).

If there are multiple eligible secondary MCSs to all group members then either one MCS is arbitrarily chosen or distance metrics can be used to select one.[15] In the case that there is no eligible secondary MCS to all group members, a secondary MCS that it will restore the maximum number of group members is selected. When there are no eligible secondary MCS for all group members such that 100% MCS survivability is not guaranteed, a warning should be signaled to the network manager and group members. This implementation to provide MCS survivability can be extended beyond a secondary MCS - each group can have multiple prioritized backup MCSs (for the case of multiple failures) and one MCS can be designated as the global "Default MCS" for the network if no prioritized backup MCS(s) are available due to fault conditions.

For completeness, there is a Byzantine failure scenario which could lead to different group members acting on different secondary MCS information. This may occur when either a faulty primary MCS intermittently transmits inaccurate messages or a faulty group member intermittently changes received messages. This can be addressed by a Byzantine agreement protocol with signed messages. In this protocol all group members exchange the selected secondary MCS message they received from the primary MCS with all other group members, but each message has a digital signature such that a group member cannot change the content of messages without being detected. If the transmitting primary MCS is faulty, then comparison of messages between group members will reveal multiple values for the selected secondary MCS. In the case of a faulty transmitting primary MCS, all group members can either (1) roll-back to the previous checkpoint and reroute to the specified secondary MCS; (2) reroute to a specified "Default MCS"; or (3) use an algorithm to choose from among the secondary MCS values (i.e., a majority voting algorithm). If a group member is faulty, this will either be revealed by a corrupted digital signature or a timeout because it cannot send a faulty message without being detected.

#### 5.2.3 MCS Outage Identification

In considering the single-point-of-failure vulnerability in the MCS Model, distinguishing an MCS outage from other events is non-trivial. Identification of an MCS outage will require signaling beyond attempting to reestablish a connection to the MCS. The fault location within an ATM point-to-point or point-to-multipoint VC is identified by ATM OAM signaling cells, but in the MCS Model the MCS concatenates a point-to-point VC with a point-to-multipoint VC such that fault information cannot be assumed through the MCS (from the sender side of the MCS to the receiver side of the MCS or vice versa). To a sender, an MCS outage will terminate any reliability mechanism feedback from receivers such that the sender cannot distinguish whether all receivers have left the group or there is an MCS fault. To a receiver, an MCS outage will terminate any group communication such the receiver cannot distinguish whether all senders have left the group or there is an MCS fault.

One solution to eliminate the uncertainty of distinguishing between ungraceful exits and an MCS fault is a "session announcement" mechanism such that when a member joins or leaves a group an announcement is broadcast from the MCS to the group. Thus when a group member no longer responds to signaling and no session announcement of the member leaving the group has been broadcast then a fault has occurred to disconnect that member. If a sender has received no point-to-point ATM fault signals and all the receivers have been disconnected by a fault for a time greater than a worst-case restoration time then an MCS fault has occurred. If a receiver receives no point-to-multipoint ATM fault signals and all the senders have been disconnected by a fault for a time greater than a worst-case restoration time then an MCS fault has occurred.

#### 5.2.4 Restoring Isolated Nodes in the MCS Model

If group members are disconnected by a fault and an MCS fault has not occurred then this introduces the new problem of restoring disconnected nodes to an established group communication. Figure 6 shows examples of single fault restoration scenarios for isolated nodes in the MCS Model. The restoration techniques can be categorized as: server side (sender-side or receiver-side), MCS-participation (MCS-involved or not MCS-involved), circuit type (point-to-point or point-to-multipoint), and backup route (preplanned backup or dynamic search). Link capacity sharing between working and backup VCs or among backup VCs from disjoint working VCs is discussed separately. The fault location is identified by special OAM ATM signaling either on the sender-side of the MCS (point-to-point VC) or the receiver-side of the MCS (point-to-multipoint VC).

[15] examples of distance metrics that can be used to select a secondary MCS among multiple eligible secondary MCSs include the secondary MCS with the shortest path to the failed primary MCS or the secondary MCS which has the shortest aggregate distance to all the group members of the failed primary MCS

Figure 6A shows a sender-side, MCS-involved, point-to-point, preplanned restoration scheme. This restoration scheme is analogous to the VC Mesh restoration scheme described in Figure 3A with similar survivability characteristics. It requires a point-to-point backup VC between the sender and the MCS be identified and reserved when each sender joins the group. In summary, the sender-side, MCS-involved, point-to-point, preplanned restoration scheme is fast, cost-effective (including link capacity sharing) and guarantees single fault tolerance.

Figure 6B shows a receiver-side, MCS-involved, point-to-point, preplanned restoration scheme. This restoration scheme is also analogous to the VC Mesh restoration scheme described in Figure 3A with similar survivability characteristics. It requires a point-to-point backup VC between the MCS and the receiver be identified and reserved when each receiver joins the group. If a single fault occurs in the trunk of the point-to-multipoint VC emanating from the MCS then every receiver will need their own disjoint point-to-point VC for restoration and any link capacity sharing will result in less than 100% restoration. In summary, the receiver-side, MCS-involved, point-to-point, preplanned restoration scheme is fast, suited for infrequent single-receiver disconnection events, well-suited for a small number of receivers, and costly to provide single fault tolerance (due to the possibility of receiver-side "trunk" faults).

Figure 6C shows a receiver-side, not MCS-involved, point-to-point, preplanned restoration scheme. This restoration scheme is analogous to the VC Mesh restoration scheme described in Figure 3B with similar survivability characteristics. This scheme requires an additional backup point-to-point VC with a disjoint route be reserved between an receiver-side upstream node (Figure 6C shows an upstream "branch node") and a receiver when each receiver joins the group. In summary, the receiver-side, not MCS-involved, point-to-point, preplanned restoration scheme is fast, suited for single receiver disconnection events, and allows link capacity sharing, but can only provide guaranteed restoration from a single fault at a high cost.

Figure 6D shows a receiver-side, not MCS-involved, point-to-point, dynamic search restoration scheme. This restoration scheme is analogous to the VC Mesh restoration scheme described in Figure 3C with similar survivability characteristics. In this scheme when a receiver is disconnected, a dynamic search occurs to identify a point-to-point VC to the nearest non-disconnected node in the point-to-multipoint VC emanating from the MCS. It may be the case that the nearest node in the point-to-multipoint circuit from the MCS is either disconnected itself by the same fault event or a route with the necessary residual bandwidth for a VC may not be available. In both cases the next nearest node in the point-to-multipoint circuit from the MCS is considered and so forth. Dynamic search is necessary because the location of the fault determines eligible reconnection nodes and it would be very costly to reserve preplanned backups for the entire multitude of fault location possibilities on the receiver-side of the MCS. Although it is point-to-point, this scheme is slow because of the signaling involved in a dynamic search and the worst case scenario of a trunk fault near the MCS. If a fault occurs in the trunk near the MCS then this scheme tends toward the point-to-point VC restoration scheme described in Figure 6B (as the fault occurs closer and closer to the MCS). In summary, the receiver-side, not MCS-involved, point-to-point, dynamic search scheme is slow, relatively costly, well-suited for frequent group dynamics, and cannot guarantee single fault tolerance since the residual bandwidth for a restoration VC is not reserved and may not be available when needed.

Figure 6E shows a receiver-side, MCS-involved, point-to-multipoint, preplanned restoration scheme. This restoration scheme is analogous to the VC Mesh restoration scheme described in Figure 3D with similar survivability characteristics (including the same three unique properties). This scheme requires a disjoint, point-to-multipoint backup VC be identified and reserved from the MCS to all receivers when each receiver joins or leaves the group. In summary, this receiver-side, MCS-involved, point-to-multipoint, preplanned restoration scheme is fast, and performance degrades with increasing group dynamics, but it is guaranteed to provide single fault tolerance on the receiver-side of the MCS and moderately cost-effective (due to cost decreasing effects of link capacity sharing).

Figure 6F shows a receiver-side, end-to-end, point-to-multipoint, preplanned restoration scheme very similar to Figure 6E but with one difference: the working circuits and preplanned backup circuits are not disjoint. This restoration scheme is analogous to the VC Mesh restoration scheme described in Figure 3E with similar survivability characteristics. Figure 6F shows a link fault in the receiver-side trunk of the point-to-multipoint VC emanating from the MCS which is restored by a point-to-multipoint backup VC which has every link in common with the working VC except for the link where the fault occurs. To provide such single link fault tolerance, there needs to be a separate point-to-multipoint backup VC for every link in the working point-to-multipoint VC. Figure 6F is an example where single node fault tolerance cannot be provided with this technique because a failure of the branch node defaults the restoration to the point-to-point technique of Figure 6B. Thus this scheme can provide a single link fault guarantee and not a single node fault guarantee. In summary, this receiver-side, end-to-end, point-to-multipoint, preplanned, non-disjoint restoration scheme is fast, is moderately costly, performance degrades with increasing group dynamics, and is only guaranteed to provide single link fault tolerance.

## 6 Summary

A comprehensive structure to describe the survivability issues relevant to ATM group communications has been presented. Since the objectives of ATM multipoint survivability are a superset of the objectives of ATM point-to-point survivability, this paper has extended restoration techniques from circuit-switched communications to the proposed models for ATM group communications considering the additional issues specific to multipoint communications (group dynamics and scalable reliability).

After analyzing all possible single fault scenarios, our conclusion is that different restoration techniques are appropriate for different implementations of ATM group communications depending on several factors: proposed model; availability and connectivity to backup servers; single fault tolerance guarantees; restoration time requirements; group dynamics; and cost. These factors are summarized in Table 3 for the case of single fault restoration for isolated nodes.

The determining factor for the survivability of ATM group communications is the model choice - the VC Mesh Model or the MCS Model. The MCS Model is dominated by the MCS single-point-of-failure vulnerability while the VC Mesh Model has increased exposure to link and node failures and scalability problems. The availability and connectivity to backup servers applies only to the MCS Model - backup servers can be adjacent to the primary server via a direct connection or be located remote from the primary server across a network. Appropriate survivability techniques for a primary MCS with an adjacent backup MCS are multihoming and replication, which can maintain VCs without performance degradation at the cost of overhead needed to keep mirror MCSs current and consistent. An appropriate survivability technique for a primary MCS with a remote backup MCS is the distributed backup scheme approach of primary and secondary MCSs which does not need to maintain state information but VCs may be disrupted during rerouting. Since either model can be utilized on a group-by-group basis, it is likely that different multipoint survivability techniques will be used simultaneously on the same ATM network. It is also likely that ATM multipoint survivability will be driven by strategic decision-making based on threat analysis given that there may be several possible techniques which provide similar survivability.

While research into network survivability for scenarios of random failure and malicious attack is absolutely vital, achieving survivable group communications may become most important to maintaining the operational stability of networks introducing real-time services with its corresponding expectations of reliability/survivability and to transitioning to efficient bandwidth management techniques (e. g. , multicasting) given economic market pressures.

Research is needed to further quantify the tradeoffs described in this paper. Addressing the survivability of ATM group communications has increasing urgency with commercial networks announcing new IP multicast retail services.[18] Several of these commercial networks have ATM backbones.

## References


[1] G. Armitage, "Support for Multicast Over UNI 3.0/3.1 Based ATM Networks," *IETF RFC 2022*, Nov. 1996.

[2] ATM Forum Technical Committee, *ATM User-Network Interface (UNI) Signaling Specification ver. 4.0*, ATM Forum Document af-sig-0061.000, July 1996.

[3] ATM Forum, *ATM User-Network Interface (UNI) Specification ver. 3.1*, Prentice Hall, June 1995.

[4] ATM Forum Technical Committee, *B-ISDN Inter Carrier Interface (B-ICI) Specification ver. 2.0 (integrated).* ATM Forum Doc. af-bici-0013.003, Dec. 1995.

[5] T. Chen et. al., "Monitoring and Control of ATM Networks Using Special Cells," *IEEE Network Mag.*, Sept./Oct. 1996, pp. 28-38.

[6] R. Cotter, D. Medhi, and D. Tipper, "Traffic Backlog and Impact on Network Dimensioning for Survivability for Wide-Area VP-Based ATM Networks," *15th Intl. Teletraffic Conf.* (ITC 15), 1997, pp. 691-700.

[7] R. Doverspike and B. Wilson, "Comparison of Capacity Efficiency of DCS Network Restoration Routing Techniques," *J. of Net. and Sys. Mgmt.*, Vol. 2, No. 2, 1994, pp. 95-123.

[8] O. Gerstel, I. Cidron, and S. Zaks, "The Layout of Virtual Paths in ATM Networks," *IEEE/ACM Trans. on Net.*, Vol. 4, No. 6, December 1996, pp. 873-884.

[9] B. Jaeger and D. Tipper, "On Fault Recovery Priority in ATM Networks," *IEEE Intl. Conf. on Comm. (ICC'98)*, June 1998.

[10] R. Kawamura, K. Sato, and I. Tokizawa, "Self-Healing ATM Networks Based on Virtual Path Concept," *IEEE J. on Sel. Areas in Comm.*, Jan. 1994, pp. 120-127.

[11] K.R. Krishnan, R.D. Doverspike, and C. Pack, "Improved Survivability with Multi-Layer Dynamic Routing," *IEEE Comm. Mag.*, July 1995, pp. 62-68.

[12] D. Medhi, "Multi-Hour, Multi-Traffic Class Network Design for Virtual Path-Based Dynamically Reconfigurable Wide-Area ATM Networks," *IEEE/ACM Trans. on Net.*, Vol. 3, No. 6, Dec. 1995, pp. 809-818.

[13] K. Murakami and H.S. Kim, "Virtual Path Routing for Survivable ATM Networks," *IEEE/ACM Trans. on Net.*, Vol. 4, No. 1, Feb. 1996, pp. 22-39.

[14] D. Niehaus et. al., "Performance Benchmarking of Signaling in ATM Networks," *IEEE Comm. Mag.*, Aug. 1997, pp. 134-143.

[15] T. H. Noh, D.R. Vaman, X. Gu, "Reconfiguration For Service and Self-Healing in ATM Networks based on Virtual Paths," *Comp. Net. and ISDN Sys.*, Vol. 29, pp. 1857-1867.

[16] E. Oki, N. Yamanaka, and F. Pitcho, "Multiple-Availability-Level ATM Network Architecture," *IEEE Comm. Mag.*, Sept. 1995, pp. 80-88.

[17] A. Orda and R. Rom, "Multihoming in Computer Networks: A Topology-Design Approach," in *IEEE INFOCOM'88*, pp. 941-945.

[18] D. Pappalardo, "UUNET Targets Content Providers with New Dial-In Multicast Service," *Network World*, September 29, 1997, p. 18.

[19] G. N. Rouskas and I. Baldine, "Multicast Routing with End-to-End Delay and Delay Variation Constraints," *IEEE J. on Sel. Areas in Comm.*, Vol. 15, No. 3, Apr. 1997, pp. 346-356.

[20] J. Sosnosky, "Service Applications for SONET DCS Distributed Restoration," *IEEE J. on Sel. Areas in Comm.*, Vol. 12, No. 1, Jan. 1994, pp. 59-68.

[21] A. Vakhutinsky and M. O. Ball, "Fault-Tolerant Virtual Path Layout in ATM Networks," in *15th Intl. Teletraffic Conf.* (ITC 15), 1997, pp. 1031-1041.

[22] W-P. Wang, D. Tipper, B. Jaeger, and D. Medhi, "Fault Recovery Routing in Wide Area Packet Networks," *15th Intl. Teletraffic Conf.* (ITC 15), 1997, pp. 1077-1086.

[23] B. Waxman, "Routing of Multipoint Connections," *IEEE J. on Sel. Areas in Comm.*, Vol. 6, No. 9, Dec. 1988, pp. 1617-1622.

[24] C.S. Wu, S.W. Lee and Y.T. Hou, "Backup VP Preplanning Strategies for Survivable Multicast ATM Networks," *IEEE Intl. Conf. on Comm. 1997* (ICC'97), pp. 267-271.

[25] Y. Xie, M.J. Lee, and T.N. Saadawi, "Multicasting Over ATM Using Connection Server," *IEEE Intl. Conf. on Comm. 1997* (ICC'97), pp. 1376-1380.

| *Objective* | *Technique* |
|---|---|
| minimize cost | <u>minimize</u> number of links, nodes, and bandwidth provisioned to provide survivability |
| adequate service continuity | specific services have connectivity, bandwidth, and QoS guarantees *AND/OR* <u>maximize</u> total demand restored *AND/OR* <u>minimize</u> blocked/unrestored connections |
| user transparency | <u>minimize</u> effect on established connections not directly associated with network restoration |
| congestion constraint | congestion magnitude/duration below outage thresholds *AND/OR* contained to an acceptable region |
| reliability probability constraint | reliability of service/network above required thresholds |
| restoration time constraint | restoration time below pertinent service/network outage thresholds |
| ⇒ group dynamics | mechanism which allows members to join/leave a group frequently |
| ⇒ scalable reliability | reliability mechanism which avoids the "ACK implosion effect" |

⇒ issue unique to multipoint communications

Table 1: *Objectives of Survivable Group Communications*

| *VC Mesh Model* | *Multicast Server Model* |
|---|---|
| uses a source-based distribution tree | uses a source-based distribution tree |
| can optimize paths for each source | part of path chosen by server out of source's control |
| less efficient network resource utilization:<br>more bandwidth required;<br>more signaling required;<br>end hosts have one VC termination per source per each ATM interface<br>(number of VCs increases linearly with number of sources within each participating group) | more efficient network resource utilization:<br>less bandwidth required;<br>less signaling required;<br>end hosts have only two VCs per server per each ATM interface<br>(number of VCs increases linearly with number of participating group servers) |
| more expensive if ATM provider charges per VC | less expensive if ATM provider charges per VC |
| non-scalable group membership signaling:<br>every source needs to modify their point-to-multipoint VCs | scalable, efficient, and flexible group membership control: only one point-to-multipoint VC needs to be updated |
| less stable:<br>number and frequency of group membership changes may swamp network's signaling capacity depending on number of sources | more stable: period of time after an arrival or departure from the group when all sources do not have a path to the actual set of group members is short |
| lower delays due to optimal paths and no queueing at a server | higher delays due to longer paths and assembly/queueing at a server |
| no buffering | buffering constraint: all cells belonging to a packet must be stored until the entire packet is received |
| sources do not receive their own "reflected" messages | sources receive their own reflected messages unless server is reconfigured to prevent this |
| distributes traffic and cell replication load across many links and switches | server is a traffic concentration point and potential bottleneck |
| group dynamics: well-suited for large number of listen-only members in a group of infrequent joins and leaves | group dynamics: well-suited for large number of sources in a group of frequent joins and leaves |
| ⇒ potentially more VCs with more sources<br>⇒ means an increased exposure to link and<br>⇒ node failures | ⇒ potentially less VCs means a decreased<br>⇒ exposure to link and node failures |
| ⇒ no single point of failure will disable<br>⇒ a group's communication | ⇒ server is a single point of failure for<br>⇒ a group's communication |

⇒ directly relevant to survivability

Table 2: *A Comparison of ATM Forum Multicast Models: VC Mesh versus Multicast Server*

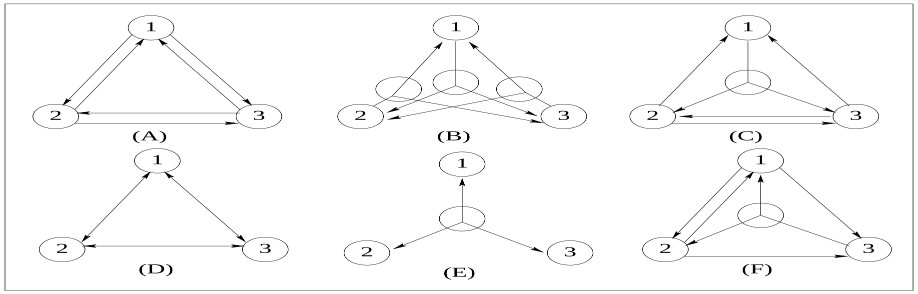


Figure 1: *The Formation of a VC Mesh: (A) a 3-switch group formed by 6 point-to-point asymmetrical circuits; (B) a 3-switch group formed by 3 point-to-multipoint asymmetrical circuits; (C) a 3-switch group formed by 1 point-to-multipoint and 4 point-to-point asymmetrical circuits; (D) a 3-switch group formed by 6 point-to-point symmetrical circuits; (E) a 3-switch group formed by 3 point-to-multipoint symmetrical circuits; (F) a 3-switch group formed by 1 point-to-multipoint and 4 point-to-point circuits (permutation of 2C)*

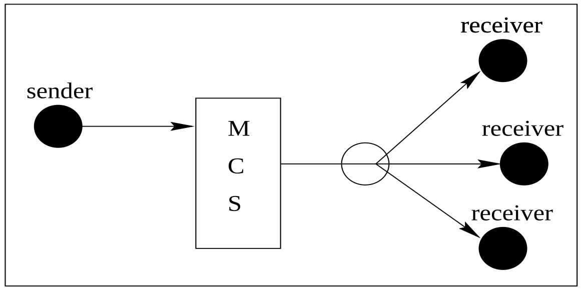


Figure 2: *The MCS Model - each sender establishes a point-to-point VC to the MCS and the MCS establishes a point-to-multipoint VC to all group members*

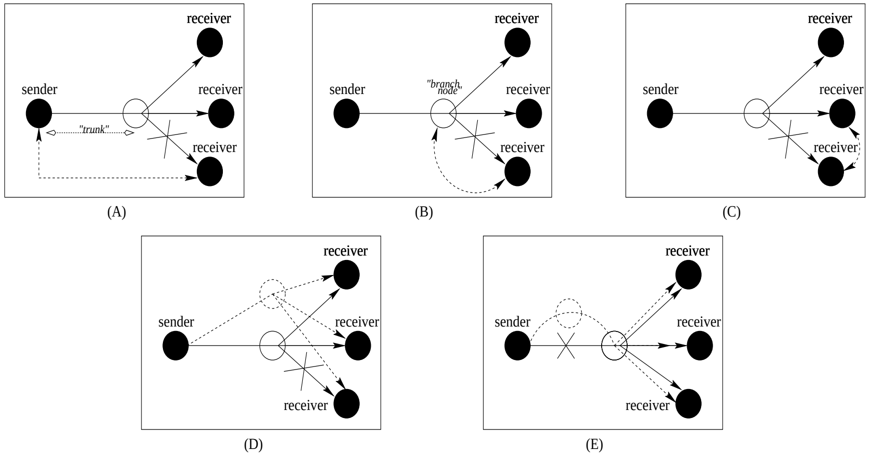


Figure 3: *Single Fault Restoration Scenarios of VC Mesh Model: (A) a fault is restored by a disjoint point-to-point VC between sender and disconnected receiver; (B) a fault is restored by a disjoint point-to-point VC between an upstream node (branch node in this case) and the disconnected receiver; (C) a fault is restored by a disjoint point-to-point VC between the disconnected receiver and the nearest non-disconnected node in the established point-to-multipoint VC; (D) a fault is restored by reserving a disjoint point-to-multipoint VC from sender to all receivers; (E) a fault in the trunk is restored by a non-disjoint point-to-multipoint circuit from sender to all receivers*

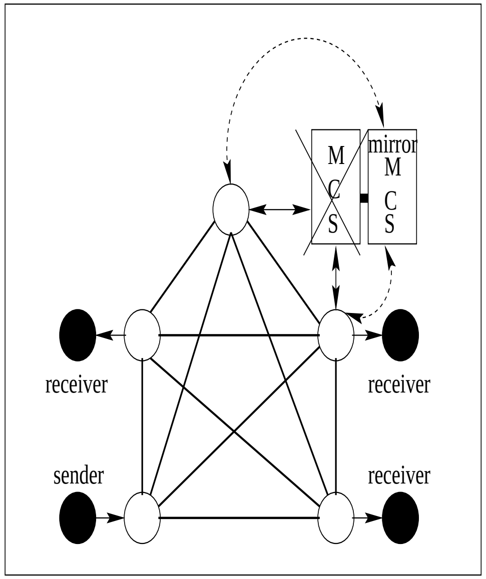


Figure 4: *MCS Restoration with an Adjacent Backup Server: a primary MCS fault will be restored by an adjacent backup MCS (mirror MCS) with the same multihomed network attachment (Note the dedicated connection between the primary MCS and the adjacent backup MCS which is isolated from the network)*

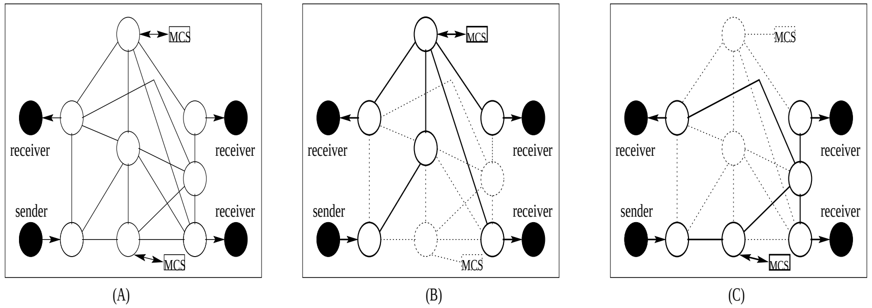


Figure 5: *Restoration of an MCS via the Remote Backup Scheme: (A) primary and secondary MCS devices attached to the network; (B) operation of the primary MCS; (C) operation of the secondary MCS*

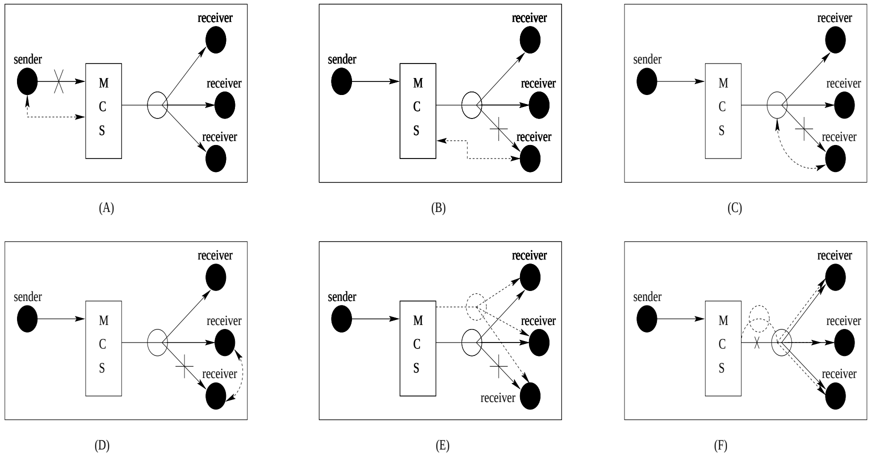


Figure 6: *Single Fault Restoration Scenarios For Isolated Nodes in the MCS Model: (A) a fault isolating a sender is restored by a disjoint point-to-point VC from the sender to the MCS; (B) a fault isolating a receiver is restored by a disjoint point-to-point VC from the MCS to the receiver; (C) a fault is restored by a disjoint point-to-point VC between an upstream node (a branch node in this case) and the disconnected receiver; (D) a fault isolating a receiver is restored by a disjoint point-to-point VC with the nearest non-disconnected node in the established point-to-multipoint VC from the MCS to receivers; (E) a fault isolating a receiver is restored by establishing a new disjoint point-to-multipoint VC from the MCS to receivers; (F) a fault in the trunk is restored by a non-disjoint point-to-multipoint circuit from MCS to all receivers*

| *VC Mesh Model* | *MCS Model (isolated node)* |
|---|---|
| **Single Fault Tolerant Guarantee** | |
| Figure 3A - yes | Figure 6A - yes |
| Figure 3B - yes | Figure 6B - yes |
| Figure 3C - no | Figure 6C - yes |
| Figure 3D - yes | Figure 6D - no |
| Figure 3E - no | Figure 6E - yes |
| (only single link) | Figure 6F - no (only single link) |
| **Restoration Speed** | |
| Figure 3A - fast | Figure 6A - fast |
| Figure 3B - fast | Figure 6B - fast |
| Figure 3C - slow | Figure 6C - fast |
| Figure 3D - fast | Figure 6D - slow |
| Figure 3E - fast | Figure 6E - fast |
| | Figure 6F - fast |
| **Level of Acceptable Group Dynamics** | |
| Figure 3A - moderate dynamics | Figure 6A - moderate dynamics |
| Figure 3B - moderate dynamics | Figure 6B - moderate dynamics |
| Figure 3C - frequent dynamics | Figure 6C - moderate dynamics |
| Figure 3D - infrequent dynamics | Figure 6D - frequent dynamics |
| Figure 3E - infrequent dynamics | Figure 6E - infrequent dynamics |
| | Figure 6F - infrequent dynamics |
| **Cost (in terms of network resources)** | |
| Figure 3A - high cost | Figure 6A - low cost |
| ( if single fault guarantee) | (cost scales with senders) |
| Figure 3B - high cost | Figure 6B - high cost |
| (if single fault guarantee) | (if single fault guarantee) |
| Figure 3C - moderate cost | Figure 6C - high cost |
| | (if single fault guarantee) |
| Figure 3D - low cost | Figure 6D - moderate cost |
| Figure 3E - moderate cost | Figure 6E - low cost |
| | Figure 6F - moderate cost |

Note: MCS as single-point-of-failure addressed by:
(1) multihoming/server replication for adjacent backup MCS(s) and
(2) primary/secondary servers for remote backup MCS(s)

Table 3: *Factors in Providing Survivability to ATM Group Communications*