# On the Impact of Route Processing and MRAI Timers on BGP Convergence Times

Shivani Deshpande and Biplab Sikdar
Department of ECSE, Rensselaer Polytechnic Institute, Troy, NY 12180

***Abstract*— Fast convergence of BGP routes, coupled with reduced message complexity form one of the key factors determining the stability and performance of inter-domain routing in the Internet. This paper characterizes the impact of topology and the message handling procedure of BGP on its convergence time. For BGP router networks characterized by random graphs, we obtain analytic expressions to evaluate the convergence times in terms of the number of Minimum Route Advertisement Interval (MRAI) timer rounds. We then isolate artifacts in the BGP message handling procedure which lead to redundant MRAI timer instances and propose minor but effective modifications which significantly reduce the convergence times. Simulation results show that the proposed changes can successfully eliminate all redundant instantiations of the MRAI timer, even in the worst case convergence scenarios resulting from route withdrawals.**

## I. Introduction

The dynamics and convergence times of Border Gateway Protocol (BGP) play an important role in the end user's perception of network performance, arising from the fact that BGP controls the flow of inter-domain traffic. Recent studies have shown that the establishment of stable routes after a node failure can take on the order of 3 to 15 minutes [4]. In [4] it was also shown that in a fully connected network, the lower bound on convergence time is given by $(n-3)minRouteAdver$ in a $n$ node network where $minRouteAdver$ is the MRAI interval (usually 30 seconds). This paper focuses on relating the causes of the MRAI timer instances to the network topology and BGP's message handling procedure. We then propose techniques to eliminate these causes.

Research on BGP convergence times has focused on both evaluating the bounds on the convergence times as well as techniques to improve it. In general, BGP does not have any bounds on the convergence times due to its immense flexibility and adaptability achieved from its open-ended structure and policy based routing possibilities [3], [5], [8]. In [2] it was shown that the optimal value of $minRouteAdver$ depends on the specific topology and thus cannot serve as a general mechanism to reduce convergence times. While consistency assertions may reduce the convergence times [6], this requires extra computation resources and additional message overheads and runs into difficulties in some pathological cases. Reducing convergence times through improved ghost flushing has been proposed in [1]. However, this does not eliminate all cases of unnecessary MRAI timer invocations.

We present here a detailed analysis of BGP's message handling and route establishment procedures and model the impact of the same and of the topology on the occurrence of MRAI instances. This allows us to identify different classes of scenarios where redundant MRAI timers may occur and we then propose minor changes to specifically target these factors. While the proposed changes are quite minor, our simulations show that they successfully eliminate the instances of redundant invocations of the MRAI timer originating from two causes. Depending on the topology, the reductions on the convergence times are of the order of 30 seconds to a few minutes. Additionally, the proposed changes achieve their objective of low convergence times without increasing BGP's message handling complexity. Our techniques are also effective at reducing convergence times for events initiated by route withdrawals. We specifically show the effectiveness of our techniques for the worst case scenarios identified in [4].

The rest of the paper is organized as follows. Sections II and III present simulation and analytic studies of how the topology and BGP's message handling process lead to unnecessary MRAI timer instances, respectively. Sections IV and V present the modifications proposed to BGP and simulation results to show their effectiveness. Finally, Section VI presents the concluding remarks.

## II. BGP Message Handling and Convergence Times

The dynamics of the convergence is tied closely to the interaction of the message handling process and the MRAI timer and also the topology of the network. In this section, we present a detailed description of how certain nuances of these factors can lead to unnecessarily large convergence times.

### A. BGP Overview

BGP is based on the path vector routing mechanism. It involves exchange of a path vector i.e. a list of Autonomous Systems (ASes) that need to be traversed in order to reach a particular destination. Inter-domain routes are distributed using BGP UPDATE messages which contain the list of currently reachable IP-prefixes to which a route is being advertised or withdrawn, list of ASes that are traversed to reach a particular network destination as well as various route attributes. In order to control the rate of BGP messages pushed into the network, each router uses the MRAI timer. Two UPDATE messages sent from a single BGP speaker that advertise feasible routes to some common set of destinations must be separated by at least one MRAI. If new routes are selected multiple times while awaiting the expiration of the MRAI timer, the last route selected is advertised at the end of that Interval. With inter-AS rounds trip times being much smaller than the MRAI (default value 30 seconds), BGP convergence delay can be characterized in terms of MRAI rounds.

The MRAI timer is instantiated at a BGP speaker when the route to a particular destination changes due to the explicit withdrawal of some pre-existing route at that speaker or due to its implicit withdrawal. An implicit withdrawal is caused when a route of higher degree of preference is received by a BGP router. In this case the currently installed route is withdrawn from service and the new route is installed instead. These implicit or explicit withdrawals lead to a sequence of UPDATE messages and we now describe how they can lead to redundant instantiations of the MRAI timer thereby increasing the convergence times.

### *B. Some Observations on Convergence Times*

In this section we use simulations with the SSFNet simulator [7] to analyze the generation of redundant MRAI instances in BGP routers. Topologies were generated for various network sizes and link existence probabilities between arbitrary pairs of nodes. For all the simulations the simulation time was $100$seconds. Also, each AS had one BGP speaker and the default MRAI timer value implemented was 30 seconds. Also the shortest path-length criteria is considered before the local preference criteria for route selection as is prevalent in the Internet today [4].

From the results obtained for the set of experiments run, the following prominent observations were made:

- Most of the random topologies generated required at least one MRAI round for convergence.
- The presence of multiple routes going to the same destination leads to the formation of MRAI instances.
- Though their occurrences are rare, the presence of “nested” multiple routes for a given destination can lead to convergence times with multiple MRAI instances. For an example of a nested path, consider ASes $AS2$ and $AS10$ in Figure 1. Note that $AS4$ has two 2-hop paths to $AS10$: 4-3-10 and 4-9-10. Now, these multiple paths at $AS4$ form nested multiple 3-hop paths at $AS2$ when we consider the following possible paths: 2-4-3-10, 2-4-9-10 and 2-1-9-10. Note, however that in this topology multiple MRAI instances are not observed as none of the nested routes are installed due to the presence of a shorter path $2-3-10$. In general, the invocation of the MRAI timer for the two hop paths can delay the invocation of the MRAI timer for the three hops paths and so on, leading to a cascade of MRAI timer instances.

We now take the topology shown in Figure 1 and elaborate on the causes that lead to redundant MRAI instances.

### *C. Redundant MRAI Instances*

Consider the case where a BGP router currently has a MRAI timer instantiated for one of its peers for a given destination. Now, if this peer advertises a route of larger length (i.e. lower preference) for the same destination to the BGP router, it is clear that retaining the current MRAI timer instantiation will only delay the eventual convergence to the shorter route. However, according to the current BGP route processing rules, when a longer route is received from one of the BGP peers for the same destination for which currently a MRAI timer is installed, the MRAI timer is still not cancelled. For an example of such an event, consider again Figure 1 and the four routes received by $AS6$ to reach $AS10$: (1) AS-path: $9-10$ from $AS9$, (2) AS-path: $7-10$ from $AS7$, (3) AS-path: $4-9-10$ from $AS4$ and (4) AS-path $5-9-10$ from $AS5$. In the simulation, the path with the lower degree of preference i.e. $7-10$ was received first at $AS6$, then established and advertised. The receipt of the path $9-10$ (which has higher degree of preference (DOP)) immediately afterwards, resulted in an implicit withdrawal of this route. However, this updated route cannot be advertised immediately because a MinRouteAdvertisement Interval has not passed since the last route advertisement. Also, much before the MRAI timer expires and $AS6$ can advertise its route for $AS10$ to these peers, it receives the routes $4-9-10$ and $5-9-10$ from $AS4$ and $AS5$ respectively. These routes are longer than the route to $AS10$ that has been currently installed. This can serve as an indication at $AS6$ that an invalid MRAI instance has been installed for these two peers.

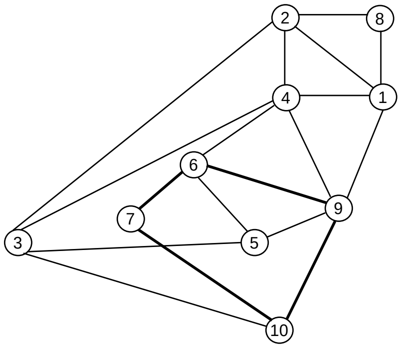


Fig. 1. Topology showing multiple paths between ASes: *Paths 6-7-10 and 6-9-10 are highlighted*

## III. A MODEL FOR THE MRAI TIMER

We now determine the probability that the convergence phase, either the initial or when a new address prefix is advertised, involves $k$ MRAI instances. Consider a network of $N$ BGP routers where a BGP session exists between any of the peers with probability $p$ and the delays on each link are drawn from an identical distribution. We also assume that paths with the shortest hop count are preferred over other paths and for paths with the same length, any one is chosen arbitrarily.

We first find the probability that there is at least one MRAI timer instance at an arbitrary peer. From our discussion in the last section, the MRAI timer is instantiated when the route with the highest DOP arrives at the peer after other route advertisements to the destination are received and incorporated in LOC-RIBS. Now, given any two peers, the number of possible $i$ hop paths with distinct nodes between them in a given topology is given by $n(i, N) = N - i - 4$, $i \geq 2$, and each of these paths exists with probability $p^i$. Now, the longest path possible in the network has $N - 1$ hops. However, the longest path possible with at least 2 paths of that length is given by $\lambda(N) = N - 2,\ N \geq 4$.

Given $j$ paths of length $i$ hops, the probability that a path with the highest DOP arrives first at the node is $1/j$ and in this case, the MRAI timer is not instantiated. In all other cases, a path with a lower DOP is established as the path to the destination and the MRAI timer will be instantiated and the probability of the event is given by

$$\binom{n(i,N)}{j}(p^i)^j(1-p^i)^{n(i,N)-j}\frac{j-1}{j} \tag{1}$$

The probability of a MRAI timer instantiation at the BGP router is then obtained by summing the above expression for all possible values of $j$ and path length $i$. Thus

$$P[\text{MRAI=1}] = \sum_{i=2}^{\lambda(N)} \sum_{j=2}^{n(i,N)} \left[\binom{n(i,N)}{j}\frac{j-1}{j}(p^i)^j(1-p^i)^{n(i,N)-j}\right] \tag{2}$$

The above expression assumes that paths of all possible lengths are advertised and propagated inside the network. However, it is more likely that paths longer in length than those already received are not propagated in the network (though strictly, it depends on the specified local policy at the BGP router). In such cases, we only need to consider the possibility of an $i$ hop path resulting in a MRAI instance only if it is one of the shortest paths. To account for such cases, we first note that for a network with $N$ routers, the maximum possible number of distinct $i$ hop paths is given by

$$\eta(i,N) = \frac{(N-2)!}{(N-i-1)!} \tag{3}$$

Since each $i$ hop path exists with probability $p^i$, the probability that no $i$ hop path exists is given by

$$P[\text{no } i \text{ hop path exists}] = (1-p^i)^{\eta(i,N)} \tag{4}$$

Then, the probability that the shortest path between any two routers is at least $i$ hops is given by

$$P[\text{shortest path} \geq i \text{ hops}] = \prod_{j=1}^{i-1}(1-p^j)^{\eta(j,N)} \tag{5}$$

Combining the above expression with Eqn. (2), the probability of a MRAI timer instance when we only forward the shortest path received so far is given by

$$P'[\text{MRAI=1}] = \sum_{i=2}^{\lambda(N)} \sum_{j=2}^{n(i,N)} \left[\binom{n(i,N)}{j}\frac{j-1}{j}(p^i)^j (1-p^i)^{n(i.N)-j}\prod_{k=1}^{i-1}(1-p^k)^{\eta(k,N)}\right] \tag{6}$$

The above expression considers only one pair of routers in the network. Accounting for all the $N(N-1)/2$ possible source destination pairs in the network, the probability of an MRAI timer instance is given by

$$P[\text{MRAI=1}] = 1 - \Big(1 - P'[\text{MRAI=1}]\Big)^{N(N-1)/2} \tag{7}$$

Note that the expression obtained above is an approximation of the probability of a MRAI instance in the network. However, they match closely with simulation results and these results are presented later in this section.

We now evaluate the probabilities of multiple MRAI expirations, starting with the case of 2 consecutive MRAI instances. Note that in order to have two consecutive MRAI instances, we first need a case with a single MRAI timer instance with a subset of the nodes in the network, which can be obtained using Eqn. (6). Then, among the remaining nodes, we need to find the probability of existence of parallel paths to the same destination. Consider the case when the longer of the nested paths is of length $i$ hops. The inner nested paths can thus be at most $i-1$ hops. If there are $j$ such $i$ hop paths between the two routers under consideration, the maximum possible number of nodes available to form the inner nested paths is $N' = N - 1 - (i-1)j$. The possibility of an MRAI timer in these $N'$ nodes with a maximum path length of $i_m = i - 1$ can be obtained from Eqn. (6) as

$$P[\text{MRAI} = 1|N', i_m] = \sum_{i=2}^{\min\{\lambda(N'), i_m\}} \sum_{j=2}^{n(i,N')} \left[\binom{n(i,N')}{j} (p^i)^j(1-p^i)^{n(i,N')-j}\frac{j-1}{j}\prod_{k=1}^{i-1}(1-p^k)^{\eta(k,N')}\right] \tag{8}$$

Now, the inner nested paths have a minimum length of at least two hops and after accounting for the source and destination nodes and the common node of the inner nested paths, we have at most $N-5$ nodes available for the outer paths. Then, using similar logic as in the single MRAI case, the probability of two MRAI timer instances for a given pair of routers is

$$P'[\text{MRAI} = 2] = \sum_{i=3}^{\lambda(N-5)} \sum_{j=2}^{n(i,N-5)} \left[P[\text{MRAI} = 1|N', i_m] (p^i)^j \prod_{k=1}^{i-1}(1-p^k)^{\eta(k,(i-1)j+2)}\right] \tag{9}$$

where $N' = N - 1 - (i-1)j$ and $i_m = i - 1$. Accounting for all pairs of routers, the probability of two consecutive MRAI timer instances in the network can be expressed as

$$P[\text{MRAI} = 2] = 1 - \Big(1 - P'[\text{MRAI} = 2]\Big)^{N(N-1)/2} \tag{10}$$

The expressions for even greater sequences of MRAI timer instances can be obtained similarly. Note that the above model assumes that all link delays are approximately the same, and thus a path of longer length does not arrive at a node before any shorter paths.

In Table I we compare the results obtained for $P[\text{MRAI}{=}\ 1]$ from Equation (7) with the simulation results for $N = 10$ and note that the two results are in close agreement. Similar results were also obtained for other network sizes.

The observations of this section show that redundant MRAI instances occur with probabilities close to 1 for most cases underlying the need to address the cause behind them.

| Edge probability | Simulation | Analysis |
|---|---|---|
| $p = 0.15$ | 0.350000 | 0.225896 |
| $p = 0.30$ | 0.923077 | 0.941511 |
| $p = 0.45$ | 1.000000 | 0.999729 |
| $p = 0.60$ | 1.000000 | 0.999990 |
| $p = 0.75$ | 1.000000 | 0.999907 |

TABLE I

$P$[MRAI= 1] in a random network with $10$ nodes

| Number of nodes | Same link delays | | Different link delays | |
|---|---|---|---|---|
| | p = 0.3 | p = 0.6 | p = 0.3 | p = 0.6 |
| 10 | 66.13 | 52.64 | 42.04 | 53.40 |
| 15 | 59.89 | 47.47 | 49.09 | 57.81 |
| 20 | 63.94 | 43.99 | 53.86 | 62.74 |
| 25 | 67.89 | 43.20 | 58.85 | 64.87 |

TABLE II

Average percent reduction in number of MRAI rounds after implementing longer route receipt detection

| Number of nodes | Same link delays | | Different link delays | |
|---|---|---|---|---|
| | p = 0.3 | p = 0.6 | p = 0.3 | p = 0.6 |
| 10 | 100 | 100 | 100 | 100 |
| 15 | 100 | 100 | 100 | 100 |
| 20 | 100 | 100 | 100 | 100 |
| 25 | 100 | 100 | 100 | 100 |

TABLE III

Average percent reduction in number of MRAI rounds after implementing adaptive MRAI timer.

## IV. Reducing Convergence Times

### A. Longer Route Receipt Detection

In order to identify and eliminate the occurrence of redundant MRAI timers on the receipt of a longer route, we propose the following changes to BGP's message handling procedure:

- On the receipt of a route advertisement check if a MRAI timer has been installed for the sender.
- Check if the route that is received is a route longer than the one currently in use for that particular destination.
- If both the above conditions are met then cancel the MRAI timer for that particular peer and send the advertisement for the current route to that peer.

We implemented this proposed change in the SSFNet simulator. For the topology of Figure 1, $66.6\%$ of the MRAI timer instances were eliminated. To further validate the effectiveness of our scheme, we performed detailed simulations for topologies with 10, 15, 20 and 25 ASes and link probabilities, $p$, between any AS pair of 0.3 and 0.6. Additionally, the simulations conducted can be classified into two categories: topologies with identical link delays of 30ms for all links and topologies with random link delays, uniformly distributed between $0$ and $100$ms. The percentage reduction in the number of MRAI timer instances for these cases are shown in Table II.

### B. Adaptive MRAI timer

Another strategy that we suggest for eliminating the redundant instances of the MRAI timer especially when the route changes are valid is to make the MRAI timer adaptive. The implementation of this strategy can be summarized as follows:

- A new variable called MRAI_THRESHOLD is used in each BGP speaker. This is the number of times the route to any particular destination has to change before the MRAI constraint is applied to the route.
- The MRAI_THRESHOLD is set statically to a particular value at the beginning of the BGP session establishment. Later, it can be dynamically updated depending on how frequently the route changes occur in the given system.
- For each reachable destination entry in a BGP router, a counter value called MRAI_COUNTER is stored.
- When a route for a particular destination is received, the MRAI_COUNTER associated with the corresponding entry is incremented. When a new route is received, the counter for the destination is initialized to one.
- The MRAI timer is started for a particular destination, only if the route to it has changed more than MRAI_THRESHOLD number of times.
- In the initial convergence phase and when a BGP router is learning new routes from its neighbors, the MRAI timer will not be installed for the first few implicit withdrawals.
- If the network is seen to exhibit frequent or large number of route oscillations, the value of MRAI_THRESHOLD can be decreased to reflect this and avoid a lot of unnecessary updates.

As a result of this strategy, each route change will be separated based on whether it is exhibiting route fluctuations or it is in the initial route learning phase. These proposed changes were also implemented in the SSFNet simulator to evaluate their effectiveness. In the simulations that we ran the MRAI_THRESHOLD is set to 4. With these changes in place, *all* the MRAI instances generated for the topology in Figure 1 were eliminated. In all the topologies simulated, the BGP convergence delay was reduced by at least $30 seconds$ as the generation of all the MRAI instances was avoided as can be seen in Table III.

## V. Explicit Route Withdrawals

We now show that the adaptive MRAI timer can be equally effective for reducing convergence times when a BGP peer explicitly withdraws its routes to a particular destination. An analysis of the BGP message handling in the worst case scenarios resulting from such withdrawals is given in [4]. The authors showed that the lower bound on the convergence time is at least $(n - 3)$ rounds of the MRAI timer for a fully connected BGP topology where $n$ is the number of ASes and that such topologies have the worst case message complexity of $(n-1)O(n-1)!$ messages in the absence of any MRAI timer constraint. The application of the MRAI timer prevents the generation of a large number of messages and thereby reduces the message complexity of the BGP system. However, the delay introduced in the final convergence is high since each MRAI round is $30 seconds$.

With the use of the adaptive MRAI timer, the convergence delay can be reduced significantly at the cost of very little

| Nodes | Convergence time | | Number of Messages | |
|---|---|---|---|---|
| | w/o adaptive MRAI timer | with adaptive MRAI timer | w/o adaptive MRAI timer | with adaptive MRAI timer |
| 6 | 65.542572 | 10.085351 | 247 | 458 |
| 8 | 125.572803 | 11.412084 | 680 | 1862 |
| 10 | 185.572760 | 14.367129 | 1424 | 5261 |
| 12 | 215.663290 | 20.609212 | 2561 | 15743 |
| 14 | 335.572764 | 27.726322 | 4294 | 28718 |
| 16 | 335.633110 | 72.360037 | 6303 | 109069 |
| 18 | 395.633104 | 202.239348 | 9083 | 28718 |

TABLE IV

CONVERGENCE TIME AND NUMBER OF UPDATE MESSAGES AFTER WITHDRAWAL OF A ROUTE WITHOUT AND WITH THE ADAPTIVE MRAI TIMER IMPLEMENTED IN FULL MESH BGP TOPOLOGIES.

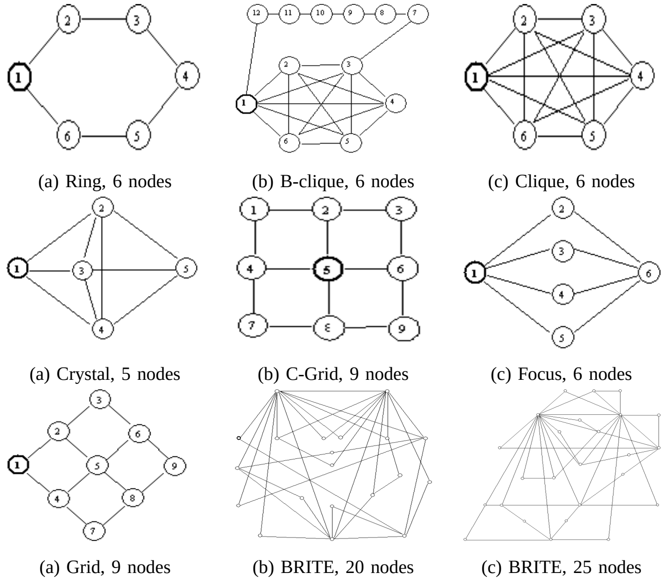
| Topology | Convergence time without adaptive MRAI timer | Convergence time with adaptive MRAI timer |
|---|---|---|
| Clique of size 6 | 65.542 | 8.326 |
| B-Clique of size 6 | 99.542 | 12.552 |
| Ring 15 | 39.48 | 9.483 |
| Focus 16 | 39.181 | 9.301 |
| Grid-16 | 69.422 | 10.447 |
| C-Grid-16 | 99.211 | 11.804 |
| BRITE-20 | 36.296 | 10.658 |
| BRITE-25 | 66.422 | 12.0758 |
| 5 node CRUSTAL | 39.211 | 9.241 |

TABLE V

CONVERGENCE TIME AFTER ROUTE WITHDRAWAL WITH AND WITHOUT THE ADAPTIVE MRAI TIMER FOR VARIOUS BGP TOPOLOGIES.

added message complexity. In this case, the number of messages exchanged at a node till it finally converges is more than that for the normal MRAI timer implementation. However, no MRAI instances are generated till the MRAI_THRESHOLD value is crossed by the number of route changes. So the final convergence of the BGP system is faster and is lower than the $(n-3)$ rounds required with the normal MRAI timer. Moreover, the message complexity is significantly lower than $(n-1)O(n-1)!$ messages required for cases without an MRAI timer. The adaptive MRAI timer thus provides a compromise between the convergence time and message complexity.

To illustrate this we performed simulations for larger fully connected topologies upto 20 ASes with constant link delays of $30ms$. The message complexity and the final convergence delay were dependent on the MRAI_THRESHOLD selected and the optimal value of MRAI_THRESHOLD is topology dependent. The results for the simulations are as shown in Table IV. We note that the adaptive MRAI timer significantly reduces the convergence times while the message complexity is higher. However, the MRAI_THRESHOLD values used in these simulations could be lowered to reduce the message complexity if that is of higher priority. Similar conclusions can be drawn from the simulations results shown in Table V for the different topologies illustrated in Figure 2. These simulations show the impact of the withdrawal of nodes of varying connectivity. In all these simulations the MRAI_THRESHOLD value is selected to be 4.


(a) Ring, 6 nodes (b) B-clique, 6 nodes (c) Clique, 6 nodes

(a) Crystal, 5 nodes (b) C-Grid, 9 nodes (c) Focus, 6 nodes

(a) Grid, 9 nodes (b) BRITE, 20 nodes (c) BRITE, 25 nodes

Fig. 2. Small scale samples of typical AS-level topologies.

## VI. CONCLUDING REMARKS

This paper investigates the impact of BGP's message handling and route processing mechanism on BGP's convergence times. We identify causes which result in redundant invocations of the MRAI timer during the route selection process, thereby significantly increasing the convergence times of BGP routers. We then propose modifications to BGP to address these issues and using simulations, show that they can remove the redundant MRAI timer instantiations resulting from these causes. The proposed solutions for reducing the convergence times are simple to implement, effective and independent of the topology. Also, the proposed modifications to the route establishment process do not have any undesirable effect on the operations of BGP and its MRAI timer during periods of transient route fluctuations. We also showed the effectiveness of our strategies at reducing the convergence times for situations resulting from explicit route withdrawals.

## REFERENCES

[1] A. Bremler-Barr, Y. Afek and S. Schwarz, "Improved BGP convergence via ghost flushing," *Proceedings of IEEE INFOCOM,* San Francisco, CA, March 2003.

[2] T. Griffin and B. Premore, "An experimental analysis of BGP convergence time," *Proceedings of IEEE ICNP,* pp. 53-61, Riverside, CA, November 2001.

[3] T. Griffin and G. Wilfong, "An analysis of BGP convergence properties," *Proceedings of ACM SIGCOMM,* pp. 277-288, Boston, MA, September 1999.

[4] C. Labovitz, A. Ahuja, A. Bose and F. Jahanian, "Delayed Internet routing convergence," *IEEE/ACM Transactions On Networking,* vol. 9, no. 3, pp. 293-306, June 2001.

[5] D. Obradovic, "Real-time model and convergence time of BGP," *Proceedings of IEEE INFOCOM,* pp. 893-901, New York, NY, June 2002.

[6] D. Pei, X. Zhao, L. Wang, D. Massey, A. Mankin, S. Wu and L. Zhang, "Improving BGP convergence through consistency assertions," *Proceedings of IEEE INFOCOM,* pp. 902-911, New York, NY, June 2002.

[7] "The SSFNet Project," `http://www.ssfnet.org`.

[8] K. Varadhan, R. Govindan and D. Estrin, "Persistent route oscillations in inter-domain routing," *Computer Networks,* vol. 32, no. 1, pp. 1-16, January 2000.