Route Flap Damping Exacerbates Internet Routing Convergence

Zhuoqing Morley Mao
UC Berkeley
zmao@cs.berkeley.edu

Ramesh Govindan
ICSI
ramesh@icsi.berkeley.edu

George Varghese
UC San Diego
varghese@cs.ucsd.edu

Randy H. Katz
UC Berkeley
randy@cs.berkeley.edu

ABSTRACT

Route flap damping is considered to be a widely deployed mechanism in core routers that limits the widespread propagation of unstable BGP routing information. Originally designed to suppress route changes caused by link *flaps*, flap damping attempts to distinguish persistently unstable routes from routes that occasionally fail. It is considered to be a major contributor to the stability of the Internet routing system.

We show in this paper that, surprisingly, route flap damping can significantly exacerbate the convergence times of relatively stable routes. For example, a route to a prefix that is withdrawn *exactly once* and re-announced can be suppressed for up to an hour (using the current RIPE recommended damping parameters). We show that such abnormal behavior fundamentally arises from the interaction of flap damping with BGP path exploration during route withdrawal. We study this interaction using a simple analytical model and understand the impact of various BGP parameters on its occurrence using simulations. Finally, we outline a preliminary proposal to modify route flap damping scheme that removes the undesired interaction in all the topologies we studied.

Categories and Subject Descriptors

C.2.2 [**Computer-Communication Networks**]: Network Protocols—*Routing Protocols*

General Terms

Performance, experimentation

Keywords

BGP, border gateway protocol, interdomain routing protocol, route flap damping, routing convergence, routing dynamics

SIGCOMM'02, August 19-23, 2002, Pittsburgh, Pennsylvania, USA.

1. INTRODUCTION

Routing mechanisms that trade-off route convergence or optimality for increased stability are often described in the routing literature. One such instance is the experience with load-based routing in the old ARPAnet, where routing system stability was achieved only by significantly *damping* link metrics [1]. Similarly, Cisco and Juniper deliberately delay route calculations in IS-IS implementations to increase stability [2]. A second instance is the default setting of Hello timers in intra-domain routing protocols. Existing implementations use fairly conservative values for these timers, resulting in slower detection of link state changes and consequently less routing update traffic [2]. In this paper, we analyze a third instance, BGP *route flap damping*.

Route flap damping is a mechanism designed to selectively limit the propagation of unstable routing information [3]. It works as follows. Each BGP-speaking router maintains a route *penalty* associated with every prefix announced by each BGP neighbor. This route penalty increments by some fixed value whenever the state of the route changes and exponentially decays with time. In effect, the penalty measures the instability of a route. The router uses locally configured thresholds to decide when to *suppress* the route (*i.e.*, not use the route because it is unstable) and when to subsequently *reuse* the route. Section 2 describes the flap damping mechanism in greater detail.

Originally proposed in the early days of the commercial Internet, route flap damping is generally assumed by the operator community to be widely deployed in today's infrastructure [4, 5]. Furthermore, it is widely held to be one of the main contributors to the overall stability of the Internet inter-domain routing system [5] by the operator community. However, there have been no rigorous studies to quantify the extent of deployment of route flap damping, nor any studies to quantify the impact route flap damping has on the stability of the Internet. We plan to pursue such studies in our future work. While the original target of route flap damping was route flaps caused either by router mis- or re-configuration, or by chronically unstable links, the mechanism can prevent the widespread propagation of other kinds of routing pathologies. These include persistent route oscillations caused by mutually incompatible policies [6], as well as route changes resulting from the repeated BGP connection tear-down and re-establishment that has been known to occur as a result of incompatible implementations.

However, as we show in this paper, route flap damping can actually *exacerbate the convergence of relatively stable routing information, sometimes by up to an hour*. The intuition for this comes from the work of Labovitz *et al.* [7], who showed that a single route withdrawal can result in other routers exploring a sequence of alternate paths before deciding that the destinations is unreachable. In this paper, we show that this kind of exploration causes what

we call *secondary flaps* that can trigger the suppression threshold of the route flap damping algorithm. This prevents the widespread propagation of a subsequent route announcement, resulting in the delayed convergence of the route. We describe this phenomenon – *withdrawal triggered suppression* – in greater detail in Section 3.

We conjecture that withdrawal triggered suppression explains the tail of the convergence distribution from the experiments of Labovitz *et al.* [7, 8]. Even though their experiments injected route changes roughly once every two hours (and therefore should not have triggered route flap damping), they found that routes took nearly fifteen minutes to converge (a time constant that is consistent with at least one set of route flap damping parameter values [9]). Furthermore, as more and more Autonomous Systems are multi-homed today [5], one can expect greater levels of path exploration, resulting in greater likelihood of route suppression.

In addition to describing the withdrawal triggered suppression, we gain insight into the phenomenon both through analysis (Section 4) and simulation (Section 5). Analysis characterizes the progress of secondary flaps and their impact on convergence in simple topologies. Simulation in SSFNet [10] studies how, if at all, various proposed BGP features (such as sender-side loop detection and withdrawal rate-limiting [7]) impact withdrawal triggered suppression. To our surprise, topologies with more alternate paths do not necessarily have a greater likelihood of exhibiting withdrawal triggered suppression. We also find that in some topologies, sender-side loop detection is effective in eliminating this phenomenon. In Section 6 we analyze real traces to show that such flaps that can cause long convergence delays occur frequently.

Finally, we evaluate the effectiveness of a simple modification to route flap damping called *selective flap damping*. It eliminates withdrawal triggered suppression in all the topologies we studied (Section 7). The key new idea is to ignore monotonic route changes (as is typical in path explorations after failure) as flap damping triggers. Section 8 describes related work, and Section 9 concludes with some directions for future work.

2. BACKGROUND

Route flap damping, which we abbreviate as *RFD*, was designed and deployed on the Internet in the mid 1990s, primarily in response to frequent route flapping. This phenomenon, usually thought to be caused by router re-configuration or by links with intermittent connectivity, manifests itself as frequent BGP route changes. Each such route change causes route recomputation and increases the computation load on the route processor. At the time when RFD was deployed, route processors were significantly less powerful than they are today, and its deployment led to a significantly more stable routing system.

RFD has been shown to be effective in ameliorating the effects of routing instabilities other than those for which it was originally designed. One kind of routing instability is that resulting from the repeated tear-down and re-establishment of a BGP peering session (a peering session flap) that was a hallmark of some early BGP implementations. Peering session flap occurs when these BGP implementations receive BGP routing tables that exceeded the router's memory or receive an incorrectly formulated BGP update. Such flaps can result in frequent route changes for a large collection of routes. RFD can suppress these until the peering flap is resolved by operator intervention. Implementations have now been largely fixed to avoid peering session flaps, but route flap damping remains an important safeguard against future implementation errors that lead to large-scale repeated propagation of routing information. A second kind of routing instability that RFD can[1] suppress are persistent route oscillations caused by mutually conflicting routing policies [6]. These oscillations manifest themselves at a router as repeated route changes. RFD can significantly reduce the frequency of these oscillations.

Today, route flap damping is widely regarded as an important contributor to the overall stability of the Internet routing system by the operator community. To quote Geoff Huston [5]:

> ...coupled with widespread adoption of BGP route flap damping, has been very effective in reducing the short-term instability in the routing space.

In what follows, we describe the route flap damping mechanism in some detail. To do this, it helps to have a simple model of the way a BGP router processes routing information. We describe a simplification of the route processing model in the BGP RFC [11]. Each BGP router has several *peers* (neighbors) from each it receives *routes* to IP address prefixes over a transport connection. Conceptually, routes received from each peer are stored in a peer-specific database called the *Adj-RIB-In*. For a given prefix, the router's BGP decision process computes the most preferred route to the prefix from all the Adj-RIB-Ins and stores it in the *Loc-RIB*. The decision process then determines what subset of the *Loc-RIB* should be advertised to each peer. This subset is stored in a per-peer database called *Adj-RIB-Out* and advertised to the peer.

An important feature of BGP implementations is a hold-down timer on routes advertised to peers. This timer, called the `MinRouteAdvertisement` timer (or MRAI timer as defined in [12]) has a default value of 30 seconds. After a route to a prefix has just been advertised to a peer, subsequent changes to the route are held down until the MRAI timer expires (some vendors implement MRAI on a per-peer, rather than a per-route basis [7]). In doing so, the MRAI timer reduces routing instability during route convergence. As Labovitz *et al.* have shown, it also qualitatively affects the convergence process by limiting the exploration of alternate routes after route withdrawal.

While the MRAI timer was designed to reduce route changes during convergence, it clearly cannot suppress route instabilities caused by extraneous factors (such as unstable links) that cause flaps on larger time scales. Route flap damping was designed for this and works as follows. For each prefix P and for each peer or neighbor N, a BGP router maintains a penalty $p_{[P,N]}$. The penalty changes according to two simple rules:

- Whenever a peer N's route to prefix P changes (either the route transitions from being available to being unavailable, vice versa, or from one route to a better route, or vice versa), the router increments $p_{[P,N]}$. This increment is fixed, dependent on the type of the change.

- $p_{[P,N]}$ decays exponentially with time according to the equation

$$p_{[P,N]}(t') \;=\; p_{[P,N]}(t)e^{-\lambda(t'-t)} \qquad (1)$$

  where λ is a configurable parameter.

Intuitively, the penalty maintains an exponentially decaying instability history of a particular route from a particular peer.

When a router receives a route from N to prefix P, it first updates the penalty $p_{[P,N]}$ according to the rules described above. It

[1] In theory at least. The authors are unaware of actual observations of this kind of routing instability.

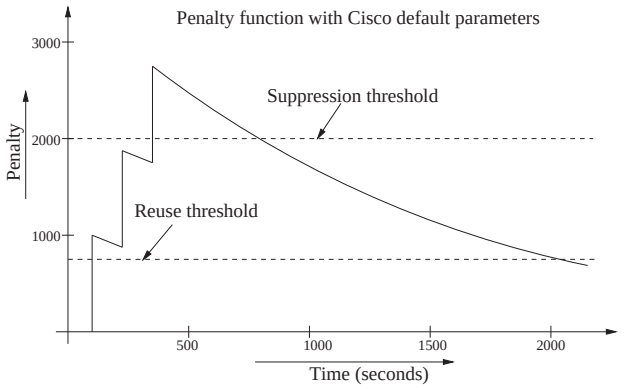


Figure 1: RFD penalty function with Cisco default parameters

then determines whether $p_{[P,N]}$ has crossed a configurable threshold, called the *suppression threshold*. If so, it marks the route as suppressed and inserts it into N's Adj-RIB-In. Suppressed routes are not used to compute the Loc-RIB. When it marks a route as suppressed, it also sets a timer for the time at which the current penalty would decay to below a *reuse threshold*. If the route's state changes before the reuse timer expires, the router cancels the reuse timer, recomputes the penalty, and starts a new reuse timer. When the reuse timer expires, the BGP decision process is invoked to compute the new best route to the prefix. Based on the default Cisco parameter setting (Table 1), Figure 1 pictorially depicts a route's penalty as a function of time and the times at which the route is suppressed and reused for a route that flaps three times with a 2 minute interval. In this case, the route flaps is suppressed for more than 28 minutes.

A typical implementation of route flap damping supports several parameters, all of which are in principle configurable:

- A value of λ, usually expressed using a *half-life* parameter H – the time for the penalty to decay to half its value.[2]
- A suppression threshold, which is the value of the penalty above which the route is suppressed.
- A reuse threshold, which is the value below which the route is considered reusable.

In addition to the above, implementations also have a parameter that limits the duration a route is suppressed. This is achieved either using a configurable maximum penalty or a configurable maximum suppress time. Some implementations also support different penalty increments for route withdrawals, route readvertisements, and route attribute[3] changes.

Despite the richness of the parameter set, deployment experience has shown that connectivity problems can be hard to debug if different routers use different sets of RFD parameters [3]. Consider the case where a customer's upstream provider is multi-homed and the provider's backup path applies less aggressive damping than the primary path. In this case, when the customer's route flaps, traffic to the customer might flow in through the upstream provider's backup path which does not suppress the customer's route, even when the primary path is available.

[2]Using Equation 1, we can obtain λ from H using the equation $e^{-(\lambda H)} = 0.5$

[3]Recall that BGP routes carry several attributes, the AS path being one of them.

Table 1: Default route flap damping parameter settings

RFD parameter	Cisco	Juniper
Withdrawal penalty	1000	1000
Readvertisement penalty	0	1000
Attributes change penalty	500	500
Cutoff threshold	2000	3000
Half-life (min)	15	15
Reuse threshold	750	750
Max suppress time (min)	60	60

For this reason, the operator community has recommended a standard set of flap damping parameters [9]. Three salient features of this recommendation are worth pointing out. First, the recommendation calls for different parameter sets for different prefix lengths, a recommendation called "progressive" flap damping. The intuition behind this is simply that smaller prefix lengths should be less aggressively suppressed because they represent a larger address space. Second, to prevent route suppression of relatively stable routes, it specifies that route should not be dampened until at least the fourth flap. Third, the recommended parameters are fairly aggressive. Even the least aggressive parameter set, governing prefixes of length 20 and lower, has a minimum outage time of 10 minutes and a maximum of 30 minutes. Longer prefixes can be suppressed for up to an hour if they flap at least four times.

It is not clear to what extent the recommendations for the flap damping parameters are followed by operators. We note that different vendors have different default parameters (Table 1), and we suspect that most ISPs simply use these parameters.

3. WITHDRAWAL AND ANNOUNCEMENT TRIGGERED SUPPRESSION

Route flap damping was designed to limit the propagation of unstable routing information. In this section, we show by working through two simple topologies that route flap damping can actually suppress relatively stable information. In particular, a single announcement of a route or a single withdrawal of a route followed by an announcement can cause route penalties to accumulate beyond the suppression threshold, causing the route to be suppressed. We call the former *announcement triggered suppression*, and the latter *withdrawal triggered suppression*.

For simplicity of exposition, we assume the following BGP model: (a) Route selection is based shortest AS paths. In case of ties, the route starting with the lower router ID is chosen. (b) The MRAI timer is 30 seconds and only applies to route announcements not withdrawals as recommended by the BGP RFC [11]. (c) No sender-side loop detection (SSLD) is used.[4] (d) Message propagation and processing delay are both bounded and negligible relative to the MRAI value. (e) We show how route suppression can occur for both the Cisco and Juniper parameters in Table 1 even when following the RIPE recommendation [9] of not suppressing a route until at least four flaps are received. In Section 5 we explore how variations on this model impact withdrawal triggered suppression.

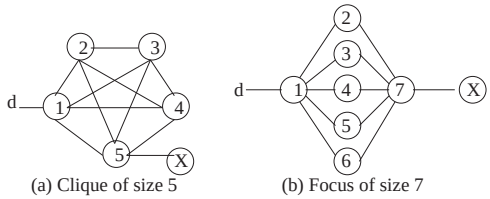
3.1 Withdrawal Triggered Suppression

To illustrate withdrawal triggered suppression, we use a clique of size 5, shown in Figure 2(a). The clique topology is a canonical topology that has been used to explain pathological route conver-

[4]At the time of this writing, at least one major router vendor does not yet implement SSLD. In Section 5, we also show that even with SSLD enabled, withdrawal triggered suppression can happen.

Table 2: Example of withdrawal triggered suppression in a 5-node clique

Stage	Time	Routing Tables	Messages Processed	Messages Queued in System
0	N/A	steady state 2(*1, 31, 41, 51) 3(21, *1, 41, 51) 4(21, 31, *1, 51) 5(21, 31, 41, *1)		steady state
1	N/A	1 withdraws the route 2(-, *31, 41, 51) 3(*21, -, 41, 51) 4(*21, 31, -, 51) 5(*21, 31, 41, -)	1→{2,3,4,5}W	2→{1,3,4,5} [231], 3→{1,2,4,5} [321], 4→{1,2,3,5} [421], 5→{**1,2,3,4,X**} **[521]**
2	N/A	announcement from 2 2(-, *31, 41, 51) 3(-, -, *41, 51) 4(231, *31, -, 51) 5(231, *31, 41, -)	2→{1,3,4,5} [231]	3→{1,2,4,5} [321], 4→{1,2,3,5} [421], 5→{1,2,3,4,X} [521]
3	N/A	announcement from 3 2(-, -, *41, 51) 3(-, -, *41, 51) 4(231, 321, -, *51) 5(231, 321, *41, -)	3→{1,2,4,5} [321]	4→{1,2,3,5} [421], 5→{1,2,3,4,X} [521]
4	N/A	announcement from 4 2(-, -, -, *51) 3(-, -, 421, *51) 4(231, 321, -, *51) 5(*231, 321, 421, -)	4→{1,2,3,5} [421]	5→{1,2,3,4,X} [521]
MRAI timer expires				
5	30	announcement from 5 2(-, -, -, -) 3(-, -, *421, 521) 4(*231, 321, -, 521) 5(*231, 321, 421, -)	5→{1,2,3,4,X} [521]	2→{1,3,4,5} W, 3→{1,2,4,5} [3421], 4→{1,2,3,5} [4231], 5→{**1,2,3,4,X**} **[5231]**
6	N/A	withdrawal from 2 2(-, -, -, -) 3(-, -, *421, 521) 4(-, *321, -, 521) 5(-, *321, 421, -)	2→{1,3,4,5} W	3→{1,2,4,5} [3421], 4→{1,2,3,5} [4231], 5→{1,2,3,4,X} [5231]
7	N/A	announcement from 3 2(-, -, -, -) 3(-, -, *421, 521) 4(-, -, -, *521) 5(-, 3421, *421, -)	3→{1,2,4,5} [3421]	4→{1,2,3,5} [4231], 5→{1,2,3,4,X} [5231]
8	N/A	announcement from 4 2(-, -, -, -) 3(-, -, -, *521) 4(-, -, -, *521) 5(-, *3421, 4231, -)	4→{1,2,3,5} [4231]	5→{1,2,3,4,X} [5231]
MRAI timer expires				
9	60	announcement from 5 2(-, -, -, -) 3(-, -, -, -) 4(-, -, -, *5231) 5(-, *3421, 4231, -)	5→{1,2,3,4,X} [5231]	3→{1,2,4,5} W, 4→{1,2,3,5} [45231], 5→{**1,2,3,4,X**} **[53421]**
10	N/A	withdrawal from 3 2(-, -, -, -) 3(-, -, -, -) 4(-, -, -, *5231) 5(-, -, *4231, -)	3→{1,2,4,5} W	4→{1,2,3,5} [45231], 5→{1,2,3,4,X} [53421]
11	N/A	announcement from 4 2(-, -, -, -) 3(-, -, -, -) 4(-, -, -, *5231) 5(-, -, -, -)	4→{1,2,3,5} [45231]	5→{1,2,3,4,X} [53421], 5→{**1,2,3,4,X**} **W**
12	N/A	announcement from 5 2(-, -, -, -) 3(-, -, -, -) 4(-, -, -, -) 5(-, -, -, -)	5→{1,2,3,4,X} [53421]	5→{1,2,3,4,X} W, 4→{1,2,3,5} W



(a) Clique of size 5 (b) Focus of size 7

Figure 2: 5-node clique and 7-node focus: node 1 announces route to d, route changes are observed at node X.

gence in BGP [7]. Note, we have verified the occurrence of withdrawal triggered suppression in a 4-node clique in a testbed for both Cisco and Juniper routers with default parameter settings [13]. In Section 5, we show that withdrawal triggered suppression is not unique to the clique, but the extent to which it occurs does depend on the topology.

Our example from Table 2 starts at the point after node 1 has announced a route to destination d, and all nodes have reached steady state. We now show if node 1 flaps *just twice*, by first withdrawing and then re-announcing the route to d, node X will suppress the route. Table 2 illustrates the convergence process corresponding to a single route withdrawal by node 1, following the notation in [7]. Each stage denotes the processing of a single set of messages from a node to all its peers. The "Routing Table" column shows the state of routing tables of nodes 2, 3, 4, and 5. The active route is denoted with an asterisk, and an invalid path with a dash. Thus, $4(231, *31, -, 51)$ means that node 4 currently uses route $[3\ 1]$ and has a backup route going through nodes 2 and 5. As an example, in stage 1 node 2 sends the route $[2\ 3\ 1]$ to its neighbors. When this message is processed in stage 2, node 3 realizes that this route goes through itself and so records the route from node 2 as invalid, and switches to the route from node 4.[5]

The "Message Processed" column shows the message processed at a given step, and the messages waiting to be processed are indicated in the last column. Messages from each peer are processed in the order they are received; messages from different peers can be processed in any order. We use $i \rightarrow \{j_1 \ldots j_n\}[path]$ to describe that node i sends to nodes $j_1 \ldots j_n$ a route of the ASpath, $path$. Withdrawal is indicated by W.

Consider the messages sent by node 5 to node X (indicated in Table 2 in bold font). Four messages are received by X (three announcements and one withdrawal), which account for four flaps. At X, the penalty value associated with the route to d is slightly less than 2500, depending on the precise message propagation delays. Using Cisco's setting, the penalty already exceeds the suppression threshold–2000, causing route suppression. For Juniper's setting, the subsequent announcement by node 1 accounts for another flap, causing the penalty to be close to 3500, also exceeding the suppression threshold–3000. And since X can only reach d through 5, its connectivity is affected because of route flap damping! In our example, it takes *at least 15 minutes* for the route to be restored.[6]

[5]The reader may wonder why this problem cannot be entirely avoided by simply invalidating all routes that contain a node i when node i sends a withdrawal. For instance, in stage 1, when node 2 receives a withdrawal from node 1, it seems intuitive to invalidate the routes $[3\ 1]$ and $[4\ 1]$ as well. Sadly, this is not possible in general because policies may require invalidating direct routes without invalidating indirect routes. This is the basis of a recent proposal [14], but it does not eliminate such path explorations due to withdrawals caused by policy changes.

[6]The penalty value is above 2000 and it has to decay to 750 before the route can be re-used. This requires that that penalty be halved at least once. Since the half life time is 15 minutes, the route is suppressed for at least 15 minutes.

Note that the batching effect of the MRAI timer improves the convergence time in this example by preventing extra updates. For example, when node 3 gets the announcement from node 2 in stage 2, node 3 switches to [4 1] but cannot announce it till the timer expires. But before this happens, node 3 changes its route again to [5 1] in stage 4.

This example illustrates an interaction, which has not been previously well studied, between two BGP mechanisms: the route withdrawal process that has been shown [7] to involve path exploration of successively increasing lengths (in cliques with no policy) and the mechanism to ensure the stability of the overall infrastructure. The rest of the paper is devoted to analyzing this interaction in detail for various topologies and BGP configuration settings and to evaluating a possible solution.

3.2 Announcement Triggered Suppression

A companion phenomenon is announcement triggered suppression. We show that in some topologies, a *single* route announcement can result in the route being suppressed at some node in the topology.

Consider the so-called *focus* [12] topology of size 7, shown in Figure 2(b). We use the same set of assumptions as in the clique case, except that instead of withdrawing the route, node 1 announces a new route to all its peers. In this case, node 7 has five routes to d of ASpath length 2. Suppose 7 prefers routes going through larger router IDs. Suppose also that the route announcements to node 7 arrive in the following order: [2 1], [3 1], [4 1], [5 1], [6 1], separated by time intervals at least as large as the MRAI value. This means that node 7 will also announce to X these five routes in the order they are received, because the succeeding route is always preferred over the preceding one. By a similar argument to the above, when node X receives five announcements in sequence, it suppresses the route to d.

In this paper, we do not explore announcement triggered suppression further. Its very occurrence depends on topology and very precise timing of update propagation. We believe it is unlikely to occur *frequently* in practice.[7] Withdrawal triggered suppression, on the other hand, depends less on precise timing, and therefore is more likely to occur. Thus, we explore the latter phenomenon exhaustively in this paper.

4. A SIMPLE ANALYTICAL MODEL

In this section, we explore route flap damping in an n-node clique (Figure 2(a)) using a simple analytical model. Our goal is to predict the minimum clique size for which withdrawal triggered suppression can be consistently observed.[8] We analytically evaluate the route penalty in the clique as a function of time, $p(t)$.

Suppose that $p(0) = 0$, and that the route penalty increment is 1. We assume a simplified BGP model in which each node processes messages in lock-step order. That is, at each time step, every node processes all the routes received from all its neighbors in the previous step, selects its best route, and re-advertises that route to all its neighbors. This model approximates BGP processing where each time tick corresponds to one MRAI time interval. Labovitz *et al.* showed that in this model, at least $(n-1)$ steps are needed for the clique, before the route is withdrawn [7].

Consider a node X attached to some clique node i. We compute the penalty $p(t)$ for route d announced by i to X. Now, by our model above, at each time tick node i in the clique picks a new route and advertises it. Thus, at each time tick, node X's penalty progressively increases. To compute the penalty function, we can use simple induction. Clearly $p(1) = 1$; at $t = 1$, node X receives a new route from i and increments its penalty by 1. Then, $p(2) = e^{-\lambda} + 1$; in one unit of time, the previous penalty has decayed to $e^{-\lambda}$, and at $t = 2$ node X receives a single route. By the same logic, $p(3) = e^{-\lambda}(e^{-\lambda} + 1) + 1$, or, simplifying the expression, $p(3) = e^{-2\lambda} + e^{-\lambda} + 1$. This suggests that the general form of $p(t)$ is a geometric series:

$$p(t) = \sum_{j=1}^{t} e^{-\lambda(j-1)} \tag{2}$$

and a closed form for this is

$$p(t) = \frac{1 - e^{-\lambda t}}{1 - e^{-\lambda}} \tag{3}$$

For what value of t does $p(t)$ exceed the suppression threshold? Suppose we assume that the suppression threshold is 4, and at least 4 flaps are needed to suppress the route. Also, suppose that the half-life time H is 15 minutes (Table 1) and the MRAI timer is 30 seconds. Recall that in our model, one tick of time corresponds to one round of the MRAI timer; in those terms, H is 30 time ticks in our model. Now, recall that λ is the solution to the equation $e^{-\lambda H} = 0.5$; thus $\lambda = \ln(2)/H$. With our choice of parameters, then $\lambda = 0.0231$. Solving numerically, we find that the smallest value of t for which the inequalities $p(t) > (4-1)$ and $t >= (4-1)$ hold is $t = 4$. Note, we subtract 1 from the suppression threshold and maximum flap count, since the withdrawal at the end of path exploration also accounts for the additional flap with penalty of 1.[9] We also know that after the $(n-1)$'th MRAI round, each node receives the longest path in the clique, which will cause it (at the next computation step) to withdraw that route [7]. Thus, to explore four MRAI rounds, we need a clique of size at least $five$. Hence, the smallest clique in which withdrawal triggers suppression is a clique of size five.

5. SIMULATION

While our analytic results give us some intuition for the interaction between route flap damping and BGP convergence, they cannot reveal the subtle variations that may arise from differences in BGP features (such as sender-side loop detection), topology effects, or from variations in message propagation latency. Simulation gives us more insight into the conditions under which withdrawal triggers suppression. In this section, we discuss results obtained using the SSFNet simulator [10], a Java-based simulation package with a built-in BGP simulator. The SSFNet BGP implementation is compliant with the BGP-4 specification in RFC 1771 [11]. We implemented route flap damping in SSFNet in compliance with RFC 2439 [3].

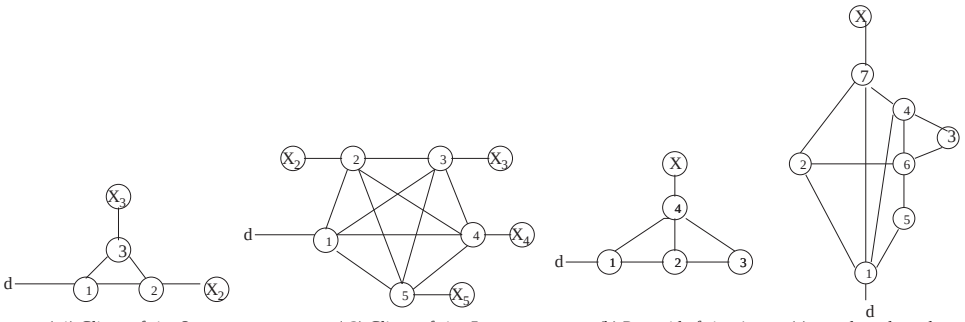
5.1 Simulation Methodology and Assumptions

Our simulations explore a number of scenarios with different topologies (Section 5.2). For tractability, we study withdrawal triggered suppression for a single prefix. In all our topologies (Figure 3), the origin d for this prefix is connected to node 1, and we

[7] We validated our conjecture that announcement triggered suppression is less frequent by studying BGP update traces (Section 6).

[8] It turns out that message reordering can increase the number of messages exchanged and increase the likelihood of route suppression.

[9] Here we assume that between the last update and the final withdrawal, the penalty has not yet decreased significantly. In practice, the re-announcement of the route by node 1 after the path exploration, *i.e.*, an additional flap, will usually keep the penalty value above the suppression threshold.



(a1) Clique of size 3 (a2) Clique of size 5 (b) Pyramid of size 4 (c) sample real topology

Figure 3: Sample topologies used in simulations

study convergence of the route to d at another node X. In our experiments, d and X are connected by a single link to the rest of the topology. For ease of exposition, we assume that node d is always connected to node 1 in the topology. This simplification allows us to isolate the effect of the particular topology under study on the convergence times at X for routes to d.

Our simulation scenarios ignore route filtering due to policy. Certainly, for a given topology, route filtering can determine whether or not route flap damping is invoked by withdrawal path explorations. Labovitz *et al.* have already shown that there exist realistic policy and topology configurations in the Internet that exhibit delayed convergence [8]. We believe that, in these topologies as well, withdrawal triggered suppression can occur.

Our simulation scenarios treat individual nodes as routers. Withdrawal triggered suppression can occur among routers connected to an exchange point. More generally, it can also occur across multiple autonomous systems. In this setting, our simulations are admittedly unrealistic because they do not capture the internal topologies of ASes. However, we believe our conclusions will not be qualitatively affected by this simplification, since route flap damping is not invoked on I-BGP peering sessions. This precautionary measure prevents inconsistent routing and forwarding loops within an AS [15].

Unless otherwise specified, we study the following route change pattern in all our simulation scenarios. Node 1 announces a route to d at some time to all its neighbors. All nodes in the topology have converged to a route to d by some time t. At time t, node 1 detects a failure of the link to d and withdraws its route to d.[10] Then at time $t + \alpha$, node 1 re-announces the route to d to all its neighbors, because the transient failure has been repaired.

The choice of α affects whether withdrawal triggered suppression happens or not. If α is large enough, of course, the route penalties accumulated at the nodes as a result of the route withdrawal will have decayed below the reuse threshold. As a result, when it is re-announced, all nodes will converge relatively quickly to their route to d. Clearly, the largest value of α for which this happens depends on the topology and flap damping parameter setting. We have verified these qualitative observations for a clique topology of size 5 and for the base parameter set (described in the next section). We found that when α is greater than 1600 seconds, withdrawal triggered suppression does not occur in that topology. If α is smaller than the MRAI value, the withdrawal followed by the re-announcement will be aggregated by the MRAI timer, and withdrawal triggered suppression will not be invoked. We have also verified this in our simulator. In our simulations, we set α to 500 seconds; this is large enough for all topologies in our study to have converged after the withdrawal at time t.

In all our simulations, the link delay is set to be 0.01 seconds. Since only a single destination prefix is simulated, router workload variation is simulated using variable delay in processing updates. This delay varies uniformly from 0.01 to 1 second. In addition to this source of randomness, jitter is applied to MRAI, as suggested by RFC 1771 [11]. Each data point in our simulation results is obtained by averaging a number of simulation trials.

5.2 Simulation Scenarios and Metrics

The occurrence of withdrawal triggered suppression depends on topology as well as parameter settings for various BGP mechanisms. This section describes the topologies and parameter settings explored in this paper.

We use the topologies shown in Figure 3 in our simulations. Our goal is not to enumerate all the topologies for which route flap damping can exacerbate convergence. Rather, we study this effect for very different topologies to see if there is any qualitative difference in the interaction between RFD and convergence. We also include one real topology fragment studied in the literature [8] to demonstrate that the effect can be observed in practice.

Our topologies include (Figure 3):

- An n-node clique. The clique has been used in the literature as a canonical topology to understand withdrawal path explorations. Furthermore, cliques are not completely unrealistic topologies. Full mesh BGP peering at exchange points does occur. Whether the routing policies at these exchanges cause these path explorations is not clear.
- An n-node pyramid. This consists of $n-1$ nodes, numbered 1 through $n-1$ connected in a chain. Node n is directly connected to each one of the other nodes. The pyramid is a contrived topology. But, we chose the pyramid because it is a qualitatively different topology from the clique. The clique is highly symmetric in that every node is connected to every other node. The pyramid is highly asymmetric, with only n being connected to every other node, and all other nodes having relatively sparse connectivity. Moreover, the pyramid is a topology where we might *expect* withdrawal triggered suppression: node n has $n-1$ alternate paths of different lengths

[10] This is a simplification. The exact mechanism by which this failure is detected depends on protocol details. For example, if node d and 1 are external-BGP peers, this detection might happen because the BGP keepalive timer expires. If, instead, d is internal to node 1's AS, the failure may be detected by the failure to receive IGP Hellos from d.

to d, a property that has been shown to be at least one signature of topologies in which withdrawal path exploration can happen [7].

- A sample topology from a study done by Labovitz *et al.* [8]. This topology is a subgraph of the inter-AS topology that was actually observed in their experiments. We include this topology to show that withdrawal triggered suppression can occur in real topologies as well.

In addition to the topology, withdrawal triggered suppression depends on the parameter settings for route flap damping. It also depends on the configuration of two features in BGP implementations:

- Sender-side loop detection (SSLD): a BGP speaker avoids announcing routes to a peer if that peer would detect a loop in the route and discard it. SSLD has been shown to improve route convergence in many cases.
- Rate-limiting applied to withdrawals (WRATE): some implementations apply the MRAI timer to route withdrawals as well as updates, violating a recommendation of the specification.[11]

There is a third BGP implementation feature that can affect our findings. Some implementations set MRAI timers *per peer* instead of *per prefix*. This can reduce the likelihood of withdrawal triggered suppression by delaying announcement messages to peers. But, this in combination with WRATE can also further delay withdrawal messages, resulting in additional alternate paths explored, increasing the likelihood of triggering route suppression. We have left the study of this feature for future work since it required simulation of other prefixes in the system.

To understand whether and how these BGP features affect our findings, we explore the following sets of parameters:

Base case: This uses a "standard" set of parameters. MRAI timer of 30 seconds, no sender side loop detection, no withdrawal rate-limiting, no policies, and route flap damping are implemented at all nodes. This case uses the Cisco parameter set in the first column of Table 1, along with RIPE's recommendation of not suppressing until at least the fourth flap. The results using the Juniper parameter set are similar.

MRAI=5: This set is used to study the impact of MRAI on withdrawal triggered suppression. Here, we set MRAI to 5 secs, keeping all other parameters unchanged from the base case.

Less aggressive damping: We set the penalty increment for route attribute changes to be 250 (half the value in the base case, see Table 1), but keep other parameters unchanged. This penalizes route attribute changes less, and in this sense is less aggressive.

SSLD: In this set, we enable sender-side loop detection. All other parameters match the base case.

WRATE: In this set, we enable withdrawal rate-limiting, keeping all other parameters of the base case.

Damping disabled: Finally, we disable route flap damping in the base case. This parameter set is included for calibrating withdrawal triggered suppression.

[11] At the time of this writing, at least one major router vendor applies rate-limiting to withdrawals.

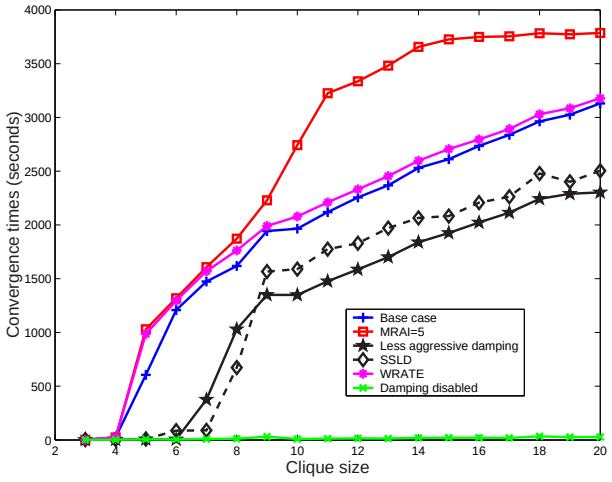


Figure 4: Convergence times of the clique topology

The primary metric for our simulations is *convergence time*. This is defined as the time between when the route to d is re-announced by node 1 till the time the node marked X sees a usable route to d. In each of the topologies depicted in Figure 3 except the clique, node X is always connected to the node n in an n-node topology. In the clique case, we connect a node X_i to each node i in the clique except node 1. We record the longest convergence time among all nodes X_i for each simulation run.

The secondary metric is the *total update count*. This is the number of update messages seen in the topology during the entire process including the initial route announcement, withdrawal, and final announcement by node 1. It helps us explain the convergence time behavior in some cases. One may argue that we should also consider instability as a metric, since RFD is aimed at reducing routing instability. However, in our experiments, we control the route changes originated at the source: only a single withdrawal followed by one announcement. We study the routing convergence behavior for such a relatively stable route.

5.3 Simulation Results

In this section, we examine the convergence time behavior of different topologies in some detail. This discussion also tells us how different parameters impact withdrawal triggered suppression.

5.3.1 Clique

Figure 4 plots the convergence time as a function of clique size, averaging 50 simulation runs. The most startling observation is that, with a *single withdrawal and announcement* from node 1, withdrawal triggered suppression can cause convergence times of up to 60 minutes (3600 seconds) for a large enough clique using our base parameter set. In the "damping disabled" case, by contrast, it takes less than 30 seconds between when the route is re-announced and when the route becomes available at each X_i connected to the clique.

Before we analyze Figure 4 in any detail, we discuss some subtle but important observations about route flap damping in the clique that are not easily learned without simulation.

Damping in Cliques: The first aspect of damping in cliques is *where* in the clique withdrawal triggered suppression is invoked. Recall that with route flap damping, suppression is per-peer. Each node in a clique is connected to every other node, but in the base

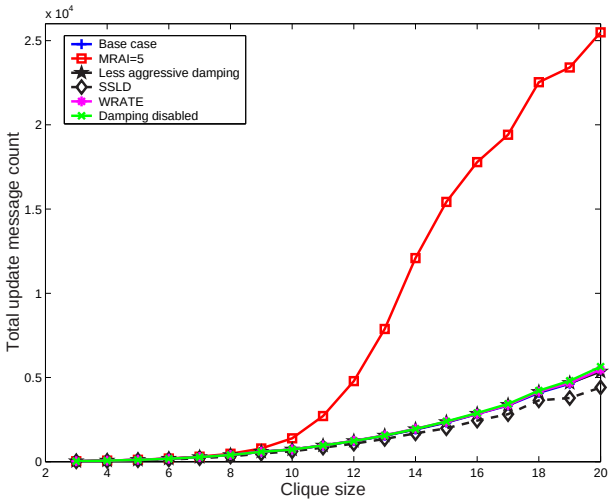


Figure 5: Total update count of the clique topology

case we find two interesting effects: (1) Some nodes do not suppress routes from any peer. (2) No node suppresses routes from all peers. In particular, since node 1 flaps only twice,[12] and all other nodes are connected to node 1, none of them suppresses node 1. Thus, when node 1 re-announces the route to d, all nodes in the clique have at least one usable route to d. But we also observed that it is not true that these nodes suppress all other neighbors either. This is a little surprising, because, from symmetry, one would have expected uniform behavior from all nodes except perhaps node 1. The reason is that in the base case, each node sends the same message to all its neighbors. However, each message is interpreted differently due to loop detection. Some updates are counted as withdrawals because the receiving node detects a loop in the AS-path. The second of two successive withdrawals is not counted as a flap. Therefore, the penalty values of different nodes accumulate differently with time. Furthermore, jitter added to the MRAI timer as well as router processing times can cause messages become reordered, resulting in different penalty values. This causes different nodes to advertise and receive routes at slightly different times. As a result, routes aggregate or "bunch" up differently. Sometimes a routing update from farther away reaches a node faster than a routing update from its neighbor.

Despite this, a node X_i that is connected to clique node i almost always (beyond cliques of a certain size) observes enough route changes that it suppresses routes from i. Thus, withdrawal triggered suppression does not manifest itself in the loss of connectivity to d from nodes in the clique, but only in nodes attached to the clique.

We also found that variable message processing and propagation delays can unexpectedly cause withdrawal triggered suppression in even a 3-node clique (Figure 3). This is in apparent contradiction to our results in Section 4, but only because our analytical model did not capture variations in message processing and propagation times. Assume that in the steady state, node X_2 has the route $[2\ 1]$ to d. When node 1 sends out a withdrawal, node X_2 first receives a withdrawal, then an alternate route $[2\ 3\ 1]$ from 2 before the final withdrawal is received. Thus, a single withdrawal results in three flaps. Now, when node 1 announces route to d again to node 2 and 3, due to variable message processing and propagation delay, node 2 sometimes announces route $[2\ 3\ 1]$ to node X_2 before announcing the preferred route $[2\ 1]$. Thus, a single announcement results in two more messages. Node X_2 thus receives a total of 5 messages from node 2, accumulating enough penalty to suppress the route from node 2.

Analysis of Results: Figure 4 plots the convergence time for each of our six scenarios as a function of clique size. We now discuss each scenario separately.

Base case: For the base case, withdrawal triggered suppression sets in with a five node clique, confirming our analysis of Section 4. This is not surprising, since four messages are required to exceed the threshold. In fact, we find from our simulations that flap damping is triggered at at least one of the X_i's in every simulation run of our five node clique. The convergence time increases monotonically as a function of clique size. The number of paths explored increases with clique size and therefore the accumulated penalty increases. As a result, for large enough cliques, convergence time increases until the maximum suppression time, which in our simulations is one hour (3600 seconds).[13]

MRAI=5: Figure 4 shows that compared to the base case, setting MRAI to be 5 seconds consistently increases the convergence times. Griffin and Premore have previously shown that reducing the MRAI timer value can result in many more routing updates [12]. Our simulations also confirm this (Figure 5). In turn, this can greatly increase the route flap penalty accumulated for each peer, and thereby the time to reuse the route. We also note that except for this scenario, the number of update messages exchanged is roughly equal for all other cases.

Less aggressive damping: Unlike decreasing the MRAI timer, this scenario exhibits a *later onset* of withdrawal triggered suppression and a lower convergence time. This scenario penalizes route attribute changes (*i.e.*, when a new route differs from the previous route only in the route attributes) by only half the regular penalty. This kind of change predominates during routing convergence. As a result, the penalty accumulates slower than in the base case. Because the thresholds are unchanged, the convergence times are lower corresponding to lower penalty values. Moreover, it takes a larger topology with more alternative routes to trigger route suppression.

SSLD: Sender-side loop detection (SSLD) consistently reduces convergence times compared to the base case. As with less aggressive damping, it also exhibits a later onset of damping. Intuitively, SSLD withdraws invalid alternate paths early and reduces the number of paths explored. This is confirmed by the update message plot (Figure 5), showing fewer number of updates. Fewer messages correspond to lower penalty values and thus faster convergence times.

WRATE: As suggested by Labovitz *et al.*, rate-limiting withdrawals can increase convergence times, since it delays the invalidation of invalid alternate paths [7]. More alternate paths are explored as a result, causing higher penalty values and thus longer convergence times. This is evident from our simulation results as well.

In summary, we observe two qualitative classes of behavior with respect to the BGP knobs we study in this section. One class is comparable to, or worse than, our base case. The second class exhibits lower convergence times and later onset of damping as a function of clique size. However, even in the second category, the

[12] Using Cisco's parameter set, node 1 only flaps once–the subsequent re-announcement after the withdrawal is not counted as a flap. Using Juniper's parameter set, it flaps twice.

[13] The convergence time can be a little higher than 3600 seconds, as shown in MRAI=5 case, since we measure the convergence time from when the announcement was sent. The route flap damping suppress timer is set some time after that.

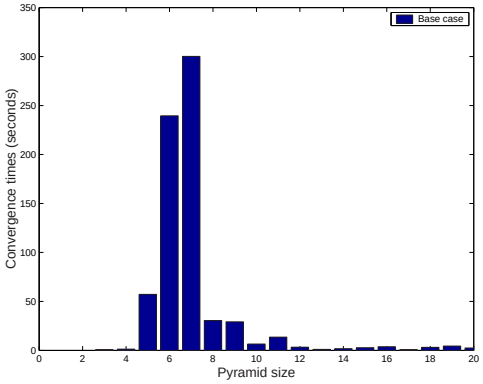


Figure 6: Convergence times of the pyramid topology (base case)

convergence times are much higher compared to the "damping disabled" case. For a clique of size 10, convergence times are more than 33 minutes. Thus, none of the BGP knobs eliminate withdrawal triggered suppression.

5.3.2 Pyramid

Having examined the clique, we now turn our attention to the pyramid. Recall that we chose to experiment with the pyramid because it was qualitatively different from the clique. Indeed the pyramid reveals significantly different behavior from the clique for many of our scenarios.

Figure 6 shows the convergence times for the base case scenario of the pyramid. These times were obtained by averaging 300 simulation runs for different sizes of pyramids. With increasing topology size, the convergence time increases and, beyond a pyramid of size seven, drops dramatically. In fact, beyond a pyramid of twelve nodes, we see almost no evidence of withdrawal triggered suppression. This is very counter-intuitive. We had assumed that since this kind of suppression was caused by BGP's exploration of different path lengths, it would be more prevalent in topologies with larger numbers of alternate paths of different lengths. In a pyramid of size n, node n has $n-1$ alternative paths of lengths from 2 to n. Thus, we expected to see monotonically increasing convergence times with the pyramid, as we did with the clique.

Non-Monotonicity in Convergence Times Explained: To understand this, consider the base case for an n-node pyramid. We evaluate the conditions that must hold for the *minimal* set of route changes to trigger flap damping at node X. We then show that this minimal set of route changes becomes increasingly unlikely due to increased message processing load on node n as the size of the pyramid increases. Note, there is one major difference between the pyramid and the clique. Although both have a large number of alternate paths of different lengths from node n to 1, all these paths in the pyramid are dependent, *i.e.*, they share common hops.

According to our parameters, to suppress a route to d, X must receive at least four route changes from node n. If we assume that the re-announcement of the route to d does not itself cause secondary flaps[14], the minimal set of routes needed to trigger a route change is as follows. In response to the withdrawal of the route to d, node n picks two alternate routes to d before withdrawing. These account for three flaps. The re-announcement of the route causes the fourth flap. Thus, the key to our explanation is understanding the circumstances under which node n *twice* announces an alternate route in response to a route withdrawal.

In steady state, all nodes i $(3 \ldots n-1)$ choose the shortest path by going through n: $[i\ n\ 1]$.[15] Now suppose node 1 sends a withdrawal to its neighbors 2 and n. When node n first receives the route withdrawal, it picks the next shortest route $r_n = [n\ 2\ 1]$ and announces it to X. This accounts for the first flap. Assuming comparable route propagation delays to node 2, at roughly the same time, node 2 picks its next shortest path $r_2 = [2\ n\ 1]$. Clearly, node n's choice and node 2's choice are mutually incompatible, so node n will never pick node 2's route. So, if node n has to pick a second alternate route (to account for the second flap), node 3 must choose route $r_3 = [3\ 2\ 1]$, because all other alternate routes go though this route. We discovered that whether node 3 chooses route r_3 is highly dependent on both the message processing delay and the message arrival order of r_n and r_2. Recall that these two routes are sent out roughly simultaneously in response to the withdrawal sent out by node 1. Note, normally the message processing order does not matter as MRAI imposes an order by preventing messages being sent out before timer expires. However, in this case, 3 has not sent out any message within the last MRAI time period and can send out an update right away in response to any route change.

The necessary and sufficient condition for node 3 to choose r_3 is that it receives r_n and announces its own choice of r_3 to node n, before receiving r_2 from node 2, and 3 does not announce another route to n before n's MRAI timer expires. We sketch a simple argument for this statement here. It is easy to see that the condition is sufficient: if that is the order of events, then n will select $[n\ 3\ 2\ 1]$ and that constitutes the second flap we have been looking for! This condition is also necessary, because if r_2 is received before node 3 processes r_n, then it can never pick $[3\ 2\ 1]$ and its only alternate route is through node n. In that case, node n will not incur a second flap to trigger flap damping at node X.

Note that it is not completely implausible for r_n to arrive at node 3 before r_2 does, since the path lengths are equal. Thus, whether r_n arrives before r_2 depends on the order in which they are sent out, and the message processing delay by nodes 2 and n. In addition, it also depends on the propagation delay (in our simulations, propagation delay is kept constant). Finally, it depends on whether node 3 processes and sends out r_3 before processing r_2. If it waits, the arrival of r_2 may invalidate r_3.[16] In our simulations, we add a randomly chosen jitter value between 0.01 to 1 seconds for processing each update message. This explains why for larger pyramids, withdrawal triggered suppression is less likely to occur. Larger sizes imply that node n is connected to more nodes, and it will take n much longer to process the announcement r_n to be sent to all other nodes. Therefore, the probability of r_n arriving before r_2 is significantly lower compared to smaller topologies. We have confirmed this explanation in our simulation results.

Examining Other Scenarios: Given our observations above, we now examine the impact of the various BGP knobs on withdrawal

[14] This is the common case in our simulations, as we rarely observe announcement triggered flaps for the pyramid.

[15] Actually, node 3 can pick either the direct path $[3\ 2\ 1]$ or the path $[3\ n\ 1]$, since they are each of the same length. Here we assume node 3 picks the latter. If it picks the former, n will never explore a second alternate route. That is because 3 will only announce a route change to n, either $[3\ 2\ n\ 1]$ or $[3\ n\ 2\ 1]$, which arrives before n can send out $[3\ 2\ 1]$. In our simulation, the tie-break rules were such that for our topologies, node 3 chooses $[3\ 2\ 1]$ over $[3\ n\ 1]$.

[16] Note, r_3 does not have to be physically sent out immediately, it can be placed in the waiting queue pending on the value of MRAI, as long as the arrival of r_2 does not cause the message to be deleted from the queue.

Table 3: 6-node pyramid convergence behavior

Parameter setting	Convergence time (second)	Update count	Damp count
Base case	239.57	93	53
MRAI=5	528.22	98	78
Less aggressive damping	195.18	92	35
SSLD	0.77	59	0
WRATE	238.51	94	34
Damping disabled	0.80	93	0

triggered suppression in a six-node pyramid (Table 3), averaging 200 simulation runs. The damp count column indicates the number of simulation runs in which withdrawal triggered suppression occurred. We notice two main differences in convergence times when compared to the behavior of the clique: (1) Sender-side loop detection completely eliminates convergence-based suppression in the 6-node pyramid! We verified that it actually does so for all other pyramid sizes for which suppression is invoked in the base case. (2) Unlike for the clique, withdrawal rate-limiting actually exhibits lower convergence time than the base case. We explain these differences below.

Table 4: 4-node pyramid convergence behavior with SSLD

Stage	Routing Tables	Msg Processed	Msg Queued
0	steady state 2(*1, 341, 41) 3(21, -, *41) 4(*1, 21, -)		steady state
1	1 withdraws route 2(-, 341, *41) 3(21, -, *41) 4(-, *21, -)	1→{2,4}W	4→{1,3}[421] 4→2 W, 2→4 W 2→{1,3}[241]
2	4's msgs 2(-, *341, -) 3(*21, -, 421) 4(-, *21, -)	4→{1,3}[421] 4→2 W	3→4[321] 3→2 W
3	2's msgs 2(-, *341, -) 3(*241, -, 421) 4(-, -, -)	2→4 W 2→{1,3}[241]	4→{1,2,3}W
4	3's msgs 2(-, -, -) 3(*241, -, 421) 4(-, -, 321) ...	3→4[321] 3→2 W ...	2→{1,3,4}W ...

SSLD: SSLD is very effective for the pyramid, because it invalidates all alternate routes within a single round of the MRAI timer. We show such an example for a 4-node pyramid in Table 4. When node 1 withdraws the route to d, node 2 picks the alternate route $[2\ n\ 1]$, but does not propagate it to n because it notices a loop. Similarly, n picks $[n\ 2\ 1]$ and does not propagate this route to 2. Instead, both node n and node 2 send withdrawals to each other (in this scenario, withdrawal rate-limiting is *not* in effect), but announce their choices to their other neighbors. When n receives node 2's withdrawal, however, n withdraws the route $[n\ 2\ 1]$ from all of its neighbors (stage 3 in Table 4). Similarly, node 2 withdraws from its neighbor 3 (stage 4). As a result, node 3 will withdraw the route from n after stage 4, so node X never sees enough flaps to exceed the suppression threshold.

WRATE: Table 3 shows that, unlike for the clique, the *WRATE* scenario can actually exhibit a lower convergence time. This is because when withdrawals are delayed by the MRAI timer, there are some cases where node n sees fewer secondary flaps compared to the base case. These cases depend on a particular sequence of route propagation. Please refer to [13] for an example of one such sequence. Intuitively, since the number of alternate routes going through n is much greater than ones that do not, withdrawal rate-limiting increases the probability of exploring the former routes.

Table 5: Convergence times of the sample real topology (Figure 3(c)) averaging 50 simulation runs

Parameter setting	Convergence time (second)	Update count	Damp count
Base case	243.45	132	11
MRAI=5	558.18	137	26
Less aggressive damping	1.73	132	0
SSLD	2.03	94	0
WRATE	410.34	135	18
Damping disabled	1.73	132	0

5.3.3 A Sample Topology

We take a sample real topology from the study done by Labovitz *et al.* [8] to test whether withdrawal triggered suppression can happen in real topologies. Table 5 shows the results, each data point denoting the average of 50 simulation runs. The damp count column indicates the number of simulation runs in which withdrawal triggered suppression occurred. Note that the impact of the various BGP knobs is consistent with our observations for the clique topology: setting the MRAI timer to a smaller value increases the number of messages and convergence times, and withdrawal rate-limiting worsens the convergence times and increases the number messages. What is interesting is that for this topology, SSLD and less aggressive damping both eliminate withdrawal triggered suppression. We found that with SSLD enabled, the number of MRAI rounds is reduced to one and thus reduces the likelihood of triggering route suppression. Note, SSLD cannot eliminate the possibility of withdrawal triggered suppression, because the route re-announcement may cause additional flaps.

5.4 Summary

In summary, our extensive simulations reveal several important observations about withdrawal triggered suppression: In many topologies, including at least one real topology fragment, BGP path explorations following withdrawal can trigger route flap damping after just a single withdrawal followed by a route re-announcement. In such cases, the route is sometimes suppressed for up to an hour. Even in topologies with a large number of alternate paths of different lengths, such as the pyramid, it is not always true that withdrawal triggered suppression is more likely to be invoked than in smaller topologies. No proposed or deployed BGP implementation features eliminate this phenomenon for all topologies. For certain topologies, *e.g.*, pyramid, sender-side loop detection can eliminate withdrawal triggered suppression.

6. TRACE ANALYSIS

We have already shown that withdrawal triggered suppression can happen in practice, by taking a realistic topology fragment from [8] and from our experiments of Cisco and Juniper routers in a 4-node clique topology [13]. How *prevalent* is withdrawal triggered suppression? This is a difficult question to answer with certainty. Instead, we get a handle on this question by performing a simple analysis of BGP update traces to determine how often we can observe an important signature of delayed convergence—successive announcements of strictly increasing path lengths. Each such sequence of length greater than four can *potentially* trigger

suppression at a damping-enabled router. For our traces analysis, we use publicly available routing update data from RIPE NCC [16] and the University of Oregon Route Views project [17].

Table 6: Withdrawal triggered flap statistics

	RIPE00 01/10/2002	Oregon RV 11/15/2001
Total instances	8533	6828
Max num announcements per instance	8	7
Total unique peers	13	20
Total unique prefixes	2768	3040
Max prefix length	30	26
Min prefix length	8	8

Our trace analysis simply counts instances of routing message sequences with strictly increasing path lengths followed by a withdrawal, ignoring path length increases caused by AS path prepending. We only recorded sequences of length four or greater, since at least four flaps are required to trigger flap damping. Table 6 shows the results of our analysis on a particular day from both data sources. We find several thousand instances of such routing message sequences in our traces. Notice also that these sequences are not restricted to a particular peer, nor from a particular prefix, and they span a wide variety of prefix lengths. This indicates that the phenomenon we describe in this paper may actually occur relatively frequently, and is therefore of considerable practical importance. As we conjectured earlier, we rarely observed update sequences indicative of announcement-triggered suppression, *i.e.*, routes of decreasing path lengths.

7. SELECTIVE ROUTE FLAP DAMPING

In this section, we consider a simple solution for both withdrawal and announcement triggered suppression. We should emphasize that our goal here is to demonstrate the existence of a relatively simple mechanism that will reduce or eliminate the occurrence of triggering route suppression during convergence. Much more evaluation and experimentation is necessary to understand the efficacy of the scheme under various topologies, as well as its incremental deployability. That is the subject of future work.

The key to our mechanism is to detect route changes due to path exploration to avoid increasing penalties. From the clique example in Section 3, one might conclude that one way to detect route changes due to path exploration is to avoid penalizing successive routes with non-decreasing path lengths. Thus, if a new route has the same or longer path length than the existing route, we do not increment the flap penalty.

While this works for the simple example we discussed above, it does not work well in general. In particular, policies at various nodes in the clique can, in theory, cause longer path lengths to be explored first than shorter ones (if they happen to be more preferred). So, a more general observation might be that each node, during convergence after withdrawal, selects routes in order of non-increasing preference until it finally withdraws the route. Thus, if the sender of a route includes its current preference for the route (a feature that BGP currently lacks for external peers), the receiver of the route can compare the sender's preference for the received route with that of the previous route from the sender. The preference value can be encoded in a specialized community attribute that is nontransitive, making our proposal incrementally deployable. The receiver can then increment the penalty for the route if the new route does not have a higher preference (at the sender) compared to the previous route.

This simple mechanism does not work perfectly. The sequence of route changes seen from a peer during withdrawal convergence can have route withdrawals interspersed with routing updates.[17] Furthermore, in some topologies such as the pyramid, this can happen even without SSLD (see [13] for an example). Thus, our mechanism has to deal with this situation as well.

Our proposed mechanism is a modification to route flap damping that we call *selective route flap damping*. It requires the sender to attach to each route announcement its local preference or the relative preference value compared to the previous route announcement. We keep two bits for each destination route from each peer. These two bits encode the comparative value of the last two announcements received. We call these two bits the *comparison bits*. 00 denotes the situation where fewer than two routes have been received. 01 denotes that the values of the two routes are the same. 10 means the latest route has higher degrees of preference than its previous route. And finally, 11 indicates the latest route is less preferred. When an announcement is received, comparison bits are recomputed based on the current announcement and the latest announcement. The newly computed comparison bits are compared with the stored comparison bits. If these two sets of comparison bits indicate that the direction of route preference change has altered, then we count the current announcement as a flap. In other words, if one set of comparison bits is 10 and the other is 11, we consider the announcement received as a flap. This heuristic is used, because secondary flaps are always of either increasing or decreasing degrees of preference.

To deal with interleaved withdrawals, selective damping temporarily ignores withdrawal messages until the next announcement is received. We keep track of the *temporary* penalty corresponding to the withdrawal message and let it decay exponentially just like the regular penalty value. This temporary penalty would have been added to the penalty in the existing scheme. If the next announcement received is considered a flap, this temporary penalty is added to the penalty value in addition to the penalty corresponding to the current flap. Otherwise, the temporary penalty is discarded. Here we add another condition under which the current route is considered a flap. If the route received has the same preference value as the previous one, we do not simply discard it as a redundant update, because the announcements could be interleaved by withdrawals. Thus, we count the current announcement as a flap if it has the same value as the previous announcement *and* is preceded by a withdrawal. The goal of this slight modification is to make sure the new scheme can contain real flaps.

Selective damping is thus designed to ignore route changes caused by withdrawal exploration, yet to mimic unmodified route flap damping. It does so, but with one caveat. Because of the way it deals with withdrawals, it penalizes true route flaps to the same extent that unmodified route flap damping would, but it might do so slightly later (because it has to wait for the announcement following the withdrawal to penalize the route). At most one extra withdrawal message is propagated under the new scheme.

Finally, selective flap damping also eliminates announcement triggered suppression, which consists of successive announcements of increasing degrees of preference. Since our scheme does not count successive monotonic route changes as flaps, both forms of suppression are eliminated.

We have validated through simulation that selective flap damping actually eliminates withdrawal triggered suppression. As shown in

[17] This can be caused by sender-side loop detection, policies, or update reordering.

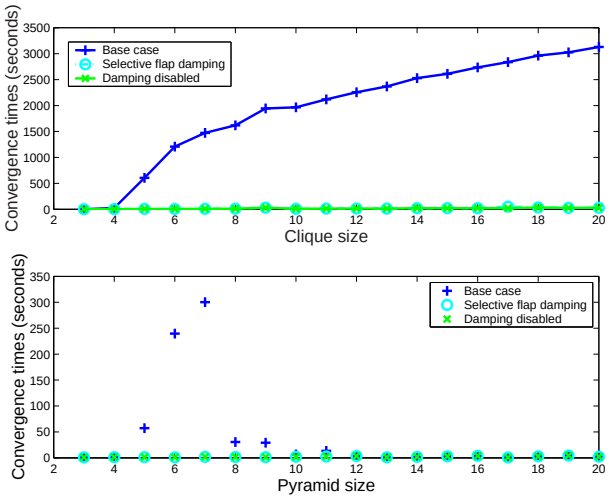


Figure 7: Convergence times of the clique and pyramid topology (averaging 50 simulation runs)

Figure 7, selective flap damping exhibits convergence times comparable to the situation when damping is disabled both for the clique and the pyramid topologies. In addition, we also verified this for our realistic topology, where selective flap damping exhibits the same convergence time and number of messages as the case when flap damping is disabled.

Furthermore, we verified that selective damping can suppress actual flaps. To do this, we simulated network failures by making node 1 in each of our topologies repeatedly flap (*i.e.*, alternately withdraw and announce the route to d) with a period of 40 seconds.[18] We then observed the number of additional messages it takes for selective damping to suppress the route compared to the unmodified route flap damping implementation. Our simulation shows it takes at most 8 additional messages for selective damping to suppress a continuously flapping route compared to the original RFD scheme. A scheme that does not use any form of damping will instead send an update every 40 seconds. For each topology size, the actual number of additional messages differs. For instance, for a clique of size 5, it takes on average 3 extra messages. For a clique of size 20, it takes on average 6 extra messages.

8. RELATED WORK

This paper has investigated the interaction between route flap damping and BGP convergence.

Route flap damping has received very little examination in the research literature. In the standards world, there are two documents most often referenced in connection with route flap damping. The route flap damping standard [3] describes the rationale for route flap damping and outlines a possible implementation strategy for the mechanism. While that document discusses some interactions between flap damping and topology, it does not discuss announcement or withdrawal triggered suppression. An associated document, the RIPE recommendations [9] tantalizingly hints that one or both of these phenomena may have been observed in practice. To quote

> ...The only explanation would be that the multiple interconnections between Ebone/AS1755 and ICM/AS1800 did multiply the flaps (advertisements/withdrawals arrived time-shifted at ICM routers through the multiple circuits). ....This would then potentially hold true for any meshed topology because of the propagation delays of advertisements/withdrawals.

However, it then proposes a solution that we do not believe addresses the problem, nor does it analyze the phenomenon in any level of detail.

Also related to route flap damping is a technique for damping link state changes. Rodeheffer *et al.* [18] proposed a filter, called a skeptic, that penalizes unstable link state information for a time that increases logarithmically with the number of flaps of the link state. The details of the algorithm are different from route flap damping, and it would be interesting to compare how the two perform on various kinds of flaps.

In the academic community, there have been two threads of prior research into the following properties of BGP: stability and convergence delays.

Stability: The first thread started with the observation that there existed certain policy configurations which could cause persistent route oscillations in BGP [19]. Later, Griffin and Wilfong [6] showed the intractability of determining a safe policy configuration for BGP. Finally, Rexford and Gao [20] proved that if BGP's policy expressiveness is confined to a simple set of policies, persistent route oscillations cannot occur. Independently, Labovitz *et al.* [21] showed that instability could occur even without policy conflicts because of implementation artifacts. Thus this first thread confirmed the value of the route flap damping standard and probably influenced the RIPE recommendations.

Convergence Delays: The second thread of BGP research is a careful analysis of the dynamics of BGP's route convergence properties [7, 8] and resulted in the interesting finding that BGP's route withdrawal process could result in a combinatorially large number of path explorations.

Thus our paper can be considered to be a convergence between these two threads of research because it shows that the RFD mechanism used to improve stability can exacerbate convergence delays.

Other more recent prior work has explored and attempted to solve delayed Internet routing convergence. Griffin *et al.* [12] explored how convergence is affected by the MRAI timer setting and addressed its impact on various topologies. In their future work, they pointed out the potential for route flap damping to be invoked by oscillations inherent in the BGP protocol. In this work, we confirm their suspicion by thoroughly studying its interaction with convergence.

More directly related is the work of Pei *et al.* [14] who attempted to avoid path exploration during route withdrawal by using consistency assertions. They showed that their approach can invalidate all paths within one MRAI round in some cases. This is an intriguing approach that might work, although it needs extensive experimentation to be widely deployed and does not work in all cases (e.g., when policy is used for say traffic engineering). It should be clear from this paper that a fix for withdrawal path exploration in BGP will reduce the occurrence of the phenomenon we see in this paper. Despite this, the value of our paper is that it provides the first analysis of the interaction between RFD and convergence and suggests an alternate solution for this interaction that is useful, if a general solution to eliminate withdrawal path exploration turns out to be hard to design and deploy.

[18] The maximum frequency is limited by MRAI timer value.

9. CONCLUSION

In this paper, we analyze a previously not well-studied interaction between BGP's route withdrawal process and its route flap damping mechanism for ensuring the overall stability of the Internet routing system. This interaction can, depending upon the topology, suppress up to one hour the propagation of a route that has been withdrawn once and re-announced. We have shown that this interaction has a number of subtle features. For instance, we found that in the pyramid topology increasing the size of the topology actually improved the rate of convergence.

We have proposed a simple fix to this withdrawal triggered suppression called selective flap damping. It relies on being able to weed out secondary flaps using a monotonicity condition which selectively avoids penalizing such secondary flaps. Our selective flap damping mechanism successfully eliminates withdrawal triggered suppression in all the topologies that we have analyzed.

We leave for further work the problem of accurately characterizing the network topologies and sizes which will induce withdrawal triggered suppression. A theoretical analysis of the properties of selective flap damping would also be desirable. Despite this, our paper together with [7, 8] makes it clear that faster convergence does require modifying BGP. This could be done by either fixing the withdrawal path exploration phenomenon (the direction followed in [14]) or by deploying a mechanism similar in spirit to selective flap damping (as in our paper). Either way, such BGP modifications could move us closer to the Holy Grail: an inter-domain routing protocol that is stable and yet reroutes traffic extremely fast after failure.

10. ACKNOWLEDGEMENTS

We thank developers of SSFnet for making the simulator available. We especially thank BJ Premore for the BGP implementation in SSFnet and prompt responses to our questions. We are grateful to Tim Griffin and Srdjan Petrovic for the implementation of SOS (Scripts for Organizing Simulations). Tim Griffin, Scott Shenker, Wilson So, and Jia Wang gave us insightful comments. We are indebted to Sprint ATL, especially to Linda Chau, Peter Lam, Bryan Lyles, and J. J. Yea, for their router testbed, so we can verify withdrawal triggered suppression using commercial routers. Finally, we also thank the SIGCOMM 2002 anonymous referees for their feedback.

This work was supported by the California MICRO Program, with matching support from Ericsson, Nokia, Siemens, and Sprint.

11. REFERENCES

[1] Atul Khanna and John Zinky, "The Revised ARPANET Routing Metric," in *Proceedings of ACM SIGCOMM 1989*.

[2] H. Yu C. Alaettinoglu, V. Jacobson, "Towards Milli-Second IGP Convergence," IETF Internet Draft: draft-alaettinoglu-ISIS-convergence-00, November 2000.

[3] C. Villamizar, R. Chandra, and R. Govindan, "BGP Route Flap Damping," RFC 2439, 1998.

[4] "From private email exchanges with Randy Bush," .

[5] Geoff Huston, "Analyzing the Internet BGP Routing Table," *The Internet Protocol Journal*, March 2001.

[6] Timothy G. Griffin and Gordon Wilfong, "An Analysis of BGP Convergence Properties," in *Proceedings of ACM SIGCOMM 1999*.

[7] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, "Delayed Internet Routing Convergence," in *Proceedings of ACM SIGCOMM 2000*.

[8] C. Labovitz, R. Wattenhofer, S. Venkatachary, and A. Ahuja, "The Impact of Internet Policy and Topology on Delayed Routing Convergence," in *Proceedings of INFOCOM 2001*.

[9] Christian Panigl, Joachim Schmitz, Philip Smith, and Cristina Vistoli, "RIPE Routing-WG Recommendations for Coordinated Route-flap Damping Parameters," October 2001, Document ID: ripe-229.

[10] "The SSFnet Project," `http://www.ssfnet.org`.

[11] Y. Rekhter and T. Li, "A Border Gateway Protocol," RFC 1771 (BGP version 4), March 1995.

[12] Timothy G. Griffin and Brian J. Premore, "An Experimental Analysis of BGP Convergence Time," in *Proceedings of ICNP 2001*.

[13] Z. Morley Mao, Ramesh Govindan, George Varghese, and Randy Katz, "Route flap damping exacerbates internet routing convergence," Tech. Rep. UCB//CSD-02-1184, U.C. Berkeley, June 2002.

[14] D. Pei, X. Zhao, L. Wang, D. Massey, A. Mankin, S. F. Wu, , and L. Zhang, "Improving BGP Convergence Through Consistency Assertions," .

[15] *BGP4 Inter-Domain Routing in the Internet*, Addison-Wesley, 1999.

[16] Ripe NCC, "Routing Information Service Raw Data," .

[17] "University of Oregon Route Views Archive Project," `www.routeviews.org`.

[18] Thomas Rodeheffer and Michael D. Schroeder, "Automatic Reconfiguration in Autonet," in *Proceedings of 13th ACM Symposium on Operating System Principles*.

[19] K. Varadhan, R. Govindan, and D. Estrin, "Persistent Route Oscillations in Inter-Domain Routing," *Computer Networks*, March 2000.

[20] Lixin Gao and Jennifer Rexford, "Stable Internet Routing Without Global Coordination," in *Proceedings of ACM SIGMETRICS 2000*.

[21] C. Labovitz, R. Malan, and F. Jahanian, "Internet Routing Stability," in *Proceedings of ACM SIGCOMM 1997*.