# IP Multicast Fault Recovery in PIM over OSPF

*Abstract*—

**Relatively little attention has been given to understanding the fault recovery characteristics and performance tuning of native IP multicast networks. This paper focuses on the interaction of the component protocols to understand their behavior in network failure and recovery scenarios. We consider a multicast environment based on the Protocol Independent Multicast (PIM) routing protocol, the Internet Group Management Protocol (IGMP) and the Open Shortest Path First (OSPF) protocol. Analytical models are presented to describe the interplay of all of these protocols in various multicast channel recovery scenarios. Quantitative results for the recovery time of IP multicast channels are given as references for network configurations, and protocol development. Simulation models are developed using the OPNET simulation tool to measure the fault recovery time and the associated protocol control overhead, and study the influence of important protocol parameters. A testbed with five Cisco routers is configured with PIM, OSPF, and IGMP to measure the multicast channel failure and recovery times for a variety of different link and router failures. In general, the failure recovery is found to be light-weight in terms of control overhead and recovery time. Failure recovery time in a WAN is found to be dominated by the unicast protocol recovery process. Failure recovery in a LAN is more complex, and strongly influenced by protocol interactions and implementation specifics. Suggestions for improvement of the failure recovery time via protocol enhancements, parameter tuning, and network configuration are provided.**

## I. Introduction

Many IP multicast applications, for example, near real-time dissemination of financial information, require high availability. This problem has not received much attention so far. One exception is STRESS [1], a tool that automates the formal evaluation of PIM sparse-mode protocol robustness. However, STRESS does not include timers, and does not consider the interaction between unicast and multicast protocols.

Multicast group membership management, unicast routing protocols, and multicast routing protocols are all required to enable end-to-end multicast capabilities. In this paper, we investigate a complete multicast routing architecture consisting of IGMP [6] for multicast group membership management in a LAN, OSPF [4] for unicast routing, and PIM sparse-mode [8] and PIM dense-mode [7] for multicast routing. OSPF is chosen because of its rapid fault recovery properties, widespread use, and its support of parametrically tuning of fault recovery time, as compared with RIP which has long, non-tunable fail-over periods. The two variants of PIM are becoming the dominant multicast routing protocols. Other multicast protocols, such as DVMRP or CBT resemble dense and sparse mode, respectively, and we thus expect that many of our results apply to these and similar protocols as well. End-to-end multicast channel fault recover is a function of the interplay of all of these protocols and is thus the focus of this paper.

We investigate how quickly the multicast channel recovers when links and routers fail in a multicast network. We define a multicast channel as the state established in routers and hosts that allows a single sender to communicate with a group of receivers. We consider single link and router faults inside the network, but we assume that sending and receiving hosts, their LANs are reliable. Since fault recovery associated with rendezvous point (RP) failures in PIM SM have been studied extensively [8], this paper focuses on other mechanisms (router, link, LAN, WAN fail-over) that are not sufficiently addressed and are less well understood by the community.

The key aims of this study are: develop a detailed understanding of the protocol interactions and sequence of events under different failure scenarios; provide quantitative insight into the effect of protocol parameters on recovery time and overhead; develop general suggestions for parametric tuning of protocols and enhancements to protocol specifications and implementation. To achieve these objectives, we combine results from analytical analysis, simulations and testbed measurements.

In the analysis, we present the interactions of the protocols (PIM, OSPF, IGMP) with end-to-end multicast channel recovery under various network failure scenarios. We also develop some quantitative results that can be used as references for network configurations and protocol development. In addition, the analysis serves as a basis for our providing recommendations on the protocol enhancement.

Simulation models for IGMP, PIM DM and support tools were constructed using the OPNET [11] simulation platform. The simulation is used to measure the control costs of the trio of protocols during steady state and failure recovery scenarios, for various random topologies and with various parametric tunings. Furthermore, the simulation is used to validate the failure recovery results derived from the analytical models.

The experimental results were supplemented by studying the operation and failure recovery of the protocols on a testbed of five Cisco routers arranged in a simple topology. This enabled a basic demonstration of failure recovery on WAN and LAN, and also allowed us to identify some implementation-related issues that affect failure recovery.

The paper is organized as follows. Section II reviews IGMP, OSPF and PIM. Section III describes the topologies and configurations we used, as well as the chain of events caused by link or router failures. We also present several analytical multicast recovery models. Section IV and V present the simulation and testbed results, respectively.

## II. Overview of Protocols

The establishment of end-to-end multicast communication channels requires several protocols to work in concert. To establish a multicast channel over a native multicast enabled WAN, a sender application needs only to send UDP packets onto the LAN using a class D IP address (group address) in the destination field of the IP header. Multicast group information on a LAN is usually maintained by the IGMP protocol. The multicast enabled routers in the network are responsible for constructing the multicast channel, and extending it to the interested receivers; in our case, this is done using PIM DM or PIM SM. The multicast protocol constructs the multicast delivery tree using the metrics and topology information maintained by the unicast routing protocol; in our case, OSPF. Below, we briefly review these protocols.

### A. IGMP

IP Multicast delivery is selective; only those hosts that have expressed interest in joining the group will become attached to the channel. The IGMP protocol manages the group interests between hosts and their first hop routers. One multicast router on each LAN serves as a *Querier* for soliciting the group membership information by periodically sending a *General Query* message at the *Query*

*Interval* (default 125 s) to all hosts on the LAN. In response, a host sends a *Host Membership Report* message to the group address for each group to which it desires to belong, within a bounded random interval *Query Response Interval* (default 10 s). When a *Querier* receives such a *Host Membership Report*, it adds the group being reported to its membership list for the LAN.

### B. OSPF

OSPF is a link state unicast routing protocol that dynamically detects topology changes and calculates new loop-free routes after a period of convergence. Each router in the network maintains a replicated database. Routers execute Dijkstra's algorithm on their database to calculate a shortest-path route to a given destination subnet. Routers flood database information periodically or when network element failures occur.

OSPF is run within an autonomous system (AS) maintained by a single administration. An AS can be further divided into OSPF areas. Within each area, routers maintain identical topology databases. Each area requires Area Border Routers (ABR) at their periphery. An ABR is connected to multiple areas and has a copy of the topological database for each area. The ABR is responsible for the propagation of inter-area routing information into the areas to which they are connected. Finally, totally stubby areas are used to reduce the storage requirements of routers within those areas for a system in which a lot of inter-AS routes are defined. Topological information is not permitted to be flooded to totally stubby area routers.

OSPF utilizes several timers that affect its rate of convergence in the event of network element failures. OSPF neighbors send *Hello* messages to each other in every *HelloInterval* (default 10 s) and will time out the neighbor if no *Hello* message is received within the *RouterDeadInterval*. The recommended ratio of the *RouterDeadInterval* to *HelloInterval* is three to one. Both the intervals must be the same for neighboring routers. In the Cisco router OSPF implementation, two timers are used to control how soon Dijkstra's algorithm is executed to update the routing database. The *Shortest Path First(SPF) Delay* timer is the timer between when OSPF receives a topology change and when it starts a shortest path calculation, after reception of an Link State Advertisement (LSA). The *SPF Holding* time is the interval between two consecutive SPF calculations, representing the minimum interval in which back-to-back Dijkstra calculations can occur.

### C. PIM

PIM operates in either *Sparse Mode (SM)* or *Dense Mode (DM)*. PIM DM is a broadcast-and-prune protocol and is best suited for networks densely populated by receivers and with plentiful bandwidth. Each router copies incoming packets to all its interfaces except the one on which the data arrived. When the multicast channel reaches a leaf router [1], the group information maintained by the router is examined and either the multicast data is forwarded onto the LAN or the router prunes back the channel. The prune state has an associated timer; the broadcast-and-prune process repeats upon its expiration. If a new member wants to join a group, the directly connected router will send a *Graft* towards the source.

PIM SM is a multicast routing protocol that dynamically builds and maintains multicast trees. PIM SM is optimized for environments where group members are sparsely distributed across a wide area. Unlike PIM DM, which has a broadcast-and-prune phase, a Designated Router (DR) [2] sends periodic Join/Prune messages towards the *Rendezvous Point (RP)*[3]. A *Join/Prune* message is also sent when a new multicast entry is created. If the data rate of the tree reaches a predefined threshold, routers with local members individually migrate from the group's shared tree to a shortest path tree (SPT) rooted at the sender's router.

When two or more routers are forwarders for a multi-access network LAN, an *Assert* process is used to elect the router with best metric to the source (DM or SM SPT) or to the RP (SM) as forwarder. All other routers remove their *oifs* towards the LAN.

Several PIM timers provide fault recovery tuning capabilities. Each PIM router periodically send *Hello* to each other every *Hello-Period* (default 30 s) and a neighbor is timed out if *Hello* messages are not received from the neighbor within *Hello-Holdtime* (default 105 s). If a DR goes down, a new DR is elected. PIM (DM and SM) also has several timers that control the maintenance of state in the routers. A timer for each *outgoing interface (oif)* is used to time out that *oif*. In DM, it is reset whenever a data packet is forwarded or a *Graft* message is received. In SM it is reset when a *Join/Prune* message is received. Both of the timers will be reset to *Prune-Holdtime*. A timer for each route entry is used to time out that entry and is reset to *Data-Timeout* (default 180 s) when receiving data packets (DM or SM SPT) and is reset to the maximum prune timer among all its outgoing interfaces once all interfaces in the *oif* list are pruned. An *Assert-timer* is also used for an entry to time out received *Asserts* after *Assert-Timeout* (default 180 s).

## III. Network Failure Scenarios

When network element failure occurs in a network, IGMP, OSPF, and PIM asynchronously interact to recover a multicast channel. The analysis of PIM SM is restricted to shared trees (not shortest path trees) and thus does not address failure during the migration period of shared tree to shortest path tree. PIM SM and DM recover from network element failures in a similar manner. However, for recovering the part of the multicast channel upstream of a router, a router running PIM SM will send a *Join* message to its Reverse Path Forwarding (RPF) [4] router, while a router running PIM DM will send a *Graft* message. From herein, "PIM" is used to refer to both the DM and SM cases, unless otherwise specified. In this section, the analytical models for the various failover scenarios are shown. For convenience, parameters used in the analysis are defined in table I.

In general the total multicast channel recovery time for a affected router can be written as:

$$T_r = T_u^{ospf} + H^{ospf} * T_p^{ospf} + T_{spf} + T_{Dijkstra} + T_u^{pim} + T_p^{pim} \quad (1)$$

The major portion of $T_r$ is contributed by $T_u^{ospf}$, $T_{spf}$, and $T_u^{pim}$, all of which have a granularity in seconds. In contrast, $T_p^{ospf}$, $T_p^{pim}$, and $T_{Dijkstra}$ are typically in milliseconds, and are thus not considered further in the model.

Single-fault network failures can be classified into four categories: link failure in the WAN, router failure in the WAN, link

[1] A network on a router interface is deemed a leaf if there is no PIM neighbor on that network.

[2] The DR is responsible for sending *Join/Prune* and *Register* messages toward the RP. When more than one router is connected to a LAN, the highest IP addressed router becomes the DR

[3] An RP is a router that acts as the root of the shared tree, and to where all joins and prunes are directed

[4] For a shared tree, the RPF interface provides the best reverse metric to the RP. For a shortest path tree, the RPF interface provides the best reverse metric to the source

| | |
|---|---|
| $T_r$ | Multicast channel failure recovery time. |
| $T_{cd}$ | The "carrier_delay" time |
| $T_{fd}^{ospf}$ | OSPF failure detection time. |
| $T_u^{ospf}$ | OSPF topology updating time |
| $T_{hi}^{ospf}$ | OSPF *HelloInterval* |
| $T_{rdi}^{ospf}$ | OSPF *RouterDeadInterval* |
| $T_p^{ospf}$ | Propagation delay of an OSPF control message on a point to point serial link |
| $H^{ospf}$ | Number of hops from the router adjacent to the network failure |
| $T_{spf}$ | SPF execution delay time after topology updating |
| $T_{Dijkstra}$ | Dijkstra execution time on the router |
| $T_{pli}^{pim}$ | The interval for PIM to poll the unicast routing table |
| $T_u^{pim}$ | The time for PIM to detect topology change |
| $T_{hhi}^{pim}$ | PIM hello holding time for detecting neighbor failure |
| $T_{nfd}^{pim}$ | PIM neighbor failure detection time |
| $T_p^{pim}$ | Propagation delay for a PIM *Join/Graft* message to recover the multicast channel |
| $T_a^{pim}$ | PIM Assert-Time. |
| $T_{qi}^{igmp}$ | IGMP *Query Interval* |
| $T_{qri}^{igmp}$ | IGMP *Query Response Interval* |

TABLE I

PARAMETERS USED IN THE ANALYSIS

failure to the client site LAN, and router failure on the client site LAN.

### *A. Protocol Interaction in WAN*

The network recovery in WAN rests solely on the interactions between OSPF and PIM [5]. In general, an OSPF implementation should feed outage information received from the data-link and physical layers into its interface state machine (Section 9.3 of RFC 2328, event *InterfaceDown*) and from there into the neighbor state machine. Most routers are able to notice within a small number of seconds that their link is down, and then they should communicate this information via an updated *router-LSA* to the rest of the OSPF routers. The speed of the recovery depends on the vendor implementations and the "carrier-delay" parameters set up for detecting an outage. Depending on type of outage, circuit, and the switch vendor, an NBMA network over ATM or FR may not give the outage indication. Even when the lower levels know that the neighbor has gone away, many networking stacks don't pass this information up to the routing protocols. In these cases, the *RouterDeadInterval* of OSPF can be used as a last resort to detect a link failure.

As soon as each router receives the new *router-LSA*, it recalculates its shortest path through Dijkstra's algorithm. PIM can learn the topology change from OSPF directly through a "notify" message (if an implementation supports it) or indirectly by periodically polling the OSPF routing table (this function is implemented in the current Cisco routers). PIM needs to determine its RPF for each source in the source-group pair (S, G) or RP.

If a new RPF has been discovered, PIM sends a *Join/Graft* message on the new RPF interface to form a new multicast channel. As specified in the PIM SM specification, the router may also send a *Prune* message out the old *input interface (iif)*, if the link is operational, to remove this part of the tree. However, a race condition may exist in this case, depending on the delay of establishing the new branch of the multicast tree. If the delay is big, removing the old *iif* may lead to packet loss, since the new multicast channel has not been established. To avoid unnecessary packet losses during the transition phase, the authors suggest keeping both the old *iif* and corresponding upstream *oifs* functional, by allowing for two *iifs* during network topology change period, at the cost of slightly increase of the overhead due to the temporary duplicate packet transmissions, during the transition period. To avoid extra overhead, a *Prune* can be sent out the old *iif* as soon as new data packets have arrived from the new *iif*, instead of waiting for the time out of the upstream *oifs*. In the following sections, representative WAN link and router failure scenarios are detailed.

[5]IGMP version 1 and 2 do not play a role in WAN multicast recovery. Comparatively, the IGMP version 3 proposal is carried beyond the leaf router into the WAN and will likely play a role in channel recovery.

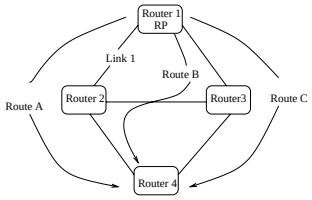



Fig. 1. WAN failure scenario

#### A.1 Link Failure in the WAN

Consider the link failure scenario shown in Fig. 1. Originally, a multicast channel exists over Route A. If Link 1 fails, Router 1 and Router 2 both immediately detect the failure since the link is directly attached to each router (not attached over a NBMA network). Each router will update the link-state database by re-originating its *router-LSA* to announce the topology change, sending it to Router 3 and Router 4. The new best metric route from Router 2 to the RP or sender is now via Router 3. PIM on Router 2 then sends a *Join/Graft* to its new RPF Router 3 to recover the failure. The multicast channel is rebuilt to Route B in Fig. 1.

While the above processing is occurring in Router 2 and 3, Router 4 will have received LSAs from Router 2 and 3 separately. Detecting its new RPF via Router 3, PIM on Router 4 triggers a *Join/Graft* to Router 3. As suggested earlier, to avoid potential packet loss due to a race condition, Router 4 may not send a *Prune* right away to Router 2. The multicast channel will migrate to Route C eventually after interfaces associated with the suboptimal path Route B time out or are pruned.

In general, the multicast channel recovery time in WAN is dependent on the "carrier-delay" time required to learn about a link outage from a lower layer, or on the OSPF *RouterDeadInterval* if link failure can not be detected earlier at lower layers. Every OSPF Hello message resets the OSPF *Inactivity Timer*, with a link failure occurring (on average) at the mid-point of the hello interval. Hence the average OSPF failure detection time is:

$$T_{fd}^{ospf} = min\{T_{rdi}^{ospf} - 0.5 * T_{hi}^{ospf}, T_{cd}\} \quad (2)$$

The worst-case time for OSPF to detect a failure is:

$$T_{fd}^{ospf,w} = min\{T_{rdi}^{ospf}, T_{cd}\} \quad (3)$$

After detecting the topology change, OSPF starts a shortest path calculation after *SPF Delay* time. We can then represent the average OSPF topology database updating time as:

$$T_u^{ospf} = T_{fd}^{ospf} + T_{spf} \quad (4)$$

The worst case OSPF topology database updating time is:

$$T_u^{ospf,w} = T_{fd}^{ospf,w} + T_{spf} \quad (5)$$

If PIM is notified of the unicast table change by OSPF, multicast channel recovery can be initiated immediately after OSPF updates the topology. If instead, PIM polls the unicast table to learn of changes, an additional delay of $0.5 * T_{pli}^{pim}$ is incurred on average. In general, we represent the average time for PIM to detect the topology change as $T_u^{pim}$, and corresponding worst-case time as $T_u^{pim,w}$. The multicast channel recovery time can now be written as:

$$T_r = T_u^{ospf} + T_u^{pim} \quad (6)$$

The worst-case multicast channel recovery time can be represented as:

$$T_r^w = T_u^{ospf,w} + T_u^{pim,w} \quad (7)$$

A.2 Router Failure in the WAN

Router failure in the WAN is similar to multiple simultaneous link failures. Assume a multicast channel is instantiated via Route A, as shown in Fig. 1. If Router 2 fails, Router 4 immediately detects its interface to Router 2 is torn down. Router 4 updates its OSPF database, executes the Dijkstra's algorithm to update its network topology, and floods an OSPF LSA. When PIM on Router 4 finds that its best reverse path metric to the RP or sender is now via Router 3, it sends a *Join* to Router 3 to recover the multicast channel via Route C. Router 1 takes no proactive action during the recovery. The channel recovery is triggered by those routers further downstream from the failed router.

*B. Protocol Interaction on a LAN*

Multicast channel recovery in LAN is more complicated than that in WAN. In addition to the interaction of OSPF and PIM protocols as presented in Section III-A, IGMP plays a role in LAN multicast channel recovery.

The OSPF failure detection time on LAN may depend more critically on the *RouterDeadInterval*. When routers are on an Ethernet, for example, the fact that router X's cable has fallen out will not lead the other routers on the Ethernet to destroy their adjacencies with Router X until OSPF *RouterDeadInterval* has expired. However, as long as they can receive Router X's new router-LSA (that is, as long as the Ethernet is not a single point of failure), the other routers on the Ethernet will update their routing tables well before the adjacency is destroyed.

On the LAN, PIM routers can act in two important roles: Designated Router (DR) and last-hop router. A DR in PIM SM is responsible for initially drawing-down the multicast channel to the LAN (Section II-C). The last-hop router is the last router to receive multicast data packets before they are delivered to directly-connected member hosts. The DR is normally the last hop router. However, a router on the LAN that is not the DR but has a lower metric route to the data source or to the group's RP may take over the role of the last-hop router.

When the DR receives an IGMP *Membership Report*, it adds the receiving interface to its *oif* list. It may also send *Join* messages to its RPF router (if the existing entry had no active *oif*s). If the DR is not the last-hop router, this may trigger a new *Assert* process.

In our case, PIM DM does not need a DR, although it was required on a LAN running IGMP v1. Its multicast channel formation and failure recovery are therefore a little different from PIM SM.

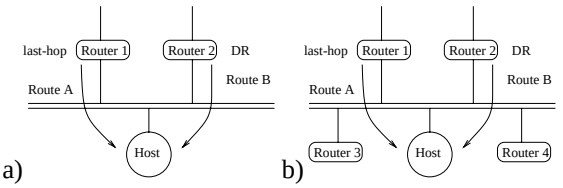



Fig. 2. LAN failure scenario, DR and last-hop router are different routers

B.1 LAN Failure Recovery - PIM-SM

1. **Scenario 1:** last-hop router and DR are separate routers (Fig. 2). Since the DR is not the last hop router, it does not have an *oif* towards the LAN. In this case, when the link immediately upstream or downstream of the DR fails, the multicast channel for LAN stays alive in either case since the failure point is not on the multicast path. For completeness, we present the transient behavior in either case.

(a) The DR's upstream link fails. The DR will detect the outage right away if the failed link is a serial link, or at most wait for *RouterDeadInterval* if the lower layer cannot convey the outage information to OSPF. When DR has active *oifs* in addition to the one towards the LAN, it may send *Join* to the new RPF immediately after the failure is detected. However a multicast branch that goes through the DR towards the LAN can be recovered only when the IGMP *Membership Report* reactivates the pruned *oif* after the unicast channel recovery. The average time for the DR to recover its multicast branch is:

$$T_r = T_u^{ospf} + T_u^{pim} + 0.5 * (T_{qi}^{igmp} + T_{qri}^{igmp}) \quad (8)$$

The worst multicast channel recovery time is:

$$T_r^w = T_u^{ospf,w} + T_u^{pim,w} + T_{qi}^{igmp} + T_{qri}^{igmp} \quad (9)$$

(b) The link between the DR and the LAN fails. On average, the time to detect a neighboring router failure (DR failure) is about $T_{nfd}^{pim} = T_{hhi}^{pim} - 0.5 * T_{hi}^{pim}$. After the failure detection, the router on the LAN with the next highest Ethernet interface IP address becomes the DR. Subsequently, the DR must acquire the IGMP group membership information, and this contributes a term (as in the previous case) of $0.5 * (T_{qi}^{igmp} + T_{qri}^{igmp})$. The average recovery time is therefore given by:

$$T_r = T_{nfd}^{pim} + 0.5 * (T_{qi}^{igmp} + T_{qri}^{igmp}) \quad (10)$$

The worst case recovery time is:

$$T_r^w = T_{hhi}^{pim} + T_{qi}^{igmp} + T_{qri}^{igmp} \quad (11)$$

(c) The upstream link of last-hop router fails.
If there is an alternative link, the last-hop router will *Join* to the new RPF upon detecting the change in the unicast table. In this case, the average and worst case recovery time will be the same as in equation 6 and 7 If, as a result, the affected router no longer remains the last hop router, the *Assert* process will lead to a new last-hop router being elected and a new optimal multicast channel established.
If there is no alternative link from last-hop router towards the RP or sender, the multicast channel is recovered through the DR by sending a *Join* message when a new IGMP *Host Membership Report* is received from a host on the LAN. The recovery time in this case is as given by:

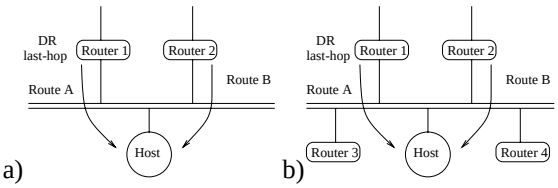



Fig. 3. LAN failure scenario, DR and last-hop router are the same router

$$T_r = 0.5 * (T_{qi}^{igmp} + T_{qri}^{igmp}) \quad (12)$$

The worst case recovery time is:

$$T_r^w = T_{qi}^{igmp} + T_{qri}^{igmp} \quad (13)$$

(d) The link between the last-hop router and the LAN fails.
The DR may be informed of the topology change through a router-LSA quickly. However, if no routers exist downstream of the current last-hop router, the DR will not reactivate the multicast channel until it receives the new IGMP *Membership Report*. The average recovery time will be the same as equation 12 and 13.
If the DR regards the affected last-hop router as RPF router, it needs to detect the failure and graft to the new RPF. The average and worst case recovery time for the multicast channel are therefore as in equation 8 and 9.
If there are routers downstream of the affected last-hop router (Fig. 2 b), they will detect the topology change through OSPF router-LSAs. The routers previously considering the affected last-hop router as the RPF router will send *Join* to the new RPF once a new RPF is detected. The multicast channel recovery time in this case depends critically on the topology change detection time and on average is as equation 6.
The downstream routers with a different RPF neighbor (according to the original unicast table) from the last-hop router may need to wait for the *Assert-Timer* to expire before they can send *Join* to the new RPF router. So the multicast channel recovery time will depend on both the *Assert-timer* value and the IGMP *Query Interval* in this case, whichever comes first. The average recovery time is:

$$T_r = 0.5 * min\{T_{qi}^{igmp} + T_{qri}^{igmp}, T_a^{pim}\} \quad (14)$$

The worst case recovery time is:

$$T_r^w = min\{T_{qi}^{igmp} + T_{qri}^{igmp}, T_a^{pim}\} \quad (15)$$

2. **Scenario 2:** last-hop router and DR are the same router. The LAN consists of two routers, with one router acting as both the last-hop router and the DR (Fig 3).

(a) The link upstream of the DR fails. Regardless of routers downstream of the DR, the DR will recover the multicast channel immediately after it determines the new RPF router, since it has active multicast entries. The average recovery time is as in equation 6.

(b) The link between the DR and the LAN fails.
If there are no routers downstream of the DR, the multicast channel will not recover until a new DR is elected and a host membership report is received by the new DR. The multicast channel recovery time is the same as equation 10 and 11. If downstream routers exist, the multicast channel can be recovered and switched to the new RPF router of the downstream routers in the same time as equation 6, on average.

### B.2 LAN Failure Recovery - PIM DM

Since PIM DM does not have a DR, some failure scenarios for PIM SM do not apply. For the multicast channel to recover, the LAN must have more than one router towards the source (Fig. 2), and the *Assert* process is used to select the forwarder (or last-hop router) for the LAN (Router 1). We refer to the router that loses the *Assert* as Router-Other (Router 2).

1. Router-Other's upstream link fails. If Router-Other has an active entry (on-tree oifs other than the one towards LAN), it sends a *Graft* to its new RPF upon failure detection. Otherwise, Router-Other will pull down the multicast channel towards LAN again if it receives a new IGMP report. The average recovery time is

$$T_r = max\{T_u^{ospf} + T_u^{pim}, 0.5 * (T_{qi}^{igmp} + T_{qri}^{igmp})\}. \quad (16)$$

   Note that the recovery time is different from equation 8, since in PIM DM, the RPF neighbor will acknowledge the *Graft* by sending *GraftAck*. If failure is detected after arrival of a new IGMP report, the *Graft* message will be lost and the sender will periodically (default 3 s) retransmit the *Graft* message, until a new RPF is found. On the other hand, if a IGMP report arrives first, the resulting active entry allows the multicast channel to be recovered immediately after the new RPF is detected.
   In addition, PIM DM can recover from data packet flooding when the Router-Other's pruned interface towards LAN times out before a new IGMP report is received. When no multicast entry exists in Router-Other, a new entry will be created when a data packet arrives and the channel through Router-Other can recover quickly.
2. The upstream link of last-hop router fails. The multicast channel will either recover quickly as in equation 6 and 7 if there is an alternative link towards the source, or recover through Router-Other in a time given by equation 12 and 13.
3. The downstream link of last-hop router fails. The recovery scenario is similar to the corresponding case in PIM SM.

In addition to the failure scenarios presented above, a failure of the IGMP *Querier* will increase the next IGMP group membership report interval. As long as this does not happen in coincidence with the failure of other components that are more critical to a multicast channel, it is not a concern.

Router failures in the LAN is similar to the downstream link failure cases. From the presentation above, we can see that depending on the failure scenario, the multicast channel recovery for a LAN may critically depend on several parameters, the most important of which are OSPF *RouterDeadInterval*, PIM *Hello-Holdtime*, *Assert-Time* as well as IGMP *Query Interval*.

### *C. Totally Stubby Area Considerations*

In additional to protocol behavior, the network configuration can also influence the failure recovery. For example, if OSPF totally stubby areas are configured in the network, the final migrated multicast channel may not necessarily have the best metrics to the source or RP. Furthermore, the multicast channel might not be recovered at all in some totally stubby area configurations.

Consider the hypothetical network in Fig. 4. Originally, the multicast channel traverses Route A: Router 1 → Router 2 → Router 4 → Router 6. If WAN Link 1 fails, for example, Router 2 sends

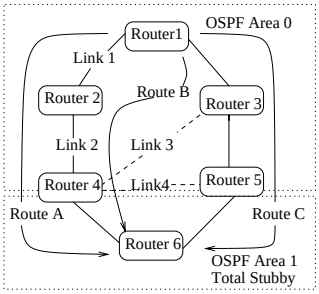



Fig. 4. OSPF Stubby area failure recovery

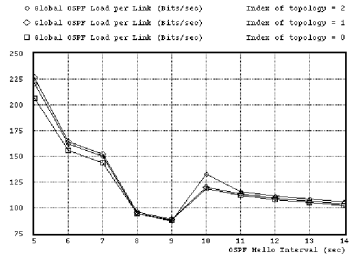



Fig. 5. OSPF load change with the variation of Hello interval

a *Join/Graft* to Router 3 to rebuild the multicast channel via Route B. The multicast channel will not migrate to Route C even though Route C may have better metrics than Route B. Since OSPF Area 1 is configured as totally stubby, OSPF LSAs are not flooded into Area 1 by either OSPF Area Border Routers (ABR) Router 4 or Router 5.

Now consider the case Link 3 and Link 4 do not exist. If Link 2 fails, Router 4 learns of the failure but it cannot recover the multicast channel since it only has Router 6 as its neighbor in Area 1. Router 6 has a potential route to the RP or sender via Route C but has no reachability knowledge concerning other OSPF areas via Router 5. Thus, Router 6 does not migrate the channel to its other upstream link. The network failure, in this scenario, causes the multicast channel to Router 6 to be unrecoverable using PIM SM. In PIM DM, the next rebroadcast period will cause the channel to be re-established via Route C. If the network is redesigned to add Link 3 or Link 4, Router 4 could then build the multicast channel via Router 3 or Router 5. When using OSPF totally stubby areas, the OSPF area border routers should always have an alternative upstream link within the OSPF Area to the RP or sender, to provide for multicast channel recovery.

If Router 4 were to fail, instead of a backbone link, as described above, then Router 6 would send a Join/Graft on its other upstream link to Router 5 (new RPF) to recover the channel. The recovery occurs because Router 4 is co-located with Router 6 in the same OSPF area.

## IV. Simulation and Results Analysis

Simulation models for an IP multicast system have been developed for the investigation of end-to-end multicast failure recovery behavior and performance by using OPNET [14]. The models include IGMP, PIM DM, modifications to the IP forwarding engine, a random topology generator ported from the TIERS topology generator [10], a multicast sender and receiver application model, a link fault injector, as well as several probes to acquire simulation statistics. More detailed descriptions of the design and implementation of these models can be found in [14]. In addition to the study of end-to-end multicast failure recovery time, we also calculate the traffic control loads generated by the different protocols under normal network conditions and in network failure recovery scenarios.

### *A. Simulation Parameters and Design Decisions*

We simulated each combination of network topology, and protocol parameters. The parameters that were varied in our simulation are as follows:

1. Network topology. In order to be able to generalize the results, multiple random topologies were created and used in our experiments. In the majority of the simulations (unless otherwise specified), three random topologies, each consisting of 36 nodes, were used. In each topology, the default redundancy factor was 4, and the percentage of receivers was set to 80% for the (single) group. For any particular topology, depending on the experiment, we varied the number of routers in the network, the redundancy factor (2, 3, 4), and the percentage of receivers relative to the total number of nodes in a network.
2. OSPF parameters. In order to study the failure recovery time, the OSPF *Hello* and *Dead* timers are tuned. The *RouterDeadInterval* is set to three times the *HelloInterval* in all the simulations. In addition, the SPF calculation time was reduced from its default value of 10 seconds to 1 second.
3. PIM parameters. In the PIM implementations of some of the router vendors, such as Cisco, the unicast routing table is polled periodically to allow PIM to detect the network topology changes. To minimize the influence of the polling interval on the simulation failure recovery and focus on the protocol interactions themselves, the polling interval was set to a small value (0.2 s).
4. Application layer parameters. To study the end-to-end multicast channel failure recovery behavior, the end to end recovery time is measured. The arrival traffic was generated using a CBR model. Using this model, receivers detect when they have become disconnected from the multicast channel if they fail to receive the next expected packet. The data rate is set to a low value (two per second) to reduce the simulation time due to the handling of large number of events, while keeping the multicast channel alive. As a result, there is no packet loss due to buffer overflow in the simulation environment.

### *B. The Control Load of OSPF and PIM*

From the analysis results in Section III, we have seen that the failure recovery time is closely related to the OSPF *Hello* interval. We first study the change in OSPF control load due to the variation of OSPF Hello interval. Subsequently, we discuss the effects of the network redundancy factor and the receiver population on the PIM DM control load.

#### B.1 OSPF Control Load versus OSPF Hello Interval

The OSPF *HelloInterval* was varied from 5 s to 15 s in 1 s steps, and correspondingly the *RouterDeadInterval* was varied from 15 s to 45 s, in 3 s steps. The effect on the OSPF control load was studied for 3 random 36-node network topologies with redundancy factor 4. Intuitively, the OSPF control load will decrease as the OSPF hello interval increases. The results in Fig. 5 show that this is true for a hello interval of less than 9 s, with the load varying almost inversely with the *HelloInterval*. When the hello interval is greater than 9 s, the overhead due to variations in the hello interval appear to be negligible. This is because the average OSPF control load

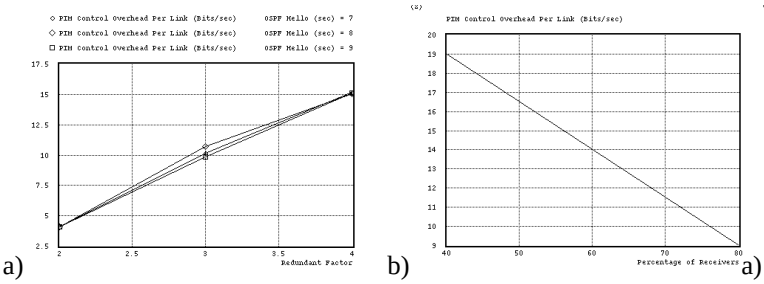



Fig. 6. PIM DM load change with the variation of network redundancy factor a) and receiver percentage b)

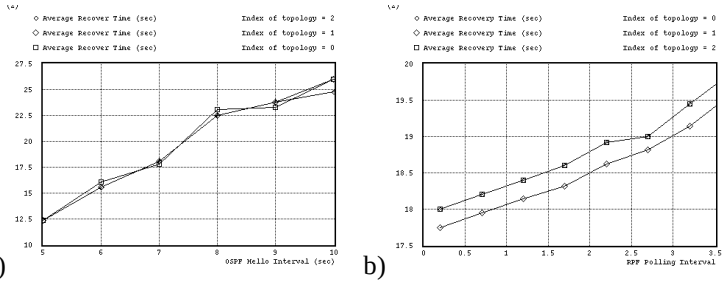



Fig. 7. a) Variation of multicast channel recovery time with the OSPF Hello interval (PIM polling interval set to 0.2 s) b) Variation of multicast channel recovery time with the PIM polling interval

per link is no longer dominated by the overhead of hello messages when the hello interval is large. Over the entire range, the load is not strongly affected by the network topology. Overall, the average OSPF control overhead in a link is very small (less than 250 bps in all cases).

### B.2 PIM DM Control Load versus Network Topology

In the first simulation, the PIM control load was measured on three different 36-node networks, with redundancy factors 2, 3 ,a) and 4 respectively. Fig. 6 a) shows that the total PIM control overhead is directly proportional to the network redundancy factor. This can be understood as follows. PIM DM control load is dominated by the periodical *Hello* and *Prune* messages. The *Hello* load will not be influenced by the network topology. The *Prune* will increase as the network redundancy factor increases, since the data packets are flooded across more links and trigger more PIM *Prunes*.

In the second simulation, the PIM control load was measured on a single 36-node network, while varying the number of receivers on this network. Fig. 6 b) shows that when the percentage of the receivers (relative to the number of nodes) increases, the PIM DM control load actually decreases. This is because as the receiver population increases, the number of links branches on the multicast tree increases, and fewer *Prunes* will be sent out. This indicates that PIM DM efficiency increases in a network densely populated with receivers, which is the primary design goal of PIM DM.

### *C. Single Multicast Channel Recovery Time*

As discussed in Section III-A, the recovery time from a link or router failure in a WAN is strongly dependent on the speed with which lower layers of the protocol stack in neighboring routers learn of the outage and how quickly they inform the OSPF protocol. Accordingly, the vendor implementation dependent "carrier delay" parameters have a strong influence. In case the OSPF routers are not able to learn of the outage through the lower layers, the expiry of the OSPF *Inactivity Timer* is used as a last resort.

However, in our simulations, since the OSPF implementation in OPNET does not send the link layer failure information to the network layer, failure can be detected only when the OSPF *Inactivity Timer* expires. Hence, we only study the influence of the *RouterDeadInterval* interval (or equivalently, the proportional *Hello Timer* interval) on the failure recovery time. In fact, as will be seen later in Section V, the experiments using the Cisco testbed show that a data link layer outage can provide much quicker failure recovery time in the WAN, depending on the setting of the "carrier detection" parameter value.

As before, failures were simulated on a randomly generated 36 node network of redundancy factor 4, identified in the graphs.

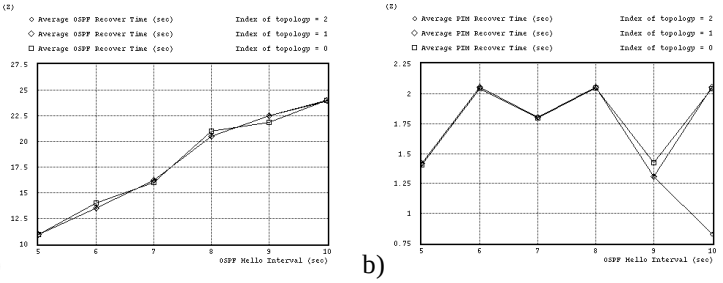



Fig. 8. a) Variation of OSPF topology updation time with the OSPF Hello interval. b) The time between topology updation by OSPF and multicast channel recovery using PIM, as a function of the OSPF Hello interval

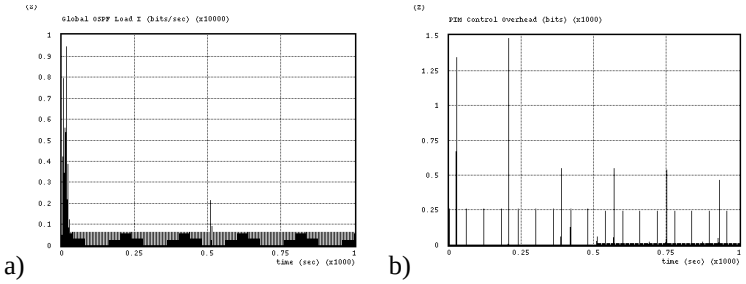
Faults were injected singly at randomly selected links, and also at randomly (uniformly) distributed times. As mentioned in the Section IV-A, packet loss in the simulated network only happens if there is a network failure. Accordingly, a failure is detected for a group if a receiver in the group detects a missed packet. The recovery time for an individual receiver is defined as the time interval between the packets received immediately before and immediately after the missing packet(s).

Each data point in Fig. 7 and 8 is the failure recovery time averaged over all receivers for a particular fault, and also averaged over approximately 100 single faults at random links.

Fig. 7 a) shows that the failure recovery time increases with the OSPF Hello interval and does not depend on the network topology. The comparison between Fig. 7 a) and Fig. 8 a) shows that the failure recovery time is dominated by the OSPF recovery time. This is approximately 2.5 times the OSPF Hello interval as predicted by the analysis (equation 2).

Since triggered *Graft/Join* is used to recover a multicast channel, PIM does not have a major contribution to the failure recovery time. After a unicast routing table is updated by OSPF, PIM takes at most a polling interval (which is set to 0.2 second for the experiments of Fig. 7 a) and 8 a)) to find out the topology change, and triggers the *Join/Graft* to a new RPF router, thus migrating to the new multicast channel. However, the recovery time after topology updation, shown in Fig. 8 b), is larger than the expected PIM recovery time. This is because it takes about extra SPF Delay (= 1 s) for OSPF to start a new topology calculation after the topology updation. The end-to-end packet loss detection method (with data packets interval 2 s) also contributes to the apparent PIM recovery time, and also makes it somewhat random.

As expected, the component of the recovery time after topology updation does not change with OSPF Hello interval. The failure recovery time does not change very much with the network topology either, since the propagation delay of the control messages is




a) b)

Fig. 9. OSPF load change (a) and PIM DM load change (b) during failure recovery, beginning at t=500 seconds

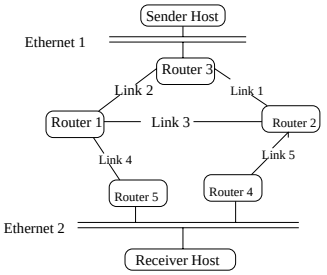



Fig. 10. Testbed topology

negligible (of the order of micro-seconds) as compared to the delay due to the protocol (tens of seconds). Therefore, to reduce the multicast channel recovery time, the OSPF hello interval and SPF calculation interval should be set as small as possible.

The average overhead due to polling of unicast table is approximately half of the polling interval. Fig. 7 b) shows the variation of the PIM recovery time with the polling interval, with the OSPF Hello interval set at 7 seconds. As expected, there is a linearly increase in this component of the recovery time with the polling interval. To allow a fast recovery, the polling interval should be set as small as possible. The recovery time is nearly the same for all the network topologies shown.

### D. Network Load Change during Failure Recovery

During failure recovery, the number of control messages increases - this includes PIM join, prune messages and OSPF link state update messages. In this section, we compare the average link control load during failure recovery with the control load during steady state. Network topologies were generated as in the previous section. The OSPF Hello Interval was set to a default value of 10 seconds, and the PIM unicast routing table polling interval is set to 0.2 second. At simulation time 500 seconds, a fault was injected.

Figure 9 a) shows that at the beginning of the simulation, the OSPF control load is higher than the load in steady state. This is due to the flooding of LSAs by all nodes in the network. As OSPF reaches steady state, the control load becomes smaller but increases periodically every 10 seconds and 30 minutes. The small load shown in the lighter area every 10 s is due the periodical OSPF Hello load. LSAs are flooded periodically every 30 minutes.

At time 500 s, when the fault is injected, the load increases due to the flooding of updated LSA as a result of a topology change. However, the increase in the control load is minimal, compared to the increase due to the half-hourly flooding of LSA's.

Similar to the OSPF case, Figure 9 b) over the same time period shows that the PIM DM control load is higher during the establishment of the PIM neighbor relationships between PIM enabled routers. The load shown consists mainly of PIM *Hello*, *Graft* and *Prune* messages. In steady state, PIM hello messages are sent periodically every 30 seconds and prune messages every 180 seconds. During the failure event at 500 seconds, the PIM control load, unlike that of OSPF, does not increase, but remains flat. This is because the PIM channel recovery is highly localized and the extent of localization depends on the network topology and redundancy factor. If short, alternative paths exist, the multicast channel can be recovered with minimal additional PIM control loading.

## V. Experimental Results

Experiments are executed to verify the behavior of the component protocols under the LAN and WAN failure scenarios described in Section III. Measurements of the multicast channel recovery times are provided, given a set of tightly tuned parameters.

### A. Testbed Set-up

A testbed was constructed as shown in Fig. 10. Routers 1, 2, and 3 were CISCO 4700 routers and routers 4 and 5 were CISCO 2503 routers. The routers implemented OSPF as the unicast routing protocol, and PIM DM and SM and IGMP protocols. The hosts ran Microsoft Windows NT 4.0. The link speeds were T1 on Link 1, 2, and 3, and 64 Kb/s on Links 4 and 5.

### B. Test Procedure

Since the study was focused on the interaction amongst the component protocols during fail recovery, only a single sender and single receiver were required for the testing. In all the test cases, a multicast tree was first established from sender to receiver. To simulate a link failure, a selected link on the multicast tree was manually open-circuited. To simulate a router failure, the selected router was manually powered off. The sender application generated multicast data at the rate of 1 KB/s. The receiver logged the received multicast packets into a file, allowing the detection of missing packets and packets received out of sequence.

The overall failure recovery process was monitored using several mechanisms. Router debug messages were monitored and logged via telnet sessions into the respective routers. The router debug messages contained a time stamp, which was synchronized among all the routers in the network using the Network Time Protocol [12]. The debug messages provided causal ordering of routing protocol operations. Four W&G LAN and WAN network analyzers [13] were used to analyze data traffic and IGMP, OSPF, and PIM control messages on the multicast channels. The analyzers operate with a synchronous clock, and thus packet delays could be measured accurately within $10^{-5}$ seconds.

### C. Parametric Tuning

In the Cisco router implementation, several protocol parameters can be tuned for the purpose of failure recovery. The IGMP parameters that may be tuned include the *Query Interval*, the *Query Response Interval* and the *Other Querier Present Interval*. By reducing these parameter intervals, new group information may be discovered more rapidly by the router, and querier failure can be detected faster. They are set to default value 125 s, 10 s, and 255 s respectively. In most implementations, including the one by Cisco, the other non-querier routers on a LAN shadow the IGMP database maintained at the router acting as the querier. When the

querier fails, a new querier router is elected and the transition occurs rapidly since IGMP information is already on-hand.

The OSPF *HelloInterval* and *RouterDeadInterval* were set to one and three seconds, respectively, and the "carrier delay" time was set to two seconds. The SPF delay time and holding time were set to zero and ten seconds, respectively. This meant that the network failure could be detected within three seconds in the worst case, and the SPF calculation would be immediately processed after the detection.

The PIM Hello message interval was tuned to two seconds, so that routers on a LAN would detect the DR failure on average in approximately five seconds (2.5 times PIM hello interval), as in equation 6. PIM SM sends periodic *Join/Prune* messages to its upstream routers to keep the multicast channel alive (soft state). The default period is 60 seconds. As explained in section III, the PIM polling interval determines how often PIM polls the unicast routing table, and therefore dictates how quickly it initiates recovery of the multicast channel upon detecting a change. In the implementation of PIM used on the testbed, the PIM polling interval was not tunable and was fixed at five seconds.

### *D. Results*

Tables II and III summarize the experimental results. The total recovery time consists of three components. The OSPF recovery time was measured as the time from when the network element failure occurred to when the affected router received the corresponding LSA. The PIM recovery time was measured as the time from when the affected router received the corresponding LSA till the time a PIM *Join/Graft* message was sent. The *Join* Latency was the time taken by the affected router to process a received *Join/Graft* message and forward it to an upstream router, plus the transmission time of the *Join/Graft* message itself.

The first set of tests was conducted with OSPF configured as a totally stubby area at the client site. The OSPF areas are configured in Figure 10 such that: Link 1, 2, and 3 are in Area 0, and Link 4, 5 and Ethernet 2 are in totally stubby Area 1. The individual protocol component recover times under the various multicast channel failure scenarios are shown in Table II. The initial route of the multicast channel, prior to the failure, the corresponding failure event, and subsequent component recovery times are listed from the perspective of the identified router.

Table III shows the measured results where Links 4, 5 and Ethernet 2 are in non-stubby Area 1. In the first failure event (Link 1 failure) under this configuration, the multicast channel recovery occurs in two steps. In Step 1, Router 2 recovers from the Link 1 failure by constructing the multicast tree through Router 1 to Router 3. When Router 4 determines that the better metric to the RP or source is through Router 5 (Router 3 $\rightarrow$ Router 1 $\rightarrow$ Router 5), the Step 2 migration takes place at Router 4.

In both the OSPF totally stubby area and non-stubby area cases, the average and worst case fail-over time, as given by equations 6 and 7 for failure of link 1, link 5, Router 2 respectively, is measured to be approximately 5 and 8 seconds, respectively, plus a few hundred additional milliseconds. It is noted that when Router 4 (acting as both the DR and last-hop router) fails, the multicast channel can be recovered in about five seconds after the DR failure is detected. This is much shorter than the 65 seconds predicted by equation 10, which is based on the protocol that the DR needs to wait either for the IGMP report to reactivate its *oif* towards LAN or for the periodic flooding of data packets in PIM DM (whichever happens first) before it can reactivate its *oif* towards the LAN. The rapid recovery of the multicast channel occurs in the Cisco implementation because all routers on the LAN cache the multicast group membership information, and the multicast channel is recovered as soon as the new DR is elected. However, when the last hop router and DR are not co-located and the last hop router (Router 5) fails, the DR does need to wait either for the IGMP report (SM) or for the periodic flooding of data (DM), as will be observed from the following experiments.

In the case of the Router 5 failure, the results are related to the PIM protocol specifications and the specifics of the vendor's PIM protocol implementation. For PIM SM, recovery requires approximately 60 seconds. The DR did not prune its interface towards the LAN in Cisco's implementation and the multicast channel recovered when the periodical *Join* was sent upstream by the DR (every 60 seconds). Rather than waiting for the next periodic *Join* interval, the router implementation could be changed to immediately send a *Join* upstream, once the DR detects failure of the last-hop router. PIM DM requires 1.5 and 3 minutes in the average and worst cases. PIM DM recovers when the data is rebroadcast at every 3 minutes interval as expected in Section III-B.2. Similar to the PIM SM case, some improvement in the protocol specifications can lead to much faster failure recovery process as explored in Section VI.

## VI. Discussion

In this section, we present some general insights and design guidelines on the basis of our analysis, simulations, and experiments, and understanding of the protocol behavior in the various failure scenarios.

### *A. General observations*

1. In general, multicast channel recovery time is dominated by the time required to re-construct the unicast routing table. Although the test-bed results show a substantial recovery time attributed to PIM, in most cases this was due to large polling interval with which PIM looked up the unicast routing table. Trigger based active joining of multicast trees (as used in PIM) allows the multicast channel to be recovered quickly thereafter.
2. The simulation results for control overhead and recovery time yielded similar results for all randomly generated topologies with the same number of nodes and the same redundancy. This indicates that our results are generally representative for networks of a given size and complexity.
3. Protocol control loads: The PIM DM control load increases proportionally with the redundancy factor and decreases inversely with the percentage of receivers. The OSPF load increases proportionally as OSPF *Hello* interval decreases and is acceptable in the simulated parameters range (10 s - 5 s). In general, the default assignment of protocol timers appears to be conservative, and the tightening of these parameters for speeding up the failure recovery does not lead to excessive overhead. If possible, the unicast routing parameters should be tuned to allow rapid detection of topology changes and prompt updating of the routing table.
   Neither PIM nor OSPF has high control traffic during failure recovery, and the combined overhead for each link is always less than 1 kbps in all simulation cases.

### *B. Effect of Network Configuration on Fault Recovery*

Network configuration can potentially influence the failure recovery.

| Failure Event | OSPF Recovery | PIM Recovery | Join Latency | Total Recovery | Router Perspective | Initial Route before failure |
|---|---|---|---|---|---|---|
| link 1 | 2.11853 | 2.87677 | 0.05926 | 5.05456 | R2 | R3→R2→R4 |
| link 5 | 2.02733 | 3.38755 | 0.05251 | 5.46739 | R4 | R3→R2→R4 |
| Router 2 | 2.06035 | 4.60794 | 0.06246 | 6.73075 | R4 | R3→R2→R4 |
| Router 4 (FWD&DR) | 3.012 | 4.176 | 0.006 | 7.194 | R5 | R3→R2→R4 |
| Router 5 (FWD) SM | 2.470 | 64.027 | 0.128 | 66.625 | R4 | R3→R1→R5 |
| Router 5 (FWD) DM | 2.470 | 95.025 | 0.128 | 97.623 | R4 | R3→R1→R5 |

TABLE II
FAIL-OVER TIME (IN SECONDS) WITH OSPF TOTALLY STUBBY AREA

| Failure Event | OSPF Recovery | PIM Recovery | Join Latency | Total Recovery | Router Perspective | Initial Route before failure |
|---|---|---|---|---|---|---|
| link 1 (step1) | 2.1431 | 4.32362 | 0.01918 | 6.4859 | R2 | R3→R2→R4 |
| (step2) | 0 | 3.28387 | 0.01574 | 3.29961 | R4 | R3→R2→R4 |
| link 5 | 2.65603 | 3.40131 | 0.08288 | 6.14022 | R4 | R3→R2→R4 |
| Router 2 | 2.12218 | 4.16531 | 0.04512 | 6.33261 | R4 | R3→R2→R4 |
| Router 4 (FWD&DR) | 2.563 | 4.001 | 0.007 | 6.971 | R5 | R3→R2→R4 |
| Router 5 (FWD) SM | 2.638 | 60.024 | 0.023 | 62.685 | R4 | R3→R1→R5 |
| Router 5 (FWD) DM | 2.638 | 92.012 | 0.023 | 94.673 | R4 | R3→R1→R5 |

TABLE III
FAIL-OVER TIME (IN SECONDS) WITH OSPF NON-STUBBY AREA

1. If there are OSPF totally stubby areas, the OSPF area border routers should always have an alternative upstream link to the OSPF area backbone. Channel recovery is driven from the affected receiver(s) upstream towards toward the RP or source. If there is only a single link from the area border router to the backbone, and that link fails, the failure information is not propagated to the stubby area. Thus, the routers in the stubby area are not able to take action to find an alternative or better route to the RP or source. In this case, the channel may never recover.
2. When establishing static routes from client site router(s) towards the backbone, the router closest to the backbone terminating the static link should always have an alternative upstream link to the RP or sender. The motivation is identical to that for the totally stubby area.

### C. PIM Enhancement for Fault Recovery

1. Fast recovery from DR failure. On a LAN, DR reliability of the PIM SM is critical, and it is necessary to detect the inaccessibility or failure of the DR quickly for prompt recovery of the multicast channel. One possibility is for the DR to reduce its *Hello Interval* to inform other routers of its presence more frequently, and for other routers to correspondingly reduce the *Hello-Holdtime* for the DR, so that it is timed out sooner in case of failure. Also, as discussed earlier, a backup DR could be introduced to allow PIM to more quickly recover from a DR failure without the necessity of waiting for the new DR to reload the group membership database. Alternatively, all LAN routers could maintain a cache of IGMP group information, regardless of their current role.
2. Fast recovery from last-hop router failure. Based on PIM SM specification, a DR will only send a *Join* message upon receiving a new IGMP group information message after it loses the *Assert* to the last-hop router. As a result, the affected multicast channel due to the failure of the last-hop router may take long time to recover as observed in the testbed. To allow PIM SM to recover quickly after the last hop router becomes inaccessible via the LAN, the DR could record the last-hop router address, obtained from the assert process. If the last-hop router becomes inaccessible through the LAN, the DR would not need to wait for an IGMP report to reactivate its *oif* to the LAN. Similarly, a backup router can be used in PIM DM to take the responsibility of the DR for rapid detection of the last-hop router failure. With these improvement, the large recovery delay for PIM SM and DM detected in the testbed could potentially be avoided.
3. Reducing extra delay due to polling. In the Cisco implementation, PIM periodically polls the unicast routing table to discover changes in the unicast topology, which can subsequently trigger changes in the multicast channels. A potentially more efficient way in which protocol independence could be achieved, is via interrupts. When a unicast route changes, the unicast routing entity could inform the multicast routing component of the change in state.

Some of these improvements can be made in either the implementation or architecture to reduce the fail-over time of multicast channels. With the various suggested improvements and parameter tunings, the multicast channel can be made to recover within a few seconds. The improvements mainly allow the unicast and multicast modules to more rapidly update their states, rather than waiting several minutes, as is done in the current default protocol behaviors or specific implementations. Finally, it may be possible to apply policy to multicast routing protocols to improve upon the multicast channel availability.

## VII. CONCLUSIONS

The fault recovery behavior of end-to-end IP Multicast channels is a function of several protocols, including IGMP, unicast and multicast routing protocols. In this paper, the recovery behavior and interactions of three protocols, IGMP, OSPF, and PIM are studied. Analytical models are developed that provide the expected IP multicast channel recovery time. Simulation models are

developed to measure the control overhead of PIM and the failure recovery time of IP Multicast channels, using various random topologies and with different protocol tuning parameter settings. Furthermore, an experimental testbed is used to measure the failure recovery of IP multicast channels in the event of link and router failures. Simulations for WANs show multicast channel recovery to be relatively robust and light weight, in terms of protocol control overhead and recovery latency. It is shown that most of the failure recovery time is attributed to the unicast routing protocol recovery process, in this case OSPF. Failure recovery in a LAN is found to be more complex. It is strongly influenced by protocol interactions and implementation decisions. Experiments show that it is also light-weight in terms of recovery latency and overhead, except for a couple of cases which are discussed. Finally, suggestions for improvement of the failure recovery time via protocol enhancements, parameter tuning, and network configuration are provided.

## REFERENCES


[1] A., Helmy, D., Estrin, "Simulation-based 'STRESS' Testing Case Study: A Multicast Routing Protocol," in *Sixth International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)*, July '98, Montreal, Canada.

[2] A., Helmy, D., Estrin, S., Gupta, "Fault-oriented Test Generation for Multicast Routing Protocol Design," in *Formal Description Techniques & Protocol Specification, Testing, and Verification (FORTE/PSTV) IFIP*, Nov. '98 Paris, France.

[3] G. Malkin, "RIP Version 2," Request for Comments (Standard) 2453, Internet Engineering Task Force, Nov., 1998.

[4] J. Moy, "OSPF Version 2," Request for Comments (Draft Standard) 2328, Internet Engineering Task Force, Apr. 1998.

[5] J. Moy, "Multicast Extensions to OSPF," Request for Comments (Proposed Standard) 1584, Internet Engineering Task Force, Mar. 1994.

[6] W. Fenner, "Internet Group Management Protocol, Version 2," Request for Comments (Proposed Standard) 2236, Internet Engineering Task Force, Nov. 1997.

[7] S. Deering, D. Estrin, D. Farinacci, V. Jacobson, A. Helmy, D. Meyer, and L. Wei, "Protocol Independent Multicast Version 2 Dense Mode Specification," Internet Draft, Internet Engineering Task Force, June 1999. Work in progress.

[8] D. Estrin, D. Farinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, V. Jacobson, C. Liu, P. Sharma, and L. Wei, "Protocol Independent Multicast-Sparse mode (PIM-SM): Protocol Specification," Request for Comments (Experimental) 2362, Internet Engineering Task Force, June 1998.

[9] T. Pusateri, "Distance Vector Multicast Routing "protocol," Internet Draft, Internet Engineering Task Force, Mar. 1999. Work in progress.

[10] B. Doar, Matthew "A Better Model for Generating Test Networks," in *Proceedings of Global Internet*, London, England, pp. 86-93, IEEE, Nov. 1996.

[11] Mil3 Inc., "OPNET Reference Materials," http://www.mil3.com/

[12] D. Mills, "Network Time Protocol (version 3) Specification," Internet-Draft. Mar. 4, 1998.

[13] Wandel & Goltermann Inc., "Domino Internetwork Analyzer," Operating guide, Nov., 1997.

[14] X. Wang, C. Yu, H., Schulzrinne, P. Stirpe, "IP Multicast Simulation in OPNET", submitted to OPNETWORK '99.