# An Improved PIM-SM Tree Recovery Algorithm

Tarik Čičić, Stein Gjessing
University of Oslo, Department of Informatics
P.B. 1080 Blindern, 0316 Oslo, Norway

Øivind Kure
Norwegian University of Science and Technology
P.B. 70, 2027 Kjeller, Norway

***Abstract*—Multicast tree recovery in PIM-SM networks is initiated by state changes in the underlying unicast routing protocol. A substantial part of the subsequent data loss can be attributed to the non-reductive events, e.g. link addition. In this paper we propose and evaluate an improvement to the standard PIM-SM recovery procedure, showing that the data loss can be significantly decreased regardless of the topology and session parameters.**

## I. Introduction

Protocol Independent Multicast – Sparse Mode (PIM-SM, [1]) was developed to meet the scalability requirements of sparse multicast groups in the Internet. It can run on top of any unicast routing protocol. PIM-SM creates and maintains unidirectional multicast trees based on explicit Join/Prune protocol messages, sent unicast on a node-to-node basis. The Join/Prune messages are directed towards the root of the tree using the unicast routing information. Multicast data packets originated at the root are forwarded only if they arrive on the same interface where the corresponding Join message was sent (Reverse Path Forwarding [2]). Routers maintain the input interface and other state information for known multicast groups in multicast routing tables. In stable operation, the unicast and multicast routing state is consistent.

Any event leading to a change in the unicast routing tables, such as the creation or failure of a link, also leads to a change in the PIM-SM multicast routing tables. We use the term *tree recovery* to describe the process of reestablishing the multicast routing in the network. In the transient phase, from the unicast routing change until the stabilization of the new multicast tree, multicast packet loss may occur.

PIM-SM has received a substantial attention in the research community [3][4]. Also, significant research has been done on error recovery in real-time IP multicast applications [5] and reliable multicast applications [6]. However, there has been less attention on the multicast tree recovery on the network level. Wang et al. [7] focussed on the performance of fault recovery in PIM Dense Mode running over OSPF. In addition, they analyzed the qualitative aspect of fault recovery of PIM running over OSPF. In [8], we have evaluated general PIM-SM recovery performance. In this paper, we further extend this work and show how to make PIM-SM recovery more reliable in terms of decreased packet loss.

## II. Background and Problem Statement

Based on unicast routing state and received Join/Prune messages, each router maintains a set of mappings between the input interface and the output interfaces for each known multicast group. The mappings are stored in the multicast routing table in the router. In its most basic form, the multicast routing table contains a set of entries $(S,G) : (i \rightarrow O)$ where $S$ determines the unicast address of the root, $G$ is the group address, $i$ is the input interface and $O$ is the set of output interfaces. In stable operation mode, for each entry, $i$ is used to send Join/Prune packets and receive multicast packets originated at the root. PIM-SM avoids packet loops by discarding $(S,G)$-packets that arrive on interfaces other than $i$.

In case of unicast routing change, for all multicast routing entries $(S,G)$, interface pointing towards $S$ is found in the unicast routing table in order to determine the (possibly) new input interface. If the new input interface differs from the old one, the multicast routing entry is updated: the new input interface is set instead of the old one and the new input interface is removed from the output interface list, if it was in it. Finally, the control messages are sent to the neighboring routers: Join-$(S,G)$ at the new input interface and Prune-$(S,G)$ at the old input interface, if it is operational. This recovery process needs time to stabilize. Multicast group members downstream from the recovery-affected routers will often experience interrupts in data delivery during this transient phase.

The unicast routing changes are caused by events belonging to three broad classes: *Topology Reduction*, e.g. link failure, removal or node failure, *Topology Enrichment*, e.g. link recovery or adding a new link and *Dynamic Routing Change*, e.g. link metric change. If the topology reduction has occurred, the packet loss is often inevitable, since it takes time to reconstruct the multicast tree using alternative links. Intuitively, events belonging to the other two classes, called "*benign events*" in the rest of this paper, should not cause any packet loss. However, the standard PIM-SM recovery procedure implies that, in the case of input interface change, the old input interface is immediately disabled. In other words, enrichment of the network by a new, operational link can cause multicast packet loss since the routers will reject the packets pending on the old, disabled input interface.

Our results [8] show that 50-80% of tested benign events lead to multicast packet loss. In this paper we discuss possible improvements of the PIM-SM recovery procedure and suggest an improved recovery algorithm aimed at reduction of the packet loss caused by the benign events. We evaluate performance tradeoffs for the suggested schemes by simulation. Finally, we provide guidelines for PIM-SM recovery implementations.

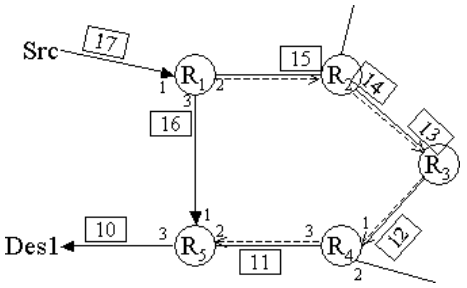


Fig. 1. Link 1-5 became operational after router 1 forwarded packet number 15. The multicast path segment is Src$-R_1-R_5-$Des1 instead of Src$-R_1-R_2-R_3-R_4-R_5-$Des1.

## III. Recovery Schemes

When a link is added to the network, the immediate switch from the old to the new input interface in the standard PIM-SM recovery scheme is the main cause of the packet loss. In the example in Fig. 1 we assume that the link $\overline{R_1 R_5}$ has just been added to the network (e.g. recovered from a failure). The source ("Src") and the destination (multicast receiver or router named "Des1") have been connected through a temporary tree, depicted using the dashed arrows. Now, the original tree is reestablished (bold arrows). The packets are arriving to router $R_5$ in sequence 11-16-12-17-13-18-14-19-15-20-21. The standard PIM-SM recovery mechanism would discard packets 11, 12, 13, 14 and 15, causing unnecessary packet loss in Des1.

We propose a modification to the standard recovery mechanism as follows:

> If the new input interface differs from the old one, the old input interface should be kept active together with the new one for a limited time period, in order to accept any pending packets that otherwise would be discarded.

We name and describe three alternative recovery schemes:

**First-on-new.** The old interface is disabled after the first packet has arrived on the new input interface.

**Complete.** The sequence number of the first packet received on the new input interface is stored. The old input interface is kept active until all pending packets with the sequence numbers lower than the stored sequence number have arrived, or a timer has elapsed (in case of upstream loss).

**Timer.** The old input interface is disabled after a configured time period.

For all three schemes, if the new input interface $i_{new}$ is in the old output interface list $O$, it has to be removed from $O$ to avoid the packet loop. This implies that our modification cannot guarantee zero loss in general. Alternatively, the Complete scheme may not remove it, but instead forward at most one copy of every packet arriving at $i_{new}$ and $i_{old}$ at the corresponding $O$-sets.

In the example in Fig. 1, "First-on-new" would discard packets 12, 13, 14 and 15. "Complete" would not discard any packets, while "Timer" would discard packets depending on the timer duration — none if the timer period is long enough.

The three alternatives vary in terms of implementation complexity and introduced overhead. The Timer scheme is simple to implement, but, if two copies of the same packet arrive to the two active input interfaces, duplicate packet is forwarded. The duplicate can be ignored by the receivers. In the network there will be an increased network load for all nodes downstream for the node with the routing update, since all duplicate packets will be forwarded. The mechanism will not cause any routing loops. The Complete scheme causes no duplicates and may eliminate packet loss due to the benign events, but it needs means for analysis of multicast flows at the transport level — the sequence numbers are available only as a part of the transport protocol, if at all.

## IV. Scheme Comparison

In this section we analyze and compare the standard PIM-SM recovery with our proposed improvements. The main tool in our evaluation is a simulation model of PIM-SM [9], which we have implemented in the Network Simulator framework [10].

### A. Simulation Environment

The simulations are performed on random networks constructed to reflect real network topologies [11, 12]. Ten random networks of 30 nodes each and average node degree of 2.5, 3.0 and 4.0 were generated. The link delay is inverse exponentially distributed between 0.5 and 13 ms with a mean value of 3 ms.

In each simulation instance, a single multicast group is created, consisting of a single, randomly placed sender and 5 randomly placed receiver nodes. The data source is a Constant Bit Rate (CBR) generator. The generated packets are 320 bytes long and sent each 2 ms.

A randomly chosen link within the multicast tree is taken down after the multicast distribution tree has stabilized and the source has started to send data. After multicast tree recovery, the link is reintroduced in the network. We count the total packet loss in all receivers caused by this benign event — from the moment the link is reintroduced until the multicast routing is stable.

Throughout this document, the per-receiver mean packet loss is used as a metric for comparison of recovery schemes and parameter sensitivity.

### B. Simulation Results

Performance of the First-on-new and Timer schemes and the standard PIM-SM recovery is compared in Fig. 2. Networks with average node degree D=3.0 were used. The Complete scheme is not simulated, since it is designed to have zero loss.

The Timer scheme reduces the standard PIM-SM loss by more than 90% for longer timer durations. The loss is not totally eliminated even for longer timer durations due to the cases where the new input interface was in the old output interface list and hence was removed (Sec. III). The Timer scheme outperforms the First-on-new scheme even for short timer durations. The figure also shows the tradeoff between the packet loss and the number of duplicate packets for the Timer scheme.

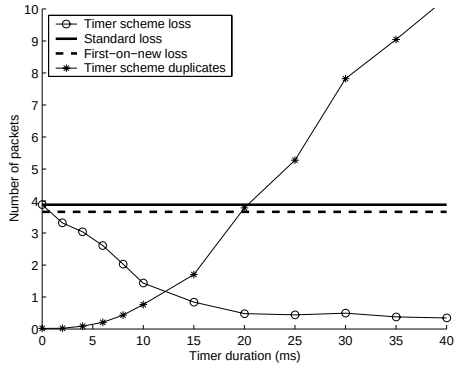


Fig. 2. Comparison of the standard PIM-SM recovery and the First-on-new and Timer schemes. Y-axis represents the per-receiver packet loss. Only the Timer scheme performance depends on the timer duration (x-axis). The Complete scheme would not cause duplicates and could have zero loss.

TABLE 1
LOSS MODEL VARIABLES

| | |
|---|---|
| $d$ | link propagation delay |
| $n$ | number of hops in pending path |
| $D$ | node degree |
| $R$ | packet rate |
| $l$ | packet length |
| $B$ | link bandwidth |
| $t_T$ | timer duration |
| $L_S$ | standard scheme loss |
| $L_T$ | Timer scheme loss |

## V. TIMER SCHEME EVALUATION

To provide better insight, we develop and validate a simple analytical model of the standard PIM-SM recovery, which can be regarded as a special case of the Timer scheme for timer duration $t_T = 0$. We then analyze the relation between the standard PIM-SM loss and the Timer scheme loss for different timer values. Finally, we evaluate and discuss the effect of different network and session parameters on the Timer scheme.

### A. Loss Model

After an occurrence of a benign event in the network, the PIM-SM routers may have to change the input interface for some of their multicast groups. In the standard recovery scheme, packets arriving at the old input interface will be discarded. For a given receiver, the network path between the closest router that has changed the input interface and the first unchanged upstream router for the given group we call the *pending path*. The length of the pending path and several traffic parameters (Table 1) determine the number of discarded packets on the old interface. In Fig. 3, receivers attached to routers $D_2$ and $D_3$ both have pending paths of four hops.

The packet loss in a receiver in a network using standard PIM-SM recovery is a product of the packet rate and the sum of the propagation and the transmission delay on each of the $n$ hops in the receiver's pending path:

$$L_S = R \cdot \sum_{i=1}^{n} (d_i + \frac{l}{B_i}) \quad (1)$$

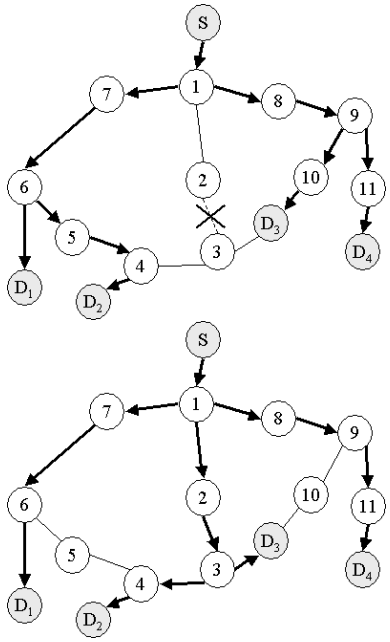


Fig. 3. Pending path. The first figure represents the situation with link $\overline{R_2R_3}$ down. In the second figure the link $\overline{R_2R_3}$ has recovered. The multicast tree from the source $S$ to the destination routers $D_{1-4}$ is depicted using bold arrows. Receivers attached to routers $D_1$ and $D_4$ are unaffected. Receivers attached to $D_2$ and $D_3$ have pending paths of 4 hops, $(R_1 - R_7 - R_6 - R_5 - R_4)$ and $(R_1 - R_8 - R_9 - R_{10} - D_3)$ respectively.

Our Timer scheme opens for keeping the old input interface alive for a fixed time period $t_T$, in hope to collect the packets pending to this interface. In any selected scenario, increasing the timer duration will imply a non-increasing packet loss and a non-decreasing number of forwarded duplicate packets.

An analytical expression for the per-receiver loss in the Timer scheme can be derived from the loss for the standard scheme (1). We have to account for the possibility that the pending path for this receiver may include a router that uses an old output interface ($o_{old} \in O_{old}$) as the new input interface ($i_{new}$) and, hence, cannot forward the packets to this receiver:

$$L_T = \begin{cases} L_S - R \cdot t_T, & \text{for } R \cdot t_T < L_S \\ & \text{and } i_{new} \neq o_{old} \\ 0, & \text{for } R \cdot t_T \geq L_S \\ & \text{and } i_{new} \neq o_{old} \\ L_S, & \text{for } i_{new} = o_{old} \end{cases} \quad (2)$$

### B. Parameter Analysis

Understanding how traffic and topology parameters influence the packet loss is needed for selecting the optimal timer period for different network configurations. In this section, we evaluate

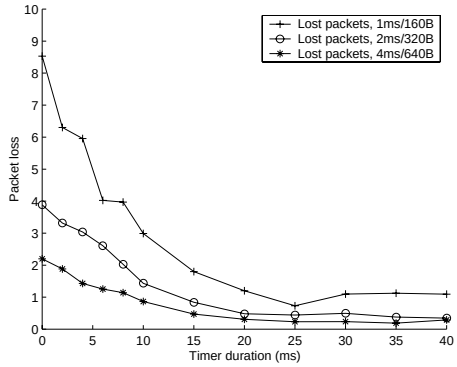


Fig. 4. The same CBT stream is divided in 160, 320 and 640 Byte long packets.

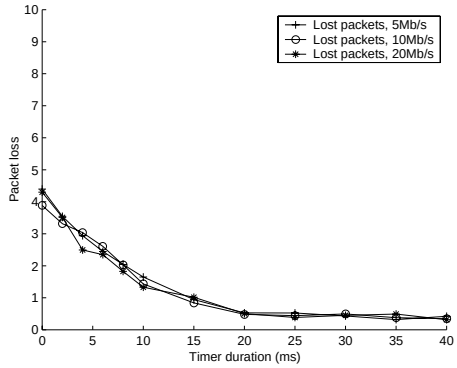


Fig. 5. Packet loss for a 1.22Mb/s stream, variable link bandwidth

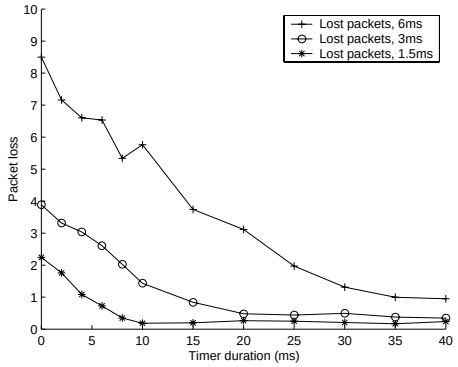


Fig. 6. Packet loss, variable mean link delay

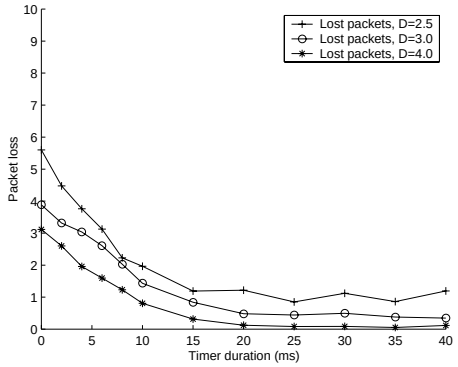


Fig. 7. Packet loss, variable node degree.

sensitivity of the Timer scheme on flow granularity, link delay, link bandwidth and average node degree using our simulator (Sec. IV).

*1) Flow Granularity* is the measure of how many packets a data quantum is divided into while preserving the mean flow bit rate. The packet loss ratio for two flows with packet rates $R_1$ and $R_2$ is expected to be roughly proportional to the $R_1/R_2$ ratio.

Our simulation results are in accordance with this assumption. In Fig. 4 the packet loss for half and double flow granularity is roughly a half and twice as high as the loss in the basic scenario.

*2) Bandwidth.* In our basic scenario, the packet transmission time is $\sim$0.26 ms (320 B packets on 10 Mb/s links). This delay stands for less than 10% of the average link delay (3 ms), therefore we expect the bandwidth to have a modest effect on the packet loss only.

In Fig. 5 we show the results for link bandwidth of 20 Mb/s, 10 Mb/s and 5 Mb/s. The packet loss is slightly higher for 5 Mb/s than for 10 Mb/s, in accordance with (1) and (2).

*3) Link Delay.* We have doubled and halved the link delay at all links in our test networks. The registered packet loss is more than doubled for the doubled link delay, and almost halved for the halved link delay (Fig. 6).

The packet loss for $t_T = 0$ and large $t_T$ is similar for the link delay tests (Fig. 6) as for the flow granularity (Fig. 4). However, in the flow granularity tests the loss value falls steeper for higher packet rate. This is in accordance with relation (2), where $R \cdot t_T$ is subtracted from the standard scheme loss (and $R$ is a parameter in the flow granularity tests). Validity of (2) is also confirmed by the fact that the flow granularity performance and the link delay performance seem to have the same value for large $t_T$.

*4) Node Degree.* The packet loss performance for networks with different average node degree follows the same pattern as the other curves, with a difference in that the values for large $t_T$ seems to be significantly higher for lower average node degree. In fact, for D=2.5 the $L_T(40)/L_T(0)$ ratio is just under 0.25, while for D=4.0 we have $L_T(40)/L_T(0) < 0.05$. This can be explained by that an old output interface on the pending path becomes the new input interface ($i_{new} \in O_{old}$) is less probable for the networks with higher average node degree.

*5) Session randomness.* Our simulation results show non-monotonic curves, even though we expect monotonically non-decreasing loss $L_T(t_T)$, according to relation (2). The large variation in the measured packet loss can be explained by the significant influence of the placement of the sender and the receivers and the choice of the faulty link in each simulation instance.

To confirm this, we conduct a separate set of simulations, where

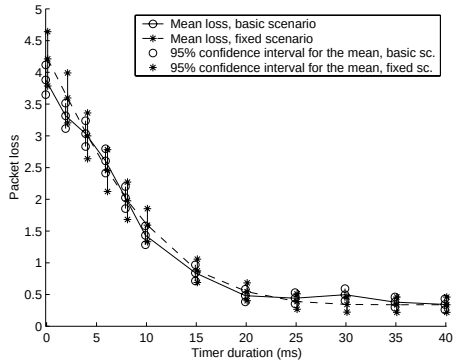


Fig. 8. 95% confidence interval for the mean, the Timer scheme. The "fixed scenario" curve represents simulations where the sender, receivers and the faulty link are fixed throughout the $t_T$ range (x-axis). Y-axis in not in the same scale as the previous graphs!

the same sender/receivers distribution and the faulty link choice is preserved throughout the timer duration range.

The mean values and the 95% confidence intervals for the mean for both simulation sets are compared in Fig. 8. The smooth, dashed line represents the results of the new, "fixed" scenario, The solid, variable line represents our original scenario. The similarity of the results indicates that the sender/receiver placement and the faulty link choice account for the variation.

### C. Discussion

We emphasis two observations deduced from our Timer scheme performance analysis:

1. It is possible to set the timer duration so that the mean packet loss is equal to any fraction of the standard PIM-SM loss larger than the minimal loss value for given average node degree in the network ($\sim$ 10% for D=3.0). The effect of the timer setting on total data loss is independent of the flow granularity.

2. The minimal value of packet loss for large timer durations is smaller for high average node degrees and low link delays.

For example, to reduce the packet loss to under 25% of the original value, the timer duration should be set to 5x average link delay in networks with average node degree D=3.0.

These observations lead us to conclude that the packet loss can be tuned down to the desired fraction based *only on the network parameters* such as the link delay and the average node degree.

## VI. Conclusion

We have presented an improved PIM-SM recovery mechanism that significantly reduces the packet loss caused by benign events.

Three alternative recovery schemes have been proposed and evaluated. The simple First-on-new scheme has a moderate effect on the packet loss. The Complete scheme may eliminate the packet loss, but is complicated to implement and requires access to the packet sequence numbers.

The proposed Timer scheme reduces the packet loss by 90% in networks with average node degree 3.0 and even more in higher connectivity networks. The Timer scheme is simple to implement and introduces overhead that is probably bearable for most networks. Furthermore, the timer duration can be selected using only network parameters such as the mean link delay and node degree.

## References

[1] D. Estrin, D. Farinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, V. Jacobson, C. gung Liu, P. Sharma, and L. Wei, "Protocol independent multicast – sparse mode (PIM-SM): Protocol specification." RFC 2362, June 1998.

[2] Y. K. Dalal and R. M. Metcalfe, "Reverse path forwarding of broadcast packets," *Communications ACM*, vol. 21, pp. 1040–1048, Dec. 1978.

[3] L. Wei and D. Estrin, "Multicast routing in dense and sparse modes: simulation study of tradeoffs and dynamics," in *Computer Communications and Networks, Fourth International Conference on*, pp. 150–157, IEEE, Sept. 1995.

[4] T. Billhartz, J. B. Cain, E. Farrey-Goudreau, D. Fieg, and S. G. Batsell, "Performance and resource cost comparison for the CBT and PIM multicast routing protocols," *IEEE J. Selec. Areas Commun.*, vol. 15, pp. 304–315, Apr. 1997.

[5] G. Carle and E. W. Biersack, "Survey of error recovery techniques for IP-based audio-visual multicast applications," *IEEE Network*, pp. 24–36, Nov. 1997.

[6] C. Diot, W. Dabbous, and J. Crowcroft, "Multipoint communication: A survey of protocols, functions, and mechanisms," *IEEE J. Selec. Areas Commun.*, vol. 15, pp. 277–290, Apr. 1997.

[7] X. Wang, C. Yu, H. Schulzrinne, P. Stirpe, and W. Wu, "IP multicast fault recovery in PIM over OSPF," in *Proceedings ACM SIGMETRICS*, June 2000.

[8] T. Čičić, S. Gjessing, and Ø. Kure, "Performance evaluation of PIM-SM recovery," in *Proceedings of International Conference on Networking ICN'01, Colmar, France*, July 2001. In press.

[9] T. Čičić, S. Gjessing, and Ø. Kure, "Tree recovery in PIM sparse mode," Research Report 293, University of Oslo, Department of Informatics, Mar. 2001. ISBN 82-7368-243-9.

[10] UCB/LBNL/VINT, "Network simulator - ns (version 2)." WWW. HTTP://WWW.ISI.EDU/NSNAM/NS/.

[11] E. W. Zegura, "Georgia tech internetwork topology models." WWW. HTTP://WWW.CC.GATECH.EDU/FAC/ELLEN.ZEGURA/GRAPHS.HTML.

[12] B. M. Waxman, "Routing of multipoint connections," *IEEE J. Selec. Areas Commun.*, vol. 6, pp. 1617–1622, Dec. 1988.