Internet Engineering Task Force   V. Van den Schrieck
Internet-Draft   O. Bonaventure
Intended status: Informational   Universite catholique de Louvain
Expires: December 31, 2007   June 29, 2007

Routing oscillations using BGP multiple paths advertisement draft-vandenschrieck-bgp-add-paths-oscillations-00.txt

Abstract

The advertisement of multiple paths for the same prefix in the Border Gateway Protocol is possible by using an extension to the attributes developed for multiprotocol transfer. This allows better path diversity in the adj-rib-ins, and can prevent some routing inconsistencies with Route Reflection. However, if not carefully used, this mechanism can also introduce routing oscillations and inconsistencies in topologies with Route Reflection that were correct without this extension.

This document describes four types of routing inconsistencies that occur in badly designed topologies when more than one path is advertised by the routers.

1. Introduction

The Border Gateway Protocol (BGP) [1] is the interdomain routing protocol currently used in the Internet. BGP-speaking routers exchange network reachability information with each other by advertising their best path to each destination prefix.

The ADD-PATH BGP capability allows multiple paths advertisement for the same prefix [2]. Using multiple nexthops for a given prefix in UPDATE messages, as explained in [4] serves similar purpose. The advertisement of all available paths to a given prefix by Route Reflectors [3] can prevent some routing oscillations, at the cost of an increased memory consumption in the Adj-Rib-Ins. If a carefully chosen subset of the paths is advertised by Route Reflectors in a topology respecting some constraints on the IGP weights, MED oscillations can also be prevented [5].

However, if the topology does not respect these constraints, routing oscillations can appear while using multiple paths advertisement, even if the system was stable with best path advertisement only.

This document presents several systems that have routing inconsistencies when a subset of the available paths are advertised by the routers when Route Reflection is used. They are inspired from systems that have routing inconsistencies with standard BGP and Route Reflection [6][7][8].

2. Topology with multiple solutions

The network described in Figure 1 represents two ASes. The edges in the schema represens eBGP or iBGP sessions, depending on the routers being in the same AS or not. AS 1 has two Route Reflectors, each of them having two clients. AS2 advertises prefix P on four eBGP sessions with AS1. We call PA the path to P via RA, PB the path via RB, etc. The IGP links between the routers in AS 1, not shown in the figure, are such that RR1's preferences on the paths to P are PC > PA > PB > PD. Similarly RR2's preferences are PB > PD > PC > PA.

If only the best path is advertised, this system converges : RR1 chooses and advertises path PA, and RR2 path PD. If all paths are advertised, they both know all the available paths, and RR1 can choose C while RR2 selects PB.

If BGP ADD-PATHS is used and the routers advertise all paths, both Route Reflectors learn the four available paths and the system also converges.

However, if only two paths are advertised, this topology has two solutions. We suppose that the two paths advertised are always the two that were ranked highest by the BGP decision process. Even if one or both of those two paths were learned from some iBGP neighbor, the next preferred paths are never advertised to that neighbor.

Depending on which route reflector advertises the paths from its client first, two different states can be reached :

- If RR1 advertises paths PA and PB before RR2, RR2 selects PB as its best path and advertises PB and PD. RR1 never learns path PC, and keeps PA as its best route.

- If RR2 is the first to send its update, RR1 chooses PC as its best path, and RR2 never learns PB.

If both Route Reflectors always send their paths to each other together, the system never converges.

System with two solutions

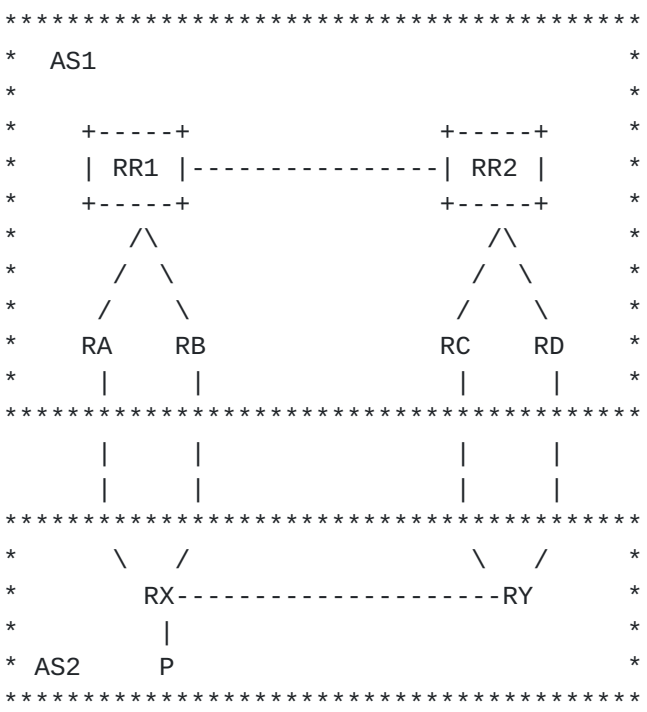

Figure 1

3. Topology with MED oscillations on the second path

In Figure 2, AS1 has two Route Reflectors. RR1 has RA and RB as clients, and RR2 has RC as client. AS 1 peers with ASX, ASY and AS0. Router RX of ASX has an eBGP session with RA, router RZ of ASY advertises the path to P to RB with a MED attribute of 1, and to RC with MED 0. Router R0 of AS0 has an eBGP session with route reflector RR1. ASX, ASZ and AS0 also peer with each other.

IGP costs in AS 1 are such that RB is the nearest router for both RRs, then RA, then RC.

If router R0 of AS0 advertises prefix P, AS1 learns 4 paths towards P : via RA, RB, RC and RR1. All routers of AS1 choose as best path the one via RR1, as it has a shorter AS path. This system converges to a unique solution if standard BGP is used.

If BGP ADD-PATHS is used and the routers advertise all paths, both Route Reflectors learn the four available paths and the system also converges.

However, if the routers of AS1 advertise only the two paths that were ranked highest by the BGP decision process, the system can oscillate on the second path. The best path is still the one via RR1, but the selection of the second path leads to the classical MED oscillation problem [6][8].

The selection of the second path of RR1 depends upon RR2 advertising PC or not. Indeed, if RR1 knows path PC, it selects PA as its second path even if PB is better in terms of IGP distances, because PB has a higher MED than PC. But if it does not know about PC, PB is selected as its second best path. Similarly, the advertisement of PC by RR2 depends upon the second best path selected by RR1. PC is advertised if RR1 selects PB, but not if it selects PA, because RR2 prefers PA over PC but not over PB because of the MED attribute.

This system is clearly inconsistent : RR1 advertises PA if RR2 selects PC, but RR2 withdraws PC if it knows about PA, which is then withdrawn by RR1, resulting in RR2 selecting PC again, and so on.

MED oscillations

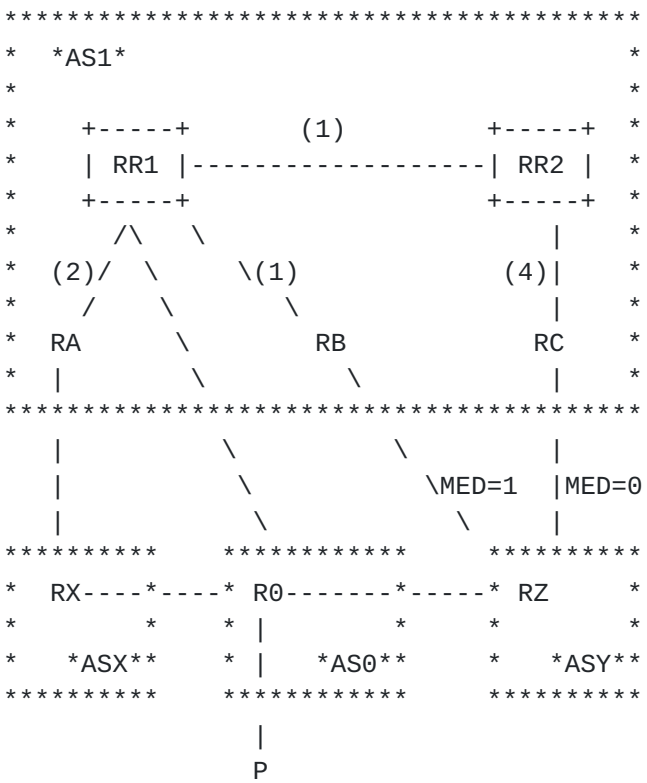

Figure 2

4. Topology with oscillations on the second path

In Figure 3, AS1 has three Route Reflectors, each of them having one client. The IGP links and the corresponding costs, not shown in the figure, are such that each Route Reflector prefers its left neighbor path over its own client path, this one being preferred over the right neighor path.

Prefix P is advertised by AS0 to AS1 and AS2. Routers of AS1 choose the path advertised on the eBGP link with AS0 as their best path, as it has the shortest AS-Path. This topology has thus coherent routing if only one path is advertised. Similarly, if all paths are learned by all Route Reflectors, they are able to choose their left neighbor path as second best path.

However, if only two paths are advertised, there are routing oscillations on the second best path chosen by each RRs. Each of them constantly advertises then withdraws its client path, depending on the right neighbor withdrawing or advertising its own client path. The system being circular, it never converges.

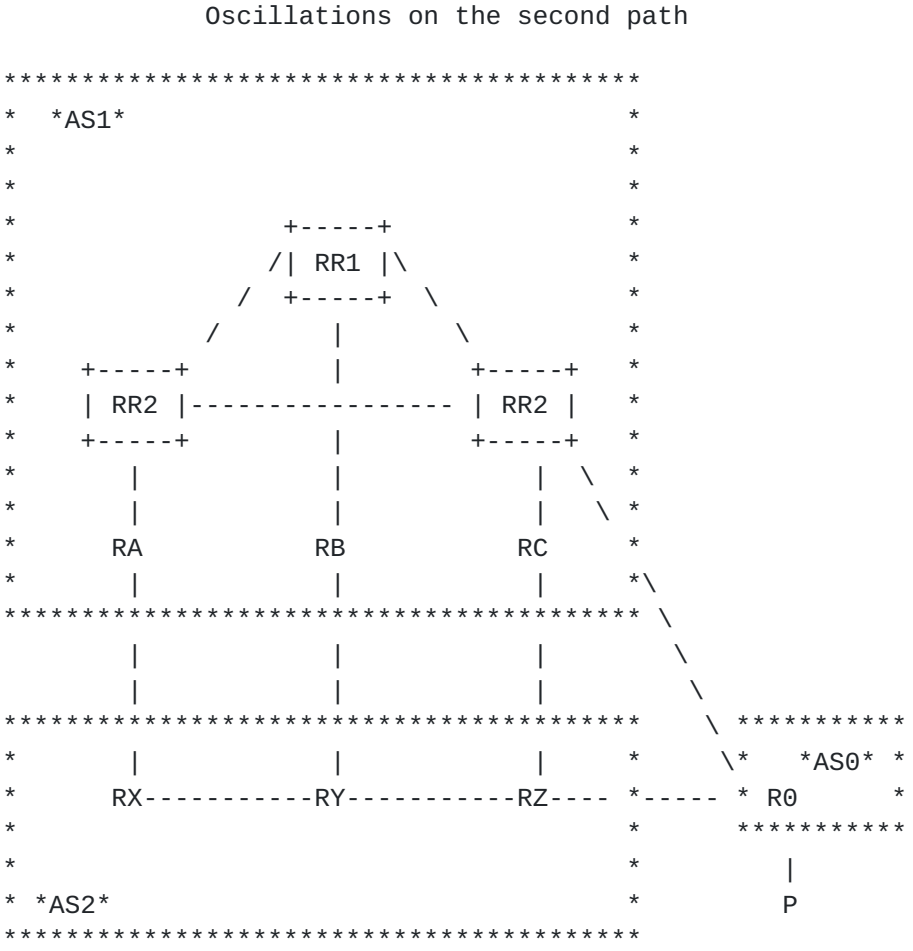

Figure 3

5. Non deterministic topology becoming oscillating

In Figure 4, AS1 has three Route Reflectors, each of them having one client. The IGP links and the corresponding costs, not shown in the figure, are such that each Route Reflector prefers its left neighbor path over its right neighbor path, this one being preferred over its own client path.

If only the best route is advertised, the system is already instable : Depending on the time on which each Route Reflector advertises the path via its client, it can converge to a common path being chosen by

all the RRs. For example, if RR1 advertises PA first, RR2 and RR3 select PA as their best path and don't advertise PB and PC. If they all advertise their client path at the same time, the system diverges : All RRs learn a better path than their own, and they thus simultaneously withdraw then re-advertise their client paths.

If two paths are advertised, things get worse : the system always diverges. Even if the initial path advertisements are not synchronised, Route Reflectors are constantly advertising and withdrawing their client path. For example, if PA is advertised first, RR2 and RR3 choose it as their best path, but still advertise respectively PB and PC as their second path. All Route Reflectors having learned two paths better than their own, they withdraw their client path then re-advertise it, and the system oscillates.

This topology is instable even with only the best path advertised, but can still reach a stable state. If the configuration of the Route Reflectors is modified so that they advertise two paths instead of one, routing oscillations appear and the system becomes inconsistent. The only way to have deterministic routing is to advertise all available paths to P.

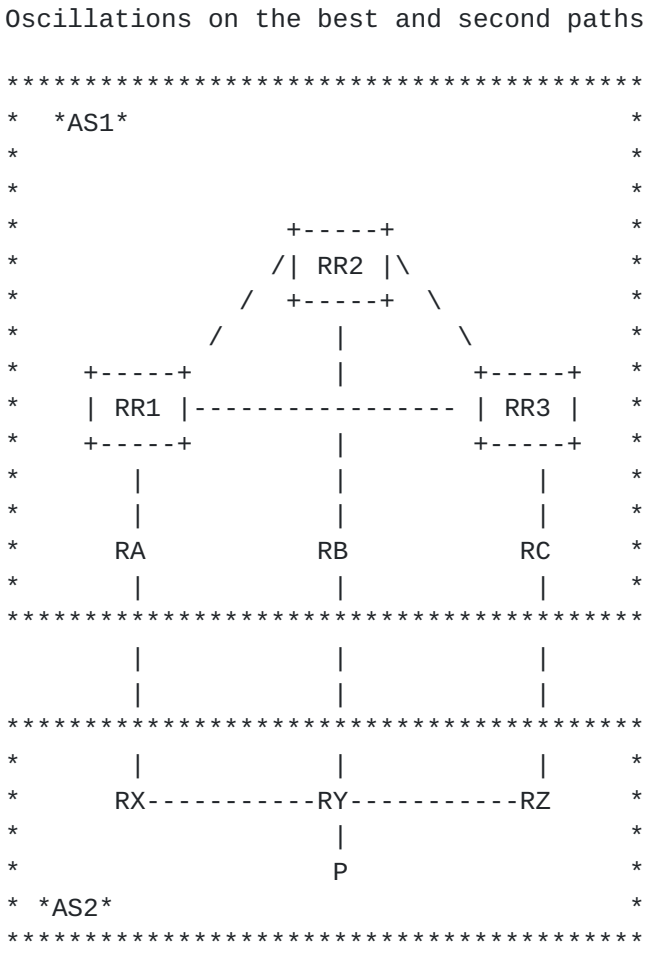

Figure 4

6. Conclusion

This document presents several situations where using the ADD-PATH BGP capability with iBGP Route Reflection introduces routing inconsistancies in systems that have coherent solutions otherwise. It also presents a bad designed system that can sometimes converge to a solution with normal BGP, but always oscillates on the best and the second route when advertising two routes instead of one.

This document shows that the ADD-PATH extension to BGP should be carefully used by operators, as routing loops can occur when

advertising only a subset of the routes. One way to prevent such routing inconsistancies is to advertise all available routes instead of a subset, such that route diversity is identical as when using Full Mesh iBGP. The drawback of this solution is the memory consumption, which can cause scalability concerns. Another possibility is to design the network such that the IGP distances between Route Reflectors and their clients are smaller that the IGP distances between Route Reflectors [5][7]. This prevents routing inconsistancies to appear, because Route Reflectors will never prefer the paths advertised by other Route Reflectors over the paths advertised by their clients. However, designing iBGP in order to respect that constraint is not always possible. Furthermore, even if the network topology respects that constraint under normal operation, it could be violated as a consequence of links or nodes failure [7].

7. IANA Considerations

This memo includes no request to IANA.

8. Security Considerations

This discussion does not introduces security concerns to BGP or any specifications referenced in this document.

9. References

9.1. Normative References

[1] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[2] Walton, D., "Advertisement of Multiple Paths in BGP", draft-walton-bgp-add-paths-05 (work in progress), March 2006.

[3] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.

9.2. Informative References

[4] Bhatia, M., "Advertising Multiple NextHop Routes in BGP", draft-bhatia-bgp-multiple-next-hops-01 (work in progress), August 2006.

[5] Walton, D., "BGP Persistent Route Oscillation Solution",

draft-walton-bgp-route-oscillation-stop-01 (work in progress), July 2005.

[6] McPherson, D., Gill, V., Walton, D., and A. Retana, "Border Gateway Protocol (BGP) Persistent Route Oscillation Condition", RFC 3345, August 2002.

[7] Griffin, T. and G. Wilfong, "On the Correctness of iBGP Configuration", In ACM SIGCOMM Computer Communication Review, Proceedings of the 2002 SIGCOMM conference, p.17-29, 2002.

[8] Griffin, T. and G. Wilfong, "Analysis of the MED oscillation problem in BGP", In Network Protocols, 2002. Proceedings. 10th IEEE International Conference on Network Protocols, 2002.

Authors' Addresses

Virginie Van den Schrieck
Universite catholique de Louvain
Louvain-la-Neuve,
Belgium

Email: firstname.lastname@uclouvain.be
URI: http://inl.info.ucl.ac.be

Olivier Bonaventure
Universite catholique de Louvain
Louvain-la-Neuve,
Belgium

Email: firstname.uclouvain@be.lastname
URI: http://inl.info.ucl.ac.be

Full Copyright Statement

Intellectual Property

Acknowledgment

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