Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: November 26, 2007

T. Li
Cisco Systems, Inc.

May 25, 2007

# BGP Stability Improvements
# draft-li-bgp-stability-00

## Status of this Memo



## Copyright Notice



## Abstract

BGP is the routing protocol used to tie the Autonomous Systems (ASes) of the Internet together. The ongoing stability of BGP in the face of arbitrary inputs, both malicious and unintentional, is of primary importance to the overall stability of the Internet. The overall issue is not a new one. Previously, one aspect of stability, known as route flap damping was originally discussed in RFC 2439. In the intervening years, a great deal of experience with flap damping and other stability concerns has been accumulated. Most recently, the

issue of BGP stability has been highlighted in RAWS. This document describes the experience that has been gained concerning stability in the intervening years, hypotheses about remaining problems, suggestions for experiments to be performed, and proposals for possible alternatives.

Table of Contents

## 1. Introduction

BGP [RFC4271] is the routing protocol used to tie the Autonomous Systems (ASes) of the Internet together. The ongoing stability of BGP in the face of arbitrary inputs, both malicious and unintentional, is of primary importance to the overall stability of the Internet. The overall issue is not a new one. Previously, one aspect of stability, known as route flap damping was originally discussed in RFC 2439 [RFC2439]. In the intervening years, a great deal of experience with flap damping and other stability concerns has been accumulated. Most recently, the issue of BGP stability has been highlighted in RAWS [I-D.iab-raws-report]. This document describes the experience that has been gained concerning stability in the intervening years, hypotheses about remaining problems, suggestions for experiments to be performed, and proposals for possible alternatives.

Please note that this document is very much a work-in-progress.

### 1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

### 1.2. History

The circuits used in computer networks have the unfortunate property that they can intermittently fail and then recover. This was an especially common failure mode for copper-based circuits. Under such circumstances, when there was a BGP speaker on both ends of the circuit, any prefixes advertised across the link would tend to oscillate at the frequency induced by the intermittent link. The oscillating prefixes would then propagate across the full Internet, causing the entire routing subsystem to churn at the rate of the prefix.

Individually, a single such prefix is not a significant issue. However, as the Internet continued to scale upwards, it became obvious that the CPU requirements to deal with the ever-increasing number of oscillating prefixes would quickly become onerous. This was aggravated by the fact that the party responsible for the flapping circuit was frequently unaware of the problem, or, worse yet, unwilling to address the issue.

Thus, the original goal of route flap damping was to protect the control plane from oscillations. This was done by determining the number of flaps and the time elapsed since the last transition. This

is fed into an exponential decay function, and, if the prefix is found to be flapping based on this data, the actual propagation of the route is suppressed. Since the frequency information must be stored even if the prefix is not currently active, there is state overhead associated with flap damping for each prefix that has been oscillating.

### 1.3. Observations

Unfortunately, flap damping isn't truly discerning about the nature of routing changes. Any routing change can easily be misinterpreted by flap damping as instability, resulting in premature damping of prefixes [Harmful].

#### 1.3.1. Path hunting

One source of path changes is BGP's normal mechanism for _path exploration_ or _path hunting_. These situations occur because BGP is a path-vector protocol, where each BGP speaker advertises the path that it is using to its neighbors, complete with the full AS path to the destination. Since the number of possible paths through even a simple topology is large, there can be many different path transitions that can possibly be advertised.

Path hunting can occur both when a prefix is first advertised or when a prefix is withdrawn. At advertisement time, the prefix may propagate through the topology at different rates, sometimes resulting in it first appearing at an AS with a suboptimal path. Over time, optimal paths will appear where suboptimal paths were before, resulting in a path change that is subsequently propagated.

Similarly, when a prefix is withdrawn from the network, each AS that receives the withdraw will select some other historical path and propagate it. If the historical path is subsequently withdrawn, the AS will again select another historical path. This will continue until the entire possible path space has been explored and eventually withdrawn.

Interestingly, the amount of path hunting can increase dramatically as the meshiness of the topology increases. It's easy to observe this if you first consider an acyclic topology (i.e., a tree). In such a topology, there is only one possible path, so there is no hunting. If a single link is added to this topology, then there is one cycle in the graph and at most two possible paths for BGP to explore. Subsequent links can add many more alternate paths, depending on their placement.

### 1.4. The wavefront model

An intuitive means of understanding the observed behavior is by analogy to a wavefront. Any change in the network triggers the dissemination of information (either updates or withdrawals) through the topology from the point of occurrence. The information travels outwards along all of the paths supported by BGP, in much the same way that a wave would propagate from a pebble dropped in a lake.

The wavefront expands at each BGP speaker, where the information is propagated to all other BGP peers, including ones that already have the information. If the newly arrived information is inferior to the existing path information, then the wave dies out at that BGP speaker. If the newly arrived path is the best path, then it continues to be the wavefront of the best information. It's easy to see from this that a single change in the network can thus generate multiple waves.

#### 1.4.1. Refraction

As can be seen from the above, information does not traverse the full BGP mesh at fixed rates. Differences in implementations, processing loads, propagation delay, damping parameters, and policy can all contribute to delaying optimal path information. Continuing the wavefront analogy, we know that waves propagate through different materials at different speeds. This phenomenon is known as _refraction_, and as seen above, can lead to the multiplication of wavefronts. Each additional wavefront represents additional processing burden on the routing subsystem.

It is interesting to note that flap damping itself may be a contributor to the creation of additional wavefronts. Since a route that is being damped will be delayed for a long time, damping is effectively delaying a wave of information, possibly creating more refractive effects.

## 2. Goals

### 2.1. Flap damping

As we reconsider the mechanisms that constitute flap damping, we need to keep in mind that the original goals of detecting and protecting the routing subsystem from noisy inputs is still a requirement. While copper circuits are now less common in the core, the overall network has expanded dramatically and there is a wide variance in the skills and experience in operational roles.

As a result, it is still possible for an errant AS to inject flapping information into the BGP mesh, either as the result of policy misconfiguration, implementation error, an intermittent circuit, or even as an intentional destructive act. Thus, it is important that there still be mechanisms that intervene and ameliorate these effects, protecting the routing subsystem.

### 2.2. Rapid convergence

While protecting the routing system is of paramount importance, it is also vital that the routing subsystem also continue to perform its primary task: providing routing. Any flap damping mechanism must continue to provide rapid convergence to some workable path so that connectivity is restored. However, this goal should not be construed to require rapid optimality. While a best path should eventually be selected and propagated, it is far more important that some connectivity be restored immediately. Most applications can survive with a sub-optimal path, while no applications can succeed if no path is selected.

### 2.3. Reduced overhead

Flap damping should also strive to deter the unnecessary exchange of information. As described above, both path hunting and refractive effects cause unnecessary churn in BGP. The flap damping mechanism should be generalized to help suppress as much of this unnecessary information as possible.

## 3. Hypotheses

In this section, we put forth a number of hypotheses about possible mechanisms to achieve the goals above. As of this writing, more investigation is needed on each of these theories, and where possible we've included some discussion of the experiments that we feel would be worthwhile. Our goal here is to examine a number of different mechanisms, understand their relative benefits, and select a small subset to become the core set of replacement mechanisms.

### 3.1. Turn it off

Given the concern about the refractive effects of path damping, [RIPE-378] recommends that path damping be disabled. While this is not unreasonable given the lack of beneficial alternatives, we feel that some of the possibilities presented here will eventually prevail and that this sentiment can be changed over time.

### 3.2. Alternate parameters

It has been suggested in [Harmful] that the default flap damping parameters in existing implementations are simply too aggressive and quickly convert normal path hunting into a damping event that precludes connectivity. Significantly increasing the parameters could permit significantly more churn to be passed by the routing subsystem while still filtering out truly periodic sources of flap.

It would be useful to test this by simply configuring numerous differing parameters and observing if there is any beneficial effect. At this time, we have no recommendations for possible alternative parameter settings.

### 3.3. Band pass filtering

Another view is that classical flap damping isn't working as well as we might like because it is measuring frequency. The current mechanism looks for a number of changes in a given period of time. If the route exceeds this threshold frequency, then it is damped. The threshold frequency is necessarily set fairly low so that it will damp true flapping circuits.

Unfortunately, path hunting creates a high frequency burst that incorrectly triggers damping. This acts as a false positive for damping that we would like to avoid. One alternative approach is to shift from looking for flapping above a given frequency and simply accept that when there is a real topological change, there will be extensive high frequency path changes. After some time, those path changes should stop and the route should then resume its stability. Subsequent path changes would then be indicative of real oscillation and would be subject to damping.

The implementation of this would be relatively straightforward. When a change is seen on a stable route, it opens an oscillation window of a fixed duration (e.g., 60s). Any changes within that window are not considered as contributing to flap damping. After the window is closed, any subsequent changes would count as significant events towards damping. Effectively, this technique creates a filter that passes very, very low frequencies and high frequencies, but will detect and deter ongoing route changes within a certain frequency band. This is normally known as a band pass filter.

### 3.4. Path length damping

The increased meshiness of the core of the Internet has significantly changed the nature of path changes that are visible in BGP. As the meshiness of the network increases, the number of parallel links

between any given pair of ASes tends to increase. This helps protect against single link failures between ASes. This also reduces the frequency of AS path changes on transit prefixes because most of the link failures in the densely meshed part of the network will not result in AS path change.

As a result, when a BGP speaker does see a change in the AS path, and in particular, when the AS path length increases, this would seem to be a good heuristic indication that there is some significant failure in the less densely meshed portion of the network. As a result, it seems likely that such failures are less likely to have alternative working paths and that the increase in path length is a harbinger of path hunting that is likely to be unsuccessful. We therefore suggest that this event could be used to trigger a flap suppression period, which would allow the prefix to oscillate arbitrarily without propagation to the remainder of the network. The obvious risk is that this would be a false negative, unnecessarily disrupting connectivity.

Again, the implementation of this would be relatively straightforward. When a BGP speaker found that it needed to change its best path for a prefix and that the new best path was longer than the previous best path, then it would issue withdraw messages to its neighbors and start a timer. Subsequent changes to the prefix would restart the timer. When the timer expired, the BGP speaker would perform a normal best path election and advertise the result, if any.

### 3.5. Optimal path hysteresis

It has been observed that the overall topology of the Internet at the AS level changes at a fairly low rate. Thus, the optimal AS path to a given prefix, ignoring transient issues, changes at a very low rate. This suggests that caching the optimal AS path and waiting for it to reappear would be another alternative heuristic to help select only the long-term optimal path.

An implementation of this technique might retain a copy of the AS path on per-prefix basis, even if it had no active path to the prefix. Because most implementations maintain a cache of AS paths, this is not necessarily prohibitively expensive. When a new AS path is received for a prefix, the new path is compared to the cached optimal path. If it matches, or it is preferable to the stored optimal path, then the new path is immediately accepted, advertised, and the cache can be updated appropriately. However, if the new AS path is inferior to the cached path, then the implementation can infer that there is some path hunting in progress and can choose to either not perform best path selection, not select the new path, or not advertise the new path. Again, after a suitable period has

elapsed, the implementation may decide that the optimal path is unlikely to appear and may process the inferior path normally.

### 3.6. Delayed best path selection

Another observation based on the discussion in section Section 1.4.1 is that the amount of flap is exacerbated by each AS selecting the best possible path each time a new path is presented. This is not strictly required by BGP, so ignoring some of the incoming paths would be perfectly acceptable. Further, an implementation could reasonably delay performing any best path analysis for an arbitrarily long time, as long as it continued to advertise the path it actually used. Thus, one possible policy would be to only perform best path selection when absolutely required. When the first path for a prefix arrives, the implementation would immediately select that path, thereby restoring connectivity. Subsequent paths from other neighbors for the same prefix would not trigger a new best path computation. Rather, they would simply start a timer that would only expire when the paths had stabilized.

## 4. Next steps

### 4.1. Call for collaboration

As can be seen from the above, there is a great deal of work yet to be done on this subject. Collaborators are most welcome in any aspect of the work.

### 4.2. Literature search

There are a number of technical articles listed below that have been published on BGP flap damping and stability that need to be reviewed and included if they prove substantive. A few known ones are listed here. There are very likely a number of other articles in the literature that are relevant.

[TON-1998]

[Infocom-1999]

[FTCS-1999]

[Sigcomm-2000]

[Infocom-2001]

[Sigcomm-2002]

[PCC-2004]

[Infocom-2005]

### 4.3. Analysis

A number of projects have collected traces of BGP update messages that demonstrate both flap and path hunting. It would be of great interest to examine the effects of some of the proposal in Section 3 in detail on these traces.

### 4.4. Prototyping, Testing and Pilot Deployment

After some analysis, it would then be helpful to actually implement the most useful possible solutions in a number of BGP implementations. Since this is a change to BGP, extensive testing is going to be necessary and a period of pilot deployment will be required. Implementers, testers, and operators could help accelerate this portion of the project.

## 5. Acknowledgments

This document builds on the work of RFC 2439 [RFC2439] and we would like to thank Curtis Villamizar, Ravi Chandra, and Ramesh Govindan for their excellent work.

## 6. IANA Considerations

This memo includes no requests to IANA.

## 7. Security Considerations

This document raises no new security issues.

## 8. References

### 8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway

Protocol 4 (BGP-4)", RFC 4271, January 2006.

## 8.2. Informative References

[Harmful] Bush, R., Griffin, T., and Z. Mao, "Route flap damping: harmful?", <http://www.nanog.org/mtg-0210/ppt/flap.pdf>.

[I-D.iab-raws-report] Meyers, D., "Report from the IAB Workshop on Routing and Addressing", draft-iab-raws-report-02 (work in progress), April 2007.

[RFC2439] Villamizar, C., Chandra, R., and R. Govindan, "BGP Route Flap Damping", RFC 2439, November 1998.

[RIPE-378] Smith, P. and C. Panigl, "RIPE Routing Working Group Recommendations on Route-flap Damping", <http://www.ripe.net/ripe/docs/ripe-378.html>.

## 8.3. Potential References

[FTCS-1999] Labovitz, C., Ahuja, A., and F. Jahanian, "Experimental Study of Internet Stability and Wide-Area Network Failures", FTCS 1999.

[Infocom-1999] Labovitz, C., Malan, G., and F. Jahanian, "Origins of Internet Routing Instability", Infocom 1999.

[Infocom-2001] Labovitz, C., Ahuja, A., Wattenhofer, R., and S. Venkatachary, "The Impact of Internet Policy and Topology on Delayed Routing Convergence", Infocom 2001.

[Infocom-2005] Chandrashekar, J., Duan, Z., Zhang, Z., and J. Krasky, "Limiting path exploration in BGP", Infocom 2005.

[PCC-2004] Duan, Z., Chandrashekar, J., Krasky, J., Xu, K., and Z. Zhang, "Damping BGP Route Flaps", IEEE International Conference on Performance, Computing, and Communications 2002.

[Sigcomm-2000] Labovitz, C., Ahuja, A., Bose, A., and F. Jahanian,

"Delayed Internet Routing Convergence", Sigcomm 2000.

[Sigcomm-2002]
Mao, Z., Govindan, R., Varghese, G., and R. Katz, "Route Flap Damping Exacerbates Internet Routing Convergence", Sigcomm 2002.

[TON-1998]
Labovitz, C., Malan, G., and F. Jahanian, "Internet Routing Instability", TON 1998.

## Author's Address

Tony Li
Cisco Systems, Inc.
170 W. Tasman Dr.
San Jose, CA 95134
US

Phone: +1 408 853 1494
Email: tli@cisco.com

## Full Copyright Statement



## Intellectual Property



## Acknowledgment

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA).