RIPE Routing-WG Recommendations for Coordinated Route-flap Damping Parameters

Christian Panigl
Joachim Schmitz
Philip Smith
Cristina Vistoli

Version 2.0

Document ID: ripe-229
Date Published: 22 October 2001
Obsoletes: ripe-210, ripe-178

Abstract

This document recommends a set of route-flap damping parameters that should be applied by all ISPs in the Internet and should be deployed as default values by BGP router vendors.

Table of Contents

1. Introduction

Route-flap damping is a mechanism for (BGP) routers that is aimed at improving the overall stability of the Internet routing table and reducing the load on the CPUs of the core routers.

1.1 Motivation for route-flap damping

In the early 1990s the accelerating growth in the number of prefixes being announced to the Internet (often due to inadequate prefix-aggregation), the denser meshing through multiple inter-provider paths, and increased instabilities started to cause significant impact on the performance and efficiency of the Internet backbone routers. Every time a routing prefix becomes unreachable because of a single line-flap, the withdrawal has to be advertised to the whole core Internet and dealt with by every single router that is carrying the full Internet routing table.

To overcome this situation a route-flap damping mechanism was invented in 1993 and has been integrated into several router software implementations since 1995 (for example, Cisco, Merit/RSd, GateD Consortium). The implementation is described in detail in RFC 2439. The flap damping mechanism
is now widely used to help keep severe instabilities under control and more localised in the Internet.

And there is a second benefit: it is raising the awareness of the existence of instabilities because severe route/line-flapping problems lead to permanent suppression of the unstable area by means of holding down the flapping prefixes.

Route-flap damping has its greatest and most consistent value if it is applied as near to the source of the problem as possible. Therefore flap-damping should be applied both at peering and upstream boundaries, as well as at customer boundaries (see 1.4 and 1.5 for details).

1.2 What is route-flap damping?

When BGP route-flap damping is enabled in a router, the router starts to collect statistics about the announcement and withdrawal of prefixes. Route-flap damping is governed by a set of parameters with vendor-supplied default values which may be modified by the router manager. The names, semantic and syntax of these parameters differ between the various implementations; however, the behaviour of the damping mechanism is basically the same.

> Each time a prefix is withdrawn, the router will increment the damping penalty by a fixed amount. When the number of withdrawals/announcements (=flap) is exceeded in a given time frame (cutoff threshold) the path is no longer used and not advertised to any BGP neighbour for a predetermined period starting from when the prefix stops flapping. Any more flaps happening after the prefix enters suppressed state will attract additional penalty. Once the prefix stops flapping, the penalty is decremented over time using a half-life parameter until the penalty is below a reuse threshold. Once below this reuse threshold the suppressed path is then re-used and re-advertised to BGP neighbours.

Pointers to some more detailed and vendor specific documents are listed in "5. References".

1.3 "Progressive" versus "flat&gentle" approach

One easy approach would be to just apply the current default-parameters which are treating all prefixes equally ("flat&gentle") everywhere. However, there is a major concern to penalise longer prefixes (=smaller aggregates) more than well aggregated short prefixes ("progressive"), because the number of short prefixes in the routing table is significantly lower and it seems in general that those are tending to be more stable and also are tending to affect more users.

Another aspect is that progressive damping might increase the awareness of aggregation needs. However, it has to be accompanied by a careful design which doesn't force a rush to request and assign more address space than needed.

A significant number of important services are sitting in long prefixes (e.g. root name servers), so the progressive approach has to exclude the strong penalisation for these so-called "golden" prefixes.

With this recommendation we are trying to make a compromise and it is therefore called "graded damping".

1.4 Motivation for coordinated parameters

There is a strong need for the coordinated use of damping parameters for several reasons:

Coordination of "progressiveness":

If penalties are not coordinated throughout the Internet, route-flap damping could lead to additional flapping or inconsistent routing because longer prefixes might already be re-announced through some parts of the Internet where shorter prefixes are still held down through other paths.

Coordination of hold-down and reuse-threshold parameters between ISPs:

If an upstream or peering provider would be damping more aggressively (e.g. triggered by less flaps or applying longer hold-down timers) than an access-provider towards his customers, it will lead to a very inconsistent situation, where a flapping network might still be able to reach "near-line" parts of the Internet. Debugging of such instabilities is then much harder because the effect for the customer leads to the assumption that there

is a problem "somewhere" in the "upstream" Internet instead of making him just call his ISP’s hot line and complain that he can’t get out any longer.

Further, after successful repair of the problem the access-provider can easily clear the flap-damping for his customer on his local router instead of needing to contact upstream Network Operation Centres (NOCs) all over the Internet to get the damping cleared.

Vendor Defaults:

As with most software implementations, there need to be some default values set when route-flap damping is enabled on routers. Vendors choosing different default values will result in a similar situation to that described above, where the more aggressive values will result in "black spots" in the Internet. Coordinated values will ensure consistency in dealing with instabilities.

1.5 Aggregation versus damping

If a customer of an ISP is only using Provider Aggregated addresses, the aggregating upstream provider doesn’t need to apply damping on these prefixes towards his customer because instabilities of such prefixes will not propagate into the Internet. However, if a customer insists on announcing prefixes which can’t be aggregated by its provider, damping should be applied. Reasons for leaking prefixes might include dual-homing (to different providers) of a customer, or customer's reluctance to renumber into the provider's aggregated address range.

1.6 "Golden Networks"

Even though damping is strongly recommended, in some cases it may make sense
to exclude certain networks or even individual hosts from damping. This is especially true if damping would cut off the access to vital infrastructure elements of the Internet. A most prominent example are the root name servers.

At least in principle, there should be enough redundancy for root name servers. However we are still facing a situation where, at least outside the USA, large parts of the Internet are seeing all of them through the same one or two backbone/upstream links (undersea cable) and any instability of those links which is triggering damping would unnecessarily prolong the inaccessibility of the root name servers for an hour (at least those sitting in a /24 or longer prefix).

Other examples of inclusions in the "Golden Networks" might be the Global Top Level Domain (gTLD) name servers, and possibly overseas or "special" networks the local ISP wishes to have continued connectivity to regardless of the instability of the infrastructure in between.

Appendix A.1 references a website which the authors believe represent an example of suitable Golden Networks. While the authors will endeavour to keep the website current, network managers are strongly encouraged to check that the networks listed are indeed still being announced and the hosts therein are still being used before implementation of route flap damping using the quoted Golden Networks. This can be done by matching BGP table announcements with the published addresses for the listed servers.

These exceptions must only be made if there are strong and identifiable needs for them - the rule should be to apply coordinated route flap damping throughout.

2. Recommended damping parameters

2.1 Motivation for recommendation

At RIPE 26 and 27 Christian Panigl presented the following network backbone maintenance example from his own experience, which was triggering flap damping in some upstream and peering ISPs routers for all his and his customers /24 prefixes for more than 3 hours because of too "aggressive" parameters:

scheduled SW upgrade of backbone router failed:

```
   - reload after SW upgrade       1 flap
   - new SW crashed                1 flap
   - reload with old SW            1 flap
                                   ------
                                   3 flaps within 10 minutes
```

which resulted in the following damping scenario at some boundaries with progressive route-flap damping enabled:

```
Prefix length:      /24     /19     /16
suppress time:      ~3h     45-60'  <30'
```

Therefore, in the Routing-WG session at RIPE 27, it was agreed that suppression should not start until the 4th flap in a row and that the maximum suppression should in no case last longer than 1 hour from the last flap.

It was agreed that a recommendation from RIPE would be desirable. Given that the current allocation policies are expected to hold for the foreseeable future, it was suggested that all /19's or shorter prefixes are not penalised harder (longer) than current Cisco default damping does. More recently, this recommendation has been altered so that only prefixes longer than a /21 are now damped more aggressively. The Local Internet Registries' minimum allocation is currently a /20, and a /21 announcement is quite feasible for
a multihoming situation.

With these suggestions in mind, Tony Barber (UUNET) designed the following set of route-flap damping parameters which have proved to work smoothly in his
environment for a couple of months prior to the publication of ripe-178 (the original version of this document).

2.2 Description of recommended damping parameters

Basically the recommended values do the following with harsher treatment for /24 and longer prefixes:

* don't start damping until the 4th flap
* /24 and longer prefixes: max=min outage 60 minutes
* /22 and /23 prefixes: max outage 45 minutes; min outage of 30 minutes
* all other prefix lengths: max outage 30 minutes; min outage 10 minutes

If a specific damping implementation does not allow configuration of prefix-dependent parameters the least aggressive set should be used:

* don't start damping before the 4th flap in a row
* max outage 30 minutes; min outage 10 minutes

Sample configurations for different vendors are referenced in Appendix A.2. These samples can be used as a basis for a configuration on other router platforms not listed there.

3. Other Features contributing to Internet Stability

3.1 BGP Route refresh

RFC 2918 describes a Route Refresh Capability for BGP-4. Prior to this, there was no mechanism to reset or refresh a BGP peering session without tearing it down and waiting for it to re-establish. This process is destructive - prefixes being exchanged between the two peering routers are withdrawn from their respective ASes, and this withdrawal can potentially pass through the whole Internet causing the burden and increased instability discussed earlier. Usually all that an ISP wishes when resetting a BGP session is to implement new or revised policy - destroying a BGP session carrying a large or the full routing table has severe impact on the ISP and his neighbours on the Internet. Furthermore, reset of a BGP session means the withdrawal of reachability information from the ISP’s customers, and they have the perception that the Internet has "vanished" -- the impression left with the end user is that of an unreliable network.

Route Refresh implements a messaging system whereby a router wishing to refresh or reset its BGP peering with its neighbour simply has to send the notification. When the neighbour receives the notification, it will send its entire announcement to its peer (obtained from BGP best path table and applicable outbound policy).

To find out if your neighbour supports Route Refresh, using Cisco IOS as an example, enter:

```
Router# sho ip bgp neigh w.x.y.c | include refresh
  Received route refresh capability(new) from peer
  Route refresh request: received 0, sent 0
```

If your router and your peer router support Route Refresh, you can use:

```
Router# clear ip bgp w.x.y.c in
```

for requesting a route refresh without clearing the BGP session.

For an outbound route refresh without clearing the BGP session use

```
Router# clear ip bgp w.x.y.c out
```

It is recommended that all users of BGP use the route refresh capability when implementing new BGP policy.

3.2 Soft-Reconfiguration

Where the neighbour does not support RFC 2918 Route Refresh, router vendors have implemented functionality to allow the alteration of BGP policy without resetting the BGP session.

In Cisco IOS this functionality is called "Soft Reconfiguration". This reserves additional memory in the router to store the BGP table exactly as it was received from the peer, prior to any inbound policy being applied. The advantage of this is that the ISP can then change any inbound policy on the router without resetting the BGP session -- the router simply uses the "raw" BGP table it has received from its peer. Disadvantage is that this functionality could potentially consume almost twice the amount of memory required for the BGP table heard from the peer.

To configure soft-reconfiguration in IOS, simply add the extra line to the BGP peer configuration as below. Soft-reconfiguration is configured on a per-neighbour basis.

```
!
router bgp 65501
  neighbor 10.0.0.2 remote-as 65502
  neighbor 10.0.0.2 soft-reconfiguration inbound
!
```

Without the keyword "soft" a "clear ip bgp x.x.x.x" will completely reset the BGP session and therefore always withdraw all announced prefixes from/to neighbour x.x.x.x and re-advertise them (=route-flap for all prefixes which are available before and after the clear). With "clear ip bgp x.x.x.x soft out" the router doesn’t reset the BGP session itself but sends an update for all its advertised prefixes. With "clear ip bgp x.x.x.x soft in" the router just compares the already received routes (stored in the "received" data structures) from the neighbour against locally configured inbound policy statements.

In Juniper’s JunOS software, all the prefixes advertised by a peer are stored on the router, allowing the router to re-evaluate new policies on the set of routes advertised by the peer. So in the event of a peer not supporting the route-refresh capability, JunOS default configuration will compensate for this in the same way the optional "soft-reconfiguration" support in IOS.

It is recommended to use soft-reconfiguration with all peers that do not support RFC 2918 Route Refresh Capability to avoid tearing down and restarting BGP peerings when new BGP policies need to be implemented.

3.3 Tuning External BGP Failover

Cisco IOS by default implements a feature known as "fast-external-fallover". This feature immediately clears the BGP session whenever the line-protocol to the external neighbour goes down. This feature is desirable so that there is fast failover in case of link failures - the router can withdraw paths as soon as the line goes down, rather than waiting for BGP keepalive timers. The drawback of this, however, is that circuits which are prone to unreliability will cause BGP sessions to drop and return (i.e. flap), resulting in instability within the ISP's network, and the potential for flap damping by upstreams or peers.

If fast-external-fallover is turned off, the BGP sessions will survive these short line-flaps as it will use the longer BGP keepalive/hold timers (default 60/180 seconds). The drawback of turning it off - and currently it has to be done for a whole router and can not be selected peer-by-peer - is that the switch-over to an alternative path will take longer.

We recommend turning off fast-external-fallover whenever possible:

```
!
router bgp 65501
  no bgp fast-external-fallover
!
```

Alternatively it might be considered acceptable to retain "fast-external-fallover" and to turn off "interface keepalives" on unreliable circuits to overcome the immediate BGP resets on any significant CRC error period.

Another potentially more satisfactory alternative would be to use a shorter per-neighbour BGP keepalive timer that has to be applied on both routers (e.g. 10 seconds that gives a hold-timer of 30 seconds):

```
!
router bgp 65501
  neighbor w.x.y.z timers 10
!
```

In JunOS, this instability can be avoided by using the following commands:

- out-delay <second>; applicable to all BGP peers, all peers in a group, or an individual peer. This implements a delay between when the routing table receives the routing information and when the information is exported to BGP peers.

- hold-time sec; applicable to all BGP peers, all peers in a group or an individual peer. This allows a shorter per neighbor holdtimer to be applied on both routers (30 sec will gives keepalives of 10 sec).

- hold-time msec; to be configured in the router interfaces where the BGP peering will be established. This delays the propagation of the interfaces-down events to the routing protocol.

4. Potential problems

4.1 Multiplication of flaps between ASes with multiple interconnections

Christian Panigl experienced the following during a circuit upgrade of an Ebone customer:

- Only ONE flap was generated as a result of the upgrade process (disconnect router-port from modem A, reconnect to modem B). Nevertheless the customer's prefix was damped in all ICM routers.

- The flap statistics in the ICM routers stated *4* flaps !!!

- 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.

There are two potential solutions to work around this problem. The first one is operational, the second one is a software configuration feature (for Cisco IOS and possibly other implementations as well).

* Schedule a downtime for at least 3-5 minutes which should be enough time for the prefix withdrawals to have propagated through all paths before reconnection and re-advertisement of the prefix. Avoid clearing BGP sessions as this also could generate a 30 minute outage through flap damping!

* Configure a permanent static route pointing to the customer interface. Even if the interface goes down, there is still an entry in the routing table for the customer network, and BGP will therefore still announce the prefix. Example, using Cisco IOS:

```
!
router bgp 65500
 network 169.254.0.0 mask 255.255.0.0
ip route 169.254.0.0 255.255.0.0 serial 5/0 permanent
!
```

  If migrating the customer from one router port to another, simply enter the second static route pointing to the new interface. Move the cable between ports - BGP continues to announce the prefix as the entry is still in the routing table.

Note: this solution only applies to customers who connect using static routes. If the customer connects using BGP, first disable fast-external-fallover on both the customer and ISP router, and then move the cable in a time period less than the BGP hold-timer.

4.2 Non-recommended flap damping parameters

There are situations where service providers would like to design their own route flap damping parameters for local needs or conditions. If this is really desired, then it is important to pay attention to how flap damping parameters are configured, whether the values are feasible or not, etc.

For example, in Cisco IOS, it is perfectly possible to configure flap damping parameters which do nothing, with IOS not giving any warning about them being "unfeasible" parameters.

* One example might be the configuration "set dampening 15 500 3000 30". Here the reuse limit is 500, maximum suppress time is 30 minutes and the half-life is 15 minutes. Using these three parameters gives a maximum possible penalty value of 2000, well below the suppress limit of 3000. So even though this can be successfully configured on the router, no damping will take place.

* Another example might be the configuration "set dampening 15 750 3000 30". Here the reuse limit is 750, maximum suppress time is 30 minutes and the half-life is 15 minutes. Using these three parameters gives a maximum possible penalty of 3000, exactly the same as the suppress-limit. In Cisco IOS, the penalty is decayed every 5 seconds, so flap damping will only take place if the update follows the withdrawal within that 5 second time frame. 99% of the time no flap damping will take place.

5. References

RIPE/Routing-WG Minutes dealing with Route Flap Damping:

http://www.ripe.net/ripe/meetings/archive/ripe-24/ripe-m-24.txt
http://www.ripe.net/ripe/meetings/archive/ripe-25/ripe-m-25.txt
http://www.ripe.net/wg/routing/r25-routing.html
http://www.ripe.net/wg/routing/r26-routing.html
http://www.ripe.net/wg/routing/r27-routing.html

Curtis Villamizar, Ravi Chandra, Ramesh Govindan
RFC2439: BGP Route Flap Damping (Proposed Standard)
ftp://ftp.ietf.org/rfc/rfc2439.txt

Enke Chen
RFC 2918: Route Refresh Capability for BGP-4 (Proposed Standard)
ftp://ftp.ietf.org/rfc/rfc2918.txt

Merit/IPMA: Internet Routing Recommendations
http://www.merit.edu/ipma/docs/help.html

Cisco BGP Case Studies: Route Flap Damping
http://www.cisco.com/warp/public/459/16.html

Cisco Documentation: Configuring BGP / Route Damping / Soft Reset
http://www.cisco.com/univercd/cc/td/doc/product/software/ios122/122cgcr/fipr_c/ipcprt2/1cfbgp.htm

ISI/RSd Configuration: Route Flap Damping
http://www.isi.edu/div7/ra/RSd/doc/dampen.html

GateD Configuration: Weighted Route Damping Statement
http://www.nexthop.com/techinfo/manuals/o_config_guide/bgp/weighted_route_dampening.shtml

Juniper Configuration: Configuring Dumping parameters
http://www.juniper.net/techpubs/software/junos44/swconfig44-routing/html/policy-damping-config.html

6. Acknowledgements

Thanks go to all the contributors to this updated version and to the RIPE NCC for hosting the "Golden Networks" website.

7. Changes over previous versions

This document is a significant rewrite and update of ripe-210. The "Golden Networks" have now been moved from this document on to a website dedicated to listing them as they are frequently changing. The authors have come across several instances of providers implementing the recommendations without actually checking that the Root Nameserver networks were still as listed in the document.

Updates to the Cisco IOS configuration have been made, and the parameters chosen for /24 networks have been corrected to make them feasible. Juniper JunOS configuration samples have been added to this document.

Examples of flap damping in operation have been added to Appendix 3.

Router configurations for the recommended route flap damping parameters have been moved out of this document to the website.

8. Authors

The authors can be contacted as follows:

Philip Smith <pfs@cisco.com>
Cristina Vistoli <cvistoli@juniper.net>
Christian Panigl <panigl@cc.univie.ac.at>
Joachim Schmitz <schmitzjo@aol.com>

Appendices

A.1 "Golden Networks"

Examples of Golden Networks can be found on a website which has been set up specifically for them. Please consult http://www.golden-networks.net/ for a sample list of current golden networks and the equivalent router configuration for these networks.

A.2 Sample Configurations

Sample Router configurations which have been contributed to this project can be found at the http://www.golden-networks.net/ website. Contributions of working configurations from other routing software should be sent to the authors for inclusion in the website.

A.3 Study of Flap Damping Operation

It is instructive to observe how route flap damping actually works on a router - doing so will help the reader understand how the particular values described in Section 2.2 were chosen. The tests were carried out using both Cisco IOS and JunOS.

A.3.1 Cisco IOS

The test bed had two Cisco routers connected to each other. One router originated prefixes, the other one had the flap damping parameters described above in the text. The router originating the prefixes would withdraw a prefix, then reannounce, then withdraw, reannounce, etc. The BGP process in IOS checks every 60 seconds for any new or withdrawn prefixes in the local configuration - so on the source router, the withdraw and announce was done by removing and adding the BGP network statement for the prefix in question. The router monitoring the flaps would see the prefix being withdrawn and then announced 60 seconds later.

A.3.1.1 For /24s

Parameters used are "set dampening 15 820 3000 30"

```
  1st flap   1000   decay to 966, 982 at update
```

2nd flap   1966   decay to 1894, 1926 at update
3rd flap   2894   decay to 2787, 2846 at update
4th flap   3280   decay to 3165, 3226 at update

Maximum possible penalty is 3280 as defined by the flap parameters, so the penalty at the 4th flap was only incremented from 2787 to 3280, not 3787 as might have been expected. At the 4th flap the prefix was marked as being suppressed for 59 minutes when the update message was received. If the update after the 4th flap was not received within 4 minutes and 20 seconds, the penalty dropped below 3000, and the prefix was not suppressed.

A.3.1.2 For /22s, /23s

Parameters used are "set dampening 15 750 3000 45"
1st flap   1000   decay to 921, 960 at update
2nd flap   1921   decay to 1777, 1850 at update
3rd flap   2777   decay to 2583, 2671 at update
4th flap   3583   decay to 3311, 3451 at update

Maximum possible penalty is 6000. At the 4th flap the prefix was marked as being suppressed for 33 minutes when the update message was received. If the update after the 4th flap was not received within 4 minutes and 40 seconds, the penalty dropped below 3000, and the prefix was not suppressed.

A.3.1.3 For remaining prefixes

Parameters used are "set dampening 10 1500 3000 30"
1st flap   1000   decay to 889, 946 at update
2nd flap   1889   decay to 1679, 1781 at update
3rd flap   2679   decay to 2367, 2526 at update
4th flap   3367   decay to 3019, 3176 at update

Maximum possible penalty is 12000. At the 4th flap the prefix was marked as being suppressed for 10 minutes when the update message was received. If the update after the 4th flap was not received within 2 minutes and 5 seconds, the penalty dropped below 3000, and the prefix was not suppressed.

A.3.2 JunOS

A similar test bed with two Juniper routers was set up using the damping parameters described in Appendix A.2.2 above. One router originated prefixes, the other router implemented the flap damping parameters. The router originating the prefixes would withdraw a prefix, then reannounce, then withdraw, reannounce, etc, with the effects being monitored on the second router.

A.3.2.1 For /24s

Parameters used are "set-high policy"

    half-life 30;
    reuse 1640;
    suppress 6000;
    max-suppress 60;

  1 up/down: decay to 1946
  2 up/down: decay to 3723
  3 up/down: decay to 5575
  4 up/down: decay to 6577

At the 4th flap the prefix was marked as being suppressed for 1 hour when the update message was received.

A.3.2.2 For /22s, /23s

Parameters used are "set-medium policy"

    half-life 15;
    reuse 1500;
    suppress 6000;
    max-suppress 45;

  1 up/down: decay to 1939
  2 up/down: decay to 3269
  3 up/down: decay to 3733
  4 up/down: decay to 4944
  5 up/down: decay to 6032

At the 5th flap the prefix was marked as being suppressed for 30 min when the update message was received

A.3.2.3 For remaining prefixes

Parameters used are "set-normal policy"

    half-life 10;
    reuse 3000;
    suppress 6000;
    max-suppress 30;

  1 up/down: decay to 1909
  2 up/down: decay to 3503
  3 up/down: decay to 5065
  4 up/down: decay to 6556

At the 4th flap the prefix was marked as being suppressed for 10 min when the update message was received

A.3.3 Summary

When analysing flap damping performance on the router or across the network, network managers should compare with the above lab tests. Note especially that slowly flapping prefixes are unlikely to be suppressed even though they show significant flapping history. A future version of this document may consider what to do in this instance.