

A “Dual-Tree” Scheme for Fault-Tolerant Multicast

Aiguo Fei, Junhong Cui, Mario Gerla, Dirceu Cavendish
Computer Science Department
University of California
Los Angeles, CA 90095
{afei,jcui,gerla,dirceu}@cs.ucla.edu

Abstract— To protect against possible network node or link failure, pre-planned failure recovery schemes are necessary to achieve higher reliability of communications. A couple of schemes have been previously reported for multicast. In this paper, we present a scheme based on a “dual-tree” structure in which a secondary tree for fault-tolerance purpose is built as a complement to the primary multicast tree. The secondary tree provides alternative delivery paths that can be activated when link or node failure is detected in the primary multicast tree. Simulation experiments show that this scheme has shorter restoration time and cause less multicast tree cost increase after restoration than some schemes proposed previously.

I. INTRODUCTION

In real networks, node or link failure is often possible. To ensure non-interruption or minimal disruption to communications required by some applications, some mechanism is needed to cope with such failures. In IP networks, node or link failure will lead to routing table update which will in turn have packets rerouted to get around the failure point. In some multicast protocol like PIM-SM[2] (Protocol Independent Multicast Sparse Mode), multicast forwarding information on a router depends on its unicast routing table. Update of unicast routing table leads to forwarding table update which essentially reorganizes the multicast tree to get around network failures. In some other protocols like CBT[1] (Core Based Tree), affected members abandon the existing connections and rejoin the tree.

Such type of fault-tolerance can be considered as fault-tolerance on-demand, it often works well enough for datagram communications. However, the long recovery latency usually experienced with such type of fault-tolerance could be undesirable for many real-time communications, and is not acceptable for some other time-critical applications which require guaranteed non-interrupted communications, such as coordination among multiple sites during NASA satellite launches. In future data network capable of QoS (Quality of Service) support, usually connection setup and resource reservation are necessary for applications which require QoS. In such an environment, dynamic recovery may lead to service interruption too long to be acceptable, and in some cases the renewed reservation involved may not succeed at all because of resource constraint.

The alternative solution to on-demand recovery is pre-planned failure restoration. For unicast, a technique called back-up channel[3] can be used. In this technique, a back-up(standby) route is setup in addition to the primary working route. The backup channel can be activated when a link or

node failure is detected on the primary route. The problem is more challenging for multicast communications since a network failure will affect a number of multicast group members. A couple of schemes were previously studied[8], [9]. One is link-protection in which a backup path is setup for each link in the multicast tree. Another one is path-protection in which a vertex-disjoint backup path from the source to each destination is setup. A variant of link-protection is also proposed in [8]. In these pre-planned schemes, restoration is more likely to succeed since backup resource can be reserved in advance; at the same time, restoration process can take less time because backup path(s) are pre-defined.

In this paper, we present a new scheme called “dual-tree” fault-tolerant multicast. In our scheme, instead of protecting each link or member individually in the multicast tree, a secondary tree is built among a subset of multicast members for fault-tolerance purpose. We will describe how our scheme works to protect communications for affected members when a network failure happens, and present simulation results of comparison with other schemes. Our scheme is simple in design and doesn’t require per-link or per-path fault-tolerance management. Simulation results show that it has shorter restoration time and results in better multicast tree after restoration.

The rest of this paper is organized as follows. In section 2, some related work is reviewed to provide necessary background. We then introduce how a dual-tree structure can be used for failure restoration in multicast in Section 3. Simulation results are presented in Section 4, followed by a short conclusion as Section 5.

II. BACKGROUND AND RELATED WORK

In IP networks, dynamic restoration is achieved through routing table update. In self-healing ATM networks, dynamic restoration can be achieved by broadcasting messages to search for restoration paths when a failure is detected. While pre-planned restoration involves setting-up back-up paths and activating back-paths when failure is detected. Dynamic restoration has the advantages of low control complexity and the flexibility to cope with topology change. However, as stated earlier, long restoration latency associated with dynamic restoration may not be acceptable for many real-time applications. On the other hand, pre-planned restoration mechanism can potentially reduce restoration time, but it requires more elaborate restoration management and more network resources in many

cases compared with dynamic approach. In this section, we briefly review some related work on pre-planned restoration as background for our work.

A. Network Model and Assumptions

The network under consideration consists of a number of nodes connected by links along which data are transmitted from a node to another. A network can be modeled as a directed graph $G(V, E)$. A node can be a router in IP networks, or an ATM switch, or any similar communication device (say, a SONET device). We are mostly concerned with communication at router/switch level, so network hosts (or other end communication devices) are not explicitly considered.

We assume all links in the network are bidirectional. Each link is also assigned a positive number as its cost (can be different on different directions). The cost of a multicast tree is the summation of the costs of all links. In this paper, we assume one-to-many or core-based multicast in which a multicast delivery tree is rooted at the source or core.

Most existing work on pre-planned restoration [4], [8], [9] has assumed single failure model in which there is only a single link or node that fails at a time. The same assumption applies in this paper.

B. Failure Restoration in Unicast Communication

SONET defines a mechanism called Automatic Protection Switching (APS) to detect failure and switch traffic from one path to another at physical layer [7]. In APS 1+1, traffic is transmitted from source to destination over two separate fibers (working and protect) simultaneously. If one fiber is cut, data is still being transmitted over the other one. This provides the fastest restoration (less than 50ms) and best possible protection, but it has the drawback of requiring twice the bandwidth. In ASP 1:1, protected traffic is only sent over the working path and the protect fiber can be used to transmit lower priority traffic. A switch-over will be initiated if the working fiber is cut. No bandwidth is wasted in case of no failure. However, if the protect path doesn't have enough bandwidth for both protected traffic and lower priority traffic in case of failure, then the lower priority traffic will be preempted.

In self-healing ATM networks[4], [5], there is an end-to-end restoration technique called path restoration, in which a backup Virtual Path (VP) is pre-computed to protect a working VP. All connections within a working VP will be switched over to use the backup VP (if bandwidth allows) through exchange of management ATM cells when a failure is detected. In another type of restoration scheme, link-protection, instead of protecting individual end-to-end VPs, all affected traffic originally carried by a failed link are re-routed to bypass failure point without the intervention of the endpoints of affected connections. They are illustrated in Fig.1. In both path restoration and link restoration, different backup VPs or paths have to share backup capacity.

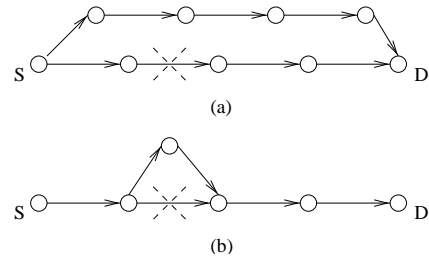


Fig. 1. (a)path restoration; (b)link restoration

C. Hot Redundancy vs. Cold Stand-by

In some restoration schemes like SONET ASP 1+1, data is transmitted over both working path and protect path. In ASP 1:1 and most ATM VP-based restoration schemes, traffic is only sent over working path and will only be sent over protect path upon failure detection and backup revocation. The former type of failure protection can be called hot redundancy while the later called cold stand-by.

Hot redundant failure protection has the advantage of very fast recovery and guaranteed (single) failure restoration, it has one main drawback: it requires twice the bandwidth. Cold stand-by method can remedy this drawback and save network resource by rerouting traffic through back-up path only when a failure happens. Normally a longer restoration time is expected. At the same time, restoration may fail if there is not enough bandwidth along the backup path (or lower priority traffic has to be preempted as in SONET ASP 1:1). Clearly spare capacity has to be reserved for backup path to improve the success probability of failure restoration. However, if every protected connection reserves the same amount of bandwidth as what reserved for working path, then it would be the same as hot redundancy. To reduce spare capacity requirement, different protect paths that go through a same link must share backup capacity (on that link) [4], [5], [10]. Backup path computation and backup capacity sharing are very important issues to consider in cold standby failure protection schemes.

D. Pre-Planned Restoration for Multicast

There had been little research reported in literature on failure restoration of multicast communications. As multi-point communications are gaining importance in modern communication networks, restoration schemes that specifically deal with multicast connections may become an important part of future self-healing communication networks.

In conventional multicast, traffic is distributed over a tree structure, a single failure will affect all members at the downstream of the failure point. Schemes similar to link restoration and path restoration in unicast can be applied to protect each working link or source-to-destination path in a multicast tree (assuming one-to-many type of multicast) [8], [9]. In link-protection, for each link ($u \rightarrow v$) in the multicast tree, another path from u to v is setup as backup. In path-protection, for each destination, a vertex-disjoint path (with the path in the

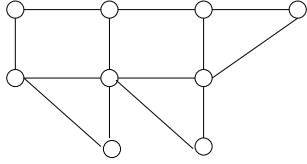


Fig. 2. A bi-connected graph.

multicast tree) from the source to that destination is setup as backup. A modified link-protection scheme is also proposed in [8], in which a backup path that protects link $(u \rightarrow v)$ can originate from u 's ancestor node or sibling node. As in unicast, backup capacity sharing strategies have to be applied to reduce backup capacity requirement and save network resource usage.

In link-protection, when a link $(u \rightarrow v)$ failure is detected, v will send a message along the protection path back to u to activate it. After that, traffic will flow over the new path from u to v given activation and necessary resource reservation are successful. In modified link-protection, v will send a message along the protection path to a node at the other end of the protection path to activate it while u doesn't need to take any action. In path protection, every member node has to send a message back to the root to activate the protection path when a failure is detected. One may note that, since the protection path(s) may share one or more links with the original multicast tree, loop may form after a restoration procedure. This important issue has been largely neglected in [8] and [9].

III. A DUAL-TREE APPROACH

A. Bi-Connected Graph and "Dual-Tree"

A bi-connected graph is a graph in which there are at least two vertex-disjoint paths between any two nodes. Given a graph G and a set of nodes S , if a bi-connected sub-graph G' can be constructed from G to include all nodes in S , then G' can be used as a multicast structure for fault-tolerance purpose for group S . One can construct a tree (rooted at the source or core) out of G' as the multicast delivery tree. In case of single node or link failure, it is always possible to find an alternative path in G' for any affected source to destination path because G' is bi-connected. Thus G' can be used for path-protection or link-protection type of fault-tolerance by systematically constructing a protection path for any core to member path in the tree or a protection path for any link in the tree. Fig.2 shows a bi-connected graph. To simplify the illustration, we only show undirected graphs in Fig.2 and others to follow and assume all links are bidirectional.

A simple algorithm that builds a bi-connected sub-graph with a vertex-disjoint "dual-tree" structure works as follows:

- (1) use an existing multicast routing algorithm (e.g., rooted shortest path tree algorithm) to build a multicast tree, this tree serves as the primary delivery tree;
- (2) identify all leaf nodes of the tree built, build a secondary tree to connect them without using any links or any inner nodes in the primary tree (vertex-disjoint).

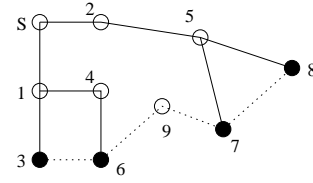


Fig. 3. A dual-tree structure.

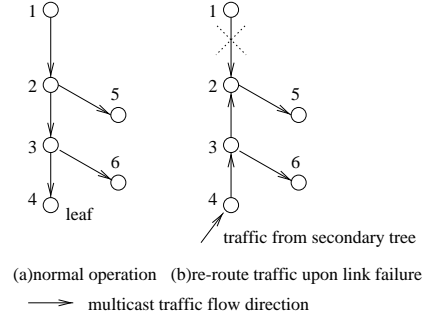


Fig. 4. Inject traffic from secondary tree to primary tree after link failure.

Theorem 1. A vertex-disjoint dual-tree built by the above algorithm is a bi-connected graph.

If the secondary tree is only link-disjoint with the primary tree, then it is an link-disjoint dual-tree. An link-disjoint dual-tree may not be bi-connected. In the rest of this paper, we refer a vertex-disjoint dual-tree simply as dual-tree except specified otherwise. A dual-tree structure is shown in Fig.3. Node S is the root, black nodes are group members, solid lines represent links used in the primary tree and dotted lines represent links in the secondary tree.

Theorem 2. In a vertex-disjoint dual-tree, if all links in the secondary tree are used bi-directionally, any leaf node in the primary tree can be reached from another leaf node through a secondary tree path vertex-disjoint with the primary tree; in link-disjoint dual-tree, that path is link-disjoint with the primary tree.

B. Failure Restoration with Dual-Tree

Though a dual-tree structure can be used for link-protection or path-protection type of failure restoration, its special structure makes it possible to do other type of failure restoration. Fig.4 shows how traffic can be injected into the primary tree from the secondary tree to reroute traffic to nodes affected by a link failure. As in Fig.4(b), when a link failure happens (link $1 \rightarrow 2$), all nodes in the downstream subtree will be affected. However, from Theorem 2, all leaf nodes in that subtree can connect to other unaffected leaf nodes through an intact path in the secondary tree. If one affected leaf node (node 4 there) is chosen to connect to an unaffected leaf node to receive traffic, then the multicast traffic can be forwarded from 4 upwards along the reverse direction of the original tree path to the immediate affected downstream node (node 2), and the whole subtree is reconnected.

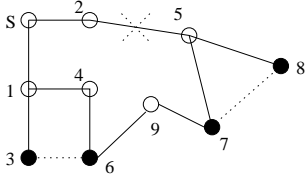


Fig. 5. Multicast tree after failure restoration.

Though multiple affected leaf nodes can start the restoration procedure simultaneously and an election procedure can be applied later to eventually select one to forward traffic upwards. However, given the tree topology information, the affected immediate node (node 2 in Fig.4) can select a leaf node and compute a restoration path and then send a message to activate the restoration path. The restoration procedure works as follows:

- (1) During normal operation, traffic is delivered through the primary tree. A leaf node in the primary tree is called a primary leaf.
- (2) If a node (x) in the primary tree detects that its parent node fails or the link connecting to its parent fails, it groups all primary leaves into two groups: that are affected (group A) by the failure and that not affected (group N).
- (3) Identify nodes in A that has a link in the secondary tree to connect to a node in N (or connect to a node in N via other intermediate node(s)), select one (y) and correspondingly the path p from y to connect node z in N .
- (4) Node x sends a Reconfig message to its child which leads to y , update its parent to be that child. Reconfig message will be forwarded in the primary tree until it reaches y , every node receiving Reconfig message changes its old parent to be a child. Reconfig message is then forwarded to z along path p and corresponding state is installed in the intermediate node(s).

Take the dual-tree in Fig.3 as an example. Now assume link from node 2 to 5 is down. Affected primary leaves are 7 and 8. Node 7 is connected through ($7 \rightarrow 9 \rightarrow 6$) to a not affected node 6, Reconfig message is sent to 6 from 5 by ($5 \rightarrow 7 \rightarrow 9 \rightarrow 6$). After restoration, the branch to reach 5 becomes ($6 \rightarrow 9 \rightarrow 7 \rightarrow 5$). Node 8 is still a child of node 5, while node 7 now becomes the parent of 5. The multicast tree after restoration is shown in Fig.5. After restoration, a new secondary tree can be constructed off-line to protect future failure.

C. Discussions

Above we presented how to construct a dual-tree structure for fault-tolerant multicast and how the restoration procedure works in case of failure. As mentioned earlier, backup capacity has to be reserved on backup links in order for restoration procedure to succeed. A key aspect of fault-tolerant multicast is how to share backup capacity among different connections to avoid reserving excessive amount of bandwidth. In this paper, we are mainly concerned with the restoration procedure. Backup capacity sharing strategies will be introduced and studied in a companion paper. In the next section, we will present

simulation comparison of restoration time and tree surcharge after restoration of our scheme and other existing schemes.

Link-protection, path-protection and modified link-protection are all designed for fault-tolerance against link failure. Path-protection and modified link-protection can also be used for fault-tolerance against node failure. For example, to compute a protect path in path-protection, if we enforce the protect path to be vertex-disjoint with the working path, then it can protect the working path against single node failure. This feature is not discussed in [8], [9], we will not go into detailed discussion here either; while it is worth pointing out that our dual-tree scheme is designed to protect against both link and node failures, as can be seen from the description of the failure restoration procedure. However, to do so it requires vertex-disjoint dual-tree which has a stronger requirement on the connectivity of the graph (than link-disjoint dual-tree). If we are only concerned with link failure, then we can use an link-disjoint dual-tree instead. The failure restoration procedure for link-disjoint dual-tree works the same way as what described earlier for vertex-disjoint dual-tree

IV. SIMULATION

The following two important metrics are evaluated in our simulation comparison of dual-tree scheme with existing schemes: (1) average restoration time, (2) tree cost increase after restoration. Restoration time is the time from when a failure is detected to when the backup path is activated. Apparently shorter the restoration time, less disruption will be caused to the communication. In link-protection, path-protection and modified link-protection, backup path is pre-computed. In our dual-tree scheme, it is computed upon failure detection from the pre-computed dual-tree structure. We assume that computation time is short and negligible compared with restoration message processing time which involves resource configuration and state installment at each node. For this reason, in our simulation, number of nodes that a restoration message travels to activate the backup path is used as the measure for restoration time. One may note that, backup path can be lengthier than the original working path and thus the multicast delivery tree can be more “expensive” in terms of tree cost than the original tree after a restoration procedure. Tree cost increase after failure restoration reflects the quality of the restoration. In simulation presented here, each link is assigned a cost of 1 (thus the cost of a multicast tree is total number of links).

We use randomly generated networks using the approach given in [6]. Network nodes are randomly chosen in a $g \times g$ grid. A link between two nodes u and v is chosen to be in the graph according to probability $p(u, v) = \lambda \exp(-d(u, v)/\rho L)$, where $d(u, v)$ is geometric distance between u and v , $L = \sqrt{2} \cdot g$ is the maximum distance between any two nodes, λ and ρ are constants ($0 < \lambda, \rho < 1$). In simulation results presented here, λ and ρ are chosen such that in average each node has a degree between 4 and 5. Networks have a fixed size of 100 nodes chosen over a 30×30 grid. All links are assumed to be bi-directional.

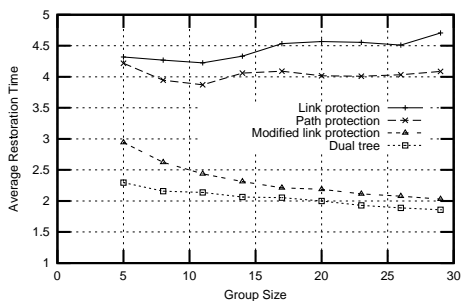


Fig. 6. Average failure restoration time vs. group size.

In addition to our dual-tree scheme, three schemes described in Section 2 are simulated: link-protection, path-protection and modified link-protection. They correspond to link restoration scheme 1, path restoration scheme 3 and link restoration scheme 2 in [9]. In the simulation experiments, we randomly generate a number of networks. For each network, a Shortest Path Tree is generated for a multicast group consisting of a specific number of nodes randomly chosen (one of them is randomly chosen as the source). An link-disjoint secondary tree is generated for our scheme, and backup paths for the other three schemes are computed with backup sharing strategies[9] applied. A link in the multicast tree is randomly chosen to go down, then four restoration procedures are applied. Restoration time is recorded for all four schemes. The path-protection may require restoration for more than one node, an average is taken as the restoration time. After restoration, percentage of tree cost increase is recorded. We run the experiments for 100 times and then take the average.

Fig.6 shows the result of average restoration time (ART) vs. group size in networks of 100 nodes. Link-protection has the longest ART, it shows that local detouring may significantly lengthen the backup path. In path-protection, an end-to-end backup path revocation is involved for each member affected by a failure, its ART comes next and close to that of link-protection. Modified link-protection has much shorter ART than link-protection and path-protection, which confirms the finding in [8]. Our dual-tree scheme has slightly shorter ART than modified link-protection. ART for both modified link-protection and dual-tree goes down with the increase of the size of the multicast group while path-protection and link-protection don't. This can be explained by the fact that more populated group members help both schemes to find shorter back-up paths, while it doesn't help path-protection or link-protection in which backup paths are constructed independent of other members. Fig.7 shows the result of average percentage increase in tree cost after restoration. All three schemes don't introduce much tree cost increase, while dual-tree scheme performs better than the other three in this metric as well.

V. CONCLUSIONS AND FUTURE WORK

Restoration of multicast connections in face of node or link failure poses new challenges to fault-tolerant communications.

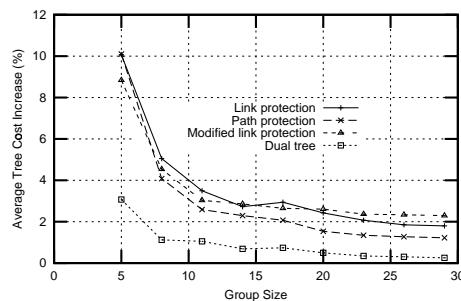


Fig. 7. Average tree cost increase after failure restoration vs. group size.

In this paper we present a scheme based on a “dual-tree” structure in which a secondary tree for fault-tolerance purpose is built as a complement to the primary multicast tree. The secondary tree provides alternative delivery paths that can be activated when link or node failure is detected in the primary multicast tree. Simulation experiments show that this scheme has shorter restoration time and cause less multicast tree cost increase after restoration than some schemes proposed previously.

We can expect applications that require such type of fault-tolerant multicasting will also have other QoS requirements such as bandwidth guarantee. To achieve high possibility of successful failure restoration, backup bandwidth should be reserved on the secondary tree in our dual-tree scheme. Backup bandwidth sharing strategies among different backup paths for the same multicast tree are discussed in [9], but they don't apply in a dual-tree scheme in which a protection tree is constructed instead of multiple backup paths. However, different multicast sessions can share backup bandwidth on their common (backup) links. Backup sharing strategies will be studied and presented in a future companion paper.

REFERENCES

- [1] A. Ballardie, P. Francis, and J. Crowcroft, “Core based trees (CBT)”, in *Proc. ACM SIGCOMM'93*, pp.85-95, September 1993.
- [2] S. Deering, D. Estrin, D. Farinacci, V. Jacobson, C. Liu, and L. Wei, “The PIM architecture for wide-area multicast routing”, *IEEE/ACM Trans. Networking*, vol.4(2), pp.153-162, April 1996.
- [3] S. Han and K. G. Shin. Fast restoration of real-time communication services from component failures in multi-hop networks. in *Proc. ACM SIGCOMM'97*, September 1997
- [4] R. Kawamura, K. Sato, and I. Tokizawa, “Self-healing ATM networks based on virtual path concept”, in *IEEE J. Select. Areas in Commun.*, vol.12(1), pp.120-127, January 1997.
- [5] K. Murakami, H. S. Kim, “Optimal capacity and flow assignment for self-healing ATM networks based on line and end-to-end restoration”, in *IEEE/ACM Trans. Networking*, vol.6(2), pp.207-221, April 1998.
- [6] B. M. Waxman, “Routing of multipoint connections”, *IEEE J. Select. Areas Commun*, vol.6(9), pp.1617-1622, December 1988.
- [7] T. Wu, *Fiber network survivability*. New York, Artech House, 1992.
- [8] C. Wu, W. Lee, Y. Hou and W. Chu, “A new preplanned self-healing scheme for multicast ATM network”, in *Proc. IEEE ICCT'96*, vol.2, pp.888-891, May 1996.
- [9] C. Wu, W. Lee, and Y. Hou, “Back-up VP preplanning strategies for survivable multicast ATM networks”, in *Proc. IEEE ICC'97*, vol.1, pp.267-271, June 1997.
- [10] Y. Xiong and L. G. Mason. “Restoration strategies and spare capacity requirements in self-healing ATM networks”, in *IEEE/ACM Trans. Networking*, vol.7(1), pp.98-110, February 1999.