Cross-Layer Techniques for Failure Restoration of IP Multicast with Applications to IPTV

M. Yuksel[1], K. K. Ramakrishnan[2], R. Doverspike[2], R. Sinha[2], G. Li[2], K. Oikonomou[2], and D. Wang[2]
yuksem@cse.unr.edu, {kkrama,rdd,sinha,gli,ko,mei}@research.att.com
[1]University of Nevada – Reno, [2]AT&T Labs - Research

Abstract— **Broadcast TV distribution over an IP network requires stringent QoS constraints, such as low latency and loss. Streaming content in IPTV is typically delivered to distribution points on an IP backbone using IP multicast protocols such as Protocol Independent Multicast Source Specific Mode (PIM-SSM). Link-restoration using MPLS or layer-2 Fast Reroute (FRR) is a proven failure restoration technique. Link-based FRR creates a pseudo-wire or tunnel in parallel to the IP adjacencies (links); and thus, single link failures are transparent to the Interior Gateway Protocol (IGP) such as OSPF. Although one may choose the back-up path's IGP link weights to avoid traffic overlap during any single physical link failure, multiple failures may still cause traffic overlap with FRR. We present a cross-layer restoration approach that combines both FRR-based restoration for single link failure and "hitless" (i.e., without loss) PIM tree reconfiguration algorithms to prevent traffic overlap when multiple failures occur.**

I. Introduction

DISTRIBUTION of real-time multimedia over an IP backbone has been gaining momentum with content and service providers [1],[2],[3],[4]. However, unlike traditional cable-based broadcast infrastructures that provide "broadcast" analog-based video (e.g., TV), using an IP backbone for real-time broadcast video distribution imposes stringent requirements for protection and restoration after a failure. Distribution of real-time 'linear' (also called broadcast) TV requires that the delay experienced by a user viewing the TV content be limited to less than a few seconds. This limits the size of the playout buffer at the receiving IPTV set-top box. Moreover, loss-recovery mechanisms have a limited capability to recover from burst packet losses. The combination of player loss-concealment algorithms, recovery through retransmissions and packet-level redundancy mechanisms such as Forward Error Control (FEC) are designed to recover from burst losses that are no more than a few tens of milliseconds. The tight QoS constraints of low latency and loss need to be met even under failures. The network challenge to meeting these constraints while providing high network availability, in turn, requires a methodology for rapid restoration [3].

Use of IP-based Protocol Independent Multicast (PIM) [5],[6],[7] to distribute the content from the source to the various distribution points on the IP backbone allows the infrastructure to be cost-competitive with the more mature cable broadcast infrastructure. However, PIM-SSM depends on a "join" and "prune" process to rebuild the multi-cast tree after a network failure. This process, when combined with the Internal Gateway Protocol (IGP) reconfiguration process (which may take several seconds or tens of seconds), takes too much time to restore the packet flow to the receivers after the failure. Thus, packet loss concealment and recovery mechanisms alone are not effective in recovering from such a long period of packet loss.

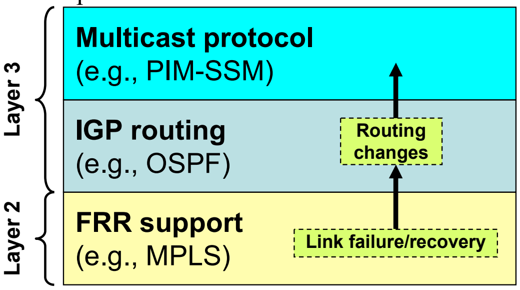


Figure 1. Cross-layer architecture for multicast failure restoration: The multicast protocol agent is notified about IGP routing changes that are triggered by link failures. The routing changes are exposed to the multicast agent without waiting for full IGP reconvergence, and the multicast agent starts reconfiguring its tree as soon as it thinks (partial) reconfiguration is

A common approach to providing the needed high availability under this IP framework is link-based Fast Reroute (FRR) [8],[9]. If the network is subject to only single link failures, enhancing the IPTV backbone with link-based FRR works well. However, in a long distance network, the probability of multiple link failures is not negligible. Analysis of concurrent failures from a large commercial IPTV deployment over a four month period revealed that in 17% of the link failure cases, at least 2 links failed concurrently, and in 2% of the cases, 3 or more links failed concurrently.

Using PIM on top of link-based FRR may experience difficulties under multiple failure situations because the IGP reconfiguration process is typically unaware when traffic is rerouted by link-based (Layer-2) FRR over alternate links. Thus, when a second failure occurs prior to repair of the first failure, traffic overlap can occur. Here, traffic overlap means that the packets of the same multicast flow (e.g., packetized video for individual 'channels' for IPTV) travel over the same directed link two or more times. Congestion may occur during traffic overlap because large-scale, real-time video streams need high bandwidth (often 2-3 Gb/s or more in aggregate across all the channels). Because of its stringent QoS constraints, loss from congestion often has the same effect on the customer's perception of service as an unprotected link failure. As failures often take several hours to repair, congestion caused by overlaps due to multiple failures can last for a long time. Therefore, to exploit the cost reduction enabled by multicast, potential traffic overlap due to multiple failures and link-based FRR needs to be avoided. The best way to achieve this is to use FRR only for a short period, then remove (or "cost-out") the failed link from the topology and then run

978-1-4244-5489-1/10/ $26.00 © 2010 IEEE

IGP reconfiguration. However, if it isn't carefully executed, this alternative can cause as many new video interruptions (due to small "hits" after single failures) as it removes, for multiple failures. This paper describes a careful multicast recovery methodology to achieve this, while avoiding the drawbacks.

A. Contributions

In [10], an intelligent IGP link weight setting algorithm was proposed in an IPTV setting that uses PIM-SSM multicast trees and link-based FRR for failure restoration. The resulting link weights ensure that the multicast tree is disjoint from backup path of every link and therefore a single link failure can be recovered using FRR without causing any traffic overlap. However multiple failures may still cause congestion and traffic overlap. Multiple failures can be close to 20% of all link failures and present an important challenge.

The research literature on dealing with multiple failures is primarily concerned with providing connectivity in unicast networks (e.g, [11]). They either ignore the problem of traffic overlap or require that certain links (with traffic overlap) have additional capacity. The considerations in an IP network carrying high-bandwidth real-time traffic is different in that multicast distribution is required and the load on individual links is significant enough that there is not a substantial amount of available bandwidth to support failures. If the FRR is successful then we know that the underlying network graph is connected so connectivity is a non-issue. In fact, without using FRR, IGP will re-converge to a new multicast tree and will have no congestion. However IGP convergence tends to be slow. Production networks rarely shorten their timers to small enough values to allow for failure recovery in the sub-second range because of the potential of false alarms [12]. So, FRR is needed to restore rapidly and it works very well in a large number scenarios (e.g., 83% of the failure cases), consisting of a single link failure. However in the remaining 17% of multiple concurrent link failure cases, it may cause traffic overlap.

We try to achieve the best of the both worlds by taking a cross-layer approach of exposing IGP routing with the information of a link failure that is restored by FRR. This transparency of link failure triggers the reconvergence of IGP routing. Then, we allow multicast routing (i.e., PIM-SSM) to see the IGP routing changes immediately (even before full IGP reconvergence) and assure multicast routing reconfigures without an additional hit. We assume the existence of a combination of video player loss-concealment algorithms, packet-level redundancy mechanisms such as Forward Error Control (FEC), and retransmission-based recovery to overcome burst losses that may be as long as 50-100 milliseconds. The mechanisms we consider here work in a complementary manner with these loss-tolerance and recovery methods.

The multicast reconfiguration protocol we develop here does not have to wait for IGP reconvergence (which can take several seconds) before it begins to reconfigure the multicast tree. Thus, it seeks to limit the time the network is exposed to the possibility of a second failure resulting in potential overlap. Moreover, the IGP reconvergence typically affects a small subset of nodes, namely nodes with a failed link in their shortest path to the root. Thus, most nodes will still have the same shortest path in the new multicast tree after a link failure.

An early version of our approach was presented in a workshop paper [13], which presented results on the performance with single failures. In this paper, we present details and correctness proofs of our framework, and extensive evaluation of its performance under multiple failures. Major contributions of this paper include:

- Protocol specification of reconfiguring multicast trees via "pending joins" without waiting for IGP reconvergence.
- Proof that the proposed protocol consistently reconfigures the multicast trees even when multiple link failures occur.
- Implementation and evaluation of the proposed protocol in a packet-based simulator under double link failures.

II. Cross-Layer Multicast Reconfiguration

As mentioned above, IPTV imposes stringent performance requirements. Even a small amount of packet loss and delay could seriously impair the video quality perceived by users. In [13], the authors report on diverse measurements from a large commercial IPTV deployment with over one million subscribers. An analysis of the customer trouble tickets indicates that nearly half are related to performance issues (e.g., video quality). Video quality monitors in the service provider's network indicate almost 79% of the alarms related the under-running of the video playout buffer. Finally, their analysis of SNMP and Syslog data from the provider network indicated that almost 55% of the performance related events were due to Layer-1 alarms and IP link flaps. In conjunction with our observation that 17% of the failure cases result in multiple link failures that has the potential of resulting in failures observed at the multicast layer, it is critical to minimize the time that a multiple link failure results in customer-observed impairments.

To understand the motivation for our approach using cross-layer failure restoration techniques, we describe in more detail the interaction of multiple failures and FRR. Consider two adjacent routers in an IP network used for IPTV distribution where the routers are part of a multicast tree using PIM-SSM [7]. Assuming the adjacency between these routers is restorable via FRR, this node pair actually consists of four different links that are visible to the IGP topology: as Figure 2 shows, two unidirectional (or ***directed***) pseudowires (dashed lines between nodes E and C for example) and two unidirectional physical-layer "PHY" links (solid lines which are the links between nodes E and C). The pseudowires are associated with a primary path and a backup path (which are typically LSPs). The primary path for each pseudowire is its corresponding PHY link and the backup path routes over other PHY links that are disjoint from the primary path. The IGP link costs (which we generically refer to as ***weights***) on the pseudowire links are lower than those for the PHY links. This causes the IGP shortest path algorithm to route over the pseudowire links rather than the PHY links in a non-failure state. Thus, when either of the PHY links fails, both links are taken out of service and the two pseudowires are switched from their primary paths to their backup paths, usually with a target switching time of 50ms or less. This switching time is sufficiently small that the higher layer packet loss concealment and recovery protocols can recover from this failure with little or no perceived video quality disruption. In addition, the back up pseudowire will switch to the backup path well before the IGP timers expire. Therefore, when the IGP Link State Advertisements (LSAs) are

978-1-4244-5489-1/10/ $26.00 © 2010 IEEE

broadcast, although they show that the PHY links are down, the pseudowire link states remain unchanged and therefore there is no change to the IGP shortest path tree and the multicast tree.

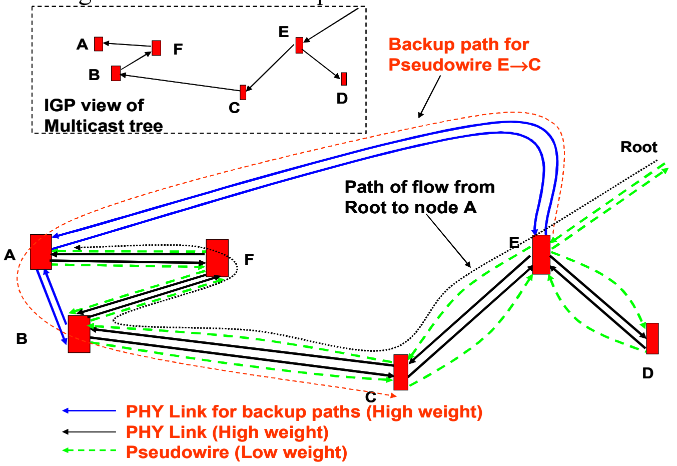


Figure 2. A network segment with pseudowires, PHY links, and a sample FRR backup path.

We give an example of this in Figure 2-Figure 4. Figure 2 depicts a network segment with 4 node pairs that have pseudowires defined. For example, node pair E-C has a PHY link in each direction and a pseudowire in each direction (a total of 4 directed links) which are for restoration. We also depict the backup path for the pseudowire E→C. Note that certain node pairs, such as E-A, are provided for restoration and, hence, have no pseudowires defined. From the point of view of the IGP topology, pseudowires can be thought of as virtual links; in fact, the IGP only differentiates these 4 links by their weights and interface IDs. Pseudowire E→C routes over a primary path, which consists of the single PHY link E→C (solid line in Figure 2). If a failure occurs to the primary path, then the router at node E attempts to switch to the backup path using FRR. Figure 3 illustrates such a single Layer-1 failure. Note the path from the Root to node A now switches to the backup path at node E (E-A-B-C), reaches node C, and then continues on its previous (primary) path to node A (C-B-F-A). Note that during the failure, although the path retraces itself between the routers B and C, because of the unidirectionality of the links the multicast traffic does NOT overlap. Also, more importantly for this paper, although the IGP view of the topology realizes that the PHY links between E-C have gone "down", the shortest path tree (and consequently, the multicast tree) remains unchanged because the pseudowire from E→C is still "up" and has lower weight. The IGP is unaware of the actual routing over the backup path.

Typically, the FRR backup path is kept active until the PHY links are repaired. Once the PHY links are back in service, the pseudowires are switched back to their primary paths rapidly (again with a target switching time of 50ms or less). This methodology works well to achieve high network availability when only non-simultaneous link failures occur (by "link" we mean the pair of unidirectional PHY links between a router pair). However, when a second failure occurs during the outage interval of the first failure (which may range from a few minutes to several hours) then, because IGP is unaware of active FRR backup paths, it is possible that traffic overlap may occur. Figure 4 depicts such a multiple failure, where both the link C-E and the router at F have failed, but the backup path for E-C is still up. IGP modifies the shortest path for the path to A, thus causing traffic overlap. This is illustrated in Figure 4, where the flows from Root to node A and the backup path of E→C used in the Root to node C flow overlap on link E→A.

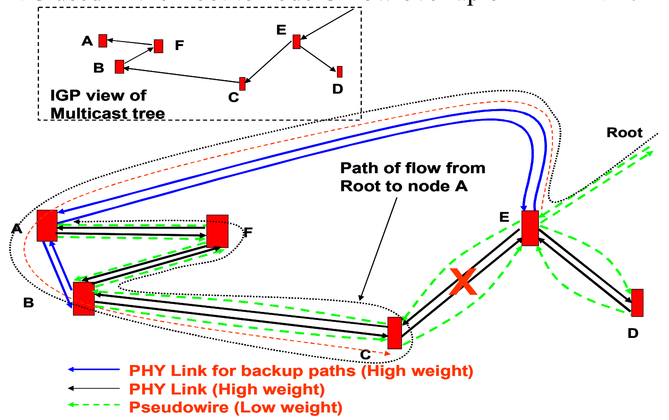


Figure 3. Single link failure: FRR backup path is put into work and IGP is unaware of the failure.

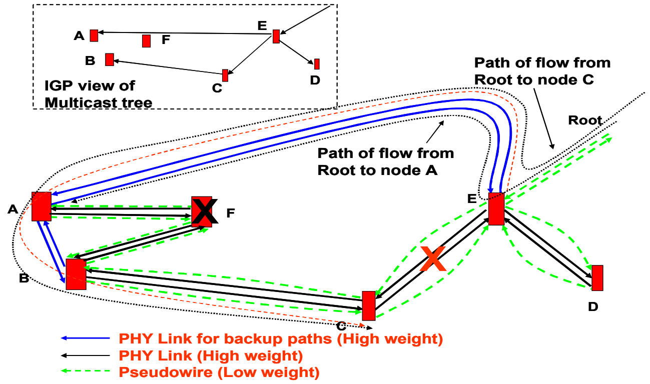


Figure 4. Multiple failures with one link and one router failing: FRR backup path overlaps with multicast traffic.

This phenomenon of traffic overlap can be avoided if we modify the approach of leaving the backup path active until the PHY link is repaired. The key is to only temporarily route the pseudowire over the FRR backup path and then have IGP converge to a new shortest path tree and a resulting multicast tree which does not use the affected pseudowire. Thus, if a second failure occurs, IGP is fully aware of the routing for the tree and can avoid the failed links during IGP reconvergence. For example, in Figure 3, not long after the failure, the pseudowires between C-E would be taken down and a new multicast tree would be re-computed. In the new multicast tree, the branch over E→C would not exist, instead routing from E→A→F→B→C. Thus, when the second failure occurs, link B→C would be pruned from the tree and no overlap would occur. However, a requirement of this new approach is to engineer a hitless (i.e., a process with no or minimal packet loss) multicast methodology, wherein we switch to the new tree after a FRR reroute without incurring packet loss. Otherwise, every time a single failure occurs and the backup FRR path is used, we will see a hit to converge to the new tree, making the new approach worse than the current approach.

The approach we describe was postulated as "Architecture 3" in [10]. The primary contributions of this present paper are to propose and describe the algorithms and protocols to perform the switchover to the new multicast tree and then evaluate its performance via simulation. Furthermore, the

978-1-4244-5489-1/10/ $26.00 © 2010 IEEE

advantage of the protocol proposed in this paper minimizes the need for the intelligent weight setting algorithm described in [10]. Details of the algorithms are described in Section IV.

III. Related Work

Most of research in IPTV focuses on architecture design [15], protocol design and selection, multimedia stream coding/decoding techniques [18], as well as potential new applications of multicast [19]. There has been extensive IP multicast research and experiments, although much less on multimedia IP network design. IETF has standardized multiple IP multicast protocols, including PIM-DM [5], PIM-SM [6], and PIM-SSM [7] and has made recommendations for reliable IP multicast [18]. To improve IP multicast performance, many models and techniques have been proposed including the overlay model [19] and MPLS P2MP [20], resource reservation and admission control to avoid congestion [21], centralized management of PIM over MPLS [22] and priority queueing or fair queueing to guarantee quality of service. These solutions have to be adapted to be useable in a carrier's multimedia distribution service. Approaches such as those proposed for reliable multicast [23] are inapplicable as these can add latency that is unacceptable in an IPTV environment.

Network restoration has also been an active research area for many decades. As video and real-time services migrate to IP based environments, IP network restoration has also increased in importance. But, little work has been done on performance evaluation of different restoration schemes in multimedia backbone design. Although simulations show that large IP networks are able to achieve sub-second convergence for the routing protocol by tuning the routing protocol's timers [24], service providers have not adapted such schemes due to concerns regarding network stability. Other schemes like failure insensitive routing [25] apply to unicast routing and are not suitable for our environment involving multicast flows.

IV. The Architecture: Hitless Multicast Reconfiguration Due to Routing Changes

First we describe today's typical carrier IP backbone that may be used for IPTV distribution. We assume that the protocols used for distributing video are PIM-SSM over IGP routing (Layer 3) with FRR at Layer-2 to recover from failures (as deployed in at least one carrier network). The underlying changes in the Layer-3 topology are first propagated to the rest of the nodes, using standard IGP (e.g., OSPF or IS-IS) techniques. This involves first the link failure detection through a lower layer indication such as SONET or Ethernet alarms or the lack of receipt of HELLOs within a *RouterDeadInterval* [12]; then, the propagation of LSAs via flooding and subsequent topology convergence, which may be a function of both computation of the SPF tree, as well as the *spfDelay and spfHoldTime* timers [26]. These timers are likely to be set to their default or conservative values in a carrier network in the interests of stable operation, resulting in an IGP convergence time on the order of several seconds. Afterwards, the PIM-SSM tree has to be reconfigured, which again may take several seconds to tens of seconds (e.g., a new join request is issued after 30 seconds, as part of the standard process of refreshing the soft-state for IP multicast).

The PIM-SSM tree is typically reconfigured after an IGP reconvergence event in a distributed manner, where each router independently computes the shortest path to the source for each multicast group. After the path recomputation process "settles", the routers independently install the new SPF tree and modify the routing and forwarding tables on their line cards. Next, the portion of the tree downstream of the failure is formally reconfigured by each router issuing a join request to attach to its new parent node, followed by a *prune* message to delete the previous path (a unidirectional link is ***upstream/downstream*** of a given router if it points towards/away-from the source). During this process, packet loss can occur (i.e., a *hit*). We wish to avoid or minimize this potential second hit after a failure (and its subsequent FRR rerouting) as we cost-out the pseudowire and reconverge to a new tree. Also, the method has to be careful to avoid another hit occurring after the link failure is repaired (e.g., to return to the original PIM tree).

In our method, after a failure, the routers coordinate the calculation of the shortest path by implementing ***local*** decisions on choosing new parent nodes (and the corresponding branches) in the new multicast tree and then send a join request to build a path from its new parent. Before describing the protocol details in the next section, we first give an intuitive feel for some of the challenges and our solutions.

A. Challenges in Designing a Robust Switchover Protocol

The **first challenge** is that the current PIM-SSM multicast does not have an explicit acknowledgement to a join request. So it is possible that the router stops receiving packets because it sent a prune request to the old parent but the join to the new parent was not successful. Our solution is a simple modification so that the prune message to remove the branch to the previous parent is not sent until the router receives PIM-SSM data packets from its new parent for the corresponding (S,G) group. The ***soft state*** approach of IP Multicast (refresh the state by periodically sending a join) is also used to ensure consistency. We use this principle to guide our tree reconfiguration process at a node in reaction to a failure. In this way, routers do not lose data packets during the switchover period. Of course, this primarily works in the PIM-SSM case, with a single source.

The **second challenge** is that the link down LSAs reach different routers at different times so that different routers may have different local views of the network topology. As an example, consider Figure-5 where we are trying to switch from the old tree to the new tree after failure of link (3,4). Router-4 needs to send a join to its new parent, router-2. Under standard rules in PIM-SSM, when router-4 sends a join request to router-2, it must stop forwarding packets to router 2, else it would form a loop. However router-2 was receiving packets from router-4 in the old multicast tree and if may not have learned of the topology change (and sent a join to its new parent, router-1) then router-2 (and consequently router-4) will stop receiving packets. The naïve solution is to force every router to wait until enough time has elapsed for all routers to learn of the changes. But this can be quite wasteful. E.g., routers 2 and 4 are not affected by router-5 so they should be able to switch to the new multicast tree without router-5 learning of the topology changes. This is especially important in a large wide area network. In order to speed up the protocol, router-4 would like to send a join as soon as router-2 learns of

978-1-4244-5489-1/10/ $26.00 © 2010 IEEE

the new topology. The key observation is that as soon as router-2 learns the new topology and starts receiving packets from its new parent router-1, it has to send a prune to its old parent router 4. This prune is the indication to router 4 that it can start receiving packets from router 2. Intuitively it is easy to see that this is the earliest router-4 can send a join request to router-2 and also notice that this decision happens independently at each router; there is no requirement that any other router has to learn of the topology changes. If one were to visualize the old tree morphing into the new tree, each branch changes as soon as it learns of the new changes. This change in protocol is accomplished by introducing a "waiting to send join" state.

The approach described in this paper potentially obviates the need for the intelligent weight setting algorithm described in [10], because any short-lived overlap will be corrected by the switchover. By lowering the priority of the packets flowing on the backup path (as described below), the short-lived overlap will also not cause data loss, as perceived by the receiver.

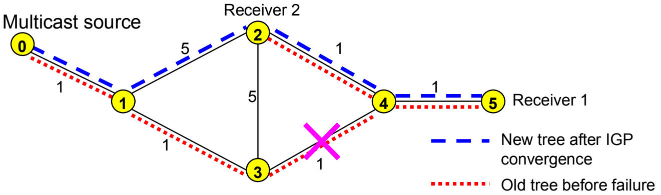


Figure 5. A sample link failure causing a previously downstream router to be the upstream one on the new multicast tree.

V. PIM Reconfiguration with Hitless Switchover

After a failure has occurred, our method uses modifications to the PIM-SSM protocol to achieve the goals stated in the previous section. We now describe assumptions and prove correctness for our multicast reconfiguration protocol.

A. Assumptions

We first make the following consistency assumptions (later, we discuss potentially relaxing these).

1. The weight of link (i,k) equals the weight (k,i) (applying to either the pseudowire or PHY link).
2. Links (i,k) and (k,i) are either both up or both down.
3. The directed least-weight path between any two routers is node and link unique. This can be always guaranteed by implementing a consistent tie-breaking rule across all routers. Note, this assumption combined with assumption 1 ensures that the least-weight path between a pair of routers $i \rightarrow k$ is the reverse of the least-weight path from $k \rightarrow i$.
4. The IGP reconfiguration process is completed in each router before a join request in response to the topology change is sent and every router has the same view of a (non-disconnected) SPT. We use this assumption to prove the correctness of our protocol. We then relax this assumption and show the protocol correctness still holds.
5. The failure is such that no packets are lost immediately after all FRR reroutes are completed (otherwise "hitless" reconvergence has no context in which to be established).
6. If a Layer 1 failure causes multiple, simultaneous failures of Layer 3 links, receiving the corresponding LSAs at each router gives a complete view of the topology prior to SPT recomputation (i.e., a single multicast tree computation after the LSAs are received reflects the new topology)
7. If multiple (non-simultaneous) failures occur, then sufficient time occurs between the failures for our algorithm to settle.

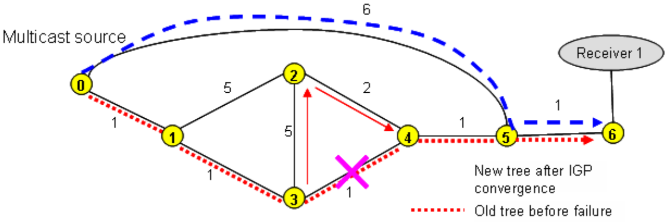


Figure 6. Protocol convergence example.

B. Major Tasks and Rules

The following tasks/rules describe our method.

Rule (a): *Expose link failure to IGP routing even though FRR backup path is in use.*

Rule (b): *Notify multicast protocol about IGP routing changes so that it can reconfigure whenever possible.* A router determines its upstream interface from its most up-to-date SPT (as described in Rule a). If this is a new upstream interface, then this router (say k) sends a join request to the upstream router (say j) ***except*** if router j is one of the current downstream interfaces for router k. In this latter case, router k locally designates the upstream interface with the pending state "waiting-to-send-join". Router k will also clear all pending "waiting-to-send-join" states on an upstream interface when it receives a subsequent LSA and computes an SPT that implies a different upstream interface.

Rule (c): *Prune old upstream interfaces only after data packets are received on the new upstream interface.*

Rule (d): *Clear all pending joins when no downstream interface is left, upon reception of prune(s) from downstream routers.* When router k receives a prune from (downstream) router j, router k prunes the interface to router j and does the following: (case d1) if router k has no remaining downstream interfaces (including receivers), then router k sends a prune message to its upstream interface and clears the "waiting-to-send-join" on any pending upstream interfaces; (case d2) else if router k has at least one remaining downstream interface (to a router other than j or receiver) and is in "waiting-to-send-join" status to router j, it sends a join request to j.

C. Algorithm Description

We now describe the major steps of the protocol via an example. Consider Figure 6:

1) Router 4 locally recognizes the failure of PHY links (3,4) and (4,3).
2) Link level FRR re-routes traffic over pseudowire (3,4) to its backup paths; pseudowire (4,3) is similarly re-routed to its backup path (not shown). Note that pseudowire (3,4) is a link of the pre-failure multicast tree.
3) Routers 3 and 4 broadcast LSAs to all their neighboring routers indicating that all the pseudowires and PHY links between {3,4} are down (or assigns them very high link weights). Note, however, that routers 3 and 4 do not take the pseudowires down yet.
4) Upon receipt of the LSA and after a hold-down period, each router with existing downstream interfaces on the old

multicast tree locally computes a SPT and determines its new upstream interface on the SPT. Recall that PIM-SSM computes paths in the reverse direction (sink-to-source) of their actual multicast packet flow. This results in router 4 needing to send a join request to router 5. However, since router 5 is also a current downstream interface of router 4, router 4 marks the interface with a "waiting-to-send-join" status. After receiving the LSA and updating SPT, Router 5 sends its join request to router 0. Since other routers with downstream interfaces (on the old tree) do not change their upstream routers, no other join requests are sent in response to the LSA update after the failure.

5) Router 0 receives the join request from router 5, adds its interface as a downstream interface, and begins to send packets to router 5.
6) When router 5 receives the first packet from router 0, it discontinues forwarding packets from router 4 and sends a prune message to router 4.
7) Since router 4 has no remaining downstream interfaces on the new multicast tree (case d1), it sends a prune message to router 3 and clears the "waiting-to-send-join" status to router 5. However if router 4 had a receiver attached to it, this would require case d2. Router 4 would have to send a join to router 5 only after receiving a prune from router 5 for the old tree once the data starts flowing to it from router 0. Subsequently, router 4 would release the "waiting-to-send-join" state and send the join to router 5. Finally, it would prune the old interface to router 3."
8) Since router 3 has no remaining downstream interfaces on the new multicast tree, it takes down both pseudowires and sends a prune request to router 1.
9) Since router 1 has no remaining downstream interfaces on the new multicast tree, it sends a prune request to router 0.

Without going through the details, once PHY link {3,4} is repaired, OSPF reconvergence is triggered when the LSAs are sent. Then, the multicast tree reconfiguration is triggered by router 5 sending a join request to router 4 and pruning router 0 once packets flow down the original tree from router 4.

Some of our key changes to create a "hitless" version of PIM-SSM [7] are contained in Rule (b) that implements the "waiting-to-send-join" and Rule (c) that sends prunes only after it receives packets from the new upstream interface. Simple as these changes appear, these rules have been carefully crafted to guarantee the process converges and is hitless.

D. Correctness Proofs

There are three major claims that establish the credibility and correctness of the algorithm described.

Claim 1: After a failure, eventually the protocol converges and the set of all links of the ***actual multicast tree*** are equivalent to the ***computed multicast tree*** (CMT). The *computed multicast tree* is defined to be the tree theoretically computed from the new SPT and the *actual multicast tree* is defined to be the tree established by the routers working independently according to our rules.

Claim 2: Except for packet loss due to potential FRR path overlap and packets "in flight", there is no packet loss due to convergence from the old tree to the new tree.

Claim 3: Packet loss will be significantly reduced due to potential FRR path overlap using a packet QoS prioritization scheme where packets sent over a pseudowire backup LSP path have lower priority than other packets.

Proof of Claim 1: We will prove that every link in the CMT also appears in the actual multicast tree. Notice that this is sufficient to prove that the two trees are identical because if the actual multicast tree contained any links not in CMT, it would imply that the actual multicast tree has all the links of a spanning tree CMT and at least one more link, implying a cycle. Assumption 4 guarantees that all routers eventually calculate the same (new) SPT. Let us start at an arbitrary leaf node contained in the CMT with a receiver interfacing on it. Rule (a) guarantees that any ***actual*** upstream link is contained in the new CMT and rule (c) guarantees that it prunes any upstream links not contained on the CMT. Recursively, as we follow the path of actual upstream links, if we encounter any upstream link not on the CMT, then this would again violate Rules (a) and (c). If we repeat this process of following the path on all leaves of the CMT, we will eventually encounter every node contained in the CMT. Thus, we have shown that for every node contained in the CMT, its actual upstream links are also contained in on the CMT. By our earlier observation, this completes the proof of Claim-1.

Proof of Claim 2: Given assumption 5, all receivers continue to receive packets before the first join or prune request is sent. Thus, if we examine a receiver and its path at any point in time from source to receiver and then examine how each join or prune request affects this path, we see a join request along this path will not cause packet loss because of Rule (c): the intermediate router must receive packets from a new upstream interface before turning off the old upstream. Rule (c) and (d) also guarantee that a prune will not be sent upstream unless 1) packets are arriving from a new upstream interface or 2) there are no more downstream interfaces. This concludes the proof.

While we don't prove Claim 3 in a formal manner, we evaluate the performance improvement through simulation given the capability for such QoS mechanisms in routers. Figure 7 is an example where the new tree might overlap on links (we call these "common links" (CL)) of the FRR path of the failed pseudowire. The dotted red tree is the original tree before the failure of link 3-4. When link 3-4 fails, the pseudowire switches to FRR backup path 3-2-4, which now overlaps with link 2-4. The inefficiency is because the same multicast packets will be carried over CL 2-4 more than once.

A possible way of solving the Congested CL problem is to assume that routers can process prioritized packets. In Figure 7, when link 3-4 fails, the IGP is then informed about the failure. Router 3 forwards packets on the backup path and marks them as lower priority packets. As the join process evolves, router 4 will send a join request to router 2 and then router 2 sends a join request to router 1. When router 1 begins to transmit to router 2, these packets will eventually arrive at router 2 along with packets from the backup path 3-2-4 at router 2. However, router 2 does not "see" any multicast packets from router 3 because the FRR path is tunneled through at Layer 2. However, Layer 2 forwarding protocols will recognize that the packets from router 3 have lower priority than those from router 1. If the link has insufficient capacity to handle the double multicast flow, then the packets from router 3 will experience loss. However, because of rule (c) described earlier, the first packet

978-1-4244-5489-1/10/ $26.00 © 2010 IEEE

from router 2 that is received at router 4 will cause the packets from router 3 (via the 3-4 pseudowire that routes over the backup FRR path 3-2-4) to be ignored. This will result in little if any packet loss. Any packet loss that occurs will be due to the switching mechanism to implement rule (c). Therefore, as stated earlier this is a function of router implementation, rather than a mathematical proof.

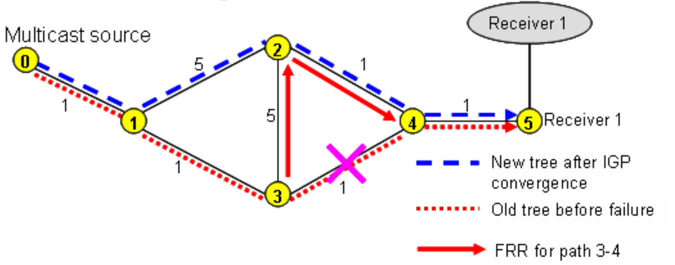


Figure 7. Example of FRR overlap.

E. Relaxation of Assumptions for Practical Operation

Now that we have established the basic validity of our approach, we discuss potential relaxations of the assumptions 1-7. We generally do not recommend relaxing assumptions 1 through 3. In particular, without assumption 3, the CMT becomes ambiguous. Finally, assumptions 6-7 are mostly provided to give clear context for the proofs. However, if one examines these proofs, they demonstrate that we can start from any "actual" multicast tree (even one that is in transition to a target state) and the process will converge to the eventual CMT once all the routers' views of the SPT settle. Thus, we feel assumptions 6-7 can be relaxed without negative impact.

The strongest assumption is #4. In reality, there may be situations where we need to relax this assumption when the variation of completion time of the IGP reconfiguration process across the routers is large. Then, it is possible a router may send a join request to an upstream interface before the upstream router has computed its new SPT. This would result in the process possibly proceeding up the wrong tree. However, it is easy to see that any router that had chosen the wrong upstream interface will, after receiving the LSA, updating its SPT and, following application of rule (a), recalculate its upstream interface per the new SPT and then send a join request to the correct upstream interface; therefore the process will eventually converge to the CMT. As demonstrated in Claim 2 earlier, no packets should be lost as this tree building process converges.

Another complication of relaxing assumption #4 is that because of potential "waiting-to-send-join" states, we need to demonstrate that the protocol does not "deadlock". Suppose we enter a state where a cycle of routers, $(i_1 ,i_2 , \ldots ,i_n ,i_{n+1})$, is deadlocked, and router i_{k-1} is "waiting to send a join" to router i_k and $i_{n+1} = i_1$. That is, router i_k is upstream of router i_{k-1}. Once the SPT has been updated consistently in all routers, rule (b) implies that these upstream links are chosen so that they are contained in the SPT in a path towards the source. However, they form a cycle, which contradicts the definition of an SPT.

Another option to is to impose a timer to insure the SPTs have all been updated before the first join is sent (thus ensuring Assumption #4 is met); however, because there so many complex timers and interactions among timers and protocols already in router-based networks (and in multicast networks) we suggest that this latter option be avoided.

Another issue regarding relaxing of assumption #4 is that due to inconsistent completion of SPT computations among the routers, it is possible for a router to receive a join request from a downstream router but then needs to send an (upstream) join request to the same router. Current PIM-SSM protocol does NOT admit the join request from the downstream router if it is not in its current SPT. Reconfiguration of the SPT or soft-state refresh messages is used to eventually correct this situation. This requirement can be accommodated in our modifications and convergence to the CMT can be established. However, we make the observation that with our new rule (b), indeed the join request from the router can be allowed (even if not on the SPT of the upstream router), but the upstream join request back to that router is put into the "waiting-to-send-join" state. Once the router updates its SPT and the correct upstream interface is calculated, the "waiting-to-send-join" state will be cleared and it will converge to the correct configuration. Strictly speaking our new protocol does not require soft-state refreshes to rectify this situation. However, as various error conditions occur and violate our assumptions, it is recommended that current PIM-SSM soft-state rules be retained to handle these situations.

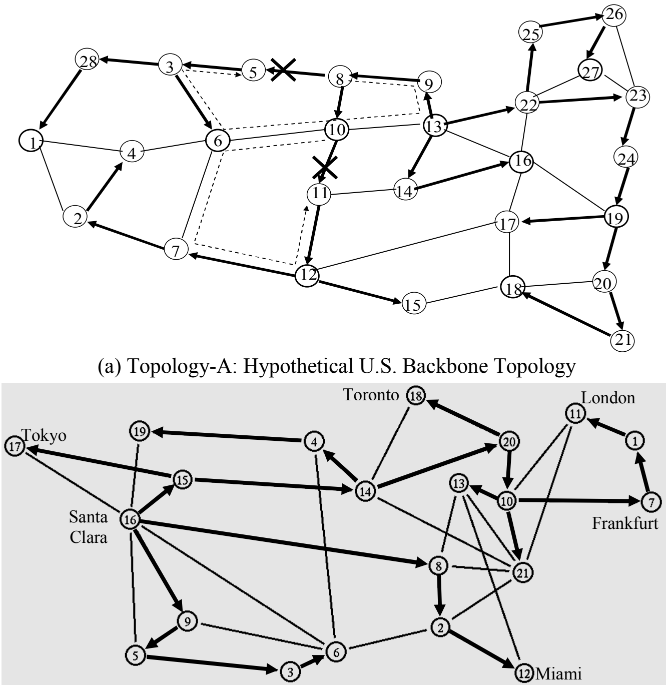
We generally do not recommend relaxing assumptions 1 through 3. In particular, without assumption 3, the CMT becomes ambiguous. Finally, assumptions 6-7 are mostly provided to give clear context for the proofs. However, if one examines these proofs, they demonstrate that we can start from any "actual" multicast tree (even one that is in transition to a target state) and the process will converge to the eventual CMT once all the routers' views of the SPT settle. Thus, we feel assumptions 6-7 can be relaxed without negative impact.

VI. Simulation Results

We evaluated our cross-layer failure restoration framework using ns-2 [30]. Specifically, our experiments aim to reveal how much *packet loss* occurs during switchover after failure. We used MPLS [31] to provide FRR support at the routers and implemented PIM-SSM over MPLS in ns-2. For the IGP routing, we used an OSPF implementation in ns-2 [12]. We made the necessary changes to the protocols so that OSPF is informed about link failures (even though MPLS forwards the data packets on the FRR path immediately after the failure) and PIM-SSM is informed about OSPF routing changes.

A. Experimental Setup

We performed our simulations on two different topologies: (1) **Topology-A** is a hypothetical US backbone topology (with 28 routers and 45 links) shown in Figure 8(a), with multicast source set at router 13 at the center of the network, and (2) **Topology-B** is the "Exodus" topology from Rocketfuel (with 21 routers and 36 links) shown in Figure 8(b), with multicast source set at router 16, which corresponds to its Point-of-Presence (PoP) in Santa Clara. We assigned equal capacities to the links but set their propagation delays proportional to their lengths. The multicast source generated UDP traffic with a packet size of 500B. The rate of the multicast traffic was 70% of the capacity of the links. We budgeted for 120ms of buffer time, i.e., a link with a 100Mb/s capacity had a buffer size of 3000 packets. We used default OSPF timer settings, e.g., *spfDelayTime*=5 secs, and *spfHoldTime*=10 secs, and a relatively short rejoin interval of 30 secs for PIM-SSM. This timer is normally set to several minutes, which would make the improvements resulting from our protocol better than is reported here.



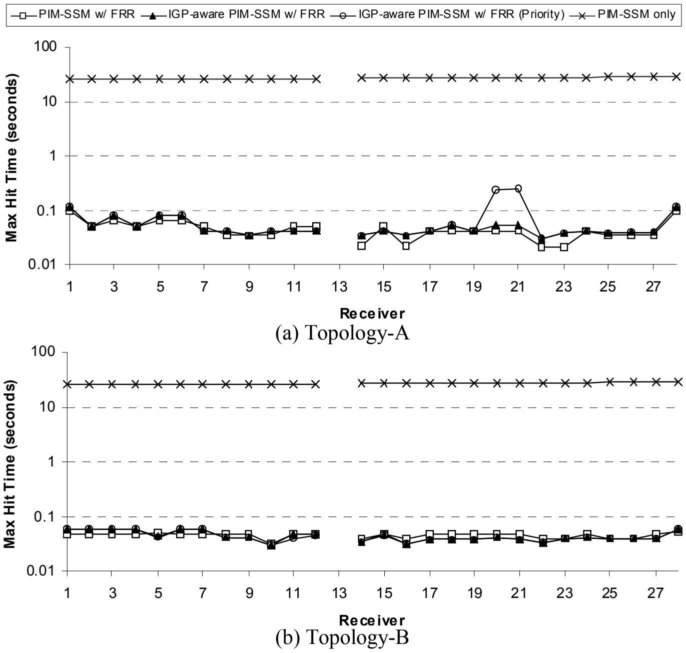
(a) Topology-A: Hypothetical U.S. Backbone Topology

(b) Topology-B: Exodus Backbone Topology (from Rocketfuel)

Figure 8. Experimental Topologies

Our simulation scenarios consisted of failing link(s) in the topology and observing the multicast data traffic during re-convergence of the protocols. We measured (a) packets lost at each receiver until the failure restoration is complete, and (b) the time taken to complete failure restoration. These two metrics, (a) and (b), quantify the "hit" caused by each protocol. We examine both the average and the maximum for the hit time and packet loss over the length of each experiment.

We compared four scenarios: (i) PIM-SSM only (i.e., no FRR support and no IGP-aware tree reconfiguration), (ii) PIM-SSM w/ FRR (i.e., FRR support exists but no IGP-aware tree reconfiguration), (iii) IGP-aware PIM-SSM w/ FRR (i.e., both FRR support and IGP-aware tree reconfiguration are used – **our proposal**), and (iv) IGP-aware PIM-SSM w/ FRR with data over the backup path being forwarded at higher priority (see Claim 3 of Section IV). For "IGP-aware PIM-SSM w/ FRR" with priority, we divided the link buffer equally between the two priority classes. For single failure cases, we evaluate the evolution of the multicast reconfiguration protocol for a period equal to the time for the PIM-SSM rejoin timer (i.e., 30 secs) after a failure occurs. As described below, for multiple failure cases, we evaluate the performance until the failed links are repaired and IGP and multicast protocols have re-converged completely. Note that the PIM-SSM convergence time is normally 3-4 times longer than the rejoin timer, as all unnecessary downstream branches will have to be pruned. However, we are conservative and use only the rejoin time in our comparison, since PIM-SSM guarantees that multicast traffic will flow to all receivers after the timer expires. We assigned OSPF link weights using two approaches: a) by using the algorithm in [10] which assures that single failures do not cause any traffic overlap for the "PIM-SSM w/ FRR" case ("**Intelligent Link Weights**" in the figures), and b) by assigning identical weights to all links ("**Equal Link Weights**" in the figures). When selecting the FRR paths, we choose the least-cost non-overlapping path that would restore the failure. This experiment setup favors the "PIM-SSM w/ FRR" case and our comparison of the four cases reveals the extra improvement our multicast reconfiguration protocol attains.



(a) Topology-A

(b) Topology-B

Figure 9. Maximum hit time observed at each receiver for single failures (Intelligent Link Weights)

B. Single Failures

To compare the four protocols for a single link failure, we failed each link in the multicast tree one by one and measured the maximum time that each of the receivers experience a hit (i.e., disconnect) as well as the maximum percentage of packets lost for each receiver during a 30 second time interval, which is the maximum failure restoration time based on the PIM-SSM rejoin timer of 30s. In Figure 9, we observe that the maximum time that a receiver experiences a hit can be quite large (10s of seconds) with just PIM-SSM. With the "IGP-aware PIM-SSM w/ FRR" schemes (both with and without prioritization of FRR traffic), receivers experience a significantly smaller amount of hit time, the same as "PIM-SSM w/ FRR", (100 ms). The minimum hit time is the time that FRR initially takes to detect the failure and switch over to the back up path. The main point here is that our cross-layer mechanisms (i.e., IGP-aware PIM-SSM) do not add any additional impairment at the receivers. In the rest of this subsection, because of lack of space, we only show the results for Topology-A. Figure 10 presents a similar result in that the maximum of the percentage of lost packets is dramatically lower with mechanisms using FRR in contrast to PIM-SSM alone. Prioritization of IGP-aware PIM-SSM w/FRR traffic helps reduce the loss for most cases.

We also examined the case where all the links have equal weights. As shown in Figure 11-a, the relative performance of the four protocols does not change significantly for the maximum hit time. But, the maximum of the percentage of lost packets grows dramatically (Figure 11-b), except when prioritization of the FRR traffic is used, i.e., “IGP-aware PIM-SSM w/ FRR (Priority)”.

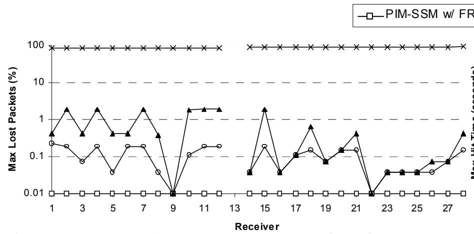


Figure 10. (Top-A) Max percentage of packets lost during single failure restoration (Intelligent Link Weights)

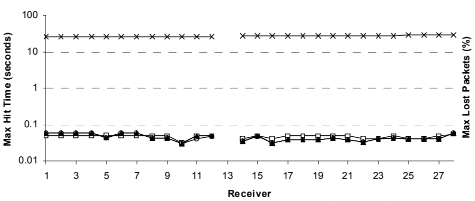


Figure 11-a. (Top-A) Max hit time observed for single failures (Equal Link Weights)

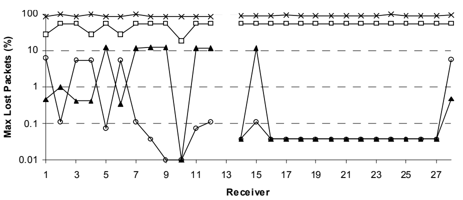


Figure 11-b. (Top-A) Max percentage of packets lost during single failure restoration (Equal Link Weights)

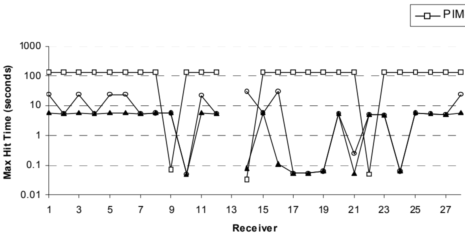


Figure 12-a. (Top-A) Max hit time observed for double failures (Intelligent Link Weights)

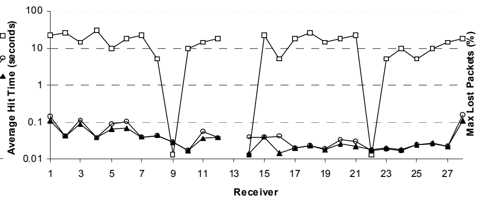


Figure 12-b. (Top-A) Average hit time observed for double failures (Intelligent Link Weights)

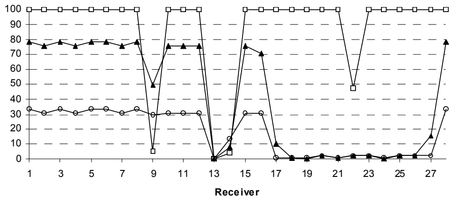


Figure 12-c. (Top-A) Max percentage of packets lost for double failures (Intelligent Link Weights)

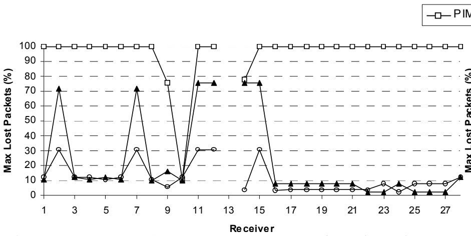


Figure13. (Top-A) Max percentage of packets lost for double failures (Equal Link Weights)

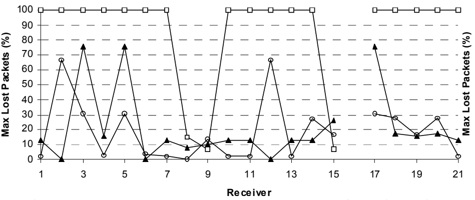


Figure 14. (Top-B) Max percentage of packets lost for double failures (Intelligent Link Weights)

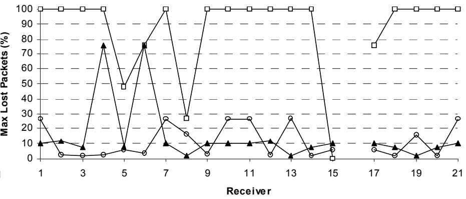


Figure 15. (Top-B) Max percentage of packets lost for double failures (Equal Link Weights)

Overall, the results for single failures show that by using the IGP-aware PIM-SSM, there is less of a requirement for the intelligent weight setting algorithm. Importantly, the IGP-aware mechanism to reconfigure PIM-SSM after FRR consistently achieves little or no packet loss beyond what PIM-SSM with FRR achieves. Indeed, the only packet loss that takes place in our scheme is due to the packets lost in transit on the failed link, and the potential congestion on the FRR path during the transient period while our scheme reconfigures the PIM tree. During this transient period, we may observe the common congested link situation, where lowering the priority of the traffic sent over the backup path reduces loss further.

C. *Double Failures*

The real benefit of IGP-aware multicast reconfiguration is when there are multiple concurrent failures in the network. To understand the significance of this benefit, we simulated the case where there are overlapping failures of two links, i.e., double link failure. The first link fails at the 10th second and the second link at the 50th second of the simulation. Then, the first and second links are repaired at the 200th and 250th seconds respectively. We measure performance from the time of the first failure until both the links are repaired and IGP and multicast protocols have re-converged completely. We simulated every possible combination of 2-link failures involving any of the link pairs causing at least one reconfiguration of the PIM tree. Figure 8(a) shows an instance of the multicast tree for a particular double link failure. Some of these link pairs actually cause some receivers to become disconnected. So, to compare the protocols, we examine the “average” hit time and packet loss in addition to the “maximum” hit time and packet loss. We do not evaluate the “PIM-SSM only” case under double failures as it performs much worse than the other three cases even for single failures.

Figures 12-15 show the impact of double failures for various intelligent or equal link weight settings on the topologies. Because of lack of space, we show only the maximum % packets lost in Figures 14 and 15 for Topology-B. These results clearly show that IGP-aware multicast reconfiguration handles multiple failures much better than the FRR-only approach. Figure 12-a shows that the IGP-aware cases can restore double failures within 5s while PIM-SSM with FRR takes more than 100s (based on the *maximum* hit time). The link weight setting method does not change this relative performance of the alternatives. Note that the reason why IGP-aware protocols still take up to 5s to restore is because the second failure may have an FRR path that utilizes the first failed link. This may only be restored by repairing the link or using a dynamic reconfiguration of the FRR paths that may be impacted by the failure of the first link (as suggested in [32]). Without such a dynamic reconfiguration, control packets (e.g., prune) may also be lost, resulting in longer restoration times.

Figures 12-b shows the *average* performance impact of double failures on each receiver. The IGP-aware protocols achieve less than 100 msecs average recovery time while PIM-SSM with FRR spends more than 10 secs for most of the receivers. For Topology-B (results not shown), the average recovery time was about 150 msecs for the IGP-aware protocol but more than 10 secs for PIM-SSM with FRR. Similarly, as shown in Figures 12-c and 13 for Topology-A and Figures 14 and 15 for Topology-B (maximum % of lost packets), the experiments show that PIM-SSM with FRR can lose up to 100% of the packets while the IGP-aware protocols up to 80%. However, looking at the average packets lost in Figure 12-b, PIM-SSM with FRR loses about 10% of the packets while the IGP-aware protocols lose less than 0.1% of the packets during restoration from a double failure. Again, the link weight setting method does not change this relative performance noticeably.

VII. Summary

The tight QoS constraints of distributing real-time multimedia require rapid restoration. Production networks rarely shorten their timers enough to achieve sub-second IGP convergence because of the potential of false alarms. The standard solution is to use layer-2 fast reroute (FRR) based restoration. However, layer-2 FRR hides the backup path from the IGP protocol. If the FRR backup path is left to stay up during the period of the link failure, this can have a negative impact if a subsequent failure occurs. It also requires very carefully chosen backup paths with intelligent weight setting to avoid link congestion due to potential path overlap.

We proposed protocol modifications to cure these shortcomings. Our methodology makes the IGP layer aware of the failure, notifies the multicast agent of the routing changes, re-converges the network quickly to the new multicast tree, and takes down the FRR backup path, all in a “hitless” manner. When the failure is subsequently repaired, it restores the normal routing, again in a “hitless” manner. Our method quickly re-converges the network and has the added advantage of working well even without intelligent weight setting for the backup path.

We demonstrate the efficacy of our schemes by running ns-2 simulations over two different network topologies. As an example of possible improvements, the average recovery time on double failures (which can be almost 20% of all link failures) is more than 10s for most of the receivers for PIM-SSM but reduces to 100ms with our enhancements.

References

[1] http://www.sbc.com/gen/press-room?pid=5838

[2] http://www22.verizon.com/FiosForHome/Channels/fios/FiosTV_comingsoon.aspx

[3] http://www.iptvnews.net

[4] http://www.sbc.com/Common/files/pdf/IPvideonetwork.pdf

[5] A. Adams, J. Nicholas, and W. Siadak, “Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol specification (revised),” *IETF RFC 3973*, 2005.

[6] D. Estrin, Ed., "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol specification", *IETF RFC 2362*, 1998.

[7] S. Bhattacharyya, Ed., “An overview of Source-Specific Multicast (SSM),” *IETF RFC 3569*, 2003.

[8] E. Rosen, A. Viswanathan, and R. Callon, “Multiprotocol Label Switching architecture,” *IETF RFC 3031*, 2001.

[9] P. Pan, G. Swallow, and A. Atlas (Editors), “Fast reroute extensions to RSVP-TE for LSP tunnels,” *IETF RFC 4090*, 2005.

[10] R. Doverspike, G. Li, K. Oikonomou, K. K. Ramakrishnan and D. Wang, “IP Backbone Design for Multimedia Distribution: Architecture and Performance,” *Proc. of IEEE INFOCOM*, April 2007.

[11] S. Kini, S. Ramasubramanian, A. Kvalbein, and A. F. Hanse, “Fast Recovery from Dual Link Failures in IP Networks,” *Proc. of IEEE INFOCOM*, 2009.

[12] M. Goyal, K. K. Ramakrishnan, and W.-C. Feng, “Achieving faster failure detection in OSPF networks”, *Proc. of IEEE ICC*, May 2003.

[13] M. Yuksel, K. K. Ramakrishnan and R. Doverspike, “Cross-Layer Failure Restoration Techniques for a Robust IPTV Service”, *Proc. of IEEE LANMAN Workshop, Sept. 2008.*

[14] A. Mahimkar et.al., “Towards Automated Performance Diagnosis in a Large IPTV Network”, *Proc. of ACM SIGCOMM*, 2009.

[15] M. Cha, W. A. Chaovalitwongse, Z. Ge, J. Yates, and S. Moon, “Path protection routing with SRLG constraints to support IPTV in WDM mesh networks,” *Proc. of IEEE Global Internet Symposium*, 2006.

[16] K. Nahrstedt and R. Steinmetz, “*Multimedia Fundamentals, Volume 1: Media Coding and Content Processing*”, 2nd Ed. Prentice Hall, 2002.

[17] B. Quinn, K. Almeroth, “IP multicast applications: Challenges and solutions,” *IETF RFC 3170*, 2001.

[18] D. Velten, R. Hinden, and J. Sax, “Reliable Data Protocol (RDP),” *IETF RFC 908*, 1984.

[19] M. Cha, S. Moon, C.-D. Park, and A. Shaikh, “Placing relay nodes for intra-domain path diversity,” *Proc. of IEEE INFOCOM*, April 2006.

[20] G. Li, D. Wang, and R. Doverspike, “Efficient distributed MPLS P2MP fast reroute,” *Proc. of IEEE INFOCOM*, April 2006.

[21] R. Braden (Ed.), L. Zhang, S. Berson, S. Herzog, and S. Jamin, “Resource ReSerVation Protocol (RSVP) -- version 1 functional specification,” *IETF RFC 2205*, 1997.

[22] C.-C. Wen, C.-S. Wu, and K.J. Chen, “Centralized Control and Management Architecture Design for PIM-SM Based IP/MPLS Multicast Networks,” *Proc. of IEEE GLOBECOM*, 2007.

[23] T. Speakman et al., “PGM Reliable Transport Protocol Specification,” *IETF RFC 3208*, 2001.

[24] P. Francois, C. Filsfils, J. Evans, and O. Bonaventure, “Achieving sub-second IGP convergence in large IP networks,” *ACM SIGCOMM Computer Communication Review*, 35(3), pp. 35-44, July 2005.

[25] S. Lee, Y. Yu, S. Nelakuditi, Z.-L. Zhang, and C.-N. Chuah., “Proactive vs. reactive approaches to failure resilient routing,” *Proc. of IEEE INFOCOM*, March 2004.

[26] J. Moy, “*OSPF Anatomy of an Internet Routing Protocol*”, Addison Wesley, 2000.

[27] K. Oikonomou, R. Sinha, and R. Doverspike, “Multi-layer network performance and reliability analysis”, *International J. of Interdisciplinary Telecommunications and Networking*, 1(3), pp 1-30, 2009.

[28] W. Tan and A. Zakhor, “Video multicast using layered FEC and scalable compression,” *IEEE Transactions on Circuits and Systems for Video Technology*, 11(3), pp. 373-386, 2001.

[29] Y. Xiong and L. Mason, “Restoration strategies and spare capacity requirements in self-healing ATM networks,” *IEEE/ACM Transactions on Networking*, 7(1), pp. 98-110, 1999.

[30] The Network Simulator, ns-2. http://www.isi.edu/nsnam/ns.

[31] G. Ahn and W. Chun, “Design and implementation of MPLS network simulator supporting LDP and CR-LDP,” *Proc. of IEEE ICON*, 2000.

[32] A. Todimala, K. K. Ramakrishnan and R. K. Sinha, “Cross-layer Reconfiguration for Surviving Multiple-link Failures in Backbone Networks”, *Proc. of OFC 2009*, San Diego, March 2009.